iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

An intelligence approach for the improvement of IT startups' social media strategy

Ana Rita Henrique Peixoto

A thesis presented for the fulfilment of the degree of Doctor in Information Science and Technology

Supervisors:
Professor Doctor Ana de Almeida
ISCTE - Instituto Universitário de Lisboa

Professor Doctor Nuno António
Nova IMS

November, 2023

iscte

TECNOLOGIAS
E ARQUITETURA

An intelligence approach for the improvement of IT startups' social media strategy

Ana Rita Henrique Peixoto

A thesis presented for the fulfilment of the degree of Doctor in Information Science and Technology

Jury President:
Professor Doctor Rúben Pereira, ISCTE - Instituto Universitário de Lisboa

Jury:
Professor Doctor Célia Maria Quitério Ramos,
Universidade do Algarve
Professor Doctor René Bohnsack,
Universidade Católica Portuguesa
Professor Doctor Sérgio Miguel Carneiro Moro,
ISCTE - Instituto Universitário de Lisboa
Professor Doctor Ana de Almeida,
ISCTE - Instituto Universitário de Lisboa

November, 2023

“Mais um saltinho de pardal.”
To my mother, always.

Abstract

Often constrained by limited resources and a restricted budget, startups struggle to acquire customers and secure vital funding. Social media platforms offer cost-effective marketing opportunities that allow startups to generate captivating content and build a community of customers and suppliers. The central aim of this dissertation is to extract knowledge that serves to improve the startups' social media strategies to attain their goals. Data for known Portuguese IT startups has been collected from X, formerly known as Twitter, and used as a social media source to analyze the startups' posting activity. The results enabled the creation of a novel startup life cycle model that allows the characterization of an IT startup's evolution. Initially, startups focus on the idea of conceptualization and prototype development. Since their priority is not social media, they post less, and their popularity levels are low. Along with the startup's growth, the social media presence expands, and the content posted mirrors the development. In addition to being more active, its popularity level increased. The attained results offer essential insights for startups to consider when building their social media strategies. Essentially, media strategies should consider the current phase and monitor what is going on with the accounts and social communities. Additionally, this dissertation also proposes new methodological processes for social media activity analysis usable on platforms like Twitter.

Keywords: Social Media; Startups Life Cycle; Twitter Data.

Resumo

Muitas vezes condicionadas com recursos limitados e um orçamento restrito, as startups lutam por angariar clientes e garantir financiamento. As redes sociais oferecem oportunidades de marketing de baixo custo que permitem às startups gerar conteúdos cativantes e criar uma comunidade de clientes e fornecedores. O objetivo central desta dissertação é extrair conhecimento para melhorar as estratégias nas redes sociais das startups e ajudá-las a atingir os seus objetivos. Analisámos dados de conhecidas startups Portuguesas de IT que foram extraídos do X, antigamente conhecido como Twitter. Os resultados permitiram a criação de um modelo para o ciclo de vida das startups que caracteriza a evolução de uma startup de IT. Inicialmente, o foco da startup é a conceptualização da ideia e o desenvolvimento do protótipo. Como a sua prioridade não são as redes sociais, publicam em menor quantidade e não são populares. À medida que cresce, as suas redes sociais expandem-se e o conteúdo publicado reflete o seu desenvolvimento. Como são mais ativas nas redes sociais, isso reflete-se no seu nível de popularidade que aumenta. Os resultados alcançados oferecem perspetivas essenciais para as startups construírem as suas estratégias de redes sociais em função da fase atual e monitorizarem o que deve acontecer nas suas contas. Por último, esta dissertação propõe metodologias a aplicar na análise de dados de redes sociais como o Twitter.

Palavras-chave: Redes Sociais; Ciclo de Vida das Startups; Dados do Twitter.

Acknowledgment

First, I want to express my grand appreciation to my supervisors, Professor Ana Almeida and Professor Nuno António, for accepting the challenge of supervising my dissertation and for their guidance and teaching through each stage of the process. But most importantly, I would like to thank them for their patience, for believing in my capacities, and for always encouraging me to do more and better. They are an example of what an excellent supervisor and teacher should be, and I will cherish all the wisdom and experiences of these last four years with them.

Besides my supervisors, I would like to thank all the professors of the doctoral classes at ISCTE. However, I want to highlight Professor Fernando Batista and Professor Ricardo Ribeiro, as their contribution was essential to my dissertation. They taught me all the principles of natural language processing, which was crucial to my research work, and mostly, they were present in these four years to help me regardless of the subject. Further, I want to show my appreciation to Professor Elsa Cardoso for all the knowledge provided regarding data visualization.

I want to thank the financial support of Fundação para a Ciência e a Tecnologia (FCT, UIDB/04466/2020), ISTAR-IUL (UIDP/04466/2020), NOVA IMS (UIDB/04152/2020), INESC-ID (UIDB/50021/2020), and CIES (UIDB/03126/2020).

Additionally, I sincerely thank all my family and friends for their unconditional support, love, and belief in my abilities. I must emphasize my gratitude to my father, Carlos, and brother, Nico, for always encouraging me to follow my dreams and showing me that I was capable of anything. They make my life so much more beautiful.

Last but not least, I need to kindly thank my boyfriend, Rúben, for his patience, for all the brainstorms we made about my research, and for the enormous reviews he did of my written work. He made my doctoral journey more fun and enjoyable, and I am massively grateful that he made part of it.

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
BI	Business Intelligence
B2B	Business to Business
CRM	Customer Relationship Management
eWOW	Electronic Word-of-Mouth
FF	Following and Followers
FFF	Family, Friends and Fools
FPEM	Funding and Product Evolution Model
FPEMv2	Funding and Product Evolution Model - extended version
IPO	Initial Public Offering
IT	Information Technology
KPI	Key Performance Indicator
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
NAV	Network Analysis and Visualization
NMF	Non-negative Matrix Factorization
PICOC	Population, Intervention, Comparison, Outcome, and Context
ROI	Return On Investment
SLR	Systematic Literature Review
SMA	Social Media Analytics
SMI	Social Media Intelligence
SMM	Social Media Marketing

CHAPTER 1

Introduction

1.1. Background and motivation

The definition of what a startup is has evolved over time. The definition introduced by Lugović and Ahmed (2015) involves two perspectives: one concerning the business dimension and the other concerning the company's characteristics. Regarding the business dimension, if a company has been established for less than one year and employs at least one person besides its founders, it can be considered a startup. As for the company characteristics, it must be an innovative and growth-oriented business. However, more recent work suggests that the startup definition depends on the actual stage of the company's life cycle (Skala, 2019). Therefore, one can say that there is no general consensus about the definition of what a startup is, but we now have different perspectives that enable the characterization of these small companies.

Compared to other companies, startups are innovative and typically, when successful, present an above-average growth in the number of customers and revenue Lugović and Ahmed (2015). Nevertheless, their resources are limited, and they usually have a restricted budget to reach customers. To thrive, startups must raise funding. Social media platforms can serve as a gateway for various communities, allowing companies to achieve their goals and grow in several dimensions (Rizvanović et al., 2023). Compared to other media, using social media platforms implies a small investment, driving startup companies to use it as a cost-effective tool to create a digital gateway for finding customers and raising funds. The last are two of the three critical startup challenges reported by Wang et al. (2016), and building the product is the third. These challenges derive from a startup company's fast pace of growth, making it difficult to identify the correct steps to take for scaling up. Gulati and DeSantola (2016) explained that startups can improve their growth and achieve their objectives by understanding the best scaling practices.

Undoubtedly, social media has become a fundamental piece of the global information ecosystem, generating large amounts of data. Social media data can provide information about clients, products, and the overall market, helping the decision-making processes (Saura et al., 2021). However, social raw data must be structured, prepared, and interpreted to infer relevant information to support decisions. Understanding social media as a tool can enhance a company's Return On Investment - ROI - while enabling better customer relationship management - CRM. Recent existing studies support this argument and consider social media data-driven projects a strategic business knowledge source (Kapoor et al., 2018; Saura et al., 2021).

Since Twitter has been an ideal platform for small businesses like startups and where they are now massively present, we have chosen this platform as our primary research data source [1]. Even think tanks, a usual birthplace for startups (Feld and Hathaway, 2020), use social media, like Twitter, to disseminate their activities and achieve funding (Castillo-Esparcia et al., 2020). Twitter differs from other social media platforms because it gives access to a global audience where users openly communicate with other users. Above all, it offers an opportunity for businesses to interact and receive instant feedback instead of acting solely as a marketing tool (Curran et al., 2011). Campos-Domínguez (2017) classified Twitter activity as spontaneous and instantaneous, which can encourage a fluid exchange of ideas. Thus, Twitter can be considered a social media tool to help a business establish a network between customers, owners, and investors - providing an environment where professional content coexists with user-generated content, that is, nonexpert content (Casero-Ripollés, 2018). Twitter has simple API access and is widely used in academic research to collect data for the analysis of online behavioral patterns (Antonakaki et al., 2021). However, in 2023, Twitter changed its name to X and removed the free API access. Nevertheless, this modification did not affect the current work since we used Twitter's "old" academic API, and for that reason, we used the name Twitter over our text.

This thesis focuses on the particular case of information technology (IT) startups founded by Portuguese executives or headquartered in Portugal as an illustrative case study. The rationale links with the fact that Portugal is renowned for having created an excellent startup ecosystem by promoting initiatives like the Startup Portugal 200M fund and several business incubators [2]. Portugal is recognized for forging high-quality engineering talents and showing a very high English proficiency index in English [3]. Since 2016, investment in Lisbon-based startups has grown 30% yearly due to several successful startups and unicorns formed in Portugal [4].

1.2. Aims and scope

Often constrained by limited resources and a restricted budget, startups need help reaching customers and securing vital funding (Wang et al., 2016). Given these constraints, social media marketing strategies can be powerful in helping startups reach their goals (Rizvanović et al., 2023). Social media platforms allow startups to generate captivating content and build a community of clients and providers. The central aim of this dissertation is to extract knowledge that serves to improve the startups' strategies to attain their goals. For that, the present dissertation has the following two principal research objectives:

(1) Explore data science instruments and methodologies to enable the extraction of valuable information on how startups use social media.

[1] https://www.thebalancesmb.com/top-reasons-why-your-small-business-should-use-twitter-2948523

[2] https://portugaldigital.gov.pt/wp-content/uploads/2022/02/Portugal_the_best_place_to_startup.pdf

[3] https://www.ef.com/wwen/epi/

[4] https://www.eu-startups.com/2022/07/10-super-promising-lisbon-based-startups-to-watch-in-2022-and-beyond/

(2) Discover the potential of the knowledge extracted from historical social media data to provide digital marketing guidelines able to improve a startup's media strategy.

As such, we have performed a social media analysis of Portuguese IT startups' content and community creation on Twitter as a case study. Recognizing the dynamic nature of startups, which can vary depending on their current stage of development (Skala, 2019), it stands to reason that their social media marketing strategies may differ accordingly. Thus, we investigated the possibility of establishing a connection between the outcomes of the social media analysis and each phase of the startup's life cycle. This link between the cycle positioning phase and the startup social media posts assists in understanding how their social media efforts evolve alongside its growth, maturation, and the resulting shifts in goals to be next attained.

In order to combine the social media analysis results with the startups' life cycle, we propose a model for the evolution of these small businesses. We conducted a systematic literature review, and the results supported the design of an integrated and comprehensive startup life cycle model presented in Chapter 2. Furthermore, from the social media historical data analysis, using the network structure and textual content of the posts (Chapters 3 and 4) allows for a data-driven tool that enables the improvement of a social media strategy by providing guidelines according to the startup's current phase.

1.3. Contributions

The primary outcome of this dissertation is showing that through data science, it is possible to extract coherent social marketing behaviors of companies, in this case, startups. Moreover, we found that not only do behaviors change according to the startup's life cycle phase, but also that their life cycle presents unique characteristics.

From a theoretical level, this dissertation provides the following contributions:

- A systematic literature review on the life cycle of IT startups.
- A novel life cycle model for startups that consolidates the review's results.
- A new methodological process based on social media analytics (SMA) that uses text mining to extract topics of the Twitter posted contents - the tweets - of IT Startups.
- A methodological process for social media community analysis based on Network Analysis and Visualization (NAV) that can be used on social media platforms like Twitter.

The managerial contributions are the following:

- A model that characterizes the life cycle of IT Startups, named FPEMv2 - Funding and Product Evolution Model, that can be applied to other types of analysis regarding elements in the startups' ecosystem.
- The characterization of the life cycle phases of the FPEMv2 based on historical data from a subset of IT startups based on a topic model over the dataset of

Twitter contents. Moreover, according to the company's current phase, this characterization indicates targets that startups should aim at when posting on Twitter.

- The results of the methodology for social media community analysis allowed the profiling of the users present in the overlapping communities in Twitter that emerge from the startups, that is, who are the *followers* and who are the different IT startups *following* in common. Examining the overlap and the type of profiles enabled us to perceive the communities' common points. Startups can expand their networks by joining the same communities and gaining a competitive advantage in their respective markets.

Lastly, during the development of this dissertation, partial results and findings were published in peer-reviewed journals and disseminated at Conference poster sessions.

The poster sessions are presented in Table 1.1.

TABLE 1.1. Poster sessions information

Year	Authors	Title	Poster Session	Appendix
2022	Peixoto, Ana Rita, Almeida, Ana de, António, Nuno	*IT Startups' Twitter content change over time according to the company life cycle*	Ciência 2022	A
2022	Peixoto, Ana Rita, Almeida, Ana de, António, Nuno	*Startups' Twitter activity analysis: the case of Portuguese IT Startups*	RecPad 2022: Portuguese Conference on Pattern Recognition	B

Two papers have been published in peer-reviewed journals directly related to the main themes in each of the papers' contents:

(1) Peixoto, A., de Almeida, A., Antonio, N., Batista, F., & Ribeiro, R. (2023). Diachronic profile of startup companies through social media. Social Network Analysis and Mining, 13, 52. http://dx.doi.org/10.1007/s13278-023-01055-2 (Peixoto et al., 2023a) (Appendix C)

(2) Peixoto, A. R., de Almeida, A., António, N., Batista, F., Ribeiro, R., & Cardoso, E. Unlocking the power of Twitter communities for startups. Applied Network Science 8, 66 (2023). https://doi.org/10.1007/s41109-023-00593-0 (Peixoto et al., 2023b) (Appendix E)

Paper (1) contributions are embedded in Chapters 2 and 3. This publication briefly describes the related work on startup's life cycle and depicts the first attempt at defining the FPEM framework. Additionally, it describes approach (A) in Chapter 3, where we employed topic modeling on a dataset of tweets from eight Portuguese-related IT startups, finding that the main topics in the tweets vary according to the current phase of the startup in its life cycle.

Most of Chapter 4 results are mainly published in (2), except for the work described in Subsection 4.4.4.

1.4. Thesis structure

The present dissertation is organized into five chapters. The first chapter aims to contextualize the IT startup ecosystem and motivates the need to explore social media content to help in the definition of a model to describe the startup business life cycle from its inception to its prime.

Chapter 2 presents a systematic literature review of the currently existing proposals for characterizing an IT startup's life cycle, as well as what are the inherent factors in this ecosystem. Moreover, based on the survey results, it ends by defining a framework for what is the startup's life cycle, named FPEMv2.

Chapter 3 explores the related work on social media textual content analysis and the relationship between startups and social media platforms. It proposes a methodology based on topic modeling techniques to extract topics from tweets. To illustrate this methodology, two datasets have been used: a smaller one relating to eight startups that acted as proof of concept - Approach (A) - and a larger dataset comprising tweets from 38 startups - Approach (B) - which is a more robust extension of the results found in (A). In fact, the results enabled a richer characterization of the life cycle phases described in FPEMv2.

After analyzing social media content data, we focus on exploring the enterprises' social network data. Chapter 4 examines the social communities created around the eight startups' dataset by the action of a startup following other users and the action of being followed. The goal was to gather information that might display patterns in Twitter of the following actions by the startups and provide social media strategies.

Chapter 5 summarizes the results, contributions, limitations, and future work on social media activity analysis of IT startups.

Notably, the present work does not follow a traditional structure for a Ph.D. dissertation. Since each chapter studied different data types, we had to apply distinct methodologies that required specific related work and different result analyses. We divided the work according to the type of analysis performed and combined all the results in the last chapter.

CHAPTER 2

A startup's life cycle

2.1. Introduction

In the field of entrepreneurship research, authors use the company growth and stages to understand the evolution of business needs in their ecosystem. Several frameworks attempt to model a startup's life cycle evolution. Some authors focus on product development (Rafiq and Wang, 2020), while others may include the global ecosystem of the startups (Freisinger et al., 2021), and yet others focus on funding matters (Paschen, 2017). Some authors recognize a life cycle comprising three phases, others four, and others propose five. Moreover, this topic is mainly disseminated by non-peer-reviewed literature, indicating the need for more systematic and empirical literature (Tripathi et al., 2019). Therefore, this chapter provides a systematic literature review (SLR) focusing on the life cycle of companies, specifically for IT startups. Since there are few scientific studies regarding the global life cycle of startups (Tripathi et al., 2019), we additionally search for the factors that influence the companies' development and growth. These factors enable the characterization of development phases affecting the scaling and, consequently, the life cycle flow. Based on the SLR, we created a model for the IT startup life cycle, termed FPEMv2 - Funding and product evolution model (version 2). This model comprises and integrates all the knowledge and conclusions from the reviewed literature. Furthermore, it includes as a novelty the fact that it uses the funding rounds that startups receive as the threshold for the phase transition and consequent company's evolution. This model is termed version 2 since it accompanies our first attempt to create a framework for the startup life cycle, the FPEM (Peixoto et al., 2023a).

The present chapter includes the systematic literature review protocol in section 2.2 and the results in section 2.3. Lastly, in section 2.4, we present the proposed IT startup life cycle model based on the SLR results.

2.2. IT startups' life cycle: A systematic literature review

We applied an SLR to explore the literature on life cycle models of IT startups to expand the knowledge about the life cycle models for IT startups defined in the literature. Consequently, we explore the factors that influence the development of these small companies. This SLR followed the guidelines proposed by Kitchenham and Charters (2007), whose methodology is illustrated in Figure 2.1.

We started by identifying the need for deploying a literature review. As previously mentioned, this review's primary goal is to investigate what are, if any, the actual models and frameworks for the life cycle of startups found in the literature. To our knowledge,

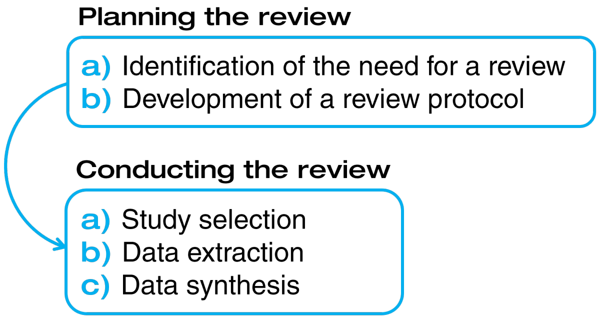


FIGURE 2.1. SLR pipeline based on the guidelines of Kitchenham and Charters (2007).

this topic has been mainly disseminated by non-peer-reviewed literature, indicating the need for a more systematic approach. To this end, we developed a review protocol, where we defined the review's research questions, the search string, the search sources, and the criteria for inclusion and exclusion of studies.

2.2.1. Review protocol

In this protocol, we defined the research questions for the systematic literature review, which led us to the consequent search string. To achieve that, we framed the research questions using the PICOC (Population, Intervention, Comparison, Outcome, and Context) criteria in Table 2.1. The population consists of IT startups and the intervention of the factors that affect the life cycle of the startup, as well as the models that define the cycle itself. The outcome is the result of the intervention criteria. We did not define comparison criteria because it does not apply to our scenario, and we did not select a context because, in our case, it is the same as the population.

TABLE 2.1. PICOC criteria to frame research questions.

Population	IT startups
Intervention	Life cycle, models, factors for development, growth
Outcome	Impact of factors of development and growth, Life cycle models

Using the PICOC criteria, this review aims to answer the following research questions:

RQ2.1: What factors influence the IT startup's development and growth along its life?

RQ2.2: What models aiming to characterize the life cycle are defined for IT startups?

To answer the questions, we selected the following search sources (electronic databases): *Science Direct*, *IEEE Xplore Digital Library*, *ACM Digital Library*, and *Springer Link*. Additionally, the search string used was: ("IT" OR "tech" OR "software" OR "technology") AND start*up AND life*cycle AND model. Table 2.2 displays the inclusion criteria of literature studies to perform the SLR, and Table 2.3 shows the exclusion criteria.

2.2.2. Conducting the review

After extracting the documents, using the search string in the four specified databases, we collected 1 550 results, excluding duplicates. The distribution of the results over the

TABLE 2.2. Studies inclusion criteria.

Inclusion criteria	Reasons for inclusion
Research focus	Studies that clearly identify the factors that influence the startups' development; Studies that define a model for the life cycle of a startup.
English language	Only English studies are considered.
Publication type	Research articles, book chapters, and conference papers.
Years	2010 to 2023.
Discipline	From journals and conferences in computer science or/and business fields.

TABLE 2.3. Studies exclusion criteria.

Exclusion criteria	Reasons for exclusion
Research focus	Studies that do not answer the research questions.

search sources is reflected in Figure 2.2. We found 920 studies from *Springer Link*, 592 from *Science Direct*, 14 from *IEEE Xplore Digital Library*, and 24 from *ACM Digital Library*. Then, we applied the inclusion and exclusion criteria, significantly reducing the number of studies to a final count of 97 results counting: one book chapter, 13 conference papers, and 83 journal articles (Table 2.4).

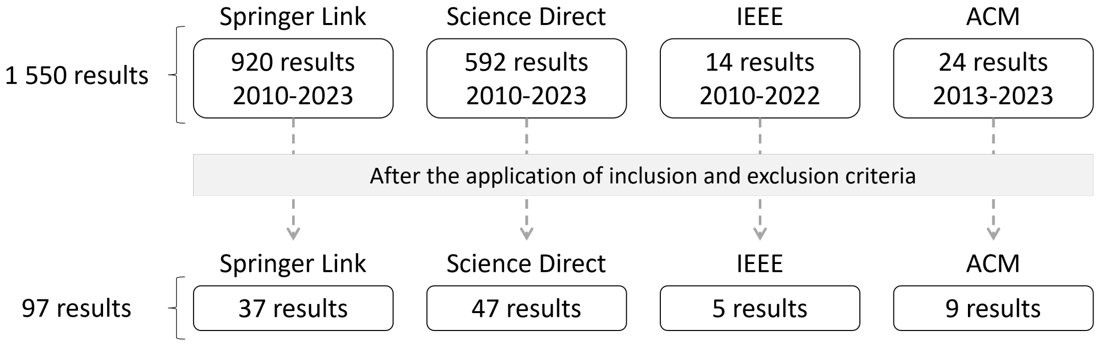


FIGURE 2.2. Databases results before and after the application of inclusion and exclusion criteria

TABLE 2.4. Publication types of the selected studies

Publication type	N	%(N=97)
Book Chapter	1	1%
Conference Paper	13	13%
Journal Article	83	86%

Figure 2.3 displays the publishing years of the selected publications and a cumulative percentage that allows us to analyze the evolution over the years. The graphic shows that the number of publications about the startup's life cycle has increased over the years and more prominently since 2019. From 2012 to 2023, at least one journal article was published yearly. In 2012, 2018, 2021, and 2023, no conference paper on the startup's life cycle subject was published. Additionally, the book chapter found is from 2018. The year

with more publications is 2022, although the data extraction was July 2023, and more documents can be published in that year.

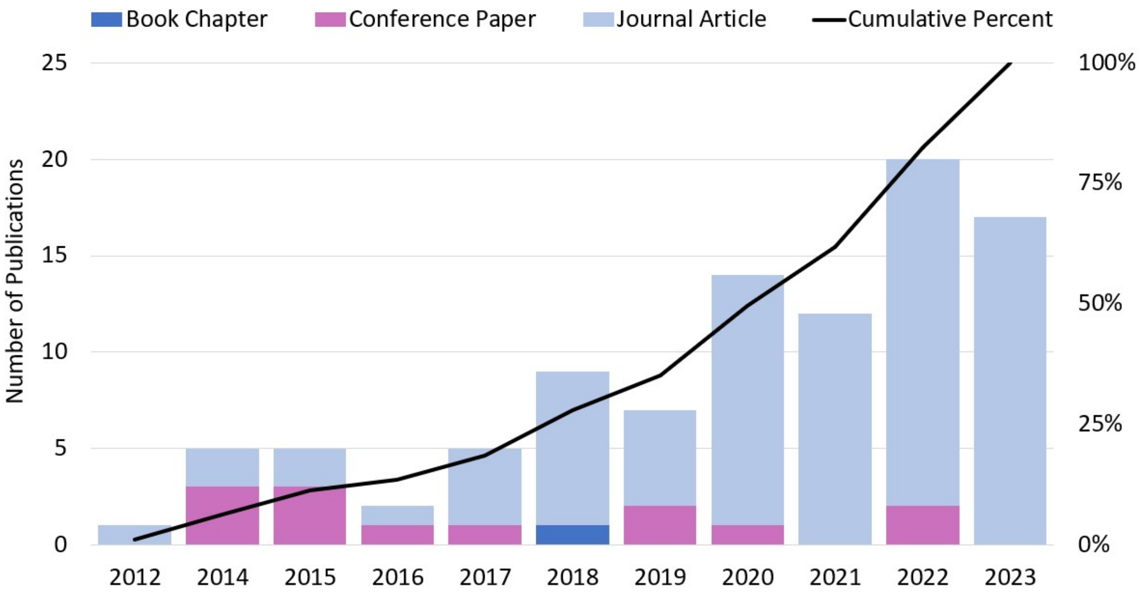


FIGURE 2.3. Publication types over the years

We also reviewed article keywords, presenting a bibliometric network with the selected publications using VOSviewer 1.6.19 (Van Eck and Waltman, 2010) as seen in Figure 2.4.

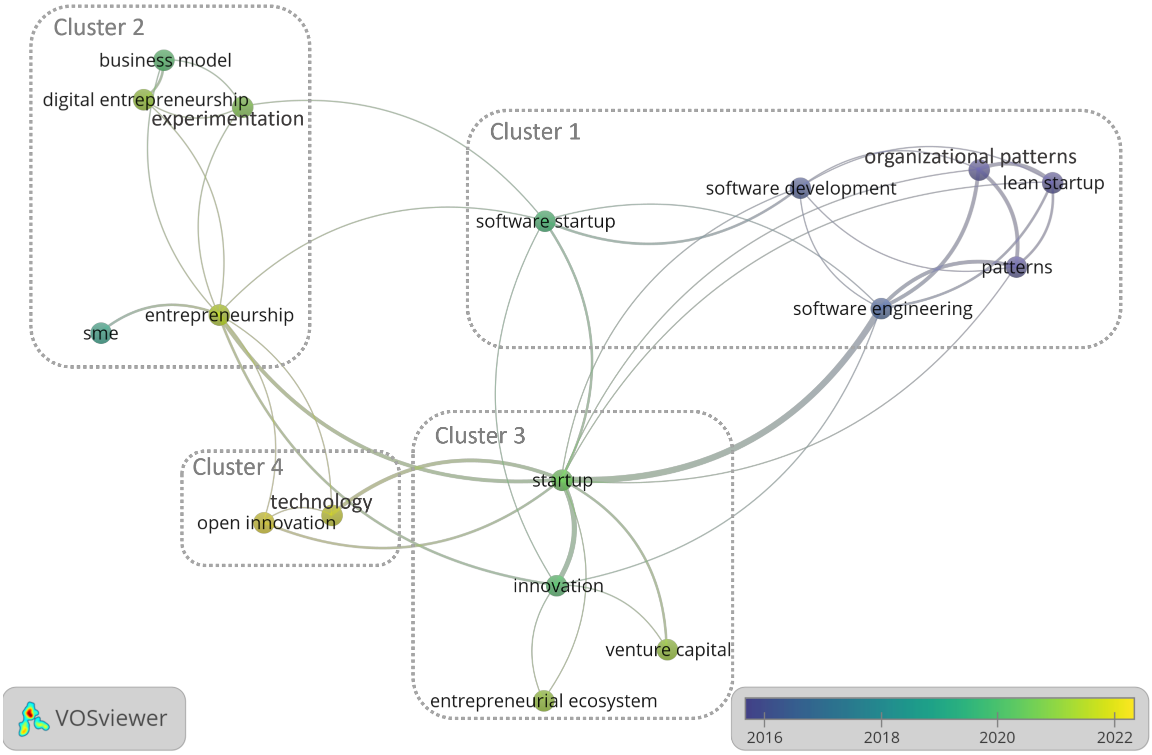


FIGURE 2.4. Bibliometric network of the co-occurrence of keywords (VOSviewer (Van Eck and Waltman, 2010))

The bibliometric network uses the keywords presented in the publications after preprocessing. We manually changed the keywords to lowercase, altered them to singular, and unified the various ways of spelling a term into a unique one. For example, the word "startup" occurred in various formats like "start-up," "startup," and "start up." The created network in Figure 2.4 contains only keywords that appear at least three times and comprise 17 out of 341 keywords. It illustrates the clusters VOSviewer found by delimiting them with dashed boxes, and represents the years of the keywords' publication using color. The links have different sizes according to the number of co-occurrences of the same two keywords in a publication, with larger sizes meaning more occurrences.

As can be observed, Cluster 1 presents the older keywords regarding software development and the consequent organization behaviors. Additionally, it presents the keyword "lean startup" due to the fact that the "lean" approach in a startup is linked to the 'agile' software development methodology. Concerning Cluster 2 it shows keywords used in publications from 2019/2020. The main topic found in this cluster is entrepreneurship and business models. Cluster 3, like Cluster 2, shows keywords found in more recent publications. It shows the word "startup" and presents terms related to the vital elements of the startup ecosystem. Lastly, Cluster 4 displays keywords in recent publications concerning innovation and technology.

In summary, older publications concerning IT startups and their life cycle focused on software development that would provide the product/service of the companies. However, more recent publications target innovation and technology.

2.3. Systematic literature review results

Reading the related literature, we were able to come up with the answers to the previously proposed RQ2.1 and RQ2.2 research questions. To answer the first question regarding the factors influencing IT startups' development and growth, we can describe their ecosystem. Understanding the startups' ecosystem elements is essential to analyzing the factors that influence the growth of startups. We concluded that this ecosystem presents nine fundamental elements: entrepreneurs, startups, established companies, funding bodies, accelerators, society and culture, education, government, and location. Next, we present a subsection concerning each of those elements.

Also, according to the reviewed literature, one vital element is funding. Startups need to obtain financing to scale the business and gain new resources. Therefore, we present a subsection regarding specifically the specification of this importance for IT startups.

Once the ecosystem is defined, we go on and describe the factors affecting the startup life. These may vary with the startup's actual stage in the life cycle. This point leads to the second question: What are the life cycle models already defined for IT startups? We then described the proposed frameworks found in the literature.

2.3.1. The ecosystem elements of IT startups

Tripathi et al. (2019) concluded that the current knowledge on the startup ecosystem is mainly shared by non-peer-reviewed literature, indicating the need for more systematic and empirical literature. However, Cantner et al. (2021) proposed a dynamic life cycle model for the entrepreneurial ecosystem, which captures the oscillations between the birth and growth of companies and the decline and re-emergence of some of those. Furthermore, Cukier and Kon (2018) presented a maturity model for software startup ecosystems that includes the following elements: entrepreneurs, startups, established companies, funding bodies, accelerators, society and culture, education, and government. This framework shows how these elements interact and influence each other to create a healthy environment for developing innovative companies. In fact, it is based on those elements that this review will characterize the IT startup ecosystem. Moreover, based on the literature found, we decided to add a new ecosystem element: location. Some works, like the one by Ross et al. (2021), use features that indirectly comprise the startups' ecosystem elements. The authors developed a model to predict startup outcomes: whether they will exit successfully through an IPO or acquisition, fail, or remain private. The model used a set of 18 features, some regarding social media use, funding, entrepreneurs, and the employees' experience.

2.3.1.1. *Entrepreneurs*

Entrepreneurs are the individuals who create and manage the startups. The life cycle of an entrepreneur consists of five periods: preparation, embarkation, exploration, expansion, and transformation (Zaheer et al., 2022). Over their life cycle, the entrepreneurs' resources increase as they gain knowledge and the company scales. Sadeghiani et al. (2022) found that the competitive intelligence of the entrepreneurs led them to pivot their early-stage business models, but the quality of the resulting models could have been better. Entrepreneurs usually face dilemmas regarding growth and profitability, which can be solved by applying a versioning or freemium strategy (Bhargava, 2014). This strategy involves offering a free or low-price version to drive mass adoption and a premium higher-price version to generate revenues. The occupational backgrounds of entrepreneurs play a significant role in venture performance (Roche et al., 2020). The study found that academic startups are less likely to achieve a liquidity event than non-academic startups. Liquidity events enable venture investors to convert their ownership stakes in a startup into cash or liquid securities, which can occur through methods such as going public, acquiring, or selling stakes on a secondary market. However, the academic ones produce as many patents and receive as much funding as non-academic startups.

2.3.1.2. *Startups*

Startups are small companies that aim to grow and scale. There are over 140,000 startups in Europe, and around a third of these have acquired at least one round of funding (Kemell et al., 2023). Most software startups fail, and up to 98% of all new product ideas fail. However, Kemell et al. (2023) highlighted that software startups drive the global economy. Nguyen-Duc et al. (2021) categorized software startups into two distinct types, effectuation-driven and mixed-logics-driven, which influence the applicability of traditional software engineering approaches. Effectuation is an approach to decision-making and action-taking often associated with entrepreneurial ventures. It starts with the resources at hand and works towards achievable goals based on those resources rather than trying to predict the future and plan accordingly. The mixed-logics-driven combines the two approaches of using the available resources and work based on future predictions. The authors defend that effectuation is the most proper framework for enabling appropriate software engineering practices within software startups. Regarding hardware startups, Berg et al. (2020) found that they follow a quality-driven approach to developing core components, where frequent user testing is a measure for early debt management. Chammassian and Sabatier (2020) stated that software startups build business models that are technology-driven, market-driven, and exit-driven.

2.3.1.3. *Established companies*

Established companies are big corporations that have activities that nurture the ecosystem, such as event organizations, local community ambassadors and mentors, acceleration programs, or local investments in startups. Some companies cooperate with startups facing some challenges. Schuh and Studerus (2022) proposed a methodology for selecting and designing such cooperation based on the explication of target systems and a comparison of objectives and corresponding requirements. On the other hand, large companies can use lean internal startups to innovate like startups (Edison et al., 2018). Lean startups prioritize rapid experimentation, customer feedback, and iterative development to adapt and succeed in uncertain markets quickly. The lean startup approach, which startups are successfully using, can also be successfully applied to larger manufacturing companies. This technique can help companies become more dynamic and flexible, react more effectively to external influences, and integrate customer needs into the product design (Jesemann et al., 2020). Regarding the lean startup approach, Ghezzi (2020) concluded that business models serve as practical guides for entrepreneurs, providing clear and specific rules that make the abstract guidelines of this approach more understandable and actionable. Hokkanen (2015) described four stages to create internal startups in a larger company: the 20% rule, incubator phase, internal startups, and exit decision. The 20% rule regards the percentage of work that should be employed weekly on new projects to enable innovation and idea creation. Next, the incubator phase will validate the idea and solution originating the internal startup. Then, the internal startup, where the company allocates resources to concretize the concept and find a new product-market fit. Lastly is the exit decision, where the parent company ends the internal startup. In that stage, they can sell the small business or close it.

2.3.1.4. *Funding bodies and Accelerators*

Funding bodies are venture capital firms, angel investors, or crowdfunding platforms.

These organizations fund the startups and are crucial for their survival and growth. Accelerators are programs that provide mentorship, resources, and funding to startups in exchange for equity.

2.3.1.5. *Society and culture*

Society and culture are the cultural values and support for entrepreneurship in the ecosystem's community. Regarding the sustainable development goals, IT startups demonstrate a statistically significant positive attitude toward pursuing economically oriented ones and a negative mindset toward socially or environmentally oriented (Lammers et al., 2022).

2.3.1.6. *Education*

The education element represents the quality of education and training programs available to entrepreneurs and startup employees. According to Wright et al. (2017), universities offer a range of support activities for student entrepreneurship, from promoting early-stage venture ideas to progressing to the next phase, which involves utilizing an incubator or accelerator for further development. Albats et al. (2018) identified the common and context-specific key performance indicators (KPIs) of university-industry collaborative projects at a micro-level. The authors determined a set of common micro-level KPIs: The number of resources allocated by partners to collaboration, the efficiency of collaboration management and clearly defined roles, the number of company innovations resulting from cooperation with a university, establishment of strategic partnerships, and renewal of business revenue structure. The authors also identified contextual micro-level KPIs, such as the number of young researchers involved, the fit between collaboration and organizational strategy, the number of joint publications, and enterprise image improvements.

2.3.1.7. *Government*

The government element concerns the initiatives and policies implemented to stimulate or accelerate the ecosystem development and the economic environment that affects the business. Regarding the economic environment, according to Cavallari et al. (2021), companies born during recessions start on a larger scale and remain larger compared to businesses created during expansions. The study also determines that when employment protection becomes stricter, it widens the employment gap in favor of startups that emerge during recessions. Concerning policy instruments, Cohen and Ernesto Amorós (2014) suggested that specific demand-side policy instruments may uniquely support technology diffusion at different stages of the technology life cycle. In the initiation stage of the technology, policies should aim at the procurement of innovation. In the growth stage of the technology, voluntary standards and incentives should be applied. Lastly, in the maturity stage, the government must regulate the technology.

2.3.1.8. *Location*

The location element concerns the place where the startups act. Kubara (2023) found that IT startups tend to co-locate and create dense clusters of business activity in urban spaces. Companies are attracted to more than just the metropolitan area but to its dense business clusters. Adler et al. (2019) claimed that startups occur at two spatial scales:

macro-geographic and micro-geographic levels. The macro level refers to the concentration of IT startup entrepreneurship in a relatively small number of global cities or metropolitan areas where a particular region specializes in a specific industry or technology. The micro level refers to the concentration of IT startups in separate sections within the leading cities and metropolitan areas, where a diverse mix of industries and technologies in a particular region or city leads to innovation and entrepreneurship.

2.3.2. The funding of startups

Startups are key players in pushing economic growth by creating jobs, increasing total factor productivity, and allocating resources efficiently. However, they face challenges securing funding and resources in Europe, which are crucial for their success (Schuh and Hamm, 2022). Nevertheless, there have been important governmental public policy initiatives to promote startup businesses. Islam et al. (2018) found that when startups win prestigious government research grants, they are 12% more likely to acquire subsequent venture capital funding. Romme et al. (2023) proposed the Depp Tech Venture approach for the limited capacity of European industry and knowledge institutions to transform scientific and technological breakthroughs into successful ventures and companies that address meaningful challenges in sustainability development goals. Joshi et al. (2022) suggested utilizing corporate social responsibility funds for impactful investments and establishing a dedicated fund to support businesses.

One way to earn funding is through venture capitalists. Gloor et al. (2020) uncovered that startups benefit from working with venture capitalists because of the opportunity to access additional funding, although their presence only sometimes translates into sales growth and operational efficiency. Alternatively, there is crowdfunding, which has experienced exponential growth in recent years (Skare et al., 2023). Crowdfunding offers financial support while providing non-financial assets, known as crowd capital. It promotes user-driven innovation and facilitates a deeper understanding of customer preferences (Paschen, 2017).

It is noteworthy that credit availability significantly impacts firm life cycles, and more companies are being created and destroyed in capital-abundant regions (Tang and Basco, 2023). Regarding the startups' life cycle, the startups have different monetary and non-monetary needs depending on their stage, influencing their selection of investors (Bauer et al., 2023). Additionally, the size and age of the companies impact the type of financing preferred, with older and larger ones choosing debt financing over equity financing (Hyun and Lee, 2022). Consequently, some investment banks have adopted a portfolio financing strategy, investing more in equity when the firm is young and more in debt as it ages.

2.3.3. Factors that influence the startups' business development

Some factors increase the risk of failure, and others increase the probability of success. The literature also shows that while some factors have a positive influence on startups development at any stage, others mostly impact the initial stage.

According to the literature, startups face a great number of challenges that increase their risk of failure, decreasing their chances of success. One critical factor is the emergence of conflicts between co-founders and/or investors, as highlighted by Bala Subrahmanya (2022). These conflicts can disrupt decision-making processes and interfere with the overall progress of the startup. Other aspects are the startup team's lack of commercial expertise and technological maturity in the chosen field (Clark and Muller, 2012). Szerb and Vörös (2021) explained that it is essential to understand the discrepancy between perceived skills and business growth expectations. Additionally, the lack of resources and difficulties in integrating knowledge also emerge as challenges (Almeida, 2021). These limitations can restrain the development and scaling of the startup. Lastly, neglecting activities related to the creation of documentation regarding software ideas and features can lead to a loss of crucial software knowledge, a concern emphasized by Maria et al. (2017). This disregard can influence the long-term viability of the startup.

Among the factors that increase the probability of success, innovation is one of the most important mechanisms for creating sustainable competitive advantage and resilience in today's dynamic environment (Mirghaderi et al., 2023). Almeida (2021) highlighted the increase in innovation capacity as a benefit. Liotino et al. (2016) emphasized that firms with innovative practices in their organizational structure tend to perform better than those without such practices. Additionally, cooperation and technological orientation play a central role in enhancing startups' innovativeness, and this contribution is significant if levels of entrepreneurial leadership and team motivation are low (Lago et al., 2023). Felicetti et al. (2023) presented a literature review on the synergies between digital innovation and entrepreneurial firms. The authors identified six relevant topics: "startups' collaboration networks," "business-model innovation," "digital platforms," "digital ventures," "the digital entrepreneur's profile," and "digital innovation ecosystems."

Furthermore, companies must display a sensitive ability to adapt to their environment, a critical insight presented by Ehsani and Osiyevskyy (2023). This adaptability allows startups to navigate challenges and seize opportunities effectively. Sales and R&D (research and development) capabilities emerge as central company-related factors (Satyanarayana et al., 2021), as these competencies support growth and innovation. Morales-trujillo and García-mireles (2019) explained that effective communication, customer interaction, teamwork, and trained management are noteworthy in startup success.

Social media ads are valuable for startups (Hervet and Guitart, 2022). Gloor et al. (2020) discovered that startups with venture capitalists on the board who are active on Twitter have attracted additional funding over the years. Leveraging these social media platforms can significantly improve visibility and customer engagement. Also, building partnerships between tech corporations and startups enables a collaborative ecosystem (Nobari and Dehkordi, 2023). As more agents act as part of the ecosystem and not in isolation, startups' obstacles to innovation tend to decrease (Noelia and Rosalia, 2020). Furthermore, promoting industry-institute partnerships, patent commercialization from higher education institutions, and focusing on the graduation and scaling of high-tech startups enable startup growth (Bala Subrahmanya, 2022). Fintech means financial technology and refers to technologies and applications created to improve and automate traditional forms of finance. Werth et al. (2023) found that, in the Fintech sector, critical factors for success include the "cost-benefit dynamic of innovation," "technology adoption," "security, privacy, and transparency," "user trust," "user-perceived quality," and "industry rivalry."

Lai (2017) found that specific business models that measure the quality of startup management can enhance the business opportunity and management skill level of startups and their survival chances. Furthermore, Ruseva (2015) explained that choosing a robust business model aligned with the specific life cycle stage is crucial, as being at a particular stage can influence the startup's trajectory toward success (Jamali et al., 2015). As shown, innovation is vital for a startup's growth. Daradkeh and Mansoor (2023) demonstrated that exploitative innovation should be applied in the initial stages, while in the growth stage, startups should use exploratory innovation to impact their performance positively. Exploitative innovation involves improving or refining existing products/services or processes. It builds on the organization's existing knowledge and capabilities. In contrast, exploratory innovation involves pursuing entirely new and often disruptive ideas, technologies, or business models.

Literature shows some critical factors for the startups' success to apply at their initial stages. In those, the startup will create a product/service business, and projects like incubators and accelerators play an essential role (Page and Holmström, 2023). Furthermore, some ideas occur in an academic environment. Santoso et al. (2023) highlighted that mentoring students to visualize their ideas as real business models helps create successful startups. Regarding the accelerators, Del Sarto et al. (2022) pointed out the different sources of external knowledge that those provide as a beneficial factor for different innovation outcomes. Additionally, accelerators encourage competitive behavior among startups through their tools, like events and co-working spaces (Moritz et al., 2022). The public financing of those activities at the regional level is positively significant for creating startups (Yusupova and Ryazantseva, 2022).

Startups begin with an idea, and the inherent risks must be considered. Akhavan et al. (2021) stated that identifying the risks before the start of the project allows the owners and investors to make accurate decisions and apply benefit-cost analysis of the alternative projects. Software startups should choose projects compatible with their maturity level and risk acceptance (Parthasarathy, 2022). After selecting the idea and creating the startup, activities to validate the idea and the product are essential to find issues related to a need for more requirements engineering (Kemell et al., 2023). The lack of knowledge has been previously pointed out as a factor that increases the risk of failure. Other helpful aspects are developing adaptive business models (Vatankhah et al., 2023) and employing user experience practices during the product specification, design, and prototyping activities (Zaina et al., 2023).

Initial stages are essential for the company's success and occur before the so-called valley of death, which corresponds to the transition between already having begun operations and starting to produce revenue. Gbadegeshin et al. (2022) concluded that startups should evaluate their operation at least every three months and re-prioritize their partnerships to overcome that valley.

2.3.4. Startup's evolution process

Various factors around the startup system depend on the different phases in their life. As previously described, even some critical factors for their success are intrinsic to the current phase. For example, Fukugawa (2018) found that to achieve better performance, incubators should adopt distinct approaches according to the life cycle phase of startups to which they aim to give support. Furthermore, Hatzijordanou et al. (2019) raised the need for future work to explore the requirements and consequences of conducting competitor analysis in the different phases of the startups. However, in the literature, the division of the life cycle phases lacks uniformity, with some authors identifying three phases, others four, and yet others five.

2.3.4.1. *Three life cycle phases*

The division of the startup life cycle into three phases is the most used in literature. This framework comprises the initial conceptualization of the idea as the first phase, the stabilization of the business as the second, and the third when they have an established in-market company. The first phase can be denominated in the literature by the following names:

- Early (Bauer et al., 2023)
- Emergence (Ganesaraman and Bala Subrahmanya, 2022)
- Nascent startups (Szerb and Vörös, 2021)
- Conception stage (Nicholls-Nixon et al., 2021)
- Creation phase (Marcon et al., 2021; Marcon and Ribeiro, 2021)
- Pre-startup stage (Nguyen-Duc et al., 2021; Paschen, 2017)
- Early stages (González-Cruz et al., 2020)
- Startup phase (Eloranta, 2014)
- Forming (Marko Leppanen, 2014)

This is the phase when the startups do not yet earn any revenue (Bauer et al., 2023), and the founders are actively involved in setting up the business (Szerb and Vörös, 2021). The focus is to elaborate and validate the business idea to demonstrate its feasibility, being a business-oriented phase (Eloranta, 2014). Additionally, it features only a small team, mostly only the founders, with the idea of creating a product/service focused on a specific customer segment. Regarding the funding in this phase, the most common source is bootstrapping, family, friends, and fools (FFF) (Nguyen-Duc et al., 2021).

The second phase can be called in the literature as:

- Mid (Bauer et al., 2023)
- Stability (Ganesaraman and Bala Subrahmanya, 2022)
- Baby business (Szerb and Vörös, 2021)
- Professionalization stage (Nicholls-Nixon et al., 2021)
- Development phase (Marcon et al., 2021; Marcon and Ribeiro, 2021)
- Startup stage (Nguyen-Duc et al., 2021; Paschen, 2017)
- Growth or expansion stages (González-Cruz et al., 2020)
- Stabilization phase (Eloranta, 2014; Marko Leppanen, 2014)

Bauer et al. (2023) argued that not having a middle stage would oversimplify the startups' development process. The authors described this second phase as the essential bridge between the loosely structured stage and the disciplined one. In this second phase, startups earn revenue but are not profitable (Ganesaraman and Bala Subrahmanya, 2022). They can pay salaries and wages for more than three but less than 42 months, and the owners actively participate in the business's management (Szerb and Vörös, 2021). This phase is product-oriented (Eloranta, 2014), where the startup has already developed and experimented with the prototypes (Nguyen-Duc et al., 2021). The startup generates some revenue but not necessarily the break-even point. Further, the founders seek support from startup ecosystem elements to accelerate their business. Concerning the financials, they fund themselves and acquire seed funding.

A third phase can be found under the following names:

- Late (Bauer et al., 2023)
- Growth stage (Ganesaraman and Bala Subrahmanya, 2022; Nicholls-Nixon et al., 2021; Paschen, 2017; Eloranta, 2014; Marko Leppanen, 2014)
- Established business (Szerb and Vörös, 2021)
- Market phase (Marcon et al., 2021; Marcon and Ribeiro, 2021)
- Pos-startup stage (Nguyen-Duc et al., 2021)
- Later stages (González-Cruz et al., 2020)

In this last phase, the startups achieve the product-market match. The typical funding is from external financing, attaining larger fund rounds. (Nguyen-Duc et al., 2021). The startups have established revenue streaming, helping them operate with a profit (Ganesaraman and Bala Subrahmanya, 2022). They can pay salaries and wages for more than 42 months (Szerb and Vörös, 2021).

2.3.4.2. *Four life cycle phases*

Some authors consider that the startups' life cycle are composed of four phases. In this format, the middle phase transforms into two different ones, giving more detail about the prototyping and the stabilization of the product in the market. The first phase only affects the idea conceptualization, and the last concerns the mature business. Di Vaio et al. (2022) called the phases as follows: pre-seed, seed, startup, and expansion. While Berg et al. (2018) named as follows: startup stage, stabilization phase, growth phase, and last stage. Following is the description of each phase in this life cycle model of four phases, based on those authors:

(1) **Pre-seed/ Startup stage**: At this phase, the startup has a small executive team with the knowledge to build the idea conceptualization product.
(2) **Seed/ Stabilization phase:** The entrepreneur develops a business model associated with the idea concept. This phase lasts until the product prototype is stable enough to be designated to a new customer without causing troubles in the development. Further, internal and external stakeholders impact the decision-making process.
(3) **Startup/ Growth phase:** This phase begins with a stable product development process and lasts until they have a well-established product in the market.
(4) **Expansion/ Last stage:** The startup is a mature organization with a robust product and predictable outcome.

2.3.4.3. *Five life cycle phases*

Other authors employ a life cycle framework of five phases in the literature. The main difference between the four phases is dividing the third phase (startup/growth phase) into two distinct phases—one for the stable development process and the other for consolidating the product in the market. In the systematic literature review, it was found that two authors proposed a five-phase life cycle. Rafiq and Wang (2020) proposed a framework based on product development while Freisinger et al. (2021) focuses on the global ecosystem of the startup evolution. Rafiq and Wang (2020) remarked that when a startup moves to the next stage, there is a need to monitor the adaptations and the learning that occurs from that change. Following are the names of each phase by Freisinger et al. (2021) and Rafiq and Wang (2020), respectively, with a brief description of each phase.

(1) **Startup conception and development / Concept in-development:** The idea creation and conceptualization characterize the first phase. The primary focus is to design the product development and secure financial resources.
(2) **Commercialization and market entry / Working prototype:** The startup has a product that meets a market need and can sell a working prototype generating few revenues. The company is not yet firmly established in the market.
(3) **Growth / Functional product with limited users:** The startup has a functional product for a limited number of users. Usually, it is experienced high growth rates in sales. The focus is on how to produce, sell, and distribute.
(4) **Consolidation / Functional product with high growth:** In this fourth phase, the focus is to attain profitability, maintaining the growth momentum.

(5) **Maturity and diversification / Mature product:** In the last phase, the startup has a mature product, and the focus shifts to diversification of the activities and product.

2.4. FPEMv2: a new proposal for a life cycle model of IT Startups

Besides the previous SLR results, we found two other studies regarding the startup's life cycle that we consider relevant: the works by Wang et al. (2016) and by Nguyen-Duc et al. (2015).

According to Wang et al. (2016), the maturity evolution of a startup goes through two stages: the learning and the growing stages. The learning stage consists of selecting a problem to solve and defining and evaluating the solution. The problem represents a real issue or obstacle for a specific target, which is solved by providing a product or service: the solution. The product concept is developed in the growing stage, followed by an implementation start leading to a working prototype. If it results, the startup obtains a functional product that evolves into a mature product. However, Wang et al. (2016) emphasizes that this is not a constant cycle, saying that a startup has to go through "multiple measure-learn loops". The loops mean reevaluating each step as being in the previously referred stages.

Concerning startups whose main product/service is software, Nguyen-Duc et al. (2015) created a conceptual model named the hunter-gatherer that, in fact, consists of two development cycles: the "hunting" cycle consists of the idea, market, and features; the "gathering" cycle features the prototype, quality, and product. The intention is that the two cycles occur at each stage, but the dimension of the cycle differs over the startup's life cycle. In the learning stage, the hunting cycle is more significant, while in the growing stage, the gathering cycle becomes prominent. Nevertheless, the cycles occur at each stage side-by-side: when the company obtains a mature product, the focus changes to quality.

These described works highlight the change of focus that occurs while the startups grow, which is aligned with the previous SLR results. These companies begin, and the focus is learning, that is, hunting for the development of an idea that solves a problem they found. As the startups mature, their preoccupation falls on gathering the knowledge to build the product and funding. Based on this characterization, we propose to divide the startup's life cycle into two main perspectives. One that follows creating a product/service based on an initial idea to solve a real problem: the *maturity evolution*. Another facet concerns the startup funding rounds: the *funding stage*. The funding rounds are financial mechanisms in which startups open or expose their shareholder structure to third parties, usually to business angels or venture capital firms, to secure investment to allow the startup to grow.

To illustrate a startup's financing milestones integrated with the startup's evolution in terms of product/service, we propose a life cycle model that is based on the previously introduced two dimensions: the funding rounds and the product maturity: the Funding and Product Evolution Model - FPEMv2 - depicted in Figure 2.5. The FPEMv2 illustrates the maturation process of a startup's life regarding time and finance in a typical success scenario. The naming of the funding rounds' is based on the Crunchbase Glossary [1].

In our first exploratory work Peixoto et al. (2023a), it was defined the first version of this model, the FPEM, depicting only four maturation phases. Since the previous work emerged from a social data analysis and only considered 8 startups during a smaller time window, the size of the dataset impacted the number of emergent phases, mainly because it only considered social media data posted after a seed funding round. As the dataset expanded, posts from startups that still had yet to receive a seed round appeared, and with it, the need to understand what type of funding they were receiving. As a result, it has been found that a prior and very initial phase must exist. This claim has been clearly supported by the SLR described. As a result, the FPEMv2 is created with more support in the literature to describe each phase.

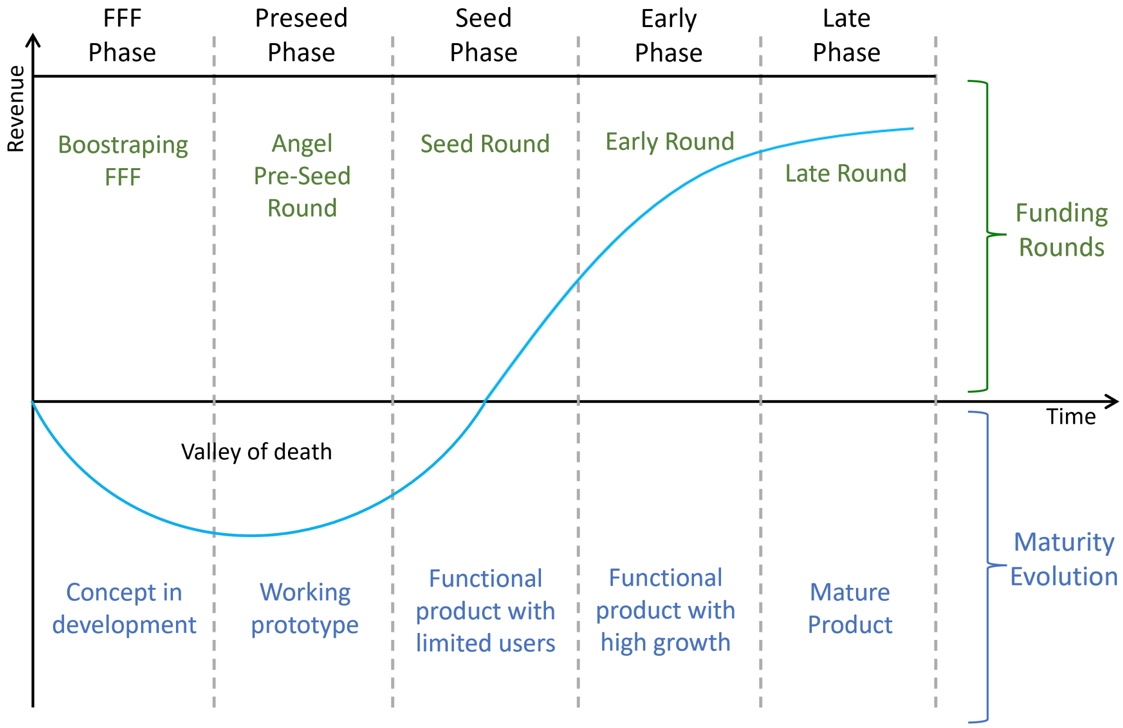


Figure 2.5. Funding and product evolution model - FPEMv2 - a five-phase startup life cycle model.

As it can be observed, the proposed model encompasses five phases. Every type of funding round can happen more than once throughout a company's life. However, a phase transition only occurs with a new funding type, implying a scale-up for the company. This measurement of phase transition is a novelty since, in the SLR results, few authors proposed criteria for the phase transition. The maturity evolution phases describe the startup's product stages based on the literature review results description.

[1] https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types

Notice that, for each phase, the association of concepts between maturity dimensions is relatively straightforward.

In the FFF phase, there is only the conceptualization of a potential solution for a concrete problem. The name of this phase is based on the initial funding of the startup. The funding is usually provided by bootstrapping or family, friends, or fools (FFF) (Paschen, 2017; Nguyen-Duc et al., 2021). In the preseed phase, funding types are usually minimal (typically below 150K USD) and are known as angel or preseed round. In this phase, the startup already has a working prototype or, at least, a proof-of-concept to gain that type of funding. In the seed phase, the startups had developed a functional product with a limited number of users, sustaining the seed funding round, which can scale to 2M USD. In the early phase, a functional product already exists that has been experiencing high-rate user growth and is ready to scale the market. In the late phase, a mature product is already established, and the correspondent funding, also called Series C round, the values may start at 10M USD with no upper limit.

The line in Figure 2.5 represents the startup revenue evolution over time. It comprises the valley of death, and it is a s-curve. The valley of death corresponds to the transition between having no earnings and starting to produce profit (Gbadegeshin et al., 2022). Usually, the company starts to have profits once it has a functional product with consumers, corresponding to the seed phase. The evolution line is an adaptation of the s-curve, like the sigmoid function, because this function can metaphorically represent growth and change (Hipkins and Cowie, 2016). In the beginning, the growth is stable until it starts to have a high growth. The stable initial corresponds to the low growth rate influenced by the difficulties of the beginning of a new startup, where the elements to work on are in high quantity. Those elements are to gain financials, develop a working prototype, conflicts between founders or investors, and the lack of expertise in commercial and technology in the chosen field, among other factors mentioned in the SLR results. Once the startup crosses the valley of death, the company experiences a high rate of growth accompanied by more users and funding. This increase corresponds to the middle of the s-curve. At the end of the s-curve, the function starts to be constant, and the startup is well-established in the market with a mature product.

2.5. Summary

Since there are few studies concerning the complete startup ecosystem, this chapter aimed to provide an IT startup life cycle model based on the existing literature. We conducted an SLR to find the factors influencing the startup's growth and which models are defined in literature for that evolution progress.

The SLR results indicated that the startup ecosystem comprises nine fundamental elements: entrepreneurs, startups, established companies, funding bodies, accelerators, society and culture, education, government, and location. Understanding these elements is essential for any life cycle startup analysis since those significantly impact the company's development and growth. Entrepreneurs are responsible for the startup's beginning, and as the company evolves, the entrepreneurship matures. These are persons whose background is crucial since their evolution will have different starting points depending on their previous occupation and consequent knowledge and resources (Roche et al., 2020). Their education can serve as a gateway for the startup birth since some universities offer a range of support activities for student entrepreneurship (Wright et al., 2017), like the creation of accelerator programs that provide mentorship and resources for startups. For example, the Portuguese university ISCTE promotes the incubator Audax [2]. On the other hand, some established companies can be nests for the startups or acquire external ones (Hokkanen, 2015).

Conflicts between the entrepreneurs, who are the founders, and the investors could also impact the startup's progress (Bala Subrahmanya, 2022) since credit availability significantly affects company life (Tang and Basco, 2023). The element funding bodies prove to be vital for these small companies, and they are the financial providers, external to the startups, that fund them throughout their growth. Some authors provided frameworks for the evolution process to model the life cycle of startups. However, there has yet to be a consensus on the actual model and the number of phases in it. Some described this process as using three, others four, and others five phases. Another limitation in the related literature concerns the transition between the phases representing the startups' scaling.

By combining the SLR results, we provided a new model for an IT startup life cycle that comprises five phases and uses the typology of funding as a transitional element, integrating it with the maturity evolution of the company's product or service. Acquiring financing is linked to the startup's definition, while the maturity evolution regards product/service development - from the concept creation to a mature and consolidated product, well established in the target market. We represented the life cycle revenue evolution using an s-curve based on the sigmoid function that metaphorically represents growth and change (Hipkins and Cowie, 2016).

Given the description of each phase, it is possible to recognize that the focus of the startup changes alongside its evolution. The evolution described opens the door to analyzing how startups use tools, like social media platforms, to achieve their shifting goals.

[2] https://audax.iscte-iul.pt/

CHAPTER 3

Characterizing life cycle phases through Twitter content

3.1. Introduction

Social media is integral to the information ecosystem, providing extensive data that can be used for informed decision-making and strategic business insights (Saura et al., 2021). Previous studies explored digital platforms where startups were present, demonstrating relevant information about companies' activity could be extracted from these platforms. Saura et al. (2019) examined tweets using "#startup" to detect indicators for success and discovered the sentiment of the most common topics in tweets about startups. A broad study by Ruggieri et al. (2018) focused on finding patterns in successful startups based on their digital platforms' presence. The authors found that newly born startups use digital platforms because it is cost-effective. Alotaibi et al. (2020) designed a framework to evaluate Twitter activity using an Arabic startup as a case study. Recent systematic literature reviews have highlighted the need for more profound research in social media intelligence (Olanrewaju et al., 2020; Smolak Lozano and Almansa-Martínez, 2021). Olanrewaju et al. (2020) proposed a set of future work themes, among which we can find the need to consider the evolution state of the company and, in consequence, its life cycle stages. We aim to fill that gap and understand how a startup's social media content changes through the different phases of its life. In other words, understand the diachronic profile that emerges from the startup's historical social media data and analyze whether it reflects its scaling evolution.

Two distinct approaches were implemented to investigate if Twitter activity evolves across the different life cycle phases. Initially, a small-scale study involving eight startups was conducted as a proof of concept (approach A). This approach was conducted and published in the form of a peer-reviewed journal article (Peixoto et al., 2023a). This study aimed to assess whether the content varied across life cycle phases. After the first study confirmed that content evolves over time, we decided to take a larger-scale approach involving 38 startups (approach B). In this approach, the tweets were grouped based on the respective phase, and the results enabled us to describe and characterize the various phases of the startup's life cycle, leading to the following research questions:

RQ3.1: Does the startups' Twitter content change according to their life cycle phases? (Approach A)

RQ3.2: If the content changes, can we use the topics created from it to characterize the startups' life cycle phases? (Approach B)

The present chapter includes in section 3.2 a related work review describing the present literature on social media analysis, focusing on the textual content and its visualization. Section 3.3 explains the methodology applied to answer the research questions. The following, section 3.4, results and discussion, presents and analyzes the outcomes. Lastly, section 3.5 outlines this chapter's main conclusions.

3.2. Related Work

Saravanakumar and Suganthalakshmi (2012) denote social media marketing (SMM) as a marketing tactic that efficiently promotes brands through social media platforms. However, how can we analyze social media content and extract relevant information demonstrating this value? This section aims to answer this question by explaining the social media analysis process and its methods and results. We focus on the social media analysis regarding the textual content by applying topic modeling techniques. Those methods bring up topics that may illustrate main themes by analyzing the textual content of documents, in this case, of social media texts, like tweets. The topics, in turn, must also be analyzed to infer relevant knowledge. Additionally, we state the relationship between the startups and their social media activities found in the literature.

3.2.1. The social media textual content analysis

Social media data has become a fundamental part of the data ecosystem and is a strategic source of knowledge for decision-making (Kapoor et al., 2018). Some paradigmatic examples can be found in extant literature. Campos-Domínguez (2017) analyzed research works in political communication on Twitter, and Godoy-Martin (2022) investigated the use of social media by communications agencies. Nevertheless, to infer relevant information from data, one must prepare and process it (Dutot and Mosconi, 2016). Social media intelligence (SMI) collects, treats, and analyzes relevant data to provide data-driven support for strategic decisions. SMI works in a data-driven cycle because social media constantly changes, with new users creating new content and generating more data for analysis. The main focus of SMI applications is product/service review analysis (Kapoor et al., 2018). The knowledge obtained by SMI is meant to describe the present state of social media. This means that if the objective is to predict outcomes and suggest future directions, a social media analytics (SMA) approach is deemed necessary (Choi et al., 2020). Comparatively, SMA and SMI present similar phases (Zeng et al., 2010), but SMA methodology and results aim for the future, while SMI concerns the present.

Social media content consists mainly of textual data, and its analysis aims to find relationships among data in text documents and extract patterns to understand the themes being addressed (Jelodar et al., 2017). This goal can be achieved by analyzing the text's sentiment/polarity or identifying the main topics in the texts. A topic is a list of words statistically defined to categorize the meaning or central theme in the text, a process termed topic modeling (Abdelrazek et al., 2022). Using topic modeling, one can find, in the literature, works addressing problems in the most varied fields, and there are several methods to conduct topic modeling (Abdelrazek et al., 2022). Among the most employed ones are Latent Dirichlet Allocation (LDA) Blei et al. (2003), Latent Semantic Analysis (LSA) Landauer et al. (2013), and Non-negative Matrix Factorization (NMF) Lee and Seung (2001), where both of the latter are based on diverse forms of matrix factorization.

LDA is one of the most popular and widespread methods for identifying latent topics in a text (Blei et al., 2003). It identifies the (relevant) topics by using generative probabilistic models. Among its application areas, we observe social media topic analysis (Saura et al., 2019; Yang and Zhang, 2018; Yu et al., 2019). While the previous studies focus on different problems, each uses topic modeling as a tool for SMA. Yu et al. (2019) developed a novel hierarchical topic modeling technique and mined the dimension hierarchy of tweets' topics over tweets of different countries. Saura et al. (2019) using the hashtag startup ("#startup") analyzed tweets and their comments. The objective was to understand the topics in those tweets and the associated sentiments. Yang and Zhang (2018) performed a similar analysis, where the authors combined topic modeling and sentiment analysis to mine the tweet's text. They concluded that the LDA algorithm makes analyzing an extensive set of tweets easy and obtains meaningful topics. Some other studies use topic modeling to explore and understand specific subjects on Twitter, like in the case of Barry et al. (2018), which analyzes alcoholic drinks advertising, or a recent study to understand how politicians tweet about climate change Yu et al. (2021). More recent works use topic modeling methods to examine Twitter information about COVID-19. For instance, Sha et al. (2020) analyzed governmental and politicians' tweets about the pandemic and inferred a set of topics describing Twitter activity in the countries under analysis. Kaila, R.P. & Prasad (2020) and Doogan et al. (2020) focused on tweets bearing hashtags related to COVID-19 to understand what non-government users tweet concerning the coronavirus pandemic and its global perception.

Topic models are abstract models that can be challenging to comprehend, and visualization is a usual method used better to understand the generated topics (Kherwa and Bansal, 2019). As previously explained, the models provide insight into the content of a document collection by grouping it into topics. The model output is the top terms of each topic in a list format bearing the respective frequency values. The more frequent visualizations of topic models encountered are word clouds or stacked/column bar charts, one of these for each topic, allowing the interpretation of each topic's meaning by emphasizing the more relevant terms.

While the former studies ascertain LDA as having achieved good results in analyzing Twitter posts, they also raise limitations about using the LDA algorithm with Twitter data. The two most common limitations are that tweets present a short text format and the need for an adequate preprocessing phase. Transforming a tweet into a document to perform topic modeling might not be adequate because it has only a few words to extract relevant topics. Therefore, most studies solve these limitations by aggregating the tweets into sets, where each collection corresponds to a document (Curiskis et al., 2020).

However, some advances appear to avoid aggregation, as in Xiong et al. (2018), where the authors propose a short-text topic model algorithm.

3.2.2. Startups and social media activity

Social media platforms have a global reach, are easy to access, and are low cost, enabling startups to use social media as a digital marketing gateway and observe the market. It allows entrepreneurs to interact with various stakeholders in the ecosystem, including partners, suppliers, universities, and resource providers, facilitating collaboration (Almotairy et al., 2020). A few studies investigate the potential relationships between startups and social media platforms. Lugović and Ahmed (2015) found a positive correlation between the ecosystem, startups' Twitter usage, and the source country's total investment. Saura et al. (2019) aimed to relate the polarity of the tweet with the topics found within the diverse sentiments. The authors classified the tweet's text and comments into positive, negative, and neutral. Then, the authors performed topic modeling for each polarity and found the related topics, enabling them to understand the Twitter audience sentiment of startup-related content. Ruggieri et al. (2018) aimed to find patterns in successful innovative startups based on their digital platforms' activity. Their study demonstrates that startups are present on digital platforms mainly because these platforms have a cost-effective performance. The authors also conclude that a community of users/providers of services is essential for the business. Such a community is fundamental for a positive impact on digital platforms, primarily on social networking websites, providing positive or negative opinions about products and companies. Concerning opinions, word-of-mouth is the everyday oral communication that creates an impression and idea about a specific subject (Keller, 2007), and online opinions are called electronic word-of-mouth (eWOM), as explained by Hennig-Thurau et al. (2004). Social media platforms are ideal tools for eWOM. Chu and Kim (2011) describe that eWOM enables the creation of a large community, which allows for increased digital engagement through social interactions, such as comments, likes, and shares. The last two represent non-verbal activities, and when their quantities are large, they might help raise a positive feeling in the social media profile in question (Wolny and Mueller, 2013). Additionally, social media activities can be used to understand the online organization's reputation (Azinhaes et al., 2021), as tweets influence the customer brand perception (Jansen and Zhang, 2009).

3.3. Methodology

This chapter's research follows the SMI steps framework described by Choi et al. (2020) for social media-based BI research. The SMI process consists of four phases: "Data collection," "Data preprocessing," "Data analysis," and "Validation & Interpretation." According to this framework, we designed a processing pipeline illustrated by Figure 3.1. Starting with data collection, which, in our case, means the extraction of tweets regarding Portuguese (or Portuguese-related) IT startups' Twitter accounts. The approach (A) aims to understand if the subject or theme of the tweets evolves according to the business growth. We have chosen eight IT startup accounts and collected tweets dating from 2015 to 2020, resulting in 15,577 tweets. The dates are chosen accordingly with the information about the startups' establishment date. Approach (B) emerged from the results obtained with exploration (A), showing evidence that the tweets' topics change throughout the startup's life cycle. As such, we devised a new goal: to understand if the tweets' topics can characterize the business scaling phases. To achieve this, we selected a set of thirty additional startups and collected data from the thirty-eight Twitter accounts, ranging between 2013 and 2022, obtaining 91,743 tweets.

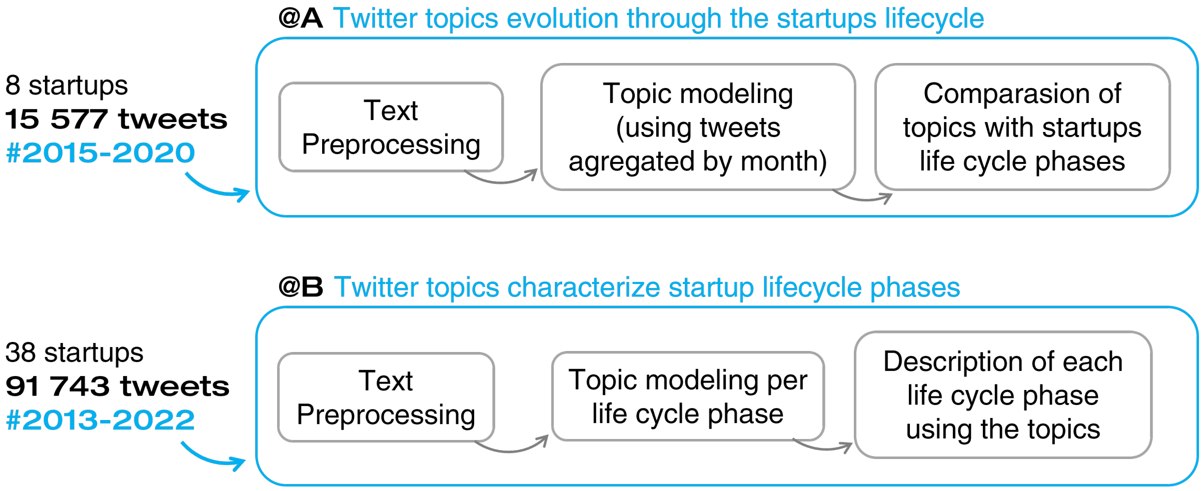

@A Twitter topics evolution through the startups lifecycle
8 startups
15 577 tweets
#2015-2020
Text Preprocessing
Topic modeling (using tweets agregated by month)
Comparasion of topics with startups life cycle phases
@B Twitter topics characterize startup lifecycle phases
38 startups
91 743 tweets
#2013-2022
Text Preprocessing
Topic modeling per life cycle phase
Description of each life cycle phase using the topics

Figure 3.1. Pipeline of the analysis of Portuguese IT startups Twitter data.

After the data extraction, the text of the tweets has been preprocessed for both approaches. In (A), the data has been aggregated by month and, in (B), by life cycle phase. Then, LDA was used for topic modeling technique (Blei et al., 2003). Finally, the model results have been validated and interpreted according to the specific purpose of each approach. In (A), this last step compares the topic modeling results with the startups' funding rounds, creating a temporal diachronic profile for each startup. That enabled us to notice the evolution of the topics in the startup tweets throughout the different phases. The previous results led us to approach (B), where the last methodological step allows us to characterize the different startup maturity phases by their topics and according to the life cycle model proposed in Chapter 2.

3.3.1. Datasets for (A) and (B)

The data that comprises the datasets were extracted from Twitter using the respective API, selecting all the tweets created by the several startup accounts for the chosen period. As mentioned, we created two different datasets for each approach, as displayed in Table 3.1. Notice that, for our study, a startup that posted at least one tweet in a particular month is said to have been *active* in that month.

TABLE 3.1. Datasets of approach (A) and (B) details.

Approach	(A)	(B)				
Time window	2015-2020	2013-2022				
Life cycle phase	All	FFF	Preseed	Seed	Early	Late
Tweets quantity	15,577	13,900	5,476	16,530	25,734	30,828
Active startups	8	26	16	31	19	8
Documents	72	104	104	104	104	104
Tweets per document (average)	273	134	53	159	247	296

Approach (A) dataset

Regarding approach (A), the extraction date is January 2021 and focuses on the tweets posted between January 2015 and December 2020. We selected this time frame because, at the extraction time, the number of posts was more concentrated between those five years due to the startups' foundation year. For this, we selected eight renowned Information Technology (IT) startups founded by Portuguese administrators or headquartered in Portugal from the *Sifted* 2020 Portugal startups list[1]. The companies were chosen because they are currently at different stages in their life cycle and are considered active on Twitter. The dataset consists of 15,577 tweets extracted from eight IT Portuguese startups acounts: *AttentiveMobile*, *Codacy*, *DefinedAi*, *Feedzai*, *Prodsmart*, *Talkdesk*, *Unbabel*, and *Virtuleap*. The tweets have been aggregated by month, resulting in 72 documents, one per month in the five years time-window, presenting an average of 273 tweets per document. These documents concern all the startups through all their life cycle phases. However, not all startups were present on Twitter throughout the time frame because some were founded later in the time range. In this approach, we did not consider five phases as stated in Chapter 2. We consider four, where the FFF is included in the preseed phase. This occurs because we formulated the startup life cycle with four phases when this approach was performed. As the research was being extended, we reformulated the life cycle and added a phase.

Approach (B) dataset

Concerning approach (B), the dataset comprises tweets from an additional set of thirty IT Portuguese startups active on Twitter on top of the eight startups used in (A). The extraction date was August 2022, and we extended the time frame to comprise tweets since

[1] https://sifted.eu/portugal-startups-top-rankings/

January 2013, resulting in data from eight years and eight months. In total, we collected a dataset of 91,743 tweets. We use a different year to extract the data to increase the number of tweets posted in each life cycle phase. We have divided the collected tweets into five subsets, one for each phase: FFF, pressed, seed, early, and late. Like what has been done in approach (A), the resulting phase documents have been aggregated by month, resulting in 104 documents per phase. Each document aggregates the tweets posted by startups on that phase in that month.

It is essential to notice that, depending on the phase, the number of tweets and startups on Twitter may differ. Of the 38 startups at the extraction date, 19 were in the seed phase, 11 in the early phase, and 8 in late. Figure 3.2 displays the 38 startups into the mentioned phases and details the last fund round type received, enabling us to place the startups into the phases.

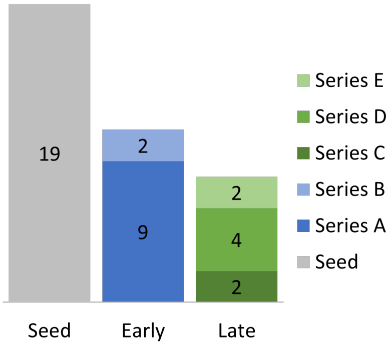


Figure 3.2. Startups' last funding round and respective life cycle phase at 1/08/2022.

In terms of the extracted tweets regarding the FFF phase and terms of the data, we found 13,900 tweets for 26 different startups, presenting an average of 134 tweets per document. The preseed phase presents fewer tweets (5,476), 16 active startups, and 53 tweets on average per document. In the seed phase, we can find more active startups (31), with 16,530 tweets and an average of 159 tweets per document. Next, in the early phase, we collected 25,734 tweets showing 19 active startups and the second-highest number of tweets per document: 247. Lastly, the late phase is the one presenting the higher volume of tweets in total (30,828) and the most extensive number of tweets per document on average, 296, presenting fewer active startups: 8. By looking into these numbers, we can envision that the startup phase influences the level of activity on the social media platform.

Figure 3.3 shows the number of tweets posted by the startups and the number of active startups distributed over our chosen time window. It is possible to see that the number of active startups in this setup has been consistently increasing over time. In the years 2013, 2014, and 2015, the quartile presenting the highest number of tweets was the fourth, first, and third, respectively. In the subsequent years, 2016 and 2017, the quartile presenting a higher number of tweets was the second. The second quartile of 2017 shows the maximum quantities of posts throughout the time window. In 2018, the quartile depicting the most significant value was the fourth. However, this is the year where the distribution of the number of tweets is the most uniform. In 2019, the first quartile presented the highest value. Interestingly enough, since this is the most problematic year for the COVID-19 pandemic 2020, we see a decrease in both the number of tweets posted and the number of active startups. In 2021 and 2022, the number of active startups rose again, although there was no corresponding increase in the tweets posted compared to pre-pandemic years. This seems to indicate that the startup phase relates to its activity.

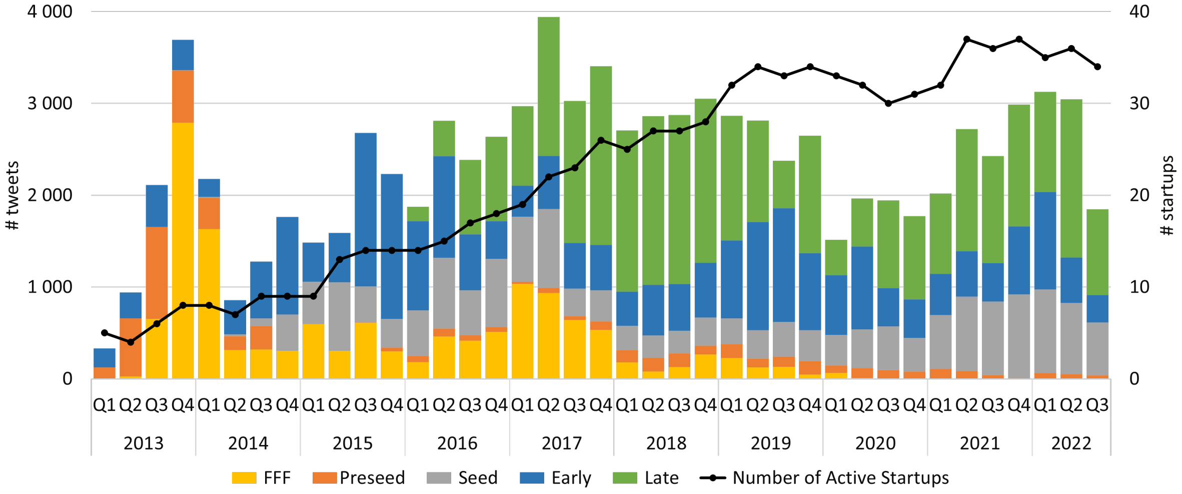


Figure 3.3. Distribution of tweets quantity over time.

Figure 3.3 displays the distribution of tweets over the years, where the bars are divided by color, regarding the correspondent phase for the startups that posted the tweet at the quartile. Since the startups have evolved over the years, from the FFF to the late phase, the colors in the graphic bar change accordingly. As previously noted, the older tweets in the dataset correspond to Twitter activity by newer startups (yellow and orange). In contrast, the recent tweets regard activity posted by more mature companies found at later phases.

3.3.2. Text preprocessing

To understand the topics of the textual tweets, datasets were aggregated by month, resulting in a corpus (a set of documents where each document has an id and the correspondent text) of 72 documents for approach (A) and 104 per phase for approach (B). The documents correspond to each of the months in the time scope of each analysis. Within a document, the id regards the month and year of the tweets. This corpus was then cleaned, retaining the vocabulary that accurately represents the startups' content to be transformed into a document-term matrix for model training.

To ensure the adequate preprocessing of tweets, we first studied the techniques applied in similar studies, thus concluding that the literature supports the need for preprocessing, enabling a preparation phase to achieve coherent topics. Table 3.2 presents the techniques found to have been applied in the existing literature. The most used techniques are URL elimination, extra white space elimination, exclusion of the terms presenting higher or lesser frequency, HTML tags elimination, and the usage of stop words.

Table 3.2. Literature preprocessing techniques usage.

Preprocessing technique	Choi and Park (2019)	Alash and Al-Sultany (2020)	Doogan et al. (2020)	Hidayatullah et al. (2018)	Yang and Zhang (2018)
Lowercase transformation	X		X		X
HTML tags elimination	X	X		X	X
URL elimination	X	X	X	X	X
Hashtag treatment	X	X			
Remove punctuation and digits			X	X	X
Remove Stop Words		X	X	X	X
Lemmatization			X		
Stemming				X	X
N-Grams		X	X		
TF-IDF		X			
Remove extra white spaces	X	X	X	X	X
Remove terms with higher frequency	X	X	X	X	X
Remove terms with less frequency	X	X	X	X	X

Since white spaces, URLs, and punctuation do not present information relevant to the topic's identification, they were removed from the documents. Next, lowercase transformation and lemmatization were performed. The lemmatization goal is to convert every word to a common base form, providing coherence to the set of words and, consequently, to the topics. Lemmatization was performed using the TextBlob library (Loria, 2018). Applying a set of stopwords, that is, a set of terms to exclude, helps to focus the model on the relevant words that might define the text's meaning. In this case, we used stopwords from the Natural Language Toolkit (Bird et al., 2009) and added the startups' names and Twitter tags - like "RT," which means it is a retweet - to the set of stopwords.

CountVectorizer from the Python library scikit-learn (Pedregosa et al., 2011) enables vectorizing the text and some preprocessing customization, like using n-grams and exclusion of terms. The n-grams used ranged from 1 to 2, uni to bigrams, to gather terms that may appear together, for example, the bigram "Machine Learning." Then, the terms that appeared less than twice were excluded to prevent possible errors and misspellings. Lastly, the exclusion of terms that appear in at least 80% of the tweets, being highly frequent terms, suggests that they are meaningless in terms of topic characterization.

3.3.3. Topic modeling

The topic modeling method employed in both of the approaches was LDA, the Latent Dirichlet Allocation method (Blei et al., 2003). The first step in the modeling is the transformation of the corpus into a document-term matrix, where each term is either a word or a bigram. For that, we used the frequency of the occurrence of the term/bigram in the document's text. We applied the LDA algorithm on the resulting matrix, employing the resources from the Python library *gensim* (Řehůřek and Sojka, 2010). Since the number of topics is an input parameter for the LDA algorithm, we performed a coherence test to understand the number of topics to use in our model construction. For each of the (A) and (B) approaches, we applied a coherence measure, the *c_v*, one of the options in *gensim*, to correctly select the number of topics.

3.4. Results and Discussion

The dataset analysis raises an important question: Do the life cycle phases influence the level of activity on Twitter by the startups? The first subsection aims to answer this question by exploring the possible relationship between phases and the number of tweets startups have posted when at that phase. The previously described approach (A) uses topic modeling to find if the content of the startups' tweets differs over the life cycle. The second subsection presents the results for (A), clearly showing that the tweets' content changed throughout the company's growth. The third subsection regards approach (B), where more data, in terms of companies and tweets, was used to characterize each one of the life cycle phases by the social media activity of the startups.

3.4.1. Activity level of startups over life cycle phases

Antonakaki et al. (2021) explains methods that can be applied to measure users' activity, popularity, and influence on social media platforms like Twitter. Here, activity means how frequently the user (a tweet account owner) interacts with the platform and the number of tweets and retweets the user has made. In this study, we have chosen to use the total number of tweets posted by each of the startups. To analyze if the startups' activity levels change over their scaling, we tested if there was a relationship between the number of tweets and the phase. We used a Kruskal-Wallis test since we have no good reason to assume that the number of tweets follows a normal distribution. This is a nonparametric method that compares the means between groups, and, in this scenario, each group will be one of the five life cycle phases. We used the *SciPy* (Virtanen et al., 2020) library for the Kruskal-Wallis test, setting the significance threshold at 0.05. The null hypothesis is that the means in each life cycle phase are the same. If the *p-value* is lower than the threshold, we reject the null hypothesis, meaning that the means on every life cycle are not the same. The test resulted in a *p-value* of $1.85E-07$, lower than the threshold, proving the existence of a statistically significant relationship between the number of tweets and the startup phases.

This relationship can be visualized in Figure 3.4, which shows the percentage of tweets made by each startup by each life cycle phase. The graph shows that the startups post more on average at the seed, early, and late phases, which means that around 40% of their content is posted on average in those three phases, corresponding to already more mature startups with prototypes or ready-to-market products. Newer startups at FFF and in the preseed phase, on average, do not post on Twitter, suggesting that they are prioritizing other business matters, such as making a problem-solving prototype of a product that adjusts to the market. In fact, by analyzing the dots of the graph, we can also tell that relevant enterprises in the dataset do not present an active Twitter profile, not posting in the first three phases of their life cycle, with several close-to-null percentages of posted tweets. This indicates that more recently founded startups take less advantage of the social media platform Twitter than more mature ones. In the seed phase, startups present different behaviors: while some do not use social media at all, others make significant use of it. Note that the dots in the line of the 100% correspond to startups that only posted in that phase because their posts appeared when it was at that exact phase, at the extraction date, thus only presenting tweeter activity for that phase.

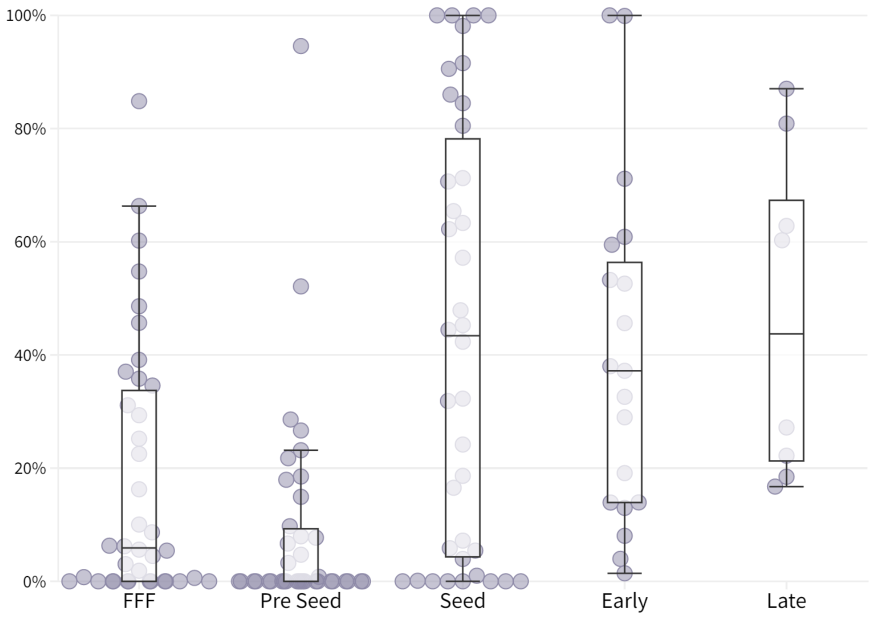


Figure 3.4. Distribution of tweets by startup over the life cycle phases - each dot represents the percentage of tweets made by a startup in the correspondent phase

3.4.2. Approach (A): Exploration of the topics on the tweets over life cycle phases

The first step with topic modeling is to transform the corpus into a document-term matrix, where each term is either a word or a bigram. For that, we use the frequency of the occurrence of the term/bigram in the document's text. We applied the LDA algorithm on the resulting matrix using the Python library *gensim* (Řehůřek and Sojka, 2010). As mentioned in subsection 3.3.3, the number of topics must be given as input for the algorithm. Thus, we performed a coherence test seeking the advisable number of topics. Figure 3.5 suggests that five might be the more reliable number of topics due to its higher coherence value.

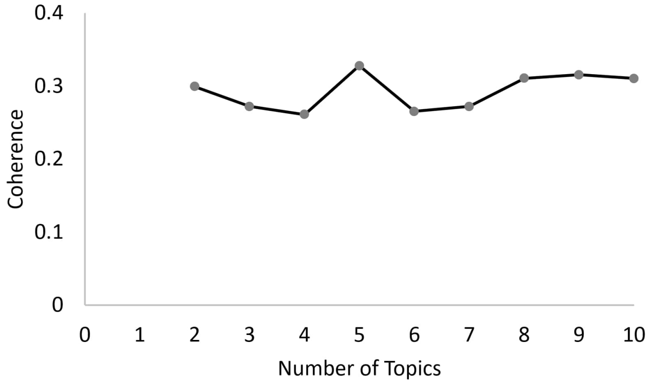


FIGURE 3.5. LDA coherence analysis.

Therefore, the topic model created by applying LDA has five topics, each characterized by the more relevant terms, with all the terms showing a similar distribution within each topic. Figure 3.6 shows an illustration presenting the five topics, termed "Fintech and ML", "Business Operations," "Bank and Funding," "Product/Service R&D" and "IT," and their relevant terms. Interestingly, the standard terms between the five topics are quite relevant for what a Portuguese IT startup might be/use nowadays: "machine learning", "lisbon," "service," "learn," and "webinar."

Following is the description of each topic:

- "Fintech and ML": we chose this name because it encapsulates terms such as "fintech", "machine learning" and "banking," as well as one important conference for this domain: "money2020";
- "Business Operations": it presents terms concerning typical companies' operations, such as "customer service," "brand," "solution," and "covid19". Additionally, it also displays "opentalk2020", a *Talkdesk*'s event regarding customer service subjects;
- "Bank and Funding": the name is supported by the terms "bank," "leader," "report," and "partner;"
- 'Product/Service R&D": this topic is sustained by terms like "innovation," "learning," and "boost.;"
- "IT" (Information Technology): is a topic associated with software, like code review and machine learning.

To understand how relevant the above-described topics are within the posting activity among the different startups, we created a heatmap that can be seen in Figure 3.7. It presents the relative percentages of each one of the topics within the content posted

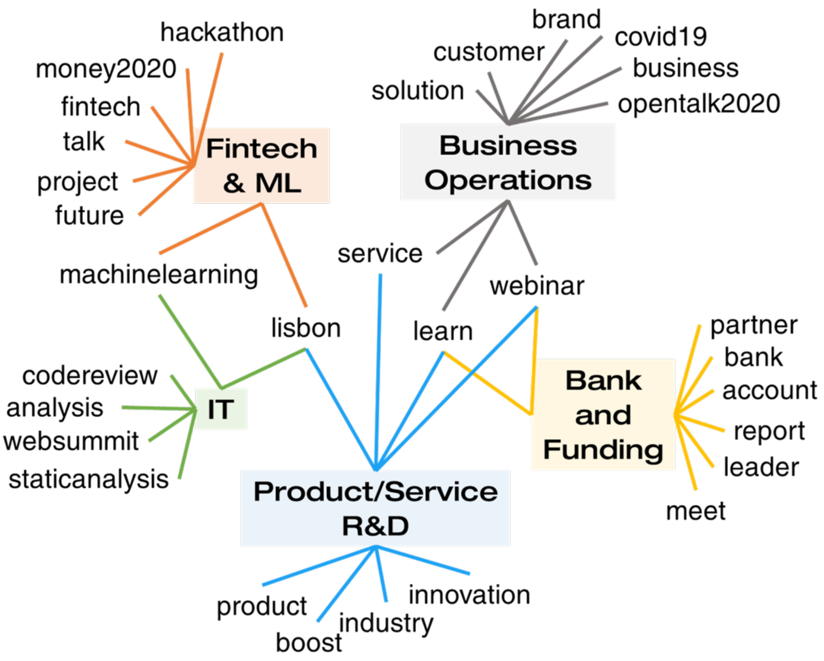


FIGURE 3.6. Approach (A) emerging topics and their more relevant respective terms.

by startups. The horizontal axis indicates the topics, and the vertical axis displays the companies. The color gradient scheme used corresponds to a percentage in the range [0, 30], from colder colors (0%) up to warmer colors (30%). The first line indicates the average percentage for each topic among the eight startups. Interestingly, approximately 50% of the content, corresponding to the combination of the topics "Fintech & ML" and "IT", directly relates to technology, which is to be expected since these are all IT companies. While half of the Twitter content focuses on startups' core business, around 20% concerns "Bank and Funding", and nearly 30% deals with enterprise matters via content related to "Business Operations" and "Product/Service R&D".

Interestingly, the way the topics split in the contents in the startup posts is similar, which is noticeable by looking at the heatmap's colors. The topics showing a higher presence in the posts and consequent warmer colors are those of "Fintech & ML", "Bank and Funding", and "IT", with the former and the latter being the ones that show the highest posting percentages, between 19.4% and 29%. This agrees both with IT being the startups core business and the fact that they are "young companies founded to develop a unique product or service, bring it to market and make it irresistible and irreplaceable for customers" [2]. Startups typically engage in a continuous iterative improvement of their products and services through feedback and usage data. With colder colors and postage percentages ranging between 9.6% and 20% are "Business Operations" and "Product/Service R&D".

[2] https://www.forbes.com/advisor/investing/what-is-a-startup/

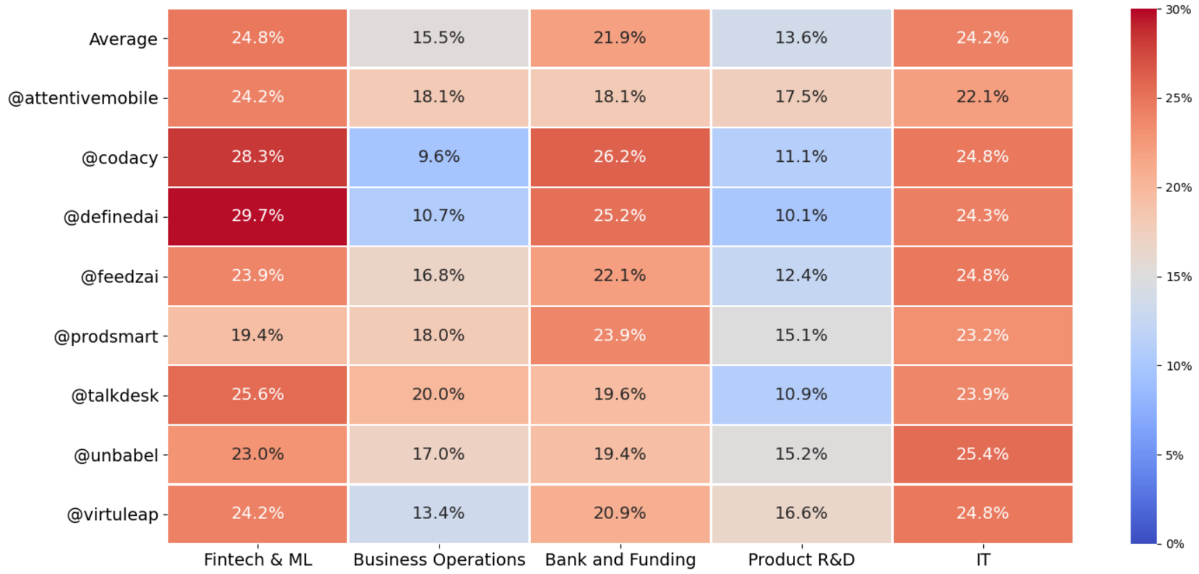


FIGURE 3.7. Topics frequency per startup.

Understanding the relative emergence of topics within tweets differs according to the four FPEM phases. Since we have no good reason to assume that the topics distribution follows a normal distribution, the Kruskal-Wallis test was employed again via *SciPy* (Virtanen et al., 2020) library setting the significance threshold at 0.05. The null hypothesis in this scenario is H0: The means for each life cycle phase are identical. If the *p-value* is lower than the threshold, we reject the null hypothesis, meaning that the means on every life cycle differ. The results for the test are presented in Table 3.3, denoting the ones with a *p-value* below the significance with (*). The topics "Product R&D", "IT," "Business Operations," and "Fintech and ML" present a *p-value* lower than the threshold, meaning that their means differ over the life cycle phases. "Bank and Funding" is the exception on the Kruskal-Wallis test, presenting a *p-value* expressively higher than the significance.

TABLE 3.3. Kruskal-Wallis tests results.

Topic	**p-value**
Product R&D	(*) 0.00545
IT	(*) $2.38E-13$
Bank and Funding	0.327
Business Operations	(*) $2.4E-06$
Fintech and ML	(*) $8.82E-08$

(*) Statistically significant ($p < 0.05$)

After the topic modeling, we divided the corpus by startup and applied the model, resulting in individual analyses representing the topics' evolution over time for each one of the companies. We combined the funding rounds' information to understand if there is a relationship between the FPEM phases and Twitter activity. In Peixoto et al. (2023a), we described the results obtained per startup, which details are in Appendix C. After the individual analysis, it became clear that there were similarities between the eight independent analyses, so we performed another study, this time using all the startups' data. Figure 3.8 displays the distributions of each topic for the different phases to understand how, as shown by the Kruskal-Wallis test, they differ throughout the FPEM phases.

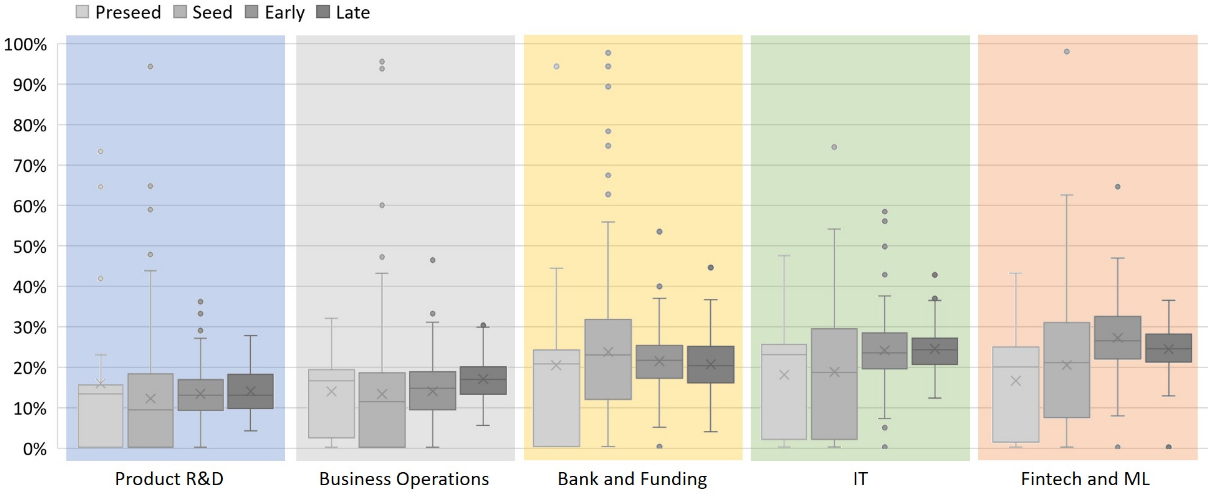


Figure 3.8. Topics distribution over life cycle phases.

The topic "Product R&D" varies through the several life cycle phases, as expected since the Kruskal-Wallis rejected the null hypothesis. It shows higher values in the preseed phase than in the subsequent ones. This variation can illustrate the higher importance of product development at the startup's beginning. That is, startups in the preseed phase are finding a solution to a problem and confirming the maturity stage stated in the life cycle description (Section 2.4). The topic "Business Operations", for which the means differ over the life cycle phases, shows lower values in preseed and increases over the following phases. This may be viewed as the opposite behavior of "Product R&D" and shows that when the startup growths, content posts about product development are substituted by business concerns. The topics "IT" and "Fintech and ML," related to the startups' core business, show a similar evolution over the phases. Both topics' appearance in the tweets increases until the early phase is achieved and slightly decreases over the late phase. Note that those differences are statistically significant to support the difference in the means over the life cycle. Lastly, the topic "Bank and Funding" is the only one that does not show any significant difference over the phases, always around 20% of appearance in the companies' tweets. The constant presence of this topic demonstrates the importance of fundraising and financial matters for startups, supporting the fact that funding rounds are an essential dimension that characterizes startups.

3.4.3. Approach (B): Topics that characterize startups' life cycle phases

Similarly to what has been done for approach (A), we started the topic modeling process by aggregating the dataset and transforming the corpus into a document-term matrix.

But now, we divided the dataset by phase, as previously explained, which resulted in five topic models. For each phase, we used the frequency of the occurrence of the term/bigram in the text of the tweets of that specific phase and applied the LDA algorithm (*gensim* Python lybrary (Řehůřek and Sojka, 2010)). A coherence test was made for each subset to choose the more reliable number of topics for building that model using the same coherence measure - c_v.

In this section, we present the topic model description with the topics and respective relevant terms, which enables the characterization of the startups' life cycle phases using their Twitter activity. Additionally, we present the frequency of the discovered topics in each subset to allow for better characterization. For the FFF and the late phase, the generated models have four topics, while in the preseed and early phases, the suitable models have two topics, and the model for the seed phase has three topics.

Figure 3.9 illustrates the FFF phase's topic model, with the relevant terms that allow for the naming of the topics: "Digital transformation solutions," "Business experience," "Real estate market," and "IT job career." The first topic, "Digital transformation solutions," refers to the terms/bigrams around the words "authentication," "solution," and "password," which relate to the digital changes in companies. The topic "Product development" joins terms about the essential core for an IT company to develop its product/service, like "team," "dev," and "engineer," and product. The topic "Market" encompasses terms like "market," "brand," and "business." Lastly, the topic "IT job career" encapsulates employment-related terms such as "experience" and "sales" terms. The common terms between the four topics emerge as a fair description for the FFF phase when the startup wants to find a "way" to create a "job" in "tech" by "developing" a "code" related product.

Table 3.4 displays the distribution of the topics within the tweets posted by the startups when at the FFF phase. The topic showing less is "Digital transformation solutions", with only 3.8%; thus, we will not use it to characterize the present phase. Instead, we will use the other three topics, each presenting above 30% frequency on Twitter content. With these insights, we can support that the startup twit contents at the FFF phase reflect the desire to work in a specific market and the need to develop an IT product by people with knowledge to leverage and kick-start the company.

TABLE 3.4. FFF phase Twitter contents' frequencies of the topics found.

FFF phase topic	Frequency
Digital transformation solutions	3.8%
Product development	30.4%
IT job career	32.9%
Market	32.9%

For the preseed topic model, we have two topics shown in Figure 3.10 that illustrate the more relevant terms. The topic "Client engagement" comprises the words "company," "client," "customer," and "marketing." The second topic, named "IT startup journey,"

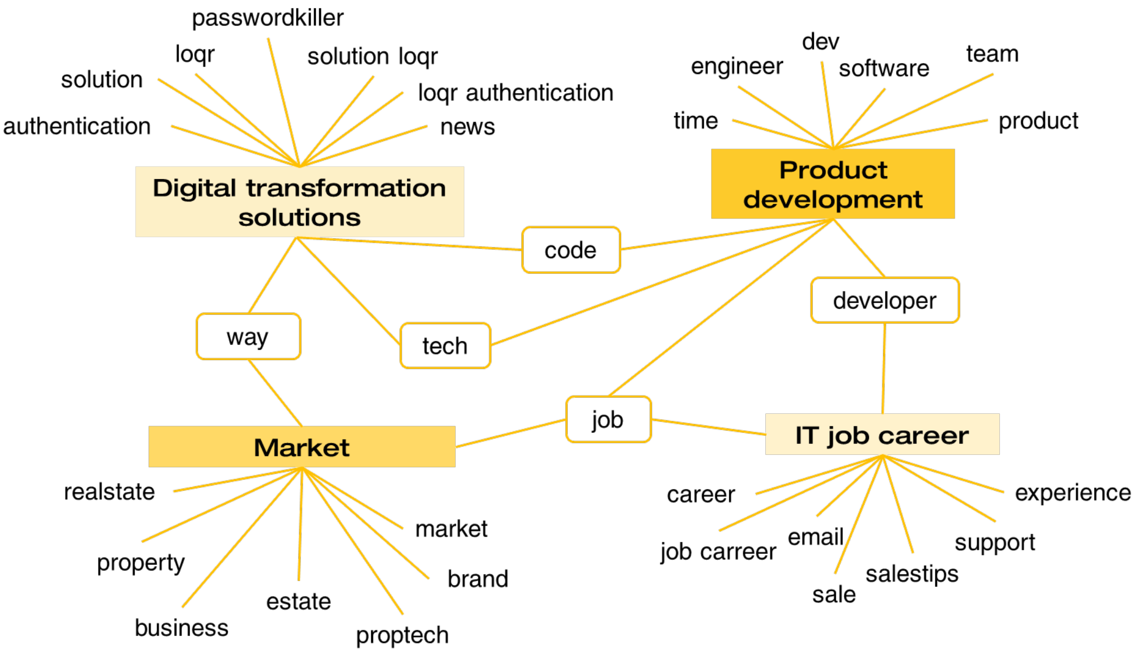


FIGURE 3.9. Approach (B): Topics found for FFF phase and respective terms.

incorporates terms like "startup," "check," "software" and "day." The word in common between the topics is "team." This phase can be characterized by the beginning of the startup journey, prioritizing the process of building a "team" and the need to reach clients. Table 3.5 shows the frequency of the topics in the tweets of the preseed phase. It demonstrates that, in this phase, more than 60% of the startup's content is about the "IT startup journey," highlighting that the day-to-day steps of building the company are crucial at this phase.

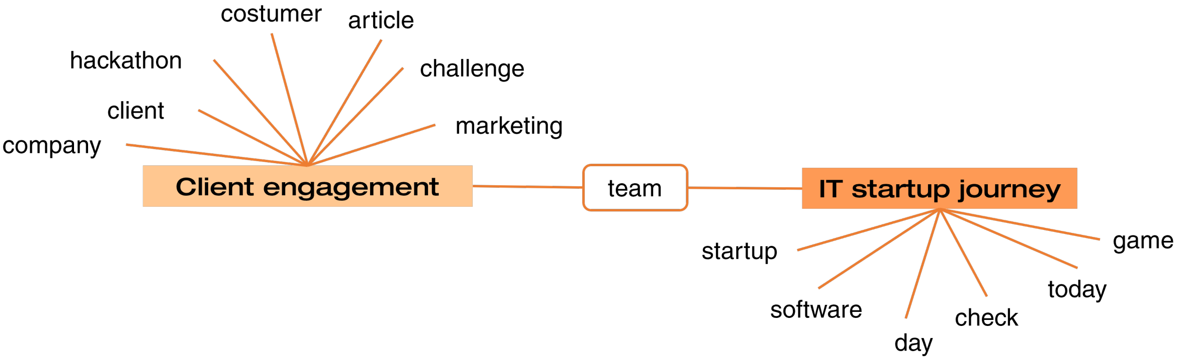


FIGURE 3.10. Approach (B): Topics found for preseed phase and respective terms.

TABLE 3.5. Preseed phase Twitter contents' frequencies of the topics found.

Preseed phase topic	Frequency
Client engagement	38.7%
IT startup journey	61.3%

Figure 3.11 shows the three topics generated from the seed phase tweets. The topic named "Technology" includes words about "experience" in domains of a technological "business." The topic "Applications security" comprises terms about defense from cyber-attacks. Lastly, a topic designated "Funding" appears to contain terms about getting a "sponsor," with positive words like "amazing" and "super." This is the first time a topic with the term "round" regarding the funding rounds occurs. There is only one word in common between the discovered topics "make," but only between "Technology" and "Funding." It highlights the importance of making it happen in the seed phase when the startup should already own a functional prototype.

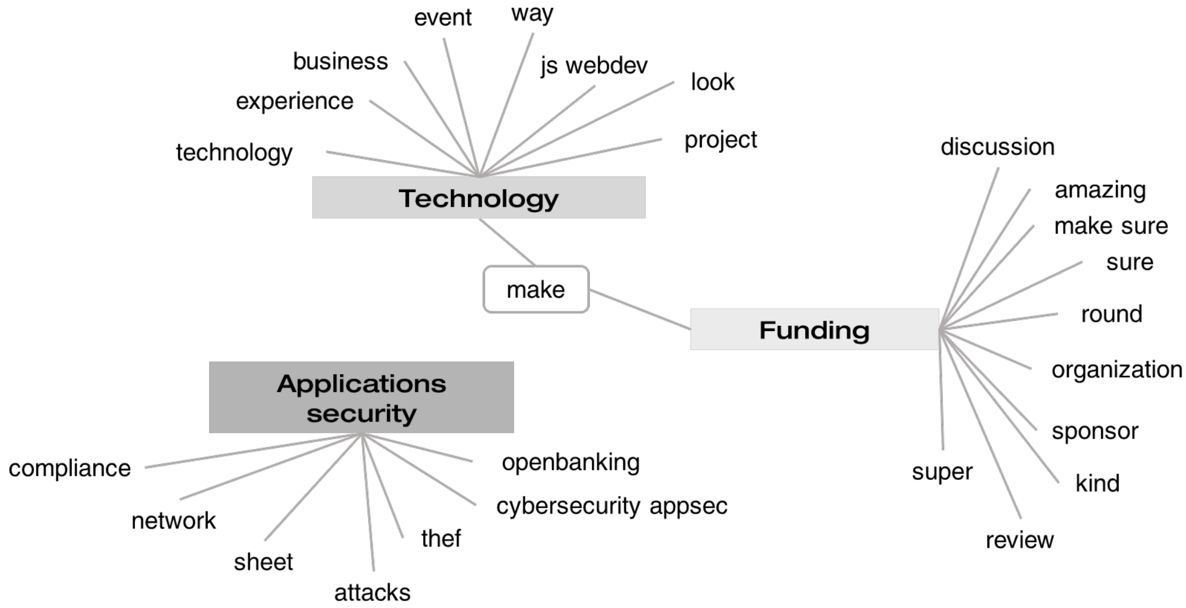


Figure 3.11. Approach (B): Topics found for seed phase and respective terms.

Table 3.6 shows the frequency of these topics' occurrence in the tweets at the seed phase. When at this phase, the startups prioritize the process of gaining field experience, with half of the Twitter content being classified as "Technology". The startups are concerned with security matters and attaining funding and sponsors to grow, with the corresponding topics appearing as 21.4% and 28.2% of the contents, respectively.

Table 3.6. Seed phase Twitter contents' frequencies of the topics found.

Seed phase topic	Frequency
Technology	50.5%
Applications security	21.4%
Funding	28.2%

The two topics created with the data subset of the early phase are represented in Figure 3.12. The topic named "Product" respects matters of day-to-day productivity and development progress in companies, with terms like "improvement," "activity," "course," and "person." The second topic, called "Funding," regards funding in innovation technology fields like artificial intelligence. The only term in common between the two topics is “future.” At this phase, the startup focuses on the future, concerned mainly with the company’s productivity and fundraising. The topic “Product” is prominent within the startup’s tweets, with 73.1% of frequency. The topic “Funding” appears in 26.8% of the tweets subjects, slightly decreasing compared to the previous phase. These frequency values can be observed in Table 3.7.

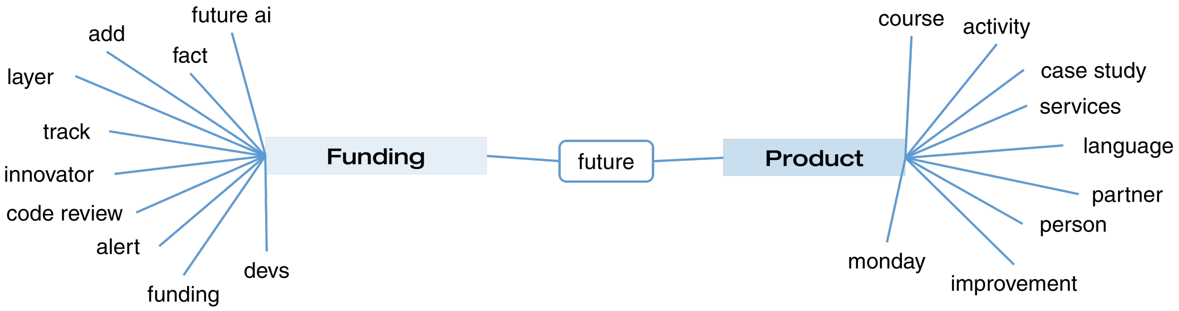


FIGURE 3.12. Approach (B): Topics found for early phase and respective terms.

TABLE 3.7. Early phase Twitter contents’ frequencies of the topics found.

Early phase topic	Frequency
Funding	26.9%
Product	73.1%

Lastly, Figure 3.13 displays the four topics regarding the late phase. One of the topics found relates to “Development solutions,” showing terms regarding technological implementation and referring to artificial intelligence and “google,” which can work as tools to develop solutions in the IT domain. A second topic is named “Partnerships” since it refers to collaborations between other technologies and companies. Both topics share a bigram, “low code,” a development solution in IT that can be strategic for fast service response. The topic “Ethical and legal practices” displays ethical issues about artificial intelligence and financial data. The last topic, “Management approaches,” encapsulates team management terminology, such as “week join,” “comment,” and “worker.” All these topics show a more even distribution in the tweets’ contents, with well-defined subjects.

Table 3.8 presents the topic frequency of the tweets made by startups in the late phase. IT startups in this phase have established products, searching for new technologies and alliances. This is reflected by their Twitter content, with more than 30% per each corresponding topic (“Development solutions” and “Partnerships”). Additionally, they are more stable companies that worry about “Ethical and legal practices” and the management of their team, with a content frequency of 13.9% and 21.5%, respectively.

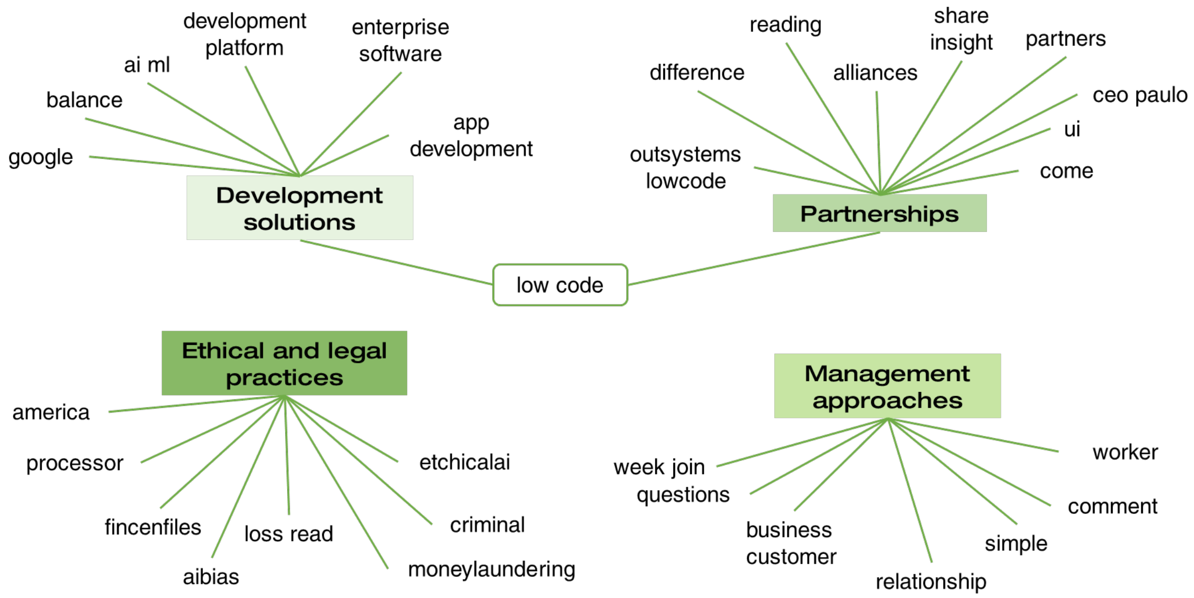


Figure 3.13. Approach (B): Topics found for late phase and respective terms.

Table 3.8. Late phase Twitter contents' frequencies of the topics found.

Late phase topic	Frequency
Development solutions	30.4%
Partnerships	34.2%
Ethical and legal practices	13.9%
Management approaches	21.5%

3.5. Summary

Startups face the challenge of raising funding and reaching customers within a restricted budget. Social media platforms like Twitter allow startups to connect with potential clients and venture capitalists. Considering this opportunity, we examine the contents shared by IT Portuguese Startups on Twitter throughout the different phases of their life cycle when considering the FPEMv2 model (Section 2.4. By investigating the essence of their posts, we aim to understand how a startup's social media content changes through the different phases of its life and analyze whether it reflects the company's evolution.

Our first research question was to apprehend if the Twitter content of IT Portuguese startups changed according to their life cycle phases. We used the content produced by eight startups from 2015 to 2020, resulting in a dataset of 15 577 tweets, the approach (A). Using that data, we performed a topic modeling from where the following five topics emerged: "Product R&D," "Bank and Funding," "IT," "Fintech & ML," and "Business Operations." These results were explored taking into account the FPEM life cycle model, creating a diachronic profile for each one of the startups, revealing that the eight IT startups' Twitter topics change over time in accordance with their current lifecycle. "Bank and funding" is the only one of the five topics present throughout the startup's life cycle, demonstrating the importance of financial investments and capital enabling the company's growth. The topic “Product R&D” appears predominantly during the preseed phase, showing that startups begin product-focused companies. In contrast, the topic “Business Operations” is prevalent in the late phase, revealing that business concerns replace the product development content with the startup’s growth. The more technological topics, “IT” and “Fintech & ML,” are more predominant in the middle phase of the life cycle, where the company’s focus is the product. Therefore, social media content evolves with these startups’ evolution and scaling stages. On a last note, this small-scale work showed that the topic frequencies per startup were similar, meaning they showed identical posting behavior.

After confirmation of the conjecture that the content changes with the companies’ evolution found in approach (A), we decided to experiment with a larger-scale study involving 38 startups (approach (B)). Our research question was to understand if the Twitter contents enable a characterization of a startup’s life cycle phase. After collecting a dataset of 91 743 tweets posted by the 38 startups from 2013 to 2022, the data were grouped based on the respective phase. We conducted a topic modeling for each one of the phases, resulting in five different models. The FFF phase resulted in a model presenting four topics: “Digital transformation solutions,” “Business experience,” “Market,” and “IT job career.” The startup content in an FFF phase reflects the goal of developing a valid IT product, and it involves people with knowledge who want to kick-start the company. The preseed topic model establishes two topics: “Client engagement” and “IT startup journey.” At this phase, more than 60% of the startup’s content is about the “IT startup journey,” highlighting the beginning of the startup creation as an entity. The seed phase resulted in three different topics: “Technology,” “Applications security,” and “Funding.” At this phase, the startups prioritize the fact that field experience must be gained, with half of the Twitter content being about “Technology.” Security matters and funding to grow are the main themes in the posts. The early phase is characterized by two topics: “Product” and “Funding.” The startup focuses on the future, mainly on the company’s productivity and fundraising. The late phase displays four topics: “Development solutions,” “Partnerships,” “Ethical and legal practices,” and “Management approaches.” Startups at this phase have already established their products and are searching for new technologies and alliances.

Figure 3.14 displays a diagram that summarizes this chapter’s conclusions, enabling us to characterize the FPEMv2 phases using the content posted on Twitter by IT Portuguese startups. It shows that approach (A) and approach (B) created a comparable topic evolution.

In approach (B), the topics portray finer details since we used a more extensive set of startups and grouped the tweets by phase. Regarding the number of tweets, the FFF and preseed phases show lower activity levels than the subsequent phases. Startups in the FFF and preseed phase, that is, newer startups, focus more on developing and researching their product. We can conclude that newer startups use the social media platform Twitter less

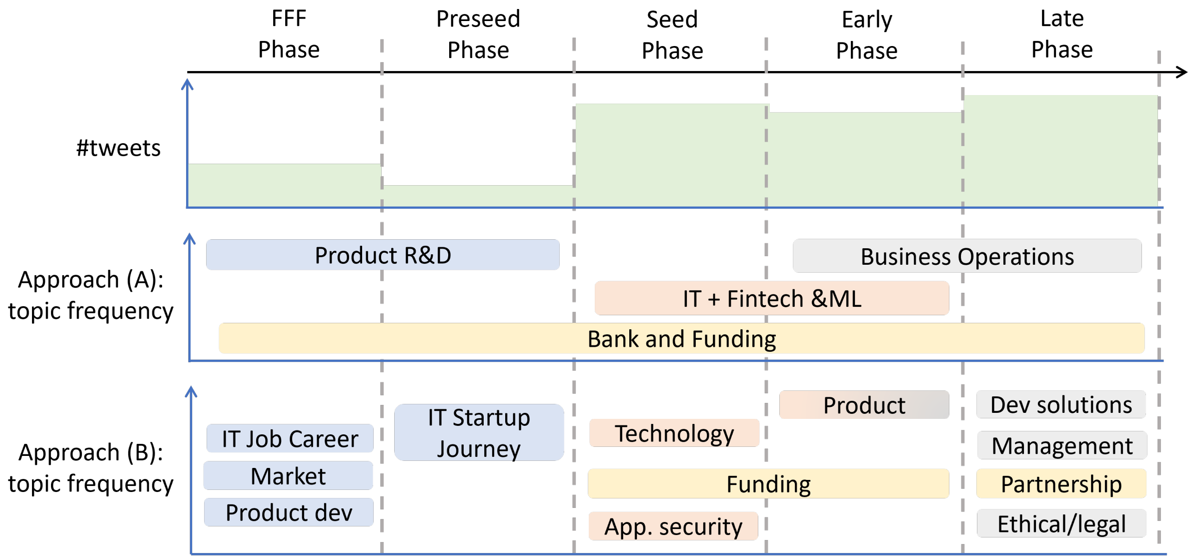


FIGURE 3.14. Life cycle phases characterized by the Twitter content.

than more mature ones. In detail, they care about finding the right persons to develop a specific product or service during the FFF phase. At the preseed phase, they value startup journey issues. When at the seed phase, they aim for a functional prototype to use as a funding promotor. Consequently, they post about the technologies to use and about funding subjects. In the early phase, startups have gained funds and display a functional product, conclusions reflected by the content they post on Twitter. Lastly, startups in the late phase have mature products. They are well-established companies, opening space to talk about business matters of IT companies, like new technologies, management approaches, and ethical/legal practices. The funding theme changes into partnership concerns, searching for sponsors and partners for their business.

CHAPTER 4

Startups Twitter communities analysis and visualization

4.1. Introduction

Startups benefit from cost-effective digital marketing opportunities through their relationship with social media platforms (Ruggieri et al., 2018). Studies have shown that active engagement on social media platforms increases digital engagement and can lead to better funding from venture capitalists or significant success in crowdfunding projects (Zhang et al., 2017; Ko and Ko, 2021; Hadley et al., 2018). Lugović and Ahmed (2015) found a positive correlation between the Twitter activity of European startups and the total investment in their country of origin. Additionally, by creating communities relating users and service providers, startups can monitor the market and take advantage of electronic word-of-mouth spread of positive opinions about their products (Ruggieri et al., 2018; Chu and Kim, 2011). Social network analysis, also known as SNA, is commonly called the process of monitoring the market and allowing for data-driven marketing strategies based on social media data (Hansen et al., 2019). Several existing studies have explored digital data using this methodology. For example, Ruggieri et al. (2018) focuses on finding startup success patterns based on their presence on digital platforms. Hingle et al. (2013) collected Twitter content to analyze dietary behavior, and the authors highlighted that data visualization allowed the identification of relationships between diet-related behavioral factors. Wu et al. (2016) show that visualization methods can help uncover social media analysis results and support data interpretation, leading to a network analysis and visualization process. Hansen et al. (2019) proposed a methodology, Network Analysis and Visualization, or NAV, to act as a design process model for enabling meaningful network analysis and extracting relevant insights.

The study documented in this chapter aims to determine the degree to which startups use social media platforms, what distinguishes these communities, and if the communities formed by each of the startups overlap. Our findings might highlight the relevance of social networks and online communities for startups. To our knowledge, none of the existing literature aims to understand how startups create communities on Twitter or if these communities intersect and create a global ecosystem. Thus, this chapter study intends to provide answers to the following research questions:

RQ4.1: Do startups form their own social communities on Twitter?

RQ4.2: In the case of community formation, are they disjoint communities, presenting different (types of) users?

To address the first research question, the study employed the previously mentioned NAV process (Hansen et al., 2019). Specifically, a social digraph was built, representing both *followers* and *following* communities of the startups under investigation. We utilized a community detection algorithm to determine if the startups form communities on Twitter based on modularity. This algorithm enabled us to visualize the communities in the digraph structure using different colors. The resulting visualization revealed that each startup had indeed formed a community. Furthermore, we were able to identify links between nodes of different communities, indicating that there was some overlap between the communities.

To address the second research question, we characterized the communities' users by analyzing their type, popularity, and activity level. This information enables the emergence of social media strategies that can be effective for startups to achieve their proposed goals, either for financial support or for product/service marketing actions.

The user characterization by popularity is a measure that uses the size of the *followers* and *following* communities. In the previous Chapter, we analyzed the social media activity content and levels over the FPEMv2 life cycle phases and discovered that the main thematics in Twitter contents vary. This observation raises the question of whether a startup's popularity level changes over the company's life cycle.

RQ4.3: Does the startup's popularity level change over its life cycle phases?

To answer the third research question, we used the dataset of 38 Portuguese IT startups studied in the previous approach (B) of Chapter 3.

After this introduction, Section 4.2 displays relevant related work, namely, the key role that informed visualizations perform for social media analysis. Section 4.3 explains the methodology that enabled answers to research questions RQ4.1 to RQ4.3. Section 4.4 presents the results and a discussion and analysis of the outputs. Lastly, Section 4.5 summarizes this chapter's main conclusions.

4.2. Related Work

Social media platforms are essential digital marketing tools for small businesses like startups (Ruggieri et al., 2018). This section explains how we can analyze communities created by startups when using social media platforms. We present a literature review of the methods and tools available for mining social media data using visualization, including network analysis and community detection algorithms, focusing on their application to study social communities on Twitter. Furthermore, we describe the role of social media in facilitating online communities. Lastly, we explain some relationships between startups and their social communities.

4.2.1. Visual analytics in social media analysis

Social media can be a valuable data source for businesses to extract digital marketing knowledge. Additionally, it can serve as a means to interact with their clients and potentially facilitate funding. Social media platforms generate two types of data: content data and interaction data. The content is usually found in a non-structured format, such as text and images. It can be retrieved from tweets and users' comments. On the other hand, the interaction data can be represented by a network structure (a graph). An example of this interaction is the *following* relation, i.e., when a user follows another. Other examples are actions, such as: *likes*, *shares*, *replies*, and *mentions*.

Social media analysis is no more than extracting information from social media data (Hansen et al., 2019). In Serrat (2017), the authors explain that such analysis can either focus on the social actors or their relationships. However, social media platforms generate high amounts of data, making it difficult to understand and analyze thoroughly. Visual analytics is a promising approach for dealing with the challenges of understanding complex data (Keim et al., 2008). It aims to explore complex data through visualization using interactive visual interfaces. Visualization techniques can uncover social media patterns and trends and support data interpretation (Wu et al., 2016). Furthermore, it helps gather insights from larger datasets, combining visualization techniques with the human dimension for enhanced data analysis. The NAV methodology arises from the need to combine network analysis with visualization, and it can be applied to each type of social media data to attain different goals.

Recent studies employ social media analysis through visualization regarding the content data. Saura et al. (2023) study tweets and apply topic modeling and sentiment analysis intending to mine the opinion of Twitter users about open innovation. They used a graph-based visualization to unveil the relation between the topics. Hu et al. (2017) also analyzed tweets and performed topic modeling. However, this study's originality lies in designing a particular technique for visualizing the content of unstructured social media text. Likewise, Smith et al. (2014) developed a new visualization to disclose the relationships between words and topics in topic models applied to unstructured social media data. Creating novel visualization methods is key since it enables improving data understanding and unveiling new insights. Hingle et al. (2013) use Twitter content to extract dietary behaviors and highlight that data visualization helped identify relationships between diet-related behavioral factors.

In the same way, the literature highlights visualization methods and intrinsic data aggregation as tools to understand and extract knowledge from social media interaction data. When a user follows another, this relation is represented by a directed link (edge), with the source being the follower and the sink on the followed one. The users connected through the *following* relation generate the so-called social graph (Gabielkov and Legout, 2012). Visualizing the social graph can reveal network features that help answer important research questions. For example, Molla et al. (2014) performed sentiment analysis of user Twitter contents and applied the results to color the graph's edges. This visualization highlighted where, in the social graph, negative, neutral, and positive opinions about a company existed. Abdelsadek et al. (2018) applied a community detection algorithm to a social graph, visually revealing the community's structure and related characteristics.

Based on the previous works, we can conclude that visualization is essential to extracting insights and knowledge from social media data. Additionally, we can improve the visualization by selecting and computing features that will be applied to the structure, such as color or format.

4.2.2. Social graphs and communities detection

As previously mentioned, the social media interaction data derived from the action of "user following another user" is encapsulated into a network structure called the social digraph (Gabielkov and Legout, 2012). In the digraph, a node represents a user, and an edge represents the user-following-user relation. For example, if user A follows user B, the graph exhibits a directed edge from node A to node B, as shown in Figure 4.1.

User A
User B
follows

FIGURE 4.1. Illustration: user A is following user B; user B is followed by user A.

Studies regarding social graphs can be performed either at the node level or at the network level. Antonakaki et al. (2021) explains methods that can be applied to node-level studies of social graphs to measure users' activity, popularity, and influence. Activity means how frequently the user interacts. In the case of Twitter, activity is measured by the number of tweets and retweets the user performs. Popularity measures how well a user is recognized, which usually can be estimated by the number of *followers*. A simple popularity measure is the *Structural Advantage* (Cappelletti and Sastry, 2012), a ratio between the number of *followers* and of *followings*. Lastly, influence estimates how a user's action influences (the actions of) other users, being the most used metric at the node level for studies involving this type of graph. Influential users are better disseminators of information through social platforms because they are more central in the graph. Consequently, graph centrality measures like *PageRank*, *betweenness centrality*, and closeness centrality are applied to evaluate the user's influence (Das et al., 2018).

Regarding the network level, network metrics enable quantitative comparison between graphs and analysis of temporal evolution (Hansen et al., 2019). Between the metrics used, we can find counts of nodes and links, average counts, or the application of concepts such as *density* and *centrality*. Antonakaki et al. (2018) used the average node degree and the average of incident edges to measure the evolution of a Twitter social graph over time. Said et al. (2019) conclude that Twitter communities have unique attributes that may impact the social media usage of their users.

Another way to dissect and extract information from a network is to apply algorithms that output some relevant structure or characteristic in the data. An essential concept in networks is that of the group or community: a set of nodes more densely connected between themselves than to others. The methods that find those groups are called community detectors and work as cluster algorithms (Hansen et al., 2019). These social media communities are essential for business, enabling a fast way to cultivate online brand awareness (Zaglia, 2013). The community detection algorithms commonly used in the literature are based on modularity optimization. Modularity measures the strength of the division of a graph. High values imply the graph has dense connections between the module's nodes and sparser connections between nodes of the different modules (Blondel et al., 2008). The modules represent the clusters and, in this case, the communities. The modularity optimization algorithms will explore every node if the modularity score increases when changing between modules. The specific steps and parameters depend on the algorithm used since, in the literature, many adaptations exist depending on the graph characteristics. Regarding social networks, Devi and Poovammal (2016) performed a complete review of the applicable options. One algorithm that stands out for social media platforms is found in the work of Leicht and Newman (2008), which considers the direction of the edges (connections). An example of an application of modularity optimization in Twitter communities can be found in Cruickshank and Carley (2020). The authors used multi-view modularity clustering to characterize and analyze hashtag COVID-19 pandemic Twitter communities.

In summary, social graphs represent social media relationships formed by users following each other. Analyzing these graphs at the node level provides insights into individual users' activity, popularity, and influence. Based on modularity, community detection algorithms can identify highly connected community nodes at the network level, acquiring valuable information for businesses looking to build brand awareness.

4.2.3. Startups and social media communities

Existing literature reveals investigations on the possible connections between startups and social media platforms. Lugović and Ahmed (2015) found a positive correlation between startups' Twitter usage and the total investment in their country of origin. Zhang et al. (2017) analyzed startups' Facebook and Twitter metrics, discovering that active engagement positively correlates with startup crowdfunding success. Ko and Ko (2021) conducted a social media analysis regarding fashion startups using Instagram as the data source. They conclude that the startups presenting a higher number of *followers* showed a higher probability of succeeding in crowdfunding projects, meaning their popularity on Instagram helps raise funds. Hadley et al. (2018) conducted a study regarding startups to analyze how their influence and popularity may affect their funding by combining US-based technology startups with venture capitalists and using Twitter as the data source. The authors found that the more central startups in the network, i.e., the most influential ones, received better funding and presented a more significant revenue.

Ruggieri et al. (2018) aimed to identify trends in thriving startups' digital activity. The study indicates that startups predominantly use digital platforms because of their cost-effective functionality. Social media platforms possess a widespread reach, are easy to access, and incur low operating costs, making them the ideal digital marketing gateway for startups to monitor the market. Furthermore, the study inferred that a community of clients and companies, as service providers, is crucial for business success.

Communities are fundamental for a positive impact on digital platforms on the startup, primarily social media communities. Since they provide positive or negative opinions on both the products and the companies. Word-of-mouth is critical in everyday oral communication, creating an impression or idea about a specific subject (Keller, 2007). In the realm of digital platforms, opinions are called electronic word-of-mouth (eWOM) (Hennig-Thurau et al., 2004), and social media are ideal tools for eWOM. Chu and Kim (2011) describe that eWOM enables the creation of a large community, which allows for increased digital engagement via social interactions, such as *comments*, *likes*, *shares*, and *followings*. The large quantity of those interactions might help raise a positive feeling in the social media profile (Wolny and Mueller, 2013).

Following the literature presented, the relationship between startups and social media platforms provides startups with cost-effective digital marketing opportunities to monitor the market and create communities of users and service providers. These communities help spread positive opinions about the products and companies via electronic word-of-mouth, increasing digital engagement through social interactions. Several studies have found that startups with active engagement on social media platforms have a higher chance of succeeding in crowdfunding projects or receiving better funding from venture capitalists.

4.3. Methodology

To visualize each startup's Twitter community, we employed a methodology based on the NAV process model (Hansen et al., 2012). This methodology stresses the need for a heavy interactive process built around the following phases: (1) Define the visualization goal, (2) Collect and structure data, (3) Interpret data, and (4) Report results. Figure 4.2 illustrates the process pipeline, describing each one of the step phases.

As we aim to understand how the *following/follower* relations create communities, in the first step, we defined our goal as that of constructing an informed visualization of Portuguese IT startups' social media communities in Twitter, in alignment with the previously proposed research questions.

The next step consisted of collecting and transforming the social media data extracted via the startups' Twitter accounts into structured data. This study case features the information technology startup active on Twitter, founded by Portuguese, or has headquarters in Portugal. The eight chosen startups are: *attentiveMobile*, *codacy*, *DefinedAi*, *feedzai*, *prodsmart*, *Talkdesk*, *Unbabel*, and *Virtuleap*. The names of the startups are presented using the Twitter account username. *attentiveMobile* is a B2B company that offers a personalized mobile messaging platform; *codacy* is an automated code review platform; *DefinedAi* is a company that develops artificial intelligence training data services and

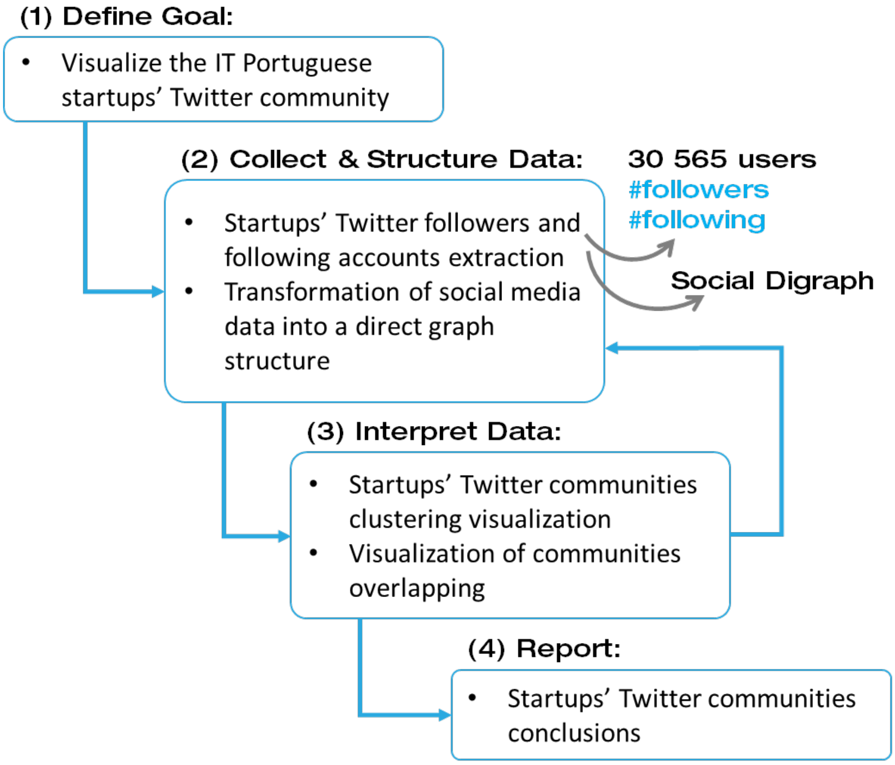

(1) Define Goal:
Visualize the IT Portuguese startups' Twitter community
(2) Collect & Structure Data:
Startups' Twitter followers and following accounts extraction
Transformation of social media data into a direct graph structure
30 565 users
#followers
#following
Social Digraph
(3) Interpret Data:
Startups' Twitter communities clustering visualization
Visualization of communities overlapping
(4) Report:
Startups' Twitter communities conclusions

FIGURE 4.2. Current project's pipeline using the NAV process model (Hansen et al., 2019).

solutions; *feedzai* is an artificial intelligence startup, and its core business is finance risk management; *prodsmart* deals with transforming factories into digital and smart ones by employing automation software to control workflows and production; *Talkdesk* is a platform to support sales teams for customer satisfaction and cost savings; *Unbabel* enables companies to serve customers in their native language with a scalable translation across digital channels; Lastly, *Virtuleap* offers a virtual reality application that promotes brain health, supported by a library of games designed by neuroscientists.

Then, we extracted data from the Twitter accounts of users who follow the companies and users whom the startups follow. The extracted data format corresponds to the Twitter user object, from which the following features have been considered: *id*, *screen_name*, *followers_count*, and *friends_count*. Subsequently, the data was structured into a social digraph, with each node representing a user and each link denoting a following relationship, thus achieving a dataset with users that follow or are followed by startups.

After the organization of the data into a digraph structure, we recurred to using different visualizations to interpret the data and enable information extraction. In order to visualize the social graph, we employed *Gephi* (Bastian et al., 2009) and defined a layout and community clustering for data interpretation. As displayed in Figure 4.2, steps (2), data structuration, and (3), interpretation, occur in an iterative fashion, where the visualization and respective interpretation may require a different organization of the data to explore emerging insights further. For this case study, this happened mainly when trying to visualize the communities' overlap. In the related literature, no specific visualization for the overlap between communities in a large social graph has been found, which led to a deeper exploration of possible visualization techniques, such as the ones presented in the coming sections, that, in turn, required different data organization.

4.3.1. Dataset

The Twitter API was used to extract relevant data, that is, data from the users that follow the eight chosen startups active on Twitter - *followers* - or users that the startups follow - *following*. The extraction occurred on May 31st, 2022, resulting in 30,565 accounts of Twitter users. Table 4.1 presents the number of *followers* and *following* users for each one of the companies, and Figure 4.3 shows an illustration of the respective percentages in terms of the total number of links (edges) for each company using a stacked bar chart.

TABLE 4.1. Comparison of the startups counts of *followers* and *following*.

Startups	Following	Followers
@attentivemobile	387	6,842
@codacy	283	5,048
@Definedai	170	1,908
@feedzai	920	3,132
@prodsmart	1,037	905
@Talkdesk	685	7,252
@Unbabel	1,116	3,627
@virtuleap	4	877

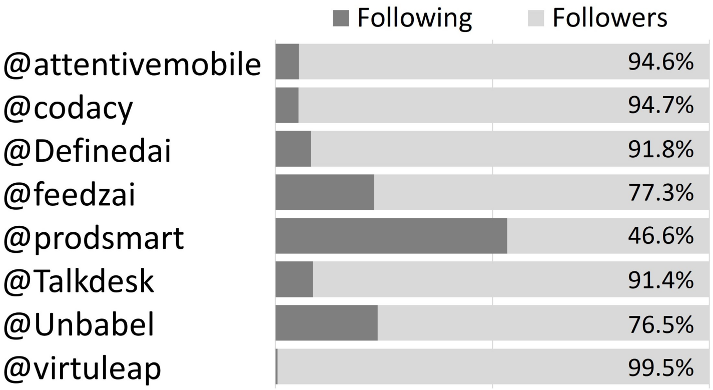


FIGURE 4.3. *Followers* and *following* distribution.

In terms of descriptive quantities, notably, all startups present a higher number of *followers* than of *followings*, meaning that their communities are mostly composed of Twitter users who follow their accounts. The only exception is *prodsmart*, for which the distribution of *followers* and *following*, although being approximately identical, shows that this company mostly follows others. Interestingly, while *Virtuleap* is the startup presenting the smaller community, it also presents an expressively higher percentage of *followers* than of *following*, being the startup showing the highest rate of *followers*, 99.5%, closely followed by *attentivemobile* and *codacy*, with 94.6% and 94.7%, respectively. *Talkdesk* is the startup showing the highest number of *followers* (7,252), followed by *attentivemobile* (6,842), and *codacy* (5,048). *Virtuleap* stands out as the startup with the smallest number of links, likely attributed to its relatively recent foundation year: 2018.

4.3.2. Social digraph creation

After extraction, data has been structured into a social digraph, that is, a directed graph, where a node represents a user and a directed edge represents the user-following-user relation. This action resulted in a graph consisting of 30,565 nodes and 34,184 directed links/edges. The graph's density, 7.32×10^{-5}, indicates that it is a very sparse graph, meaning that it presents very few edges compared to the maximum possible number of edges for this number of nodes. This sparsity was expected since the graph nodes represent mostly users who follow the startups, while information about the other nodes those users may follow or about their followings was not extracted. No weights were used since we have not extracted any quantitative information towards this end.

To enable community visualization, a community detection algorithm was used. As previously mentioned in the related work section, based on the description of the analysis performed by Devi and Poovammal (2016), we chose to use the modularity algorithm for social digraphs created by Leicht and Newman (2008). Modularity measures the density of the connections within a graph's structure and groups it into modules. As expected, the results showed eight modules, one for each startup community, which has been validated by discovering the company at the center of each founded community. To evaluate the results, we measured the modularity score, ranging between 0 and 1, with higher values indicating a stronger community structure (McDiarmid and Skerman, 2020). Our case study graph achieved a modularity value of 0.768, suggesting a robust community structure. We used the Python library *CDLib* (Rossetti et al., 2019) for the algorithm and evaluation.

For our analysis, we have selected a modularity-based algorithm due to its success in these social network scenarios, showing few large communities, even though they usually suffer from the resolution limit problem. However, community detection algorithms are known to be quite unstable, with different algorithms sometimes producing different results. To gauge the stability of our findings, we have also tested an alternate method proposed by Traag et al. (2015) to evaluate the eventual differences that may arise. This alternate algorithm uses asymptotical surprise, a metric that, like modularity, is employed to evaluate the quality of community detection in networks. This metric is a statistical approach that calculates the probability of observing at least a certain number of internal edges within the communities, given the total number of edges in the network. We choose to apply this algorithm because it is nearly unaffected by the resolution limit problem, the modularity optimization primary weakness. The results with the new algorithm are mostly identical to the previous results, described in Appendix D, and with few large communities. We decided to carry on using the results obtained by the algorithm proposed by Leicht and Newman (2008) for the analysis.

Figure 4.4 shows the number of nodes - users - for the found communities, comparing it with the respective number of *followers* plus *following* users of the startup, i.e., the total count of links of each of the companies. We have numbered each modularity class from 1 to 8, representing each startup community. Bearing in mind the quantities presented in Figure 4.4, *Talkdesk* and *attentivemobile* present the larger communities and *Virtuleap* the smallest, as expected. Interestingly, the number of nodes for each community is lower than the sum of each startup's number of *followers* and *followings*. This indicates that some users are shared between communities, meaning that the users that follow or are followed by the startups may intersect. *Virtuleap* appears as an exception, presenting an identical number of linked users and community nodes: 881 linked users and 873 nodes in the community. This means that only eight users are shared with different companies.

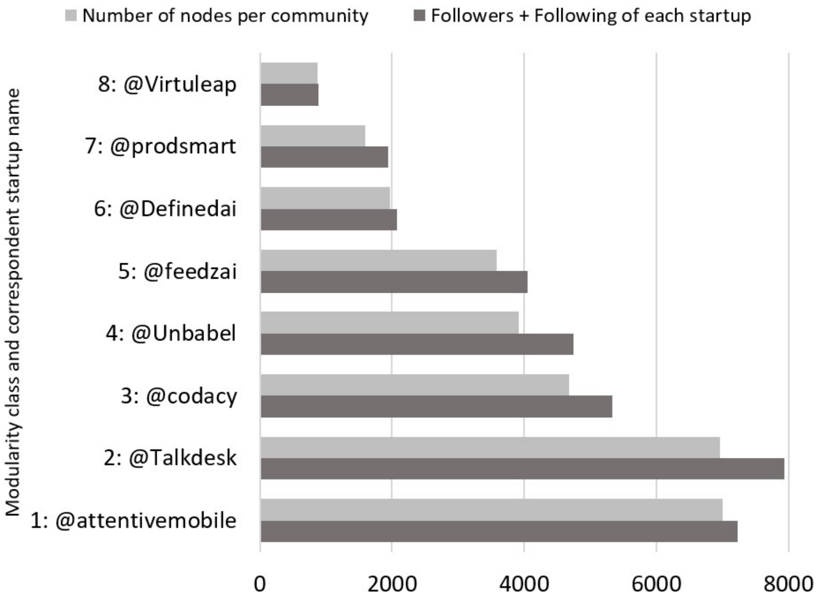


FIGURE 4.4. Communities size distribution.

4.4. Data Visualization & Interpretation

This section presents visualizations of the social Twitter communities built around the different startups, focusing on the extent and characterization of the overlap between them. Overlap in this context means that a user follows more than one of the startups or is followed by more than one startup. To extract knowledge that may inform the creation of social media marketing strategies, We examined information at the user level for the ones found in an overlap situation to understand their type profiles and characterize the general communities of *followers* and *following*.

4.4.1. Social digraph visualization

In a first step towards informative visualizations, the *circle pack layout* (Groeninger, 2015) was applied to the social digraph to perceive what and how the communities are outputted by the modularity algorithm, as described in the previous section. This layout organizes the network into circles using a selected set of network features. In this case, the only selected feature has been the modularity class, which allowed us to create a visualization where each circle represents one of the communities. To help interpret the graph, the nodes and links were colored using different colors. Since the central (centroid) of each formed cluster was found to be each of the startups, the coloring rule used a color based on the logotype of each company to color the nodes corresponding to their modularity class value. However, the links were colored using a mix of the colors from the source and the target nodes. Figure 4.5 displays the visualization thus obtained for this social digraph, where each circle represents one community. The name of the correspondent startup community central point, the startup, is also shown.

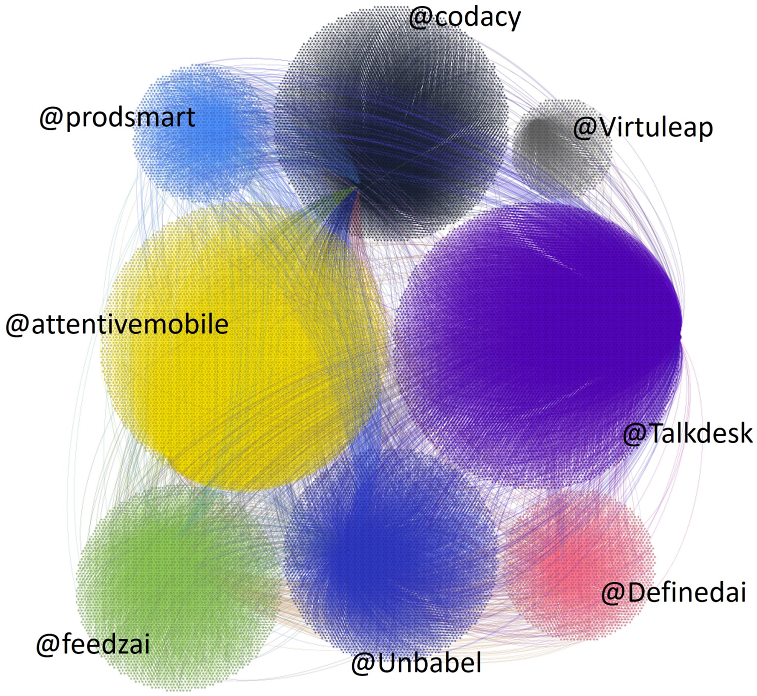


FIGURE 4.5. Social graph visualization: Circle pack layout using the modularity class.

As expected, the graph shows that the communities varying sizes, with the *Talkdesk* community showing its more significant number of members and the *Virtuleap* community the smallest, as seen in Figure 4.4. Furthermore, as anticipated, the graph shows a yet significant overlap between the several communities, with nodes (members) connected to more than one community in the social network graph. However, the exact degree of overlapping cannot be determined from this particular visualization alone, and additional analysis is needed to assess the implications of this overlap for the startups involved.

Understanding the degree of overlapping and how it may or may not differ in terms of *followers* and of *followings* is essential since this distinction may have meaningful implications regarding actions in a startup's social media strategy.

4.4.2. Overlap of communities

Understanding the communities' common points may help startups expand their networks and gain a competitive advantage in their respective markets. The following section aims to answer this question and grasp how the communities overlap. The overlap between communities in the context of social media refers to the situation where some users in the community around one startup are also part of other communities established around different startups. In other words, as depicted by the edges traversing different communities in Figure 4.5, these users follow more than one startup or are themselves followed by more than one startup.

To distinguish between the two different situations - when a startup follows a user or when a user follows a startup - we divided the network in two: one representing only the user-*follow*-startup relation and the other representing the startup-*following*-user relation. The resultant graphs' dimensions show the difference in number between these two features that have been previously noticed in the analysis of Figure 4.3: while the graph with the startups' *followers* has 27,682 nodes and 23,270 edges. The one for the users, followed by the startups, consists of 4,306 nodes and 4,225 edges.

However, the visualizations must be more comprehensive to better understand the overlap between startups. Data grouping is required since we discovered that some of the users in the overlap are shared between two or more communities. To accomplish this, we created two matrices: one for the *followers* of the startups and the other for the users the startups are *following*. The users in common for each combination of startups were counted, resulting in an overlap of 1,289 *followers* (Figure 4.6) and one with 249 *following* users (Figure 4.7). The Python library *upsetplot* (Lex et al., 2014) was used to visualize the matrices. The resultant plots represent the overlap in both of the domains: Figure 4.6 plots the *followers* the different combinations of startups have in common and how many, and Figure 4.7 regards combination and counts for startups *following* shared users.

In fact, Figure 4.6 is a subplot of the total visualization of the overlap with *followers*. Since a considerable number of users was found to overlap, for a more effective visualization, the data were filtered by applying a threshold to show values only when the number of users in common was above nine. As seen in the previous section, *Virtuleap* displays only eight users in an overlap situation, and thus *Virtuleap* appears not to be sharing *followers* with other startups in this scenario where the threshold was applied. Notably, the startup sharing most *followers* with other startups is *Unbabel*, which shares a total of 667 *followers* with all the other six startups, even if under different combinations. The one sharing the least is *attentivemobile*, with 63 shared *followers*. Furthermore, *attentivemobile* only overlaps with the startups showing the three highest numbers for overlapping. Interestingly, all these shares are duets that is these *followers* follow only two companies:

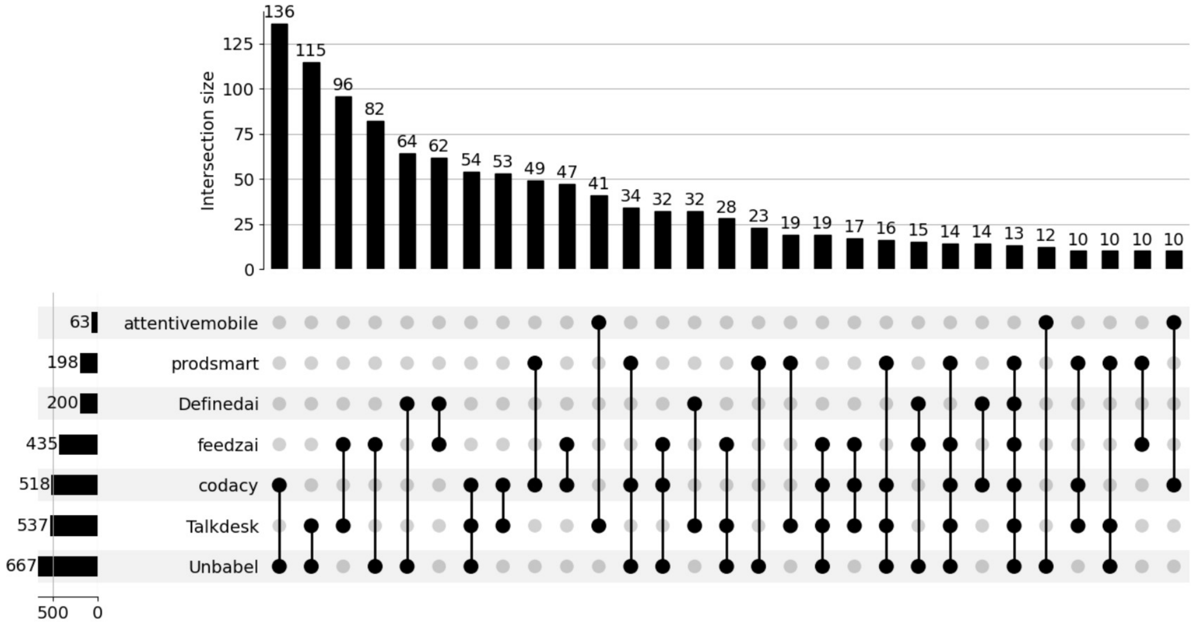


FIGURE 4.6. *Followers* overlap visualization.

Talkdesk (41 shared *followers*), *Unbabel* (12 shared *followers*) and *codacy* (10 shared *followers*). Nevertheless, this is an unexpected result since *attentivemobile* displays the biggest counts, both in its community dimension and for the total number of linked users. The pair of startups with more *followers* in common is *codacy* and *Unbabel*, showing an overlap of 136, followed by Talkdesk and *Unbabel* with 115, and *Talkdesk* and *feedzai* with 96. The visualization presents eight trios, with the one sharing more *followers* composed by *Unbabel*, *Talkdesk*, and *Codacy*, with a total of 54 users in common. Additionally, we can observe two quartets (with 19 and 16 users), one quintet (with 14 users), and one sextet (with 13 users).

Next, the shared *following* users have been analyzed. Figure 4.7 displays the overlap of the *following* by the startups, which may express coincident digital marketing options between them. Knowing the users in or not in the overlap can help to direct a digital marketing strategy. Therefore, we decided to categorize the 249 users in the overlap manually. Since the categorization was manual, we chose a set of global and vague categories of the startup ecosystem to facilitate manual categorization. The annotation procedure looked at the user's profile and bio description (usually stating the type of profile) and searched in Google for verification if needed. This annotation resulted in categories and the colored visualization shown in Figure 4.7. Similarly to what has been done with the *followers*' scenario, we applied a threshold and considered values only above two shared *following* relations. Again, *Virtuleap followings* do not appear in an overlap situation since this startup only follows four different users.

Both the startups showing the higher and the lowest levels of overlap are still the same: *Unbabel* shares more *following* users (192 users shared in total) and *attentivemobile* shares the least (26, in total). Notice that the latter has exchanged its behavior: while

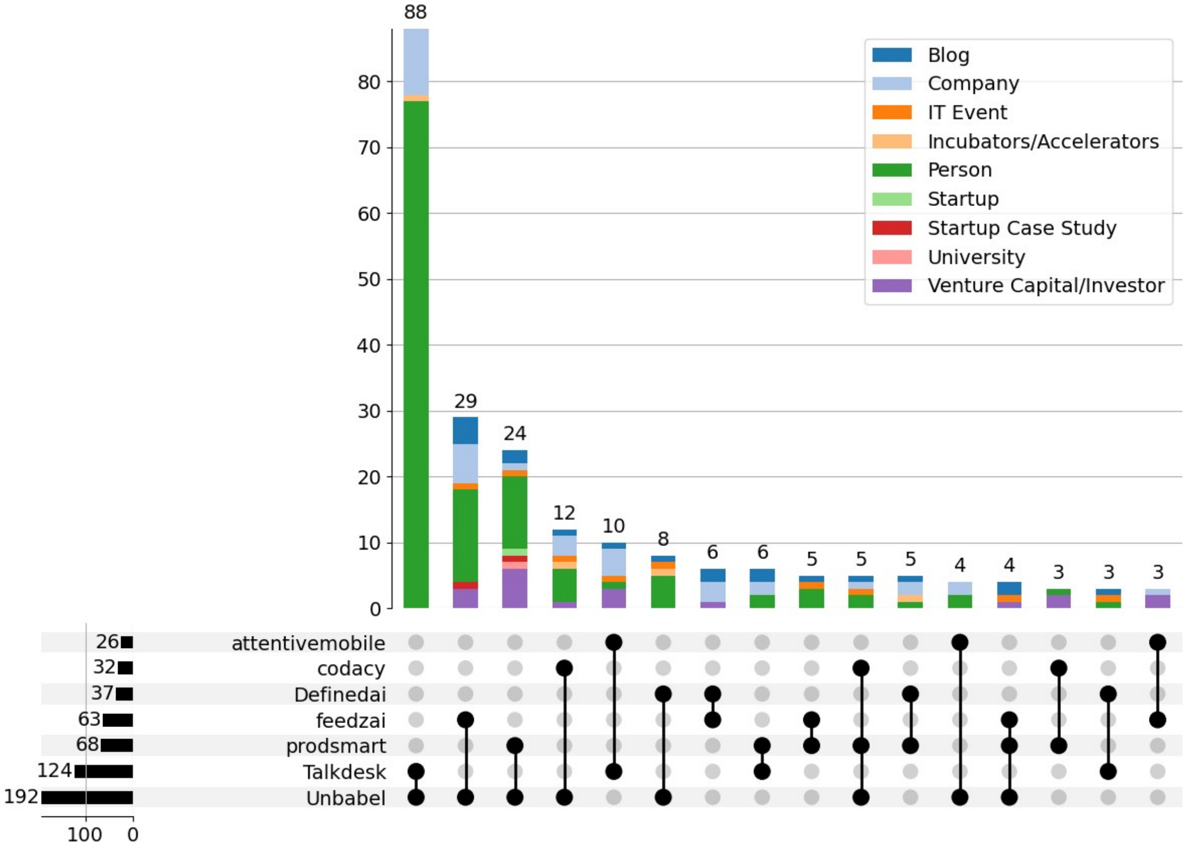


FIGURE 4.7. *Following* overlap visualization, including user categories.

it shared many *followers* with other startups in the previous analysis, it now differentiates by following different users. By the analysis of Figure 4.7, we can also conclude that *Talkdesk* and *Unbabel* are the pair presenting the highest number of *followings* in common, 88 users. Interestingly, most of the overlap occurs among Twitter users that are from the category “Person”, primarily experts in the core business field of these two startups: applications encompassing natural language processing. The next profile of common *followings* are “Company” and, naturally, “Incubators/Accelerators.” The following pairings and groups of startups show much less following in common, as can be noted by the abrupt decrease shown from the second column of Figure 4.7. We can see two trios: one consisting of five shared users and the other sharing four. The trio sharing more *following* relations - *codacy*, *prodsmart* and *Unbable* - consists of: the CEO of *codacy* (a “Person”), a Portuguese journalist (a “Person”), the Lisbon Investment Summit (an “IT Event”), *beta-i* (a “Company”), and a Portuguese blog (a “Blog”).

4.4.3. Overlapping users characterization

In this section, we study the users found in overlapping communities. Understanding who startups follow and who follows them is important to characterize the overlap better.

Regarding the *followers* overlap, we found it appealing to understand what users follow most of the startups. These users are those following four or more companies, comprising

68 users. Next, node-level metrics were applied to evaluate their activity and popularity levels. These metrics were used for the 68 users selected in the *followers* and the 249 users selected in the *following*. Concerning user activity, we retrieved the total number of tweets and retweets in their Twitter profiles. We used a version of the *Structural Advantage* for popularity, the FF ratio (Cappelletti and Sastry, 2012), that involves the number of *followers* and of *followings*:

$$\text{FF Ratio} = \frac{\#followers}{\#followers + \#following}$$

The FF ratio indicates how popular a user is on the social media platform, with higher ratios indicating that a user has more *followers* than is *following* others. Values between 0 and 0.5 indicate that the user is not particularly popular, following more users than being followed.

Concerning the *followers*' overlap, Table 4.2 shows the percentages for each type of Twitter user that has been encountered in this set of shared users and also the corresponding average values for the FF ratio and activity of each of the types.

Table 4.2. *Followers*' overlap: percentage of profiles encountered and node-level measures.

User type	n	%	FF Ratio (average)	Activity (average)
Person	37	56.9%	0.27	2,245
Blog	7	10.8%	0.44	2,450
Incubator/Accelerator	7	10.8%	0.65	3,988
Company	4	6.2%	0.38	2,216
Venture Capital/Investor	4	6.2%	0.54	989
IT Event	3	4.6%	0.60	1,184
Startup	3	4.6%	0.34	287

More than half of the users are classified as “Person,” accounting for 56.9% of the total, followed by “Blog” (10.8%), “Incubators/Accelerators” (10.8%), “Company” and “Venture capital/Investor,” both showing 6.2%. The categories showing the least are “IT event” and “Startup”, with 4.6%. The type “Incubators/Accelerators” displays the highest average FF ratio (0.65), followed by “IT event” (0.60), “Venture capital/Investor” (0.54), all above-average level of popularity. The remaining types show less favorable FF ratios, especially the type “Person”, which shows an average FF ratio of 0.27.

Regarding the activity levels, “Incubators/Accelerators” shows the highest average levels (3,988), indicating that this type of user is more engaged with their *followers* than the remaining user types. Next, we see “Blog”, with 2,450, “Person” (2,245), and “Company” (2,216), all presenting similar levels of activity. Finally, we have “IT Event” (1,184), “Venture capital/Investor” (989), and “Startup” (287).

In terms of the overlap that exists in the *following* relations, that is, the users that startups follow, Table 4.3 presents the percentage of each type of user in this overlap, as well as the average FF ratio and the activity counts. The majority of the users followed

TABLE 4.3. *Followings*' overlap: percentage of profiles encountered and node-level measures.

User type	n	%	FF Ratio (average)	Activity (average)
Person	132	53%	0.71	28,999
Company	45	18.1%	0.87	26,494
Blog	26	10.5%	0.95	107,924
Venture Capital/Investor	21	8.5%	0.84	13,563
IT Event	11	4.4%	0.74	18,759
Incubator/Accelerator	7	2.8%	0.74	3,216
Startup	3	1.2%	0.84	8,459
Startup Case Study	3	1.2%	0.88	7,752
University	1	0.4%	0.95	6,664

by the startups are "Person", accounting for 53% of the total, followed by "Company" (18.10%), "Blog" (10.50%), "Venture capital/Investor" (8.50%), "IT Event" (4.40%), "Incubators/Accelerators" (2.80%), and "Startup" (1.20%). In this profiling, we can also encounter the type "University" (0.40%), albeit showing the least number of followings. The specific university is Instituto Superior Técnico, Lisbon, Portugal, from which many of the case study startups are either spin-offs or from where their founders obtained their degrees.

Comparing these users' activity levels, we see that the type "Blog" shows the highest average activity level (107,924 tweets/retweets), consistent with this type's main function. A "Blog" engages with its audience by regularly sharing meaningful content, which also explains why this is one of the most popular user types in this overlap. When comparing this set of blog users with one of the *followers* previously discussed, the averages now are considerably higher than before, which entails that the blogs followed by startups are respected and credited blogs in this ecosystem. The following ranking position in terms of activity is occupied by the type "Person" that, with a 28,999 average count of tweets/retweets, positions itself somewhat distant from "Blog". "Company", showing an average of 26,49 showing an average of 26,494, follows closely. All the remaining types show considerably less activity when compared with any of the previous ones. It should be noticed, however, that all these user types display higher activity levels on average than those found in the *followers* set. Namely, the lowest activity count for the *followings* - 3,216 average tweets/retweets (Table 4.3) - is still higher than the highest count for the *followers* (Table 4.2).

In terms of popularity, we can see that the user types with the highest FF ratios are "Blog" and "University" (both attaining 0.95), followed by the startups of this case study (0.88), whose popularity levels are close to "Company" (0.87), again close to "Venture capital/Investors" (0.84) and "Startup" (0.84) popularity levels. On the other hand, user types "IT Event", "Incubators/Accelerators," and "Person" show the lowest FF ratios, but still, all of these users can be classified as popular since any of them shows to have more *followers* than *following* others.

Notably, our case study startups mostly share the actions of the following persons and companies relevant to the startups' ecosystem. Other *following* types are "Venture Capital/Investor" and "Blog." While both show a relatively high FF ratio, indicating popularity, the average activity levels differ. "Venture Capital/Investor" activity is not expressive, which may indicate that these users look upon Twitter more as an observational or promotional tool than with the intent of engaging with other users. "IT Event" and "Incubators/Accelerators" show the lowest average FF ratios and activity levels, suggesting these users albeit being both quite popular, are not as active on this Twitter ecosystem as the remaining types.

Significantly, the categories of users found in the studied overlaps all belong to the startup universe, highlighting the interconnectedness of this ecosystem. Upon examination of the categorization of users in both overlaps, it becomes evident that most users who follow startups and those whom startups follow are individuals. The second most common user type that startups follow is companies. Furthermore, startups follow users showing higher activity levels, with an average of 19,669 tweet/retweet count, which strikingly compares with those users that follow startups, showing expressively lower activity levels. In terms of popularity, startups follow users that are more popular than those who follow them, with an average FF ratio of 0.46 among *followers*, compared to an average of 0.84 among those whom startups follow.

4.4.4. Startup's popularity level over FPEMv2 life cycle phases

As stated by Antonakaki et al. (2021), various methods exist for measuring users' activity, popularity, and influence on social media platforms such as Twitter. The authors described that popularity measures how well other users recognize a user. To analyze if and how the popularity levels of the startups differ between the several life cycle phases in the FPEMv2 model (Chapter 2), we applied the FF ratio used in the previous section to a set of startups' Twitter profiles. We used the 38 startups from approach (B) analysis in Chapter 3. This set comprises 19 startups in the seed phase, 11 in the early phase, and 8 in the late phase. We will only analyze the three available phases since we do not own data featuring startups in the FFF or preseed phases.

To analyze if the startups' popularity levels differ according to their life cycle, we tested if there was a relationship between the FF ratio and the phase with the application of the Kruskal-Wallis test using the *SciPy* (Virtanen et al., 2020) library. We placed the significance threshold at 0.05. The null hypothesis in this scenario is H0: The means for each life cycle phase are identical. If the *p-value* is lower than the threshold, we reject the null hypothesis, meaning that the means on every life cycle differ. The test resulted in a *p-value* of 0.003, lower than the threshold, proving a statistically significant relationship between the FF ratio and the startup phases.

The graph in Figure 4.8 shows that startups are more prevalent in the late phase, with FF ratios between 0.76 and 1, with an average value of around 0.94. In this phase, the startup already has a developed and mature product and is established on the market.

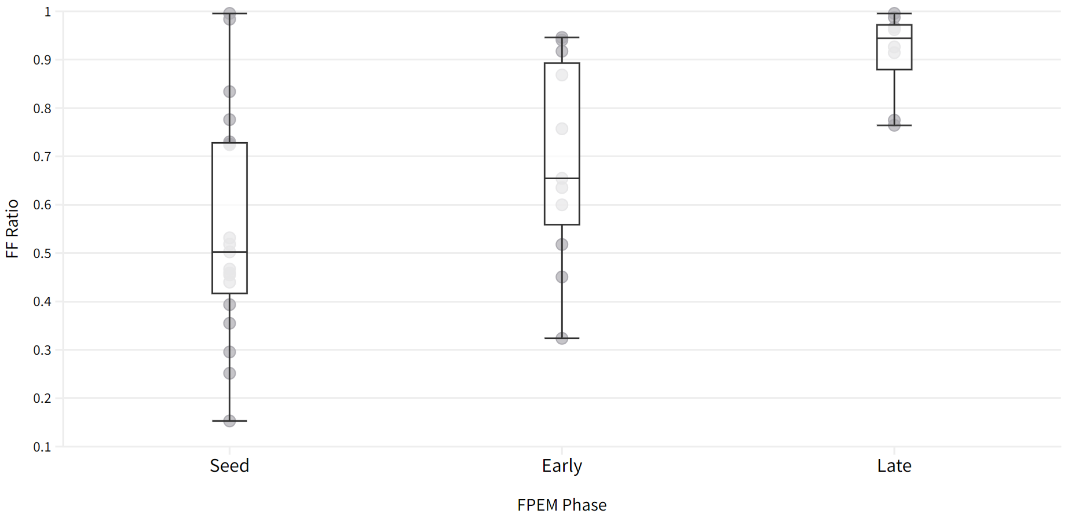


FIGURE 4.8. Level of popularity of the startup (FF ratio) on the life cycle phases.

It is natural that when startups reach the late phase, they already maintain a stabilized public and customers, which reflects on the following community and consequent value of the FF ratio. Startups in the early phase present, on average, an FF ratio of 0.65, and most startups have values between 0.56 and 0.89, the lower and upper quartile values, respectively. In this phase, startups have a functional product; some are more established on the market than others, reflecting the FF ratio values. In the previous Chapter, we concluded that newer startups post less on Twitter, suggesting they may prioritize other issues, such as making a prototype adjustable to the market. However, our case study startups display different behaviors, with some making less use of social media and others taking advantage of it. This phenomenon is reflected in social media measures like popularity, with values between 1 and 0.15 of the upper and lower extremes, respectively. Nevertheless, in the seed phase, the average value of the FF ratio is 0.5, and the upper and lower values of the quartiles are 0.73 and 0.42. Therefore, as startups grow, their Twitter account reach more users and, consequently, gain more popularity. In conclusion, these results support the claim that the startups' popularity level differs with their life cycle phases.

4.5. Summary

Active engagement on social media can lead to better funding and create communities of users and service providers, enabling startups to achieve their goals and expand their business. This work investigates how startups fare on social media platforms like Twitter and if they create their own communities. Additionally, we analyzed the social media popularity of the startups resulting from those communities, investigating if the popularity levels vary in accordance with the FPEMv2 life cycle phases.

Our primary research question was to understand if the *follower/following* relations on Twitter's social graph create social communities around startups. Using Portuguese IT startups as a case study, we collected and treated their Twitter data to create meaningful visualizations, enabling us to extract relevant knowledge about the communities. Using eight IT startups having some connection with Portugal, the case study data was organized into a social digraph, representing the users and links between the different users found in the data, resulting in a graph with over 30,000 nodes. Applying a community detection algorithm enabled the identification of communities. As expected, the results showed that the communities were built around the eight chosen startups. By encoding the color of the social graph in different colors for different startups, the created visualizations highlighted each of the startup's community of users and allowed for the characterization of these communities. With this methodology, we showed that IT Portuguese startups form their social communities on Twitter. Next, we used other types of visualizations, paired with the users' manual annotation and node-level metrics, that enabled us to characterize the found communities and find out that these communities show an interesting degree of overlap between them, either from the perspective of the startups' *followers*, as from the perspective of whom the startups are *following*.

An overlap occurs when users belong to multiple communities established by different startups. We discovered how the overlaps between communities on social media are organized. Presenting two different graphs, one for *followers* and one for *following*, we analyzed the specific overlap between the communities. The resultant plots provide a comprehensive understanding of the social digraphs of the *follower* and *following* users of the startups. For a better understanding of these communities, we built two visualizations representing each one of the overlaps. The resulting visualizations fully represent the startups' communities *followers* and *followings* overlaps. After a manual annotation of the users present on the overlaps, we discovered that all these categories of users are relevant to the startup universe, highlighting the interconnectedness of this ecosystem. Examining the user categories in both overlaps revealed that most users who follow startups and whom startups follow are persons. Companies represent the second most prevalent user category that startups choose to follow, whereas blogs and incubators/accelerators are the next categories of users that follow most startups. In addition, startups tend to follow users who post high volumes of tweets and display high popularity levels. On the other hand, those who follow startups present lower activity levels and low popularity values.

Finally, since the measure for popularity comes from a ratio involving the size of the *followers* and *following* communities, we aimed to understand if startups display different popularity values throughout the several FPEMv2 life cycle phases. The results showed that the startup's popularity level does differ between their life cycle phases. As startups grow, their Twitter accounts tend to reach more users and gain more popularity.

CHAPTER 5

Conclusion

5.1. Global conclusions

Literature states that the definition of startup depends on the actual phase it traverses (Skala, 2019). The startup's interaction with the surrounding ecosystem impacts its growth and development and varies with the company's maturity level (Cukier and Kon, 2018). As small and new businesses, startups have difficulty reaching customers and acquiring financing (Wang et al., 2016). Social media platforms can be valuable tools to help small businesses, like startups, tackle those challenges (Rizvanović et al., 2023). However, to help startups build social media strategies, it is crucial to understand what characterizes the phase where they are since the strategy to be used will depend on that phase.

A systematic literature review allowed us to perceive and describe the main startup ecosystem elements and the factors that affect the company's development, supporting a novel life cycle five-phase framework - the FPEMv2. The model presents an integrative and holistic view of all the different proposals in the literature. Using intelligence tools and social media analysis, it was possible to understand how the focus of a startup changes along its evolution and thus extract insightful knowledge about the different social media platforms' usage.

Figure 5.1 summarizes this dissertation's global conclusions by characterizing the FPEMv2 with the social media analysis data. Besides including the FPEMv2 with both dimensions (finances and product evolution), it integrates an important part of the content and the network analysis results that illustrate the different Twitter usage and startup behavior in each of the five phases.

Starting by employing a topic model approach for the social media content data analysis in text format of a small dataset of eight startups with 15,577 tweets. This initial work allowed for consistently dividing the tweets into five distinct topics. Moreover, the topic relative frequencies per startup were similar and varied with the position of the startup in FPEM. This discovery allowed us to conclude that the startups show identical posting behavior in similar life cycle phases. After this proof of concept, a larger-scale study with 38 startups, with 91,743 tweets, was performed using topic modeling divided per FPEMv2 phase. We concluded that the tweets the startups posted changed in quantity and content according to the company's growth. The startups post less in the first two phases, FFF and pressed, and the content is concerned mainly with product research and development. In the third, the seed phase, the focus goes on IT content, while in the early phase, the next one, the topic is about the product. Lastly, in the late or fifth

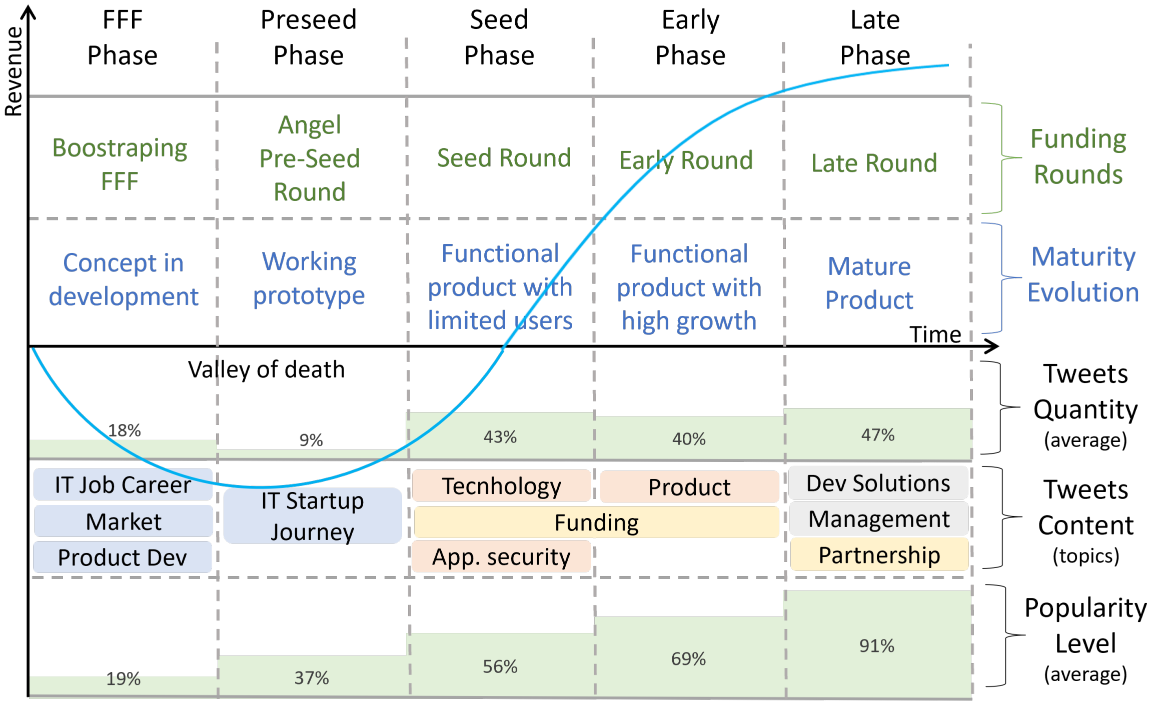


FIGURE 5.1. Characterization of the FPEMv2 life cycle phases through social media analysis.

phase, startups post mostly about management matters. Interestingly, over the life cycle, companies always post about funding matters, although with less emphasis when in their first two cycle phases.

Concerning the network data and using the smaller dataset composed of the eight startups, it was possible to perceive that startups build communities but also overlap. Using visualization techniques combined with a network modularity algorithm (Leicht and Newman, 2008) and profiling, we discovered that most of the users who follow the startups and whom startups follow are persons related to their core business or from the startups' ecosystem. Companies represent the second most prevalent user category that startups choose to follow, whereas blogs and incubators/accelerators are the next user categories that follow most startups. Notably, startups mostly follow users who post high volumes of tweets and display high popularity levels. Companies display different popularity values throughout the FPEMv2 life cycle phases, and, as expected, as startups grow, the Twitter account tends to reach more users and gain more popularity. This increase in popularity means a growth of their followers' communities, indicating that, as startups mature, their reach in social media platforms expands.

Our dissertation's first objective was to explore data science instruments and methodologies to extract valuable information on the startups' social media usage. We explored textual analysis, social community detection, and network visualization techniques that enabled the insightful characterization of the startups' life cycle phases. After applying the techniques to historical social media data, the second primary objective was to discover if providing digital marketing guidelines to improve the startup's strategies from the results is possible. Characterizing the different phases of the startups throughout their development, along with their presence on social media, enables the creation of personalized social media strategies based on the company's current stage, thus improving their strategies. The results provided new and complementary information for each of the life cycle's phases and served as validation for the novel FPEMv2 model.

5.2. Answers to the research questions

This dissertation aims to answer specific research questions per chapter since each uses different data types and applies distinct methodologies that require specific related work and different results analyses.

In Chapter 2, we conducted an SLR to discover the state of the art regarding the IT startups' life cycle. Namely, we formulated the following questions:
RQ2.1: What factors influence the IT startup's development and growth along its life?
RQ2.2: What models aiming to characterize the life cycle are defined for IT startups?

From the SLR, it is clear that startups face a significant number of challenges that increase their risk of failure and, thus, influence their growth (or their fall). Among the several factors that negatively impact the long-term viability of the startup, the relevant ones that were found are:

- Conflicts between co-founders and investors (Bala Subrahmanya, 2022).
- Lack of expertise in commercial and technological domains, essential to understanding the discrepancy between perceived skills and business growth expectations (Szerb and Vörös, 2021).
- Lack of resources and difficulties integrating knowledge (Almeida, 2021).
- Neglecting activities related to documentation creation Maria et al. (2017).

On the other hand, the following factors are essential to increase the probability of startup success:

- Apply innovation (Mirghaderi et al., 2023; Almeida, 2021; Liotino et al., 2016; Lago et al., 2023).
- Adaptation to the environment (Ehsani and Osiyevskyy, 2023).
- Effective communication internally in the company and externally with the customers (Morales-trujillo and García-mireles, 2019).
- Social media presence (Hervet and Guitart, 2022; Gloor et al., 2020).
- Partnerships (Nobari and Dehkordi, 2023; Noelia and Rosalia, 2020; Bala Subrahmanya, 2022).
- Choosing a robust business model aligned with the specific life cycle stage (Ruseva, 2015).
- Incubators, accelerators and public financing (Page and Holmström, 2023; Yusupova and Ryazantseva, 2022).

- Evaluate their operations periodically (Gbadegeshin et al., 2022).

Regarding RQ2.2, while different authors divide the startups into differently phased life cycle models, the consensus remains that startups change their goals according to the phase they are in. Beginning as small companies with no financial support and no product to sell, just an idea, if successful, they grow at a high rate until they mature and become financially stable and with a well-established product in the market. Based on historical data on funding rounds and product maturity evolution, this two-fold view of startup growth is captured and integrated into the novel FPEMv2 model.

Once the FPEMv2 life cycle model was defined and we understood that, as startups grow, their focus changes, we aimed to know if this change affects or is reflected by the company's social media activity, which led to the two questions raised in Chapter 3:

RQ3.1: Does the startups' Twitter content change according to their life cycle phases?

RQ3.2: If the content changes, can we use the topics created from it to characterize the startups' life cycle phases?

A topic modeling approach on a small-scale dataset led to the emergence of five topics. Considering the FPEM life cycle model, these results were explored, revealing that the startups' Twitter topics change over time according to their current phase, answering affirmatively to RQ3.1. A topic modeling on each life cycle model's phases allowed for an answer to RQ3.2, resulting in five different models. In this approach, since we used a more extensive set of startups and grouped the tweets by phase, the topics portray finer details and enable the characterization of the FPEMv2 model phases.

Startups actively perform two actions on social media platforms: posting content and following other accounts, raising a set of new questions. Since the following action creates a network, the social graph, in Chapter 4, we aimed to understand if this results in community formation.

RQ4.1: Do startups form their social communities on Twitter?

RQ4.2: In the case of community formation, are they disjoint communities, presenting different (types of) users?

We visualize the social graph by performing network analysis and visualization using the data from eight startups. We discovered that each startup formed a community of users, answering affirmatively to question RQ4.1. Through the visualization, we found that the diverse communities shared users, and thus the communities overlapped. Selecting only the users in the overlaps and performing a manual annotation of the users, we discovered that all these categories of users are relevant to the startup universe, highlighting the interconnectedness of this ecosystem. Examining the user categories further revealed that most users who follow startups and those whom startups follow are persons. In addition, startups follow users who post high volumes of tweets and display high popularity levels. On the other hand, those who follow startups present lower activity levels and low popularity values. This leads to the third research question:

RQ4.3: Does the startup's popularity level change over its life cycle phases?

The results showed that a startup's popularity level does differ between the several life cycle phases as observed in Figure 5.1: as startups grow, their Twitter accounts reach more users and gain popularity.

5.3. Contributions and Implications

The theoretical contributions of this dissertation are the following. Firstly, a systematic literature review on the life cycle of IT startups that reviews all the elements of the startups' ecosystem.

Secondly, it is a novel life cycle model for startups consolidating the systematic literature review's results, the FPEMv2. Other types of analysis regarding elements in the startup ecosystem can apply this proposed model to understand the startup growth's effect on them.

Thirdly, a new methodological process based on social media analytics using text mining to extract topics of tweets and a respective visualization technique that enables understanding the topics.

Fourthly, a social media community analysis methodological approach based on Network Analysis and Visualization. The presented methodology can be used on social media platforms like Twitter to analyze social communities overlap.

The managerial contributions that create implications for the startups' social media marketing are next described. Firstly, since the FPEMv2 represents the typical startup's success life cycle, it can be applied so that social media strategies are created depending on the current phase of the startup. Moreover, it can help monitor what should be happening in the company's accounts.

Secondly, the topic modeling approach previously described enabled the summarization of topics of the social media content. Based on the historical data results, this tool can be helpful for startups to understand their social media message and that of competitors.

Thirdly, using the approach employed in Chapter 3.3 startups can also monitor their popularity levels and understand what type of users they are following and that the company is following. The startup should follow users of categories relevant to the startup's ecosystem. Startups can expand their networks by following the same types of users, inserting themselves in the same communities, and gaining a competitive advantage in their respective markets.

5.4. Limitations and future work

Like any research, this dissertation had its limitations. First, the case study focus on a sample of Portugal-based (or related) IT startups. Future research should study startups from other industries and across different countries to get (I) a better general understanding of startups as a particular type of enterprise but also (ii) to perceive the specificity of the results throughout the different areas. Depending on the previous results, even for a particular area of business, a more comprehensive study (encompassing larger samples) should bring added value for the complete characterization of the respective ecosystems and their governance. Regarding the network data analysis, we developed a proof of concept. Future work should also employ a larger sample of startups encompassing companies based in other countries. Similarly, the methodology can address diverse areas for a more comprehensive characterization.

Secondly, this work relies only on what was academic publicly available Twitter data. Future studies should use data from other social media platforms, like LinkedIn, not only because Twitter no longer has free API access, but also to understand if dissemination strategies are similar whichever media platform is being used. It will enable understanding of whether they post and create communities on other platforms differently or if complementary knowledge emerges.

Thirdly, we manually performed the user categorization, enabling us to understand the types of community users overlap. However, if larger in scale, future studies should automatize the categorization process to allow the scalability of the process.

Lastly, this dissertation focuses on the content data posted by the startups and the social graph they created with the following action. This data is the one created directly by the startups. Therefore, to understand the impact of that data, future work should consider more social variables based on the interactions of other users, like replies, mentions, and likes.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2022). Topic modeling algorithms and applications: A survey. *Information Systems*, page 102131.

Abdelsadek, Y., Chelghoum, K., Herrmann, F., Kacem, I., and Otjacques, B. (2018). Community extraction and visualization in social networks applied to Twitter. *Information Sciences*, 424:204–223.

Adler, P., Florida, R., King, K., and Mellander, C. (2019). The city and high-tech startups: The spatial organization of Schumpeterian entrepreneurship. *Cities*, 87(January):121–130.

Akhavan, M., Sebt, M. V., and Ameli, M. (2021). Risk assessment modeling for knowledge based and startup projects based on feasibility studies: A Bayesian network approach. *Knowledge-Based Systems*, 222:106992.

Alash, H. M. and Al-Sultany, G. A. (2020). Improve topic modeling algorithms based on twitter hashtags. *Journal of Physics: Conference Series*, 1660:012100.

Albats, E., Fiegenbaum, I., and Cunningham, J. A. (2018). A micro level study of university industry collaborative lifecycle key performance indicators. *Journal of Technology Transfer*, 43(2):389–431.

Almeida, F. (2021). Open-innovation practices: Diversity in portuguese smes. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(3):169.

Almotairy, B., Abdullah, M., and Abbasi, R. (2020). The impact of social media adoption on entrepreneurial ecosystem. *Emerging extended reality technologies for Industry*, 4:63–79.

Alotaibi, B., Abbasi, R. A., Aslam, M. A., Saeedi, K., and Alahmadi, D. (2020). Startup Initiative Response Analysis (SIRA) Framework for Analyzing Startup Initiatives on Twitter. *IEEE Access*, 8:10718–10730.

Antonakaki, D., Fragopoulou, P., and Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164(September 2020):114006.

Antonakaki, D., Ioannidis, S., and Fragopoulou, P. (2018). Utilizing the average node degree to assess the temporal growth rate of Twitter. *Social Network Analysis and Mining*, 8(1).

Azinhaes, J., Batista, F., and Ferreira, J. C. (2021). eWOM for public institutions: application to the case of the Portuguese Army. *Social Network Analysis and Mining*, 11(1).

Bala Subrahmanya, M. H. (2022). Competitiveness of High-Tech Start-Ups and Entrepreneurial Ecosystems: An Overview. *International Journal of Global Business and Competitiveness*, 17(1):1–10.

Barry, A. E., Valdez, D., Padon, A. A., and Russell, A. M. (2018). Alcohol Advertising on Twitter—A Topic Model. *American Journal of Health Education*, 49(4):256–263.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bauer, D., Junge, S., and Reif, T. (2023). May the resources be with you: a systematic review and framework of startup funding options. *Management Review Quarterly*, (0123456789).

Berg, V., Birkeland, J., Nguyen-duc, A., Pappas, I., and Jaccheri, L. (2018). Software Startup Engineering: A Systematic Mapping Study Vebjørn. *The Journal of Systems & Software*.

Berg, V., Birkeland, J., Nguyen-Duc, A., Pappas, I., and Jaccheri, L. (2020). Achieving agility and quality in product development - an empirical study of hardware startups. *Journal of Systems and Software*, 167.

Bhargava, H. K. (2014). Platform technologies and network goods: Insights on product launch and management. *Information Technology and Management*, 15(3):199–209.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

Campos-Domínguez, E. (2017). Twitter y la comunicacíon política. In *El profesional de la información*, pages 785–794.

Cantner, U., Cunningham, J. A., Lehmann, E. E., and Menter, M. (2021). Entrepreneurial ecosystems: a dynamic lifecycle model. *Small Business Economics*, 57(1):407–423.

Cappelletti, R. and Sastry, N. (2012). Iarank: Ranking users on twitter in near real-time, based on their information amplification potential. In *2012 International Conference on Social Informatics*, pages 70–77. IEEE.

Casero-Ripollés, A. (2018). Research on political information and social media: Key points and challenges for the future. *El Profesional de la Información*, 27(5):964.

Castillo-Esparcia, A., Castillero-Ostio, E., and Castillo-Díaz, A. (2020). Los think tanks en España. Análisis de sus estrategias de comunicación digitales. *Revista Latina*, 2020(77):253–273.

Cavallari, L., Romano, S., and Naticchioni, P. (2021). The original sin: Firms' dynamics and the life-cycle consequences of economic conditions at birth. *European Economic Review*, 138(July):103844.

Chammassian, R. G. and Sabatier, V. (2020). The role of costs in business model design for early-stage technology startups. *Technological Forecasting and Social Change*, 157(April):120090.

Choi, H. J. and Park, C. H. (2019). Emerging topic detection in twitter stream based on high utility pattern mining. *Expert Systems with Applications*, 115:27–36.

Choi, J., Yoon, J., Chung, J., Coh, B. Y., and Lee, J. M. (2020). Social media analytics and business intelligence research: A systematic review. *Information Processing and Management*, 57(6):102279.

Chu, S.-C. and Kim, Y. (2011). Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International Journal of Advertising*, 30(1):47–75.

Clark, T. and Muller, P. A. (2012). Exploiting model driven technology: A tale of two startups. *Software and Systems Modeling*, 11(4):481–493.

Cohen, B. and Ernesto Amorós, J. (2014). Municipal demand-side policy tools and the strategic management of technology life cycles. *Technovation*, 34(12):797–806.

Cruickshank, I. J. and Carley, K. M. (2020). Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering. *Applied Network Science*, 5(1).

Cukier, D. and Kon, F. (2018). A maturity model for software startup ecosystems. *Journal of Innovation and Entrepreneurship*, 7(1).

Curiskis, S. A., Drake, B., Osborn, T. R., and Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing and Management*, 57(2):102034.

Curran, K., O'Hara, K., and O'Brien, S. (2011). The role of twitter in the world of business. *International Journal of Business Data Communications and Networking*, 7(3):1–15.

Daradkeh, M. and Mansoor, W. (2023). The impact of network orientation and entrepreneurial orientation on startup innovation and performance in emerging economies - The moderating role of strategic flexibility. *Journal of Open Innovation: Technology, Market, and Complexity*, 9(1):100004.

Das, K., Samanta, S., and Pal, M. (2018). Study on centrality measures in social networks: a survey.

Del Sarto, N., Cruz Cazares, C., and Di Minin, A. (2022). Startup accelerators as an open environment: The impact on startups' innovative performance. *Technovation*, 113(September 2021):102425.

Devi, J. C. and Poovammal, E. (2016). An analysis of overlapping community detection algorithms in social networks. *Procedia Computer Science*, 89:349–358.

Di Vaio, A., Hassan, R., Chhabra, M., Arrigo, E., and Palladino, R. (2022). Sustainable entrepreneurship impact and entrepreneurial venture life cycle: A systematic literature review. *Journal of Cleaner Production*, 378(August 2021):134469.

Doogan, C., Buntine, W., Linger, H., and Brunt, S. (2020). Public Perceptions and Attitudes Toward COVID-19 Nonpharmaceutical Interventions Across Six Countries: A Topic Modeling Analysis of Twitter Data. *Journal of Medical Internet Research*, 22(9):e21419.

Dutot, V. and Mosconi, E. (2016). Social media and business intelligence: defining and understanding social media intelligence. *Journal of Decision Systems*, 25(3):191–192.

Edison, H., Smørsgård, N. M., Wang, X., and Abrahamsson, P. (2018). Lean Internal Startups for Software Product Innovation in Large Companies: Enablers and Inhibitors. *Journal of Systems and Software*, 135:69–87.

Ehsani, M. and Osiyevskyy, O. (2023). Firm failure and the exploration/exploitation dilemma: The role of firm life cycle. *Long Range Planning*, 56(3):102307.

Eloranta, V.-p. (2014). Towards a Pattern Language for Software Start-Ups. In *19th European Conference on Pattern Languages of Programs*, pages 1–11.

Feld, B. and Hathaway, I. (2020). *The startup community way: Evolving an entrepreneurial ecosystem*. John Wiley & Sons.

Felicetti, A. M., Corvello, V., and Ammirato, S. (2023). Digital innovation in entrepreneurial firms: a systematic literature review. *Review of Managerial Science*, (0123456789).

Freisinger, E., Heidenreich, S., Landau, C., and Spieth, P. (2021). Business Model Innovation Through the Lens of Time: An Empirical Study of Performance Implications Across Venture Life Cycles. *Schmalenbach Journal of Business Research*, 73(3-4):339–380.

Fukugawa, N. (2018). Is the impact of incubator's ability on incubation performance contingent on technologies and life cycle stages of startups?: evidence from Japan. *International Entrepreneurship and Management Journal*, 14(2):457–478.

Gabielkov, M. and Legout, A. (2012). The complete picture of the Twitter social graph. *CoNEXT Student 2012 - Proceedings of the ACM Conference on the 2012 CoNEXT Student Workshop*, pages 19–20.

Ganesaraman, K. and Bala Subrahmanya, M. H. (2022). How Conflicts Cause Technology Startups to Fail in India? An Empirical Analysis. *International Journal of Global Business and Competitiveness*, 17(1):40–52.

Gbadegeshin, S. A., Al, A., Ghafel, K., and Mohammed, O. (2022). Overcoming the Valley of Death : A New Model for High Technology Startups. *Sustainable Futures*, 4(April).

Ghezzi, A. (2020). How Entrepreneurs make sense of Lean Startup Approaches: Business Models as cognitive lenses to generate fast and frugal Heuristics. *Technological Forecasting and Social Change*, 161(November 2018):120324.

Gloor, P. A., Fronzetti Colladon, A., Grippa, F., Hadley, B. M., and Woerner, S. (2020). The impact of social media presence and board member composition on new venture success: Evidences from VC-backed U.S. startups. *Technological Forecasting and Social*

Change, 157(February 2019):120098.

Godoy-Martin, F. J. (2022). Las agencias de comunicación ante las nuevas redes sociales. ¿Early adopters o incorporación tardía? *Revista Internacional de Relaciones Públicas*, 12(23):225–244.

González-Cruz, T. F., Botella-Carrubi, D., and Martínez-Fuentes, C. M. (2020). The effect of firm complexity and founding team size on agile internal communication in startups. *International Entrepreneurship and Management Journal*, 16(3):1101–1121.

Groeninger, M. (2015). Gephi - circular layout.

Gulati, R. and DeSantola, A. (2016). Start-Ups that last.

Hadley, B., Gloor, P. A., Woerner, S. L., and Zhou, Y. (2018). Analyzing VC Influence on Startup Success: A People-Centric Network Theory Approach. *Studies on Entrepreneurship, Structural Change and Industrial Dynamics*, pages 3–14.

Hansen, D., Shneiderman, B., and Smith, M. A. (2019). *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann.

Hansen, D. L., Rotman, D., Bonsignore, E., Milic-Frayling, N., Rodrigues, E. M., Smith, M., and Shneiderman, B. (2012). Do you know the way to sna?: A process model for analyzing and visualizing social media network data. In *2012 International Conference on Social Informatics*, pages 304–313. IEEE.

Hatzijordanou, N., Bohn, N., and Terzidis, O. (2019). A systematic literature review on competitor analysis: status quo and start-up specifics. *Management Review Quarterly*, 69(4):415–458.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1):38–52.

Hervet, G. and Guitart, I. A. (2022). Increasing the effectiveness of display social media ads for startups: The role of different claims and executional characteristics. *Journal of Business Research*, 153(May 2021):467–478.

Hidayatullah, A. F., Pembrani, E. C., Kurniawan, W., Akbar, G., and Pranata, R. (2018). Twitter Topic Modeling on Football News. *2018 3rd International Conference on Computer and Communication Systems, ICCCS 2018*, pages 94–98.

Hingle, M., Yoon, D., Fowler, J., Kobourov, S., Schneider, M. L., Falk, D., and Burd, R. (2013). Collection and visualization of dietary behavior and reasons for eating using twitter. *Journal of Medical Internet Research*, 15(6):1–16.

Hipkins, R. and Cowie, B. (2016). The sigmoid curve as a metaphor for growth and change. *Teachers and Curriculum*, 16(2).

Hokkanen, L. (2015). Four Patterns for Internal Startups. In *20th European Conference on Pattern Languages of Programs*, pages 1–10.

Hu, M., Wongsuphasawat, K., and Stasko, J. (2017). Visualizing Social Media Content with SentenTree. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):621–630.

Hyun, S. and Lee, H. S. (2022). Positive effects of portfolio financing strategy for startups. *Economic Analysis and Policy*, 74:623–633.

Islam, M., Fremeth, A., and Marcus, A. (2018). Signaling by early stage startups: US government research grants and venture capital funding. *Journal of Business Venturing*, 33(1):35–51.

Jamali, M. A., Voghouei, H., and Mohd Nor, N. G. (2015). Information technology and survival of SMEs: an emprical study on Malaysian manufacturing sector. *Information Technology and Management*, 16(2):79–95.

Jansen, B. J. and Zhang, M. (2009). Twitter Power : Tweets as Electronic Word of Mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:183–198.

Jesemann, I., Beichter, T., Herburger, K., Constantinescu, C., and Rüger, M. (2020). Migration of the Lean-Startup approach from High-Tech startups towards product design in large manufacturing companies. *Procedia CIRP*, 91:594–599.

Joshi, K., Chandrashekar, D., Satyanarayana, K., and Srinivas, A. (2022). VC Funded Start-Ups in India: Innovation, Social Impact, and the Way Forward. *International Journal of Global Business and Competitiveness*, 17(1):104–113.

Kaila, R.P. & Prasad, A. (2020). Informational Flow on Twitter - Corona Virus Outbreak – Topic. 11(3):128–134.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., and Nerur, S. (2018). Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers*, 20(3):531–558.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In *Information Visualization*, volume 4950 LNCS, pages 154–175. Springer Berlin Heidelberg, Berlin, Heidelberg.

Keller, E. (2007). Unleashing the power of word of mouth: Creating brand advocacy to drive growth. *Journal of Advertising Research*, 47(4):448–452.

Kemell, K.-k., Nguyen-duc, A., Suoranta, M., and Abrahamsson, P. (2023). StartCards — A method for early-stage software startups. *Information and Software Technology*, 160.

Kherwa, P. and Bansal, P. (2019). Topic Modeling : A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24):1–16.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering (Software Engineering Group, Department of Computer Science, Keele .... *Technical Report EBSE 2007- 001. Keele University and Durham University Joint Report*, (January).

Ko, J. and Ko, E. (2021). What fashion startups should know before launching Crowdfunding projects: Focusing on Wadiz reward Crowdfunding. *Journal of Global Fashion Marketing*, 12(2):176–191.

Kubara, M. (2023). Spatiotemporal localization patterns of technological startups: The case for recurrent neural networks in predicting urban startup clusters. *The Annals of Regional Science*, Upcoming(0123456789).

Lago, N. C., Marcon, A., Ribeiro, J. L. D., Olteanu, Y., and Fichter, K. (2023). The role of cooperation and technological orientation on startups' innovativeness: An analysis based on the microfoundations of innovation. *Technological Forecasting and Social Change*, 192(April):122604.

Lai, K.-H. L. (2017). Implementing the Quality Startup Management System model in Hong Kong: a case study. *International Journal of Quality Innovation*, 3(1).

Lammers, T., Rashid, L., Kratzer, J., and Voinov, A. (2022). An analysis of the sustainability goals of digital technology start-ups in Berlin. *Technological Forecasting and Social Change*, 185(July 2021):122096.

Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (2013). *Handbook of Latent Semantic Analysis*. Psychology Press.

Lee, D. D. and Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, 13.

Leicht, E. A. and Newman, M. E. (2008). Community structure in directed networks. *Physical review letters*, 100(11):118703.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Liotino, K., Carvalhinha, M., Amato, J., Tromboni, P., and Yu, A. (2016). Dynamic capabilities and business model: An analysis of radical innovation inside midsized and large companies in Brazil. In *2016 Portland International Conference on Management of Engineering and Technology (PICMET)*, pages 1257–1263.

Loria, S. (2018). Textblob documentation. *Release 0.15*, 2.

Lugović, S. and Ahmed, W. (2015). An analysis of twitter usage among startups in europe. In *INFuture2015: e-Institutions-Openness, Accessibility, and Preservation-Proceedings*, pages 299–308. INFuture.

Marcon, A., Luis, J., and Ribeiro, D. (2021). How do startups manage external resources in innovation ecosystems ? A resource perspective of startups ' lifecycle. *Technological Forecasting and Social Change*, 171(December 2020):120965.

Marcon, A. and Ribeiro, J. L. D. (2021). How do startups manage external resources in innovation ecosystems? A resource perspective of startups' lifecycle. *Technological Forecasting and Social Change*, 171(December 2020).

Maria, L., Nascimento, A., and Coppe, P. (2017). Software Knowledge Registration Practices at Software Innovation Startups Results of an Exploratory Study. In *31st*

Brazilian Symposium on Software Engineering - SBES'17., pages 234–243.

Marko Leppanen (2014). Two Patterns for Minimizing Human Resources in a Startup. In *8th Nordic Conference on Pattern Languages of Programs*, pages 1–7.

McDiarmid, C. and Skerman, F. (2020). Modularity of erdős-rényi random graphs. *Random Structures & Algorithms*, 57(1):211–243.

Mirghaderi, S. A., Sheikh Aboumasoudi, A., and Amindoust, A. (2023). Developing an open innovation model in the startup ecosystem industries based on the attitude of organizational resilience and blue ocean strategy. *Computers and Industrial Engineering*, 181(May):109301.

Molla, A., Biadgie, Y., and Sohn, K. A. (2014). Network-based visualization of opinion mining and sentiment analysis on twitter. *2014 International Conference on IT Convergence and Security, ICITCS 2014*, pages 2012–2015.

Morales-trujillo, M. E. and García-mireles, G. A. (2019). Evolving with Patterns : A 31-Month Startup Experience Report. In *27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2019*, pages 1037–1047.

Moritz, A., Naulin, T., and Lutz, E. (2022). Accelerators as drivers of coopetition among early-stage startups. *Technovation*, 111(October 2021):102378.

Nguyen-Duc, A., Kemell, K.-k., and Abrahamsson, P. (2021). The entrepreneurial logic of startup software development: A study of 40 software startups. *Empirical Software Engineering*, 26(5).

Nguyen-Duc, A., Seppänen, P., and Abrahamsson, P. (2015). Hunter-gatherer cycle: a conceptual model of the evolution of software startups. In *Proceedings of the 2015 International Conference on Software and System Process*, pages 199–203.

Nicholls-Nixon, C. L., Valliere, D., Gedeon, S. A., and Wise, S. (2021). Entrepreneurial ecosystems and the lifecycle of university business incubators: An integrative case study. *International Entrepreneurship and Management Journal*, 17(2):809–837.

Nobari, N. and Dehkordi, A. M. (2023). Innovation intelligence in managing co-creation process between tech-enabled corporations and startups. *Technological Forecasting and Social Change*, 186(PB):122107.

Noelia, F. L. and Rosalia, D. C. (2020). A dynamic analysis of the role of entrepreneurial ecosystems in reducing innovation obstacles for startups. *Journal of Business Venturing Insights*, 14(April):e00192.

Olanrewaju, A. S. T., Hossain, M. A., Whiteside, N., and Mercieca, P. (2020). Social media and entrepreneurship research: A literature review. *International Journal of Information Management*, 50(November 2018):90–110.

Page, A. and Holmström, J. (2023). Enablers and inhibitors of digital startup evolution: a multi-case study of Swedish business incubators. *Journal of Innovation and Entrepreneurship*, 12(1).

Parthasarathy, S. (2022). A decision framework for software startups to succeed in COVID-19 environment. *Decision Analytics Journal*, 3(March):100037.

Paschen, J. (2017). Choose wisely: Crowdfunding through the stages of the startup life cycle. *Business Horizons*, 60(2):179–188.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Peixoto, A. R., de Almeida, A., António, N., Batista, F., and Ribeiro, R. (2023a). Diachronic profile of startup companies through social media. *Social Network Analysis and Mining*, 13(1):52.

Peixoto, A. R., de Almeida, A., António, N., Batista, F., Ribeiro, R., and Cardoso, E. (2023b). Unlocking the power of twitter communities for startups. *Applied Network Science*, 8(66).

Rafiq, U. and Wang, X. (2020). Continuous information monitoring in software startups. *Lecture Notes in Business Information Processing*, 396 LNBIP:280–287.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Rizvanović, B., Zutshi, A., Grilo, A., and Nodehi, T. (2023). Linking the potentials of extended digital marketing impact and start-up growth: Developing a macro-dynamic framework of start-up growth drivers supported by digital marketing. *Technological Forecasting and Social Change*, 186(October 2022).

Roche, M. P., Conti, A., and Rothaermel, F. T. (2020). Different founders, different venture outcomes: A comparative analysis of academic and non-academic startups. *Research Policy*, 49(10):104062.

Romme, A. G. L., Bell, J., and Frericks, G. (2023). Designing a deep-tech venture builder to address grand challenges and overcome the valley of death. *Journal of Organization Design*.

Ross, G., Das, S., Sciro, D., and Raza, H. (2021). ScienceDirect CapitalVX : A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7:94–114.

Rossetti, G., Milli, L., and Cazabet, R. (2019). Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):1–26.

Ruggieri, R., Savastano, M., Scalingi, A., Bala, D., and D'Ascenzo, F. (2018). The impact of Digital Platforms on Business Models: An empirical investigation on innovative start-ups. *Management and Marketing*, 13(4):1210–1225.

Ruseva, R. (2015). Patterns for startup business models. In *20th European Conference on Pattern Languages of Programs*, pages 1–11.

Sadeghiani, A., Shokouhyar, S., and Ahmadi, S. (2022). How digital startups use competitive intelligence to pivot. *Digital Business*, 2(2):100034.

Said, A., Bowman, T. D., Abbasi, R. A., Aljohani, N. R., Hassan, S. U., and Nawaz, R. (2019). Mining network-level properties of Twitter altmetrics data. *Scientometrics*, 120(1):217–235.

Santoso, R. T. P. B., Priyanto, S. H., Junaedi, I. W. R., Santoso, D. S. S., and Sunaryanto, L. T. (2023). Project-based entrepreneurial learning (PBEL): a blended model for startup creations at higher education institutions. *Journal of Innovation and Entrepreneurship*, 12(1):1–22.

Saravanakumar, M. and Suganthalakshmi, T. (2012). Social Media Marketing. *Life Science Journal*, 9(4):1097–8135.

Satyanarayana, K., Chandrashekar, D., and Mungila Hillemane, B. S. (2021). An Assessment of Competitiveness of Technology-Based Startups in India. *International Journal of Global Business and Competitiveness*, 16(1):28–38.

Saura, J. R., Palacios-Marqués, D., and Ribeiro-Soriano, D. (2023). Exploring the boundaries of open innovation: Evidence from social media mining. *Technovation*, 119(October 2021).

Saura, J. R., Palos-Sanchez, P., and Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*, 11(3):1–14.

Saura, J. R., Ribeiro-Soriano, D., and Palacios-Marqués, D. (2021). Setting B2B digital marketing in artificial intelligence-based CRMs: A review and directions for future research. *Industrial Marketing Management*, 98(August):161–178.

Schuh, G. and Hamm, C. (2022). Methodology for a Startup Lifecycle-dependent Approach of Financing for Investors and Deep Tech Startups. In *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1392–1398. IEEE.

Schuh, G. and Studerus, B. (2022). Methodology for the Startup Life Cycle-Dependent Design of Cooperation between Corporates and Startups. In *2022 Proceedings of PICMET '22: Technology Management and Leadership in Digital Transformation.*

Serrat, O. (2017). Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance. *Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance*, pages 1–1140.

Sha, H., Hasan, M. A., Mohler, G., and Brantingham, P. J. (2020). Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. *arXiv*, (2):2–7.

Skala, A. (2019). *Digital Startups in Transition Economies*. Springer.

Skare, M., Gavurova, B., and Polishchuk, V. (2023). A decision-making support model for financing start-up projects by venture capital funds on a crowdfunding platform. *Journal of Business Research*, 158(December 2022):113719.

Smith, A., Chuang, J., Hu, Y., Boyd-Graber, J., and Findlater, L. (2014). Concurrent visualization of relationships between words and topics in topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 79–82.

Smolak Lozano, E. and Almansa-Martínez, A. (2021). Estudio de la producción científica sobre social media. El caso de las revistas españolas de comunicación en JCR y SJR. *Revista de Ciencias de la Comunicación e Información*, 26:15–38.

Szerb, L. and Vörös, Z. (2021). The changing form of overconfidence and its effect on growth expectations at the early stages of startups. *Small Business Economics*, 57(1):151–165.

Tang, J. P. and Basco, S. (2023). Banks , credit supply , and the life cycle of firms : Evidence from late nineteenth century Japan. *Journal of Banking and Finance*, 154:106937.

Traag, V. A., Aldecoa, R., and Delvenne, J. C. (2015). Detecting communities using asymptotical surprise. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 92(2).

Tripathi, N., Seppänen, P., Boominathan, G., Oivo, M., and Liukkunen, K. (2019). Insights into startup ecosystems through exploration of multi-vocal literature. *Information and Software Technology*, 105(August 2018):56–77.

Van Eck, N. J. and Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538.

Vatankhah, S., Bamshad, V., Altinay, L., and De Vita, G. (2023). Understanding business model development through the lens of complexity theory: Enablers and barriers. *Journal of Business Research*, 155(PA):113350.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wang, X., Edison, H., Bajwa, S. S., Giardino, C., and Abrahamsson, P. (2016). Key challenges in software startups across life cycle stages. *Lecture Notes in Business Information Processing*, 251:169–182.

Werth, O., Cardona, D. R., Torno, A., Breitner, M. H., and Muntermann, J. (2023). What determines FinTech success?—A taxonomy-based analysis of FinTech success factors. *Electronic Markets*, 33(1):1–22.

Wolny, J. and Mueller, C. (2013). Analysis of fashion consumers' motives to engage in electronic word-of-mouth communication through social media platforms. *Journal of*

Marketing Management, 29(5-6):562–583.

Wright, M., Siegel, D. S., and Mustar, P. (2017). An emerging ecosystem for student start-ups. *Journal of Technology Transfer*, 42(4):909–922.

Wu, Y., Cao, N., Gotz, D., Tan, Y. P., and Keim, D. A. (2016). A Survey on Visual Analytics of Social Media Data. *IEEE Transactions on Multimedia*, 18(11):2135–2148.

Xiong, S., Wang, K., Ji, D., and Wang, B. (2018). A short text sentiment-topic model for product reviews. *Neurocomputing*, 297:94–102.

Yang, S. and Zhang, H. (2018). Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis. *International Journal of Computer and Information Engineering*, 12(7):525–529.

Yu, C., Margolin, D. B., Fownes, J. R., Eiseman, D. L., Chatrchyan, A. M., and Allred, S. B. (2021). Tweeting About Climate: Which Politicians Speak Up and What Do They Speak Up About? *Social Media + Society*, 7(3):205630512110338.

Yu, D., Xu, D., Wang, D., and Ni, Z. (2019). Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing. *IEEE Access*, 7:12373–12385.

Yusupova, A. T. and Ryazantseva, A. V. (2022). High-Tech Entrepreneurship in the Russian Regions: Conditions for the Emergence of New Companies. *Regional Research of Russia*, 12(2):143–153.

Zaglia, M. E. (2013). Brand communities embedded in social networks. *Journal of Business Research*, 66(2):216–223.

Zaheer, H., Breyer, Y., Dumay, J., and Enjeti, M. (2022). The entrepreneurial journeys of digital start-up founders. *Technological Forecasting and Social Change*, 179(March):121638.

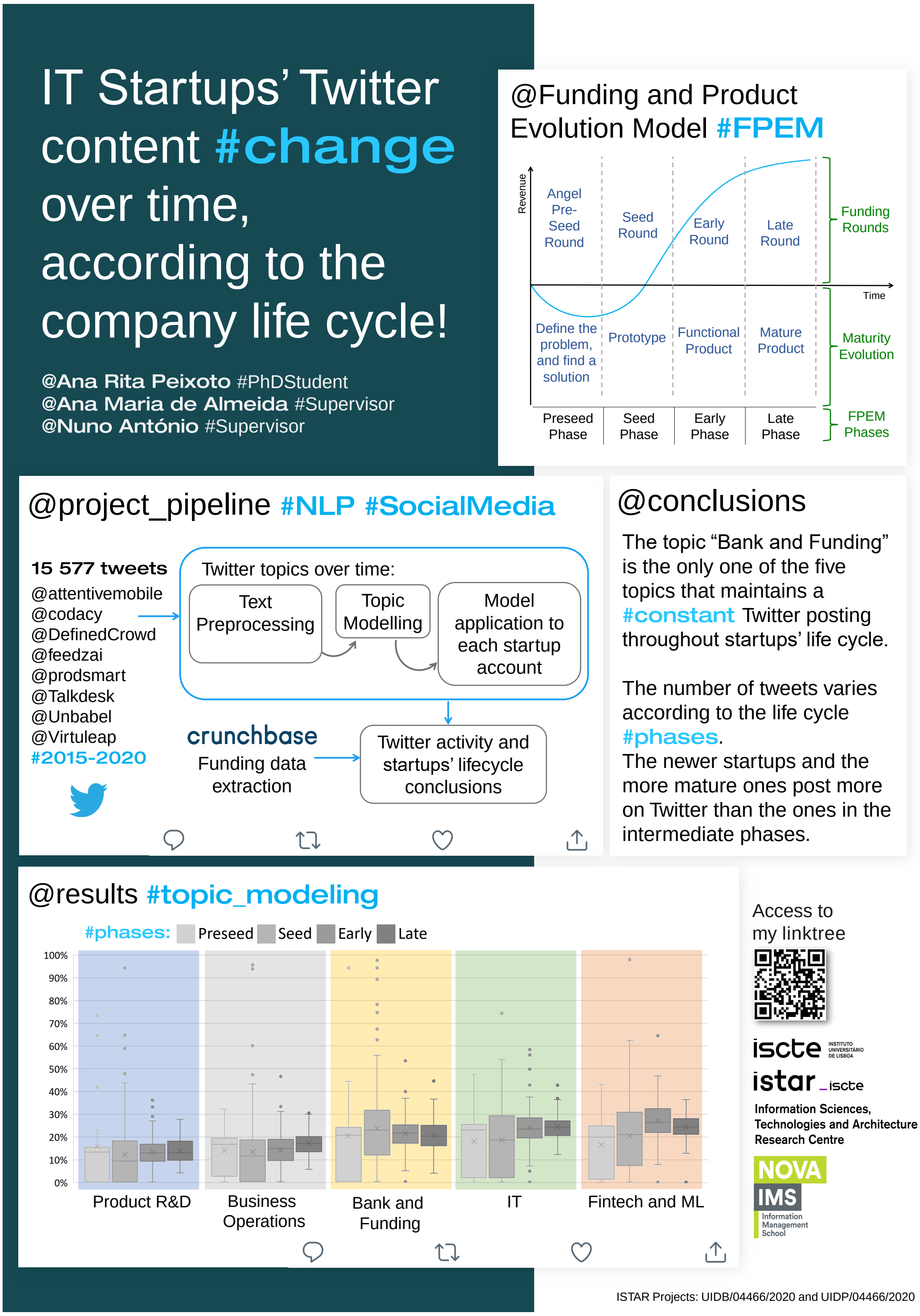
Zaina, L., Choma, J., Saad, J., Barroca, L., Sharp, H., Machado, L., and de Souza, C. R. (2023). What do software startups need from UX work? *Empirical Software Engineering*, 28(3):1–45.

Zeng, D., Chen, H., Lusch, R., and Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6):13–16.

Zhang, Q., Ye, T., Essaidi, M., Agarwal, S., Liu, V., and Loo, B. T. (2017). Predicting startup crowd funding success through longitudinal social engagement analysis. *International Conference on Information and Knowledge Management, Proceedings*, Part F1318:1937–1946.

APPENDIX A

IT Startups' Twitter content change over time, according to the company life cycle



IT Startups' Twitter content #change over time, according to the company life cycle!

@Ana Rita Peixoto #PhDStudent
@Ana Maria de Almeida #Supervisor
@Nuno António #Supervisor

@Funding and Product Evolution Model #FPEM

Revenue
Time
Angel Pre-Seed Round
Seed Round
Early Round
Late Round
Funding Rounds
Define the problem, and find a solution
Prototype
Functional Product
Mature Product
Maturity Evolution
Preseed Phase
Seed Phase
Early Phase
Late Phase
FPEM Phases

@project_pipeline #NLP #SocialMedia

15 577 tweets
@attentivemobile
@codacy
@DefinedCrowd
@feedzai
@prodsmart
@Talkdesk
@Unbabel
@Virtuleap
#2015-2020

Twitter topics over time:
Text Preprocessing
Topic Modelling
Model application to each startup account
crunchbase
Funding data extraction
Twitter activity and startups' lifecycle conclusions

@conclusions

The topic "Bank and Funding" is the only one of the five topics that maintains a #constant Twitter posting throughout startups' life cycle.

The number of tweets varies according to the life cycle #phases.
The newer startups and the more mature ones post more on Twitter than the ones in the intermediate phases.

@results #topic_modeling

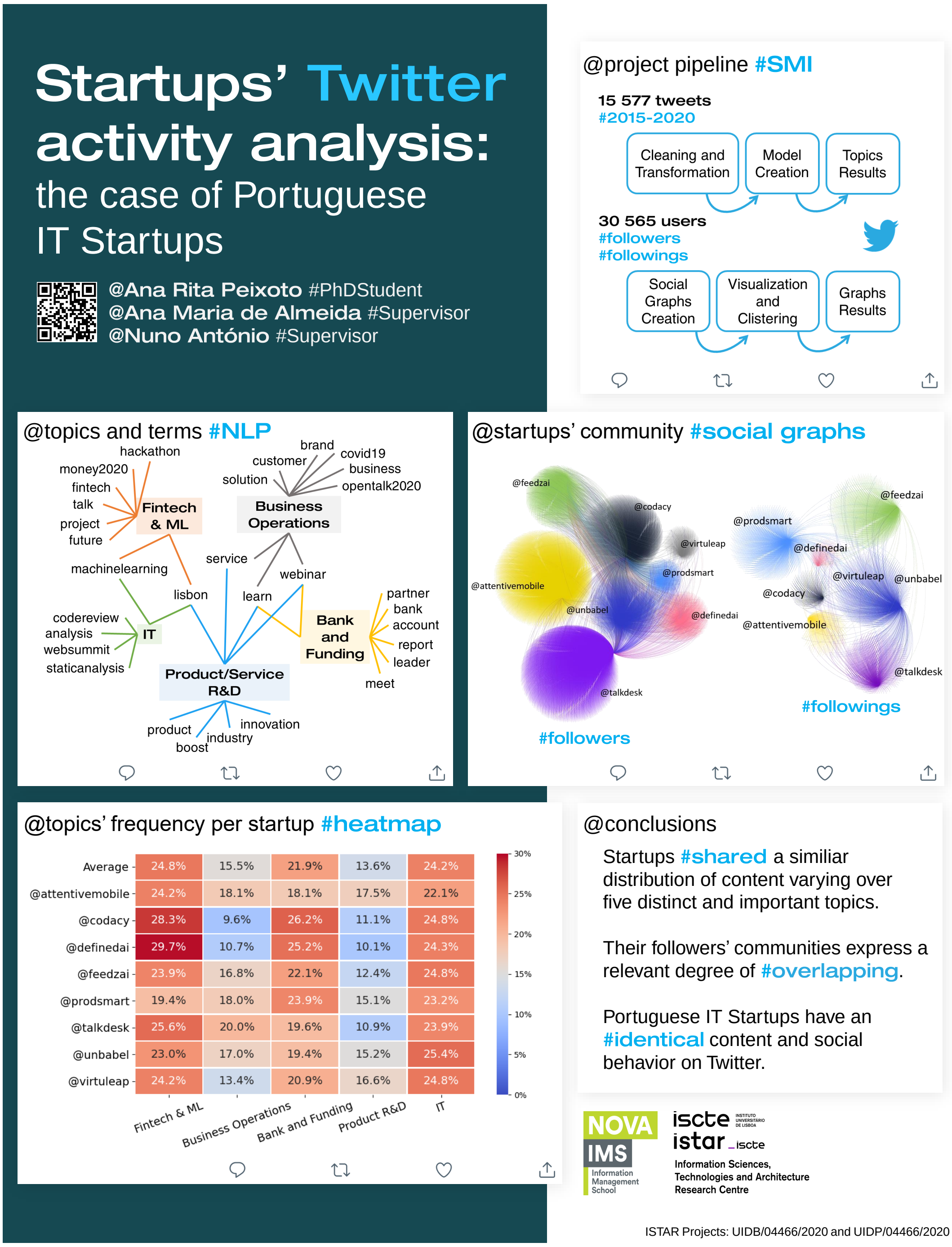
Access to my linktree

iscte INSTITUTO UNIVERSITÁRIO DE LISBOA
istar_iscte
Information Sciences, Technologies and Architecture Research Centre
NOVA IMS Information Management School

ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020

APPENDIX B

Startups' Twitter activity analysis: the case of Portuguese IT Startups



Startups' Twitter activity analysis:

the case of Portuguese IT Startups

@Ana Rita Peixoto #PhDStudent
@Ana Maria de Almeida #Supervisor
@Nuno António #Supervisor

@project pipeline #SMI

15 577 tweets
#2015-2020

Cleaning and Transformation → Model Creation → Topics Results

30 565 users
#followers
#followings

Social Graphs Creation → Visualization and Clistering → Graphs Results

@topics and terms #NLP

@startups' community #social graphs

@topics' frequency per startup #heatmap

	Fintech & ML	Business Operations	Bank and Funding	Product R&D	IT
Average	24.8%	15.5%	21.9%	13.6%	24.2%
@attentivemobile	24.2%	18.1%	18.1%	17.5%	22.1%
@codacy	28.3%	9.6%	26.2%	11.1%	24.8%
@definedai	29.7%	10.7%	25.2%	10.1%	24.3%
@feedzai	23.9%	16.8%	22.1%	12.4%	24.8%
@prodsmart	19.4%	18.0%	23.9%	15.1%	23.2%
@talkdesk	25.6%	20.0%	19.6%	10.9%	23.9%
@unbabel	23.0%	17.0%	19.4%	15.2%	25.4%
@virtuleap	24.2%	13.4%	20.9%	16.6%	24.8%

@conclusions

Startups #shared a similiar distribution of content varying over five distinct and important topics.

Their followers' communities express a relevant degree of #overlapping.

Portuguese IT Startups have an #identical content and social behavior on Twitter.

NOVA IMS Information Management School
iscte Instituto Universitário de Lisboa
istar_iscte Information Sciences, Technologies and Architecture Research Centre

ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020

APPENDIX C

Diachronic profile of startup companies through social media

Social Network Analysis and Mining (2023) 13:52
https://doi.org/10.1007/s13278-023-01055-2

ORIGINAL ARTICLE

Diachronic profile of startup companies through social media

Ana Rita Peixoto[1,4] · Ana de Almeida[1,4,5] · Nuno António[2,6] · Fernando Batista[3,4] · Ricardo Ribeiro[3,4]

Received: 19 January 2023 / Revised: 18 February 2023 / Accepted: 27 February 2023

Abstract

Social media platforms have become powerful tools for startups, helping them find customers and raise funding. In this study, we applied a social media intelligence-based methodology to analyze startups' content and to understand how their communication strategies may differ during their scaling process. To understand if a startup's social media content reflects its current business maturation position, we first defined an adequate life cycle model for startups based on funding rounds and product maturity. Using Twitter as the source of information and selecting a sample of known Portuguese IT startups at different phases of their life cycle, we analyzed their Twitter data. After preprocessing the data, using latent Dirichlet allocation, topic modeling techniques enabled the categorization of the data according to the topics arising in the published contents of the startups, making it possible to discover that contents can be grouped into five specific topics: "Fintech and ML," "IT," "Business Operations," "Product/Service R&D," and "Bank and Funding." By comparing those profiles against the startup's life cycle, we were able to understand how contents change over time. This provided a diachronic profile for each company, showing that while certain topics remain prevalent in the startup's scaling, others depend on a particular phase of the startup's cycle. Our analysis revealed that startups' social media content differs along their life cycle, highlighting the importance of understanding how startups use social media at different stages of their development.

Keywords Topic modeling · Social media · Startups · Life cycle model · Twitter data

1 Introduction

Social media platforms enable the creation of communities, provide easy access, and help companies promote their business. Their usage implies only a small investment, driving startup companies to use it as a cost-effective tool to create a digital gateway for finding customers and raising funds. The last are two of the three critical startup challenges reported by Wang et al. (2016). Building the product is the third. These challenges derive from a startup company's fast pace of growth, making it difficult to identify the correct steps to take for scaling up. Gulati and DeSantola (2016) explain that startups can improve their growth and achieve their objectives by understanding the best scaling practices.

The definition of what is a startup company has evolved over time. The definition introduced by Lugović and Ahmed (2015a) involves two perspectives: one concerning the business dimension and the other concerning the company's characteristics. Regarding the business dimension, if a company has been established for less than one year and employs at least one person besides its founders, then it can be considered a startup. As for the company characteristics,

✉ Ana Rita Peixoto
rita_peixoto@iscte-iul.pt

Ana de Almeida
ana.almeida@iscte-iul.pt

Nuno António
nantonio@novaims.unl.pt

Fernando Batista
fernando.batista@inesc-id.pt

Ricardo Ribeiro
ricardo.ribeiro@iscte-iul.pt

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal

[2] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, Lisbon, Portugal ,

[3] INESC-ID Lisboa, Lisbon, Portugal

[4] Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal

[5] CISUC – Center for Informatics and Systems of the University of Coimbra, Coimbra, Portugal

[6] CITUR, Centro de Investigação, Desenvolvimento e Inovação em Turismo, Faro, Portugal

Published online: 18 March 2023

it must be an innovative and growth-oriented business. However, more recent work suggests that the startup definition depends on the actual stage of the company's life cycle (Skala 2019). Therefore, the startup definition is not entirely settled, but some perspectives enable the characterization of these small companies.

Undoubtedly, social media has become a fundamental part of the information ecosystem, generating a large amount of data. Social media data can provide information about clients, products, and the overall market, improving the decision-making processes. However, data have to be processed, structured, and interpreted to infer relevant decision information. Understanding social media can help improve the company's investment (ROI) while enabling better customer relationship management (CRM), which is supported by recent studies that focus on social media data, considering it a strategic knowledge source for businesses (Kapoor et al. 2018). Previous studies explored digital platforms startups' data to extract relevant information about their activity. Saura et al. (2019) examined tweets using "#startup" to detect indicators for success and discovered the sentiment of the most common topics of tweets about startups. A broad study by Ruggieri et al. (2018) focused on finding patterns in successful startups based on their digital platforms' presence. The authors stated that newly born startups use digital platforms because it is cost-effective.

Nonetheless, startups' presence on digital platforms is continued since it enables the creation of a community between users and providers, which affects the scalability of the business, and opens new sources for creating value. Regarding the actual startups' activity, Alotaibi et al. (2020) designed a framework to evaluate Twitter activity using an Arabic startup as a case study. Recent systematic literature reviews have highlighted the need for deeper research in social media intelligence (Olanrewaju et al. 2020; Smolak Lozano and Almansa-Martínez 2021). Olanrewaju et al. (2020) proposed a set of future work themes, among which we can find the need to consider the evolution state of the company and, in consequence, its life cycle stages. We aim to fill that gap and understand how a startup's social media content changes through the different phases of its life. In other words, understand the diachronic profile that emerges from the startup's historical social media data and analyze whether it reflects its scaling evolution.

Since Twitter is an ideal platform for small businesses like startups and where they are now massively present, we have chosen this platform as our primary research data source.[1] In fact, even think tanks, an usual birthplace for startups (Feld and Hathaway 2020), use social media, like Twitter, to disseminate their activities and achieve funding (Castillo-Esparcia et al. 2020). Twitter differs from other social media platforms because it gives access to a global audience where users openly communicate with other users. Above all, it offers an opportunity for businesses to interact and receive instant feedback instead of acting solely as a marketing tool (Curran et al. 2011). Campos-Domínguez (2017) classifies Twitter activity as spontaneous and instantaneous, which can encourage a fluid exchange of ideas. Thus, Twitter can be looked at as a social media tool to help a business establish a network between customers, owners, and investors—providing an environment where professional content coexists with user-generated content, that is, nonexpert content (Casero-Ripollés 2018). Twitter activity is composed of tweets, which are essentially short text messages that may include images, emoticons, URLs, mentions, and hashtags. These characteristics make tweet categorization a challenging task. The textual analysis of startups' tweets was performed using a topic modeling approach. We begin by assigning a category to a tweet by uncovering the tweet's main topics and then studying the evolution of the tweets' content over the startups' life cycle. The present research differs from the existing literature by linking the results of the text analysis with each company's life cycle stage to understand if and how the startups' social media activity alters with its rise, maturity, and consequent change of goals.

This study focuses on the particular case of information technology (IT) startups founded by Portuguese executives or headquartered in Portugal as an illustrative case study. The rationale links with the fact that Portugal has created a distinctive ecosystem for IT startups over the latest years, mainly due to the Portuguese high-quality engineering talents and above-average English language fluency levels.[2] Additionally, the Portuguese government has seriously engaged in innovation policies, promoting initiatives like Startup Portugal, 200 M, and business incubators, which have fostered the creation of several startups. Since 2016, investment in Lisbon-based startups has grown 30% yearly[3] due to several successful startups and unicorns formed in Portugal. We selected eight IT startups from the Sifted 2020 Portugal startups list[4] for this work. The chosen companies are currently at different stages in their life cycle and are considered active on Twitter. The content posted by the eight startups spans five years of analysis, from 2015 to 2020, resulting in a total of 15 577 tweets.

The remainder of the paper is organized as follows: After presenting the related work, the methodology section describes our dataset, presents the methodology, and proposes a new model for the life cycle of a startup business. After presenting and discussing the results of our analysis

[1] https://www.thebalancesmb.com/top-reasons-why-your-small-business-should-use-twitter-2948523.

[2] https://www.ef.com/wwen/epi/.

[3] https://beportugal.com/startup-in-portugal/.

[4] https://sifted.eu/portugal-startups-top-rankings/.

in the results section, implications are drawn. Lastly, the conclusions section describes the main conclusions and lays the path for future work.

2 Related work

Social media platforms are essential digital marketing tools for small businesses, with half of the world's population currently using these platforms (Castillero-Ostio et al. 2021). Saravanakumar and Suganthalakshmi (2012) denote social media marketing (SMM) as a marketing tactic that efficiently promotes brands through social media platforms. However, how can we analyze social media content and extract relevant information that demonstrates this value? This section aims to answer this question by explaining the social media analysis process and its methods and results. Additionally, we describe the startups' life cycle since this constitutes the central hypothesis driving our research: the cycle of the startup's life and evolution influences its social media activity.

2.1 The social media analysis process and methods

Social media data has become a fundamental part of the data ecosystem and is a strategic knowledge source for decision-making (Kapoor et al. 2018). Some paradigmatic examples can be found in extant literature. Campos-Domínguez (2017) analyzed the research on political communication on Twitter, and Godoy-Martín (2022) investigated the use of social media by communications agencies. Nevertheless, to infer relevant information from data, one needs to prepare and process it (Dutot and Mosconi 2016). Social media intelligence (SMI) collects and analyzes relevant data to provide data-driven support for strategic decisions. SMI works as a cycle because social media constantly changes, with new users creating new content and generating more data for analysis. The main focus of SMI applications is product/service review analysis (Kapoor et al. 2018). The knowledge obtained by SMI is meant to describe the present state of social media. This focus means that if the objective is to predict outcomes and suggest future directions, a social media analytics (SMA) approach is deemed necessary (Choi et al. 2020). SMA and SMI present similar phases (Zeng et al. 2010), but the SMA methodology and results focus on the future, while SMI concerns the present.

Social media content is mainly text, and the goal of its analysis is to find relationships among data in textual documents and extract patterns to understand the themes being addressed (Jelodar et al. 2017). This goal can be achieved by analyzing the text's sentiment or identifying the main topics. A topic is a list of words defined statistically to categorize the meaning of the text, and this process is termed topic modeling. Using topic modeling, researchers in the literature address problems in the most varied fields. There are several methods to conduct topic modeling. Among the most employed ones are latent Dirichlet allocation (LDA) by Blei et al. (2002), latent semantic analysis (LSA) by Landauer et al. (2007), and non-negative matrix factorization (NMF) by Lee and Seung (2001), both based on linear algebra, namely diverse forms of matrix factorization.

LDA is one of the most popular and widespread methods for identifying latent topics in a text (Blei et al. 2002). It identifies the (relevant) topics by using generative probabilistic models. One of the areas where it is applied is in social media topic analysis, as observed in the works of Saura et al. (2019); Yang and Zhang (2018); and Yu et al. (2019). While these studies focus on different problems, each uses topic modeling as a tool for SMA. D. Yu et al. (2019) developed a novel hierarchical topic modeling technique and mined the dimension hierarchy of tweets' topics over tweets of different countries. Saura et al. (2019) analyzed tweets with the hashtag startup ("#startup") and its comments. The objective was to understand the topics in those tweets and the associated sentiments. Yang and Zhang (2018) performed a similar analysis, where the authors combined topic modeling and sentiment analysis to mine the tweet's text. They concluded that the LDA algorithm makes it easy to analyze an extensive set of tweets and obtain meaningful topics. Some other studies use topic modeling to explore and understand specific subjects on Twitter, like in the case of Barry et al. (2018), which analyzes alcoholic drinks advertising or a recent study to understand how politicians tweet about climate change by Chao et al. (2021). More recent works use topic modeling methods to examine Twitter information about COVID-19. For instance, Sha et al. (2020c) analyzed governmental and politicians' tweets about the pandemic situation and inferred a set of topics that describe Twitter activity in the countries under analysis. Kaila and Prasad (2020) and Doogan et al. (2020) focused on tweets bearing hashtags related to COVID-19 to understand what non-government users tweet concerning the coronavirus pandemic and its global perception. While the former studies ascertain LDA as having achieved good results in analyzing Twitter posts, they also raise limitations about using the LDA algorithm with Twitter data. The two most common limitations are the tweets' short text format and the need for preprocessing phase. Transforming a tweet into a document to perform a topic model might not be adequate because it has few words to extract topics. Therefore, most studies solve these limitations by aggregating the tweets into sets, where each collection corresponds to a document (Curiskis et al. 2020). However, some advances appear to avoid aggregation, as Xiong et al. (2018) demonstrated, where the authors proposed a short-text topic model algorithm.

2.2 Startups and social media

Social media platforms have a global reach, are easy to access, and are low cost, enabling startups to use social media as a digital marketing gateway and observe the market. A few studies investigate the potential relationships between startups and social media platforms. Lugović and Ahmed (2015a) found a positive correlation between startups' Twitter usage and the total investment in the source country. As previously stated, Saura et al. (2019) collected tweets presenting the hashtag startup ("#startup"). The authors aimed to relate the polarity of the tweet with the topics found within the diverse sentiments. The authors classified the tweet's text and comments into positive, negative, and neutral. Then, the authors performed topic modeling for each polarity and found the related topics, enabling them to understand the Twitter audience sentiment of startup-related content.

Ruggieri et al. (2018) aimed to find patterns in successful innovative startups based on their digital platforms' activity. Their study demonstrates that startups are present on digital platforms mainly because these platforms have a cost-effective performance. The authors also conclude that a community of users/providers of services is essential for the business. Such a community is fundamental for a positive impact on digital platforms, primarily on social networking websites, since that community provides positive or negative opinions about products and companies. Word-of-mouth is the everyday oral communication that creates an impression and idea about a specific subject (Keller 2007), and online opinions are called electronic word-of-mouth (eWOM), as explained by Hennig-Thurau et al. (2004). Social media platforms are ideal tools for eWOM. Chu and Kim (2011) describe that eWOM enables the creation of a large community, which allows for increased digital engagement with social interactions, such as comments, likes, and shares. The last two represent non-verbal activities, and when their quantities are large, they might help raise a positive feeling in the social media profile in question (Wolny and Mueller 2013). Additionally, social media activities can be used to understand the online organization's reputation (Azinhaes et al. 2021a).

2.3 The startups' life cycle and its stages

Startups are primarily defined as fast-grow innovative businesses. According to Wang et al. (2016), the maturity evolution of a startup goes through two main stages: the learning and the growing stages. The learning stage consists of selecting a problem to solve and defining and evaluating the solution. The problem represents a real issue or obstacle for a specific target, which is solved by providing a product or service: the solution. The product concept is developed in the growing stage, followed by an implementation start leading to a working prototype. In case the prototype is successful, the startup obtains a functional product that later evolves into a mature product. However, Wang et al. (2016) emphasize that this is not a constant cycle, saying that a startup has to go through "multiple measure-learn loops." The loops mean evaluating each step as being in the stages previously referred. Concerning startups whose main product/service is software, Nguyen-Duc et al. (2015) created a conceptual model named the hunter-gatherer, that in fact, consists of two development cycles: the "hunting" cycle consists of the idea, market, and features; the "gathering" cycle features the prototype, quality, and product. The intention is that the two cycles occur at each stage, but the dimension of the cycle differs over the startup's life cycle. In the learning stage, the hunting cycle is more significant, while in the growing stage, the gathering cycle becomes prominent. Nevertheless, the cycles occur at each stage side-by-side; when the company obtains a mature product, the focus changes to quality matters.

3 Methodology

This research follows the SMI steps framework described by Choi et al. (2020) for social media-based BI research. The process consists of four phases: "Data collection," "Data preprocessing," "Data analysis," and "Validation & Interpretation." According to this framework, the initial step was the extraction of the data from Twitter. As previously mentioned, a particular set of startups' accounts was targeted as a case study: information technology (IT) startups founded by Portuguese or headquartered in Portugal, selling products or services based on machine learning (ML) approaches, and presenting a B2B business model. Thus, our analysis centers on eight startups from the Sifted 2020 Portugal startups list are as follows: *AttentiveMobile*, *Codacy*, *DefinedCrowd*, *Feedzai*, *Prodsmart*, *Talkdesk*, *Unbabel*, and *Virtuleap*.

After the extraction, data was cleaned, and the corpus was prepared (data preprocessing), after which we could proceed with a topic modeling (TM) technique for the analysis (data analysis). Finally, TM results are evaluated and interpreted (validation & interpretation). The latter step is where the topic modeling results are compared with the startups' funding rounds, creating a diachronic profile for each startup. For that, the funding rounds of each startup have also been collected from Crunchbase[5]and related to the startups' life cycle phase. Our approach is illustrated in Fig. 1.

The features that define a startup differ depending on where in the life cycle phase the company is: in the beginning, these are innovative companies with limited resources, while in the growth process, they perform an above-average

[5] www.crunchbase.com

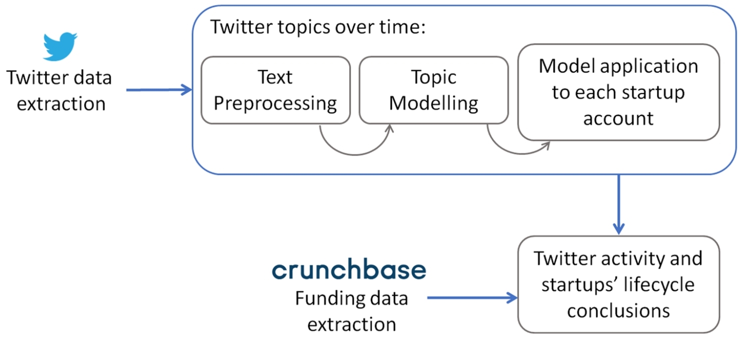

Twitter data extraction
Twitter topics over time:
Text Preprocessing
Topic Modelling
Model application to each startup account
crunchbase
Funding data extraction
Twitter activity and startups' lifecycle conclusions

Fig. 1 Project pipeline

rate increment in the number of customers and revenue; and finally, they have hyper-scalability and high company valuation, which characterizes a mature startup, demonstrating that startups change over their life cycle and that the definition of a startup depends on particular phases of company's evolution (Skala 2019). Thus, a startups' life cycle is a complex concept and, as stated by Paschen (2017), it shows two different but connected perspectives that are fundamental for the company's success: its maturity, regarding the stage of development of a product or service, and the funding rounds, that is, the fundamental investment attraction capability.

Based on the related literature, we consider that the startups' life cycle can be divided into two main perspectives. One that closely follows the concepts found in Wang et al. (2016) and regards the creation of a mature product to solve a real problem: maturity evolution. Another one concerning the startup funding rounds: funding rounds. The funding rounds are where startups open or expose their shareholder structure to third parties, usually to business angels or venture capital firms, to secure investment and allow the startup to grow (Paschen 2017). To illustrate a startup's financing milestones and evolution, we propose a life cycle model based on the previously introduced two dimensions: the funding rounds and the maturity evolution. We believe that the Funding and Product Evolution Model (FPEM), depicted in Fig. 2, illustrates the maturation process of a startup's life regarding time and revenue in a typical success scenario.

For the model, the names of the funding rounds dimension are based on the Crunchbase Glossary,[6] and in the maturity evolution, the phases describe the startup's product stages based on the work of Wang et al. (2016) and Paschen (2017). The proposed model, FPEM, encompasses four key phases named after the funding round categories: the preseed phase, the seed phase, the early phase, and the late phase. For the creation of the model, we correlated the phases with the existing funding types since these are measurable, which is essential to be able to mark when a transition occurs. Then, we connected the product maturity evolution with each of the rounds. Therefore, a phase transition occurs with a funding round of a higher rank than the previous one, implying a scale-up for the company and a product maturity evolution. Typically, startups receive new funding when their product has evolved and created value for the company. However, every type of funding round can happen more than once throughout a company's life. Notice that, for each phase, the association of concepts between maturity dimensions and funding rounds is relatively straightforward.

In the preseed phase, there is only the conceptualization of a potential and innovative solution for a concrete problem. Thus, funding is usually very limited (typically below $150 K) because it finances only an idea. These funding laps are known as angel or preseed rounds and are generally used to jump-start the company, providing financial cash to build a prototype. According to Wang et al. (2016), in this phase, the startup is in its learning stage. Next, in the seed phase, a prototype, or at least a proof-of-concept, already exists, sustaining the seed funding, which can scale up to $2 M. This round is used to build a product as market ready, incorporating the novelty proposed by the startup in the previous phase. In the early phase, the company already has a functional product and is prepared for scaling in the market. In this phase, the startup evolves for the growing stage (Wang et al. 2016). The early funding rounds, also called Series A and Series B, can have values ranging between $1 and $30 M. Lastly, in the late phase, a mature product is already established. The correspondent funding, also called Series C round, usually shows values that may start at $10 M with no upper limit.

The above-described relations between product maturity and funding rounds that represent the proposed life cycle model are validated by the topic model approach we have obtained, whose results are discussed in Sect. 4. The relations mentioned above enable us to relate each of the four

[6] https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types.

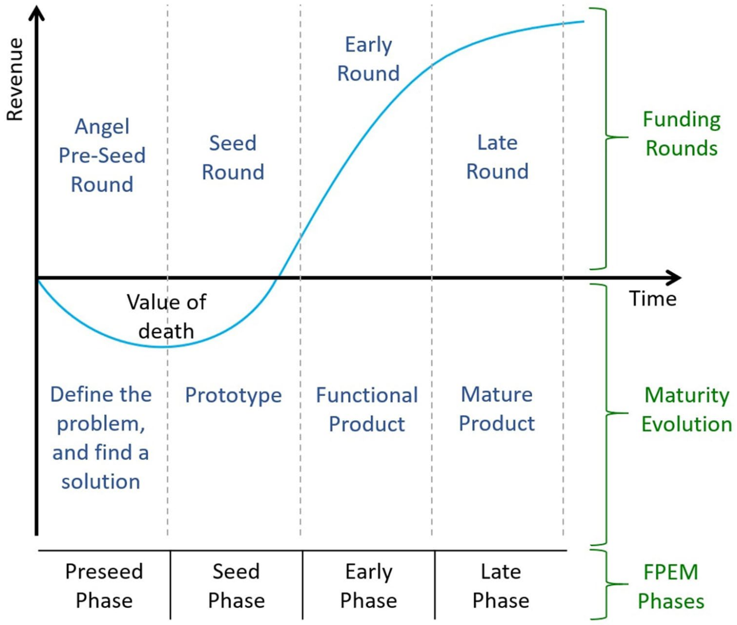


Fig. 2 Startups' life cycle model-funding and product evolution model (FPEM)

FPEM phases with the uncovered topics extracted from the tweets posted by the startups on social media during their existence.

3.1 Dataset

The dataset consists of 15 577 tweets extracted from the chosen Portuguese startups' Twitter accounts. The date of extraction date January 10, 2021, and the data covers every tweet posted by each startup since its Twitter profile creation date. The Twitter API method was employed ("GET statuses/user_timeline") to extract all the tweets posted by providing each company account's username through the library tweepy (Roesslein 2020). The analysis focuses on the last five years, where the higher quantity of posts is concentrated from January 2015 to December 2020, that is, for 72 months. To accurately examine the startups' activity over time, Table 1 shows the startup's Twitter accounts' descriptions.

It presents the company's first tweet available date, the number of followers, the number of tweets since January 2015, and the frequency per month. The last value regards the 72 months of analysis, or the number of months since the first tweet available date if it is more recent than January 2015. Additionally, the table shows the startup founding year, collected from Crunchbase.

Figure 3 shows each startup's quantity of tweets distributed over our chosen time window. It is possible to see that some startups post tweets regularly, while others present peaks with more activity. Within this context, regularly means the same temporal cadence, which is the case for half of the companies in the analysis: *AttentiveMobile*, *DefinedCrowd*, *Feedzai*, *Talkdesk*, and *Unbabel*. Particularly, *Talkdesk* account presents a higher number of tweets per month.

However, not every startup has presented tweet posts since the beginning of 2015. In the cases of *AttentiveMobile*, *DefinedCrowd*, *Feedzai*, and *Talkdesk*, the date for their first tweet available are more recent (Table 1 and Fig. 3). This inexistence of tweets may be because the company's foundation date is posterior or because more ancient tweets were voluntarily deleted. Namely, *Feedzai* and *Talkdesk* are the "oldest" startups, dating from 2011, but the overall number of postings is not that high, which might suggest that they may have deleted some of their oldest tweets.

Codacy, *Proadsmart*, and *Virtuleap* do not post regularly, and *Virtuleap* is the only company whose activity does not cover the 72 months of the analyzed time window. *Codacy* and *Virtuleap* presented a peak in 2016 and 2017, respectively. From then on, both posts regularly but used fewer tweets per month. Notably, *Proadsmart* shows a considerably lesser degree of Twitter posting activity and is the only company that does not show posts every month.

Table 1 Dataset description

Company name	Founded date	First tweet date	Followers	Number of tweets	Tweets per month
Attentive mobile	2016	08/02/2018	1115	695	20.44
Codacy	2012	02/10/2013	2796	640	22.78
Defined crowd	2015	04/02/2016	1674	1258	21.69
Feedzai	2011	23/10/2015	2630	3177	51.24
Prodsmart	2012	04/12/2012	897	211	2.93
Talkdesk	2011	26/06/2019	6586	3211	178.39
Unbabel	2013	17/11/2013	3510	2615	36.32
Virtuleap	2018	29/08/2016	791	2765	53.17

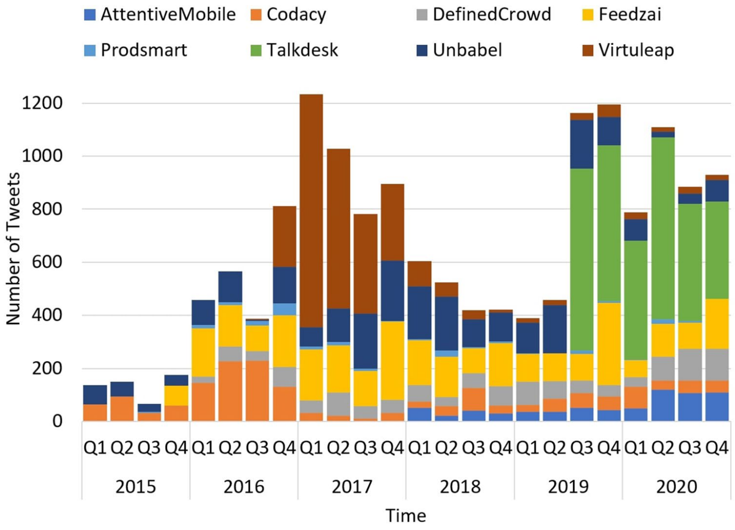


Fig. 3 Distribution of tweets quantity over time

3.2 Text preprocessing

To understand the topics of the textual tweets, we aggregated our dataset by month, resulting in a corpus (a set of documents where each document has an id and the correspondent text) of 72 documents corresponding to each month in the time-scope of the analysis. Within each document, the id regards the month and year of the tweets. This corpus was then cleaned, retaining the vocabulary that accurately represents the startups' content to be transformed into a document-term matrix for model training.

To ensure the adequate preprocessing of tweets, we first studied the techniques applied in literature's similar studies, thus concluding that the literature supports the need for a preprocessing phase enabling the preparation phase for achieving coherent topics. Table 2 presents the techniques that have been applied in the existing literature.

The most used techniques are: URL elimination, extra white spaces elimination, exclusion of the terms presenting higher or lesser frequency, HTML tags elimination, and the usage of stop words are also commonly applied.

Since white spaces, URLs, and punctuation do not present information relevant to topic's identification, they were removed from the documents. Next, lowercase transformation and lemmatization were performed. Excluding a set of stopwords, in this case, stopwords from the Natural Language Toolkit (Bird et al. 2009) help to focus the model on the relevant words that might define the text's meaning. For this, we added the startups' names and Twitter tags, like "RT," which means that it is a retweet, to the set of stopwords. The lemmatization goal is to convert every word to a common base form, providing coherence to the set of words and, consequently, to the topics. Lemmatization was done via *TextBlob* library (Loria 2020). *CountVectorizer* from the

Table 2 Literature preprocessing techniques usage

Preprocessing technique	Choi and Park (2019)	Alash and Al-sultany (2020)	Doogan et al. (2020)	Hidayatullah et al. (2018)	Yang and Zhang (2018)
Lowercase transformation	X		X		X
HTML tags elimination	X	X		X	X
URL elimination	X	X	X	X	X
Hashtag treatment	X	X			
Remove punctuation and digits			X	X	X
Remove stop words		X	X	X	X
Lemmatization			X		
Stemming				X	X
N-Grams		X	X		
TF-IDF			X		
Remove extra white spaces	X	X	X	X	X
Remove terms with higher frequency	X	X	X	X	X
Remove terms with less frequency	X	X	X	X	X

Python library *scikit-learn* (Pedregosa et al. 2011) enables vectorizing the text and having some preprocessing customization like the use of n-grams and exclusion of terms. The *n* grams used ranged from 1 to 2, uni- to bi-grams, to gather terms that may appear together, for example, the bi-gram “Machine Learning.” Then, the terms that appear less than twice were excluded to prevent possible errors and misspells. Lastly, the exclusion of terms that appear in at least 80% of the tweets. Being highly frequent terms suggests that they are meaningless in terms of topic characterization.

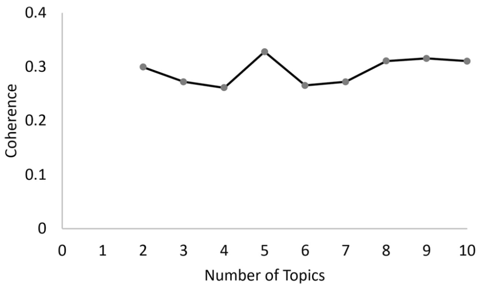


Fig. 4 LDA coherence analysis

3.3 Topic modeling

Due to its success in Twitter topic analysis-related literature, the topic modeling method here employed was LDA, latent Dirichlet allocation (Blei et al. 2002). The first step is to transform the corpus into a document-term matrix, where each term is either a word or a bi-gram. For that, we use the frequency of the occurrence of the term/bi-gram in the document's text and apply the LDA algorithm on the resulting matrix using the Python library *gensim* (Rehurek et al. 2011).

Since the number of topics must be given as input for the algorithm, we performed a coherence test for the advisable number of topics to be used in the modeling. Figure 4 suggests that five might be the more reliable number of topics due to its higher coherence value. Note that the coherence measure used here was *c_v*, one of the options in *gensim*.

Thus, the topic model created has five topics, each characterized by the relevant terms presented in Table 3, with all the terms showing a similar distribution within each topic.

The name chosen for the first topic is “Fintech and ML” because it encapsulates “fintech’’, “machine learning,’’ and “banking,” as well as one event in this domain: “money 2020.” The second topic is “Business Operations” since it presents terms correspondent concerns typical of the company's operations, such as “customer service,” “brand,” “solution,” and “covid19.” Additionally, it also displays “opentalk 2020,” a *Talkdesk*'s event regarding customer service subjects. “Bank and Funding” is the third topic, supported by the terms “bank,” “leader,” “report,” and “partner.” The fourth topic is “Product/Service R&D,” sustained by terms like “innovation,” “learning,” and “boost.” Lastly, “IT” (Information Technology) is the fifth topic associated with software, like code and security, and the more significant technological event, the Websummit.

Table 3 Topic description

Topic	Terms
Fintech and ML	Future, talk, fintech, banking, reality, money2020, lisbon, project, hackathon, machinelearning
Business operations	Business, cloud, opentalk2020, learn, covid19, service, solution, webinar, customer service, brand
Bank and funding	Bank, webinar, cloud, leader, learn, read, account, report, meet, partner
Product/service RD	Cloud, learn, product, read, industry, innovation, boost, service, webinar, lisbon
IT	Review, codereview, analysis, learning, websummit, machinelearning, machine learning, security, staticanalysis, lisbon

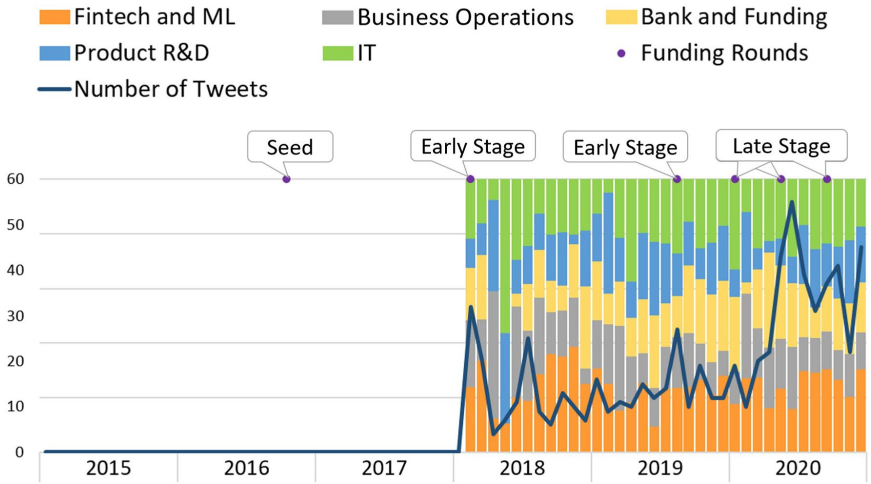


Fig. 5 Attentive mobile

4 Results and discussion

After the topic model, we divided the corpus by startup and applied the model, resulting in individual analyses representing the topics' evolution over time for each one. In order to understand if there is a relation between the FPEM phases and the Twitter activity, we combined the funding rounds' information. The first subsection describes the results obtained per startup.

After the individual analysis, it became clear that there were similarities between the independent analysis, so we performed another study using all the startups' data, whose results are outlined in the second subsection.

4.1 Topics evolution over startups life cycle

The following section regards the analysis of Twitter activity over time for each company when combined with the startup's funding rounds. Each figure shows the distribution of topics (in percentage), the number of tweets, and the funding rounds. To add context to the analysis, we provide, for each startup, a brief description of the company.

Figure 5 represents *AttentiveMobile* topics' evolution. *AttentiveMobile* is a B2B company that offers a personalized mobile messaging platform. We can see that from 2015 until February 2018, no tweets are found. Twitter social media activity started at the startup's early phase when the company already held a functional product. However, the topic "Product/Service R&D" is constantly present in their tweets over the years. In 2018, "Bank and Funding" was the topic less referred in their contents, but an increase can be seen over 2019, which may be because they needed new investment to grow. In fact, we can see that this topic increase precedes the company's late stage. Nevertheless, "Fintech and ML" and "IT" topics are always present along the years and achieve half of the content posted on Twitter, clearly related to the fact the

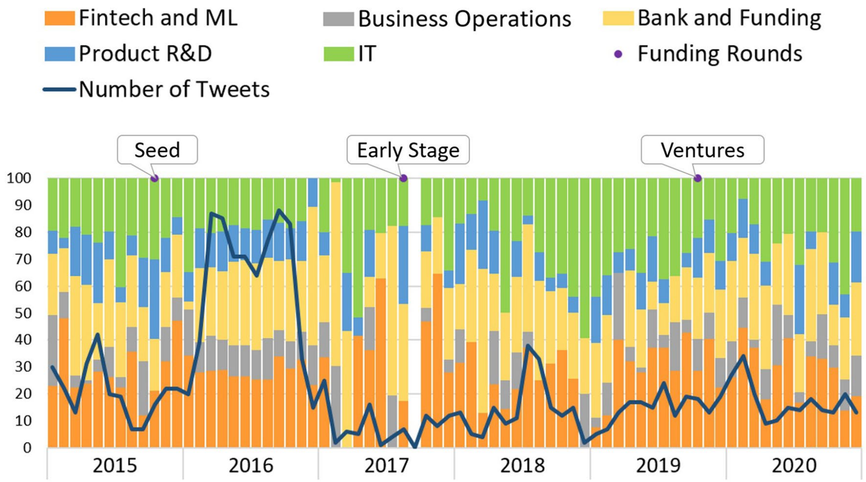


Fig. 6 Codacy

startup product is based on machine learning techniques. Since March 2020, the topics present a stationary distribution, showing a peak for the tweets' quantity in June 2020. This scenario—higher activity numbers and stable topics' distribution—emerges when the startup is in its late phase, where the company already owns a mature product.

The evolution of *Codacy* topics is depicted in Fig. 6. *Codacy* is an automated code review platform. The topics' distribution varies over the months, but it is clear that, in 2016, the number of tweets is significantly higher showing two very similar peaks. During 2016, the startup presents itself in a seed phase, meaning it should hold a prototype. The startup changes to an early phase in August 2017, and, awkwardly, in September 2017, there were no tweets. The most predominant topics in its tweets are "IT," "Fintech and ML," and "Bank and Funding." The first two may be related to the code review platform as it uses artificial intelligence methods, its core business, and the last appears associated with funding needs.

DefinedCrowd topics' evolution is shown in Fig. 7. This company develops artificial intelligence training data services and solutions. Although the startup's founding year was 2015, no tweets were available from 2015 until February 2016. From then on until July 2018, when it receives the first early round, the topic distribution variability is high over those months, both in the number of tweets and for the relative representation of topics. Once it reached its early phase, the topics presented a more structured distribution, showing an increase in the "Product/Service R&D" topic in the tweets. According to the FPEM, this is a phase where, typically, companies own a fully functional product, justifying the increment in tweets related to "Product/Service R&D." By the end of 2020, the graphic shows an increase in tweets per month, with two very similar peaks in July and in October.

Feedzai is an artificial intelligence startup whose core business is finance risk management. *Feedzai* tweet's profile evolution can be observed in Fig. 8. Notably, from January until November 2015, no tweets are available. From then on, Twitter's activity starts with the company in an early phase with an already functional product. The topics show a stationary distribution, and the number of tweets is consistent over the months, except for peaks occurring in October 2017 and October 2019, possibly because of an event occurring in October. Interestingly, in 2020, the topic "Bank and Funding" shows a decrease, and "Business Operations" has increased. The decrease may be due to the fact that in October 2017, the company reached the late phase, and raising more funds was no longer a priority. Alternatively, perhaps due to the COVID-19 ongoings, the company starts posting about the pandemic instead of financial-related tweets.

Prodsmart turns factories into digital and smart ones by employing production automation mechanisms and controlling the workflow using their software. Figure 9 represents the company's topics' evolution. Not only the presence of the company in the Twitter space varies immensely, but also the tweets' content is disparate, without any visual pattern or structure, making the distribution of the topics oscillate. During 2015, April stood out with contents relating to the topic "Bank and Funding," while in July, August, and

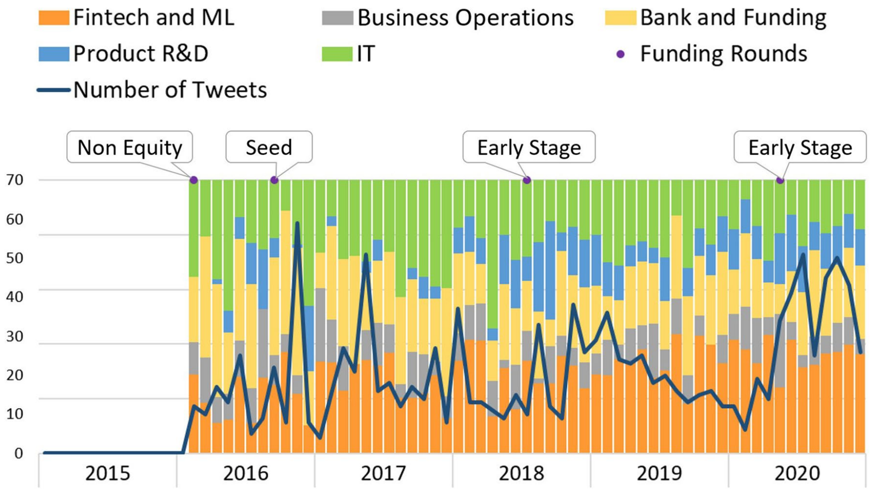


Fig. 7 Defined crowd

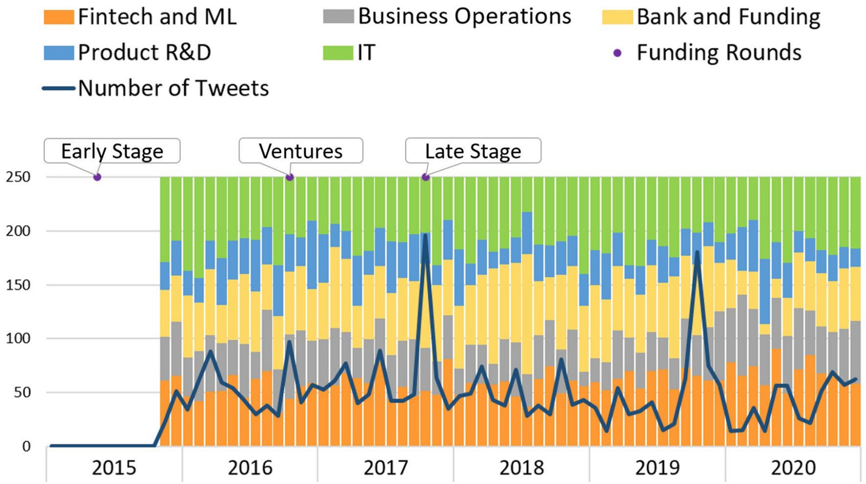


Fig. 8 Feedzai

September of the same year, the main topics in the tweets were "Product and R&D" and "Business Operations." Although the number of tweets is constantly lower compared with the other startups in the analysis, a peak occurred in November 2016. Since 2016, when the startup achieved the seed phase, the topics "Fintech and ML" and "IT," representing the technology subject, started to be present in their tweets' content. Over the years, the topic "Bank and Funding" shows a constant presence, which can be explained by the company's funding needs since *Prodsmart* did not leave the seed phase throughout the period under analysis.

Figure 10 represents the *Talkdesk* topics' evolution. *Talkdesk* is a platform to support sales teams for costumers' satisfaction and cost savings. Although founded in

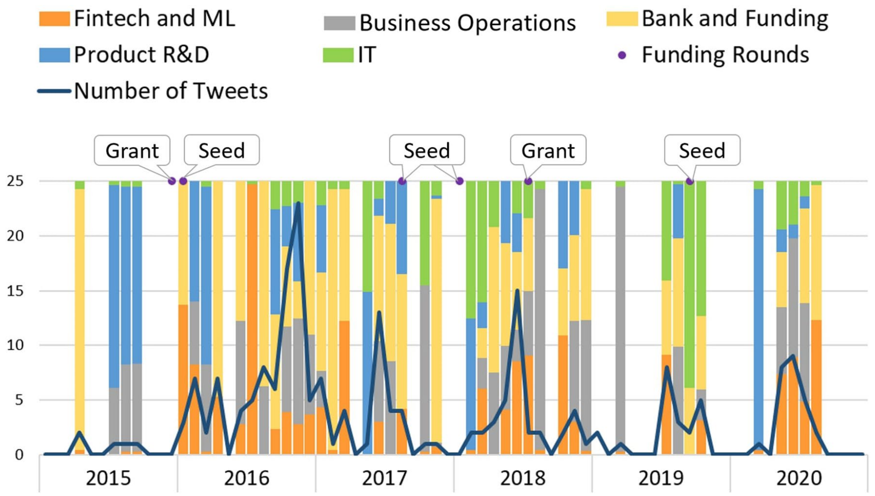


Fig. 9 Prodsmart

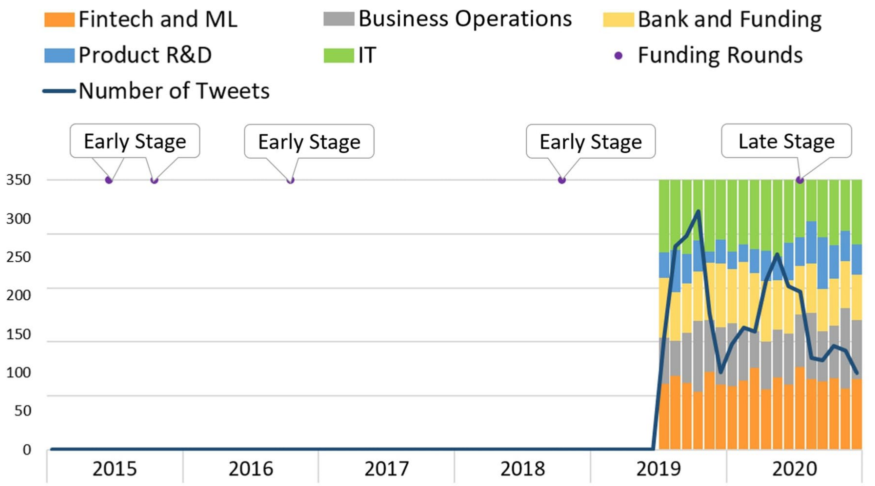


Fig. 10 Talkdesk

2011, from 2015 until July 2019, no tweets were available. However, from July 2019 forward, the number of tweets is mostly above 100/month, which suggests that the company must have decided to delete previous posts. From then on, *Talkdesk* has been at an early phase, having reached the late phase in July 2020. Regarding Twitter's activity, the topics are distributed very similarly over the months, with "Product R&D" showing the lesser number of tweets. The number of tweets shows two peaks, one in October 2019 and the other in April 2020. Since these tweets precede *Talkdesk's* entrance into a more mature phase already involving a stable product, tweeting about product development may not be between its higher priorities.

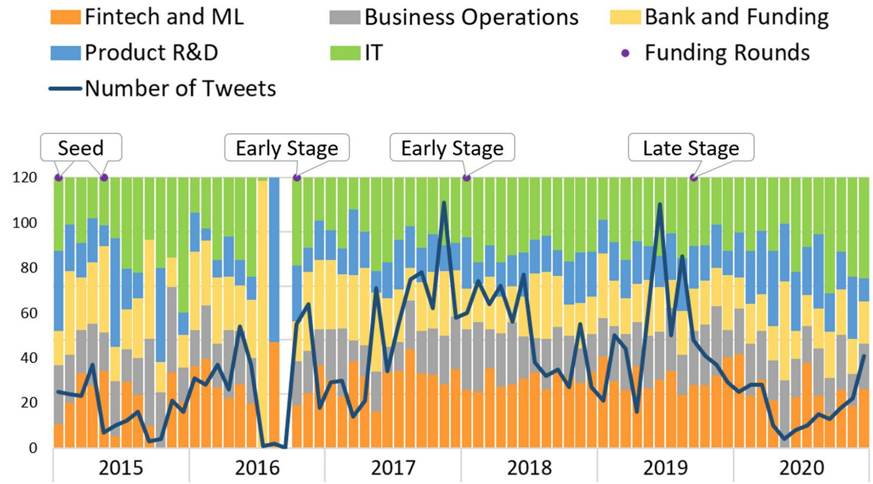


Fig. 11 Unbabel

Unbabel product enables companies to serve customers in their native language with a scalable translation across digital channels. Figure 11 represents *Unbabel* topics' evolution. The company's first seed round was in March 2014. *Unbabel* was in a seed phase until October 2016, when it reached an early phase, followed by the late phase in September 2019. In the seed phase, the tweets' topics show an oscillatory behavior, without a defined structure over the months, except for "IT" topic, which may be due to the heavy technological architecture of the company's services/products. However, since the early phase, the distribution has become stationary. In September 2016, the startup showed no posts; from September 2019 until May 2020, the number of tweets has been steadily decreasing. Maybe because of the late phase the company entered, not needing to heavily promote the new product or in need of raising extra funding.

Lastly, Fig. 12 depicts *Virtuleap* topics' evolution. This company sells a virtual reality application that promotes brain health with a library of games designed by neuroscientists. From 2015 to August 2016, there were no tweets available, and it is known that the company registry occurred in 2018 with *Virtuleap* achieving a seed round in February 2018. In fact, between 2018 and 2020, the company received five seed rounds. Tweets before 2018 can be found and a high-value peak quantity of tweets occurred in January 2017, prior to the first seed round. Additionally, the topic distribution in 2017 is mostly stationary, with the topics "Fintech and ML" and "IT" having higher representation. Since 2018, the number of posts has decreased until reaching residual values by the last quarter of 2018. Regarding the topics, by the end of 2018, the tweet content starts to show higher diversity and less structure, and the topics "Product R&D" and "Business operations" decrease when compared to the previous years.

As expected, being all of these classified as IT startups, all the companies show a good percentage of the tweet's contents addressing "IT" and "Fintech and ML." Also prevalent throughout most of the life cycle is the "Bank and Funding" theme. Thus, next section offers a more detailed analysis of the distribution of contents in terms of the phases of the FPEM.

4.2 Analysis of twitter activity in life cycle phases

The previous observations suggest that the content and the number of tweets posted by the startups may differ over their FPEM life cycle phases. It is possible to see (Fig. 13) that in terms of life cycle phases, the percentage of topics differs.

As it can be seen, the topic "Product R&D" is slightly higher in the preseed phase, and "Business Operations" is more eminent in the late phase. Newer companies need to focus on product development and in its promotion, while more mature startups already hold a final product in the market, allowing them to prioritize business concerns. The topics "Fintech and ML" and "IT" have similar distribution over all the life cycle phases, although showing a higher percentage in early and late phases. Lastly, the topic "Bank and Funding" shows to be the more constant theme, averaging about 20% for all posts. Concerning newer companies, in preseed and seed, those post more about the topic "Bank

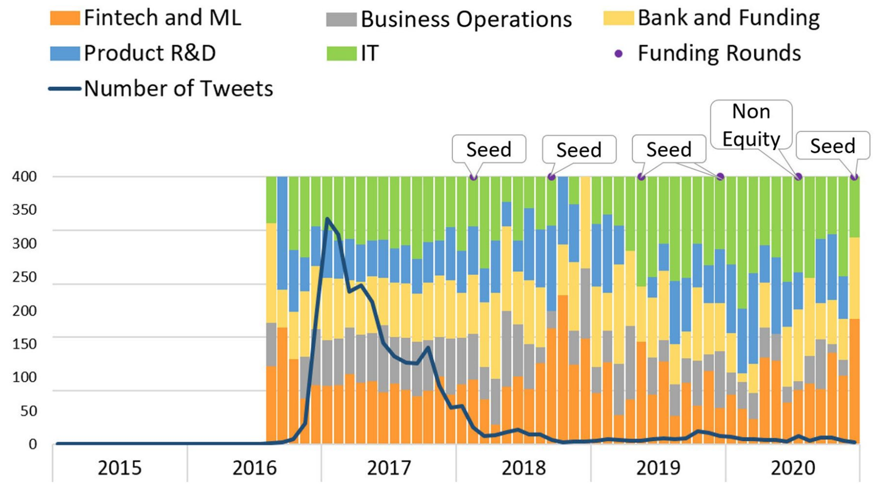


Fig. 12 Virtuleap

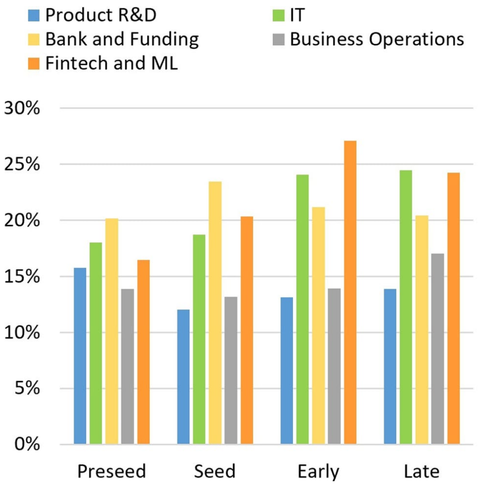


Fig. 13 Average of the topics' predominance per phase

and Funding," demonstrating the importance that financing has for their growth. In contrast, companies in early and late phases post more about the technology applied in their product, corresponding to the topics "Fintech and ML" and "IT." Additionally, the preseed phase is the one with minor variance between the topics' percentages over the phases, showing that for companies that at in this stage of their life

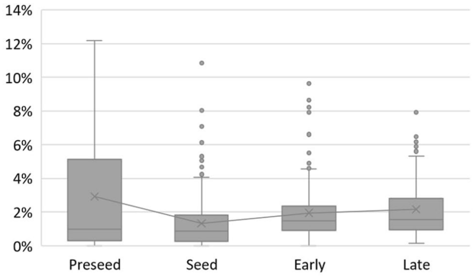


Fig. 14 Tweets quantity over life cycle phases

cycle may have a specific focus for their Twitter content, since they tend to post more (Fig. 14) and more consistently.

To understand if the relative emergence of topics within tweets differs according to each of the four FPEM phases, and since we have no good reason to assume that the topics distribution follows a normal distribution, the Kruskal–Wallis test was used. This is a nonparametric method that compares the means between groups, which in this scenario will be the four life cycle phases. We used the *SciPy* (Virtanen et al. 2020) library for implementing the Kruskal–Wallis test, setting the significance threshold at 0.05. The null hypothesis states that the means in each life cycle phase are the same. If the *p-value* is lower than the threshold, we reject the null hypothesis, meaning that the means on every life cycle are not the same. The results are presented in Table 4,

Table 4 Kruskal–Wallis tests results

	p-value
Topic: product R&D	(*)0.00545
Topic: IT	(*)2.38E − 13
Topic: bank and funding	0.327
Topic: business operations	(*)$2.4E - 06$
Topic: fintech and ML	(*)$8.82E - 08$
Number of tweets	(*)$2.72E - 08$

(*) Statistically significant ($p < 0.05$)

denoting the ones with a p-value below the significance with (*).

The topics "Product R&D," "IT," "Business Operations," and "Fintech and ML" present a *p-value* lower than the threshold, meaning that their means differ over the life cycle phases. "Bank and Funding" is the exception on the Kruskal–Wallis test, presenting a *p-value* expressively higher than the significance. This might imply that it remains more stable over the life cycle, which is consistent with the analysis of the information depicted in Fig. 13.

The results also prove a statistically significant relationship between the number of tweets and the startup phases. This relationship can be visualized in Fig. 14, which shows the proportion of tweets posted per month and distributed into the life cycle phases. The graph shows that all the startups have posted more on average when traversing the preseed phase.

Additionally, the higher variation in the preseed phase may be due to the fact that some of the startups in the analysis have been in this phase through a big part of the data time window. However, posts from some other startups at a preseed phase were not available (or not included in the case where it occurred before 2015). Notoriously, once a seed phase is achieved, startups' number of posts is notably less. This decrease in posting may be because they have received a funding round and are now more focused on product development. Nevertheless, the number of tweets slightly increases through the early and late phases.

Figure 15 displays the distributions of each topic to understand how, accordingly to the Kruskal–Wallis test, they differ throughout the FPEM phases. The topic "Product R&D," with the rejection of the null hypothesis, means that the distribution varies through the life cycle phases. In fact, this topic presents means higher values for the preseed phase and lower ones in the subsequent ones. This change can illustrate the importance of product development in the startups' beginning and confirms the maturity stage correspondent state in the life cycle description of FPEM. That is, startups in the preseed phase are finding a solution to a problem. The topic "Business Operations," which means they differ over the life cycle phases, has lower values in preseed and increases over the following phases. Having the opposite behavior of "Product R&D" and showing that with the startup growth, content about product development is exchanged by business concerns. The topics "IT" and "Fintech and ML," related to the startups' core business in the analysis, have a similar evolution over the phases. Both topics increase until the early phase and lightly decrease in the late phase. Note that those have a statistical significance to support the mean difference over the life cycle. Lastly, the topic "Bank and Funding" is the only means that do not differ over the phases, always staying around 20% value. The

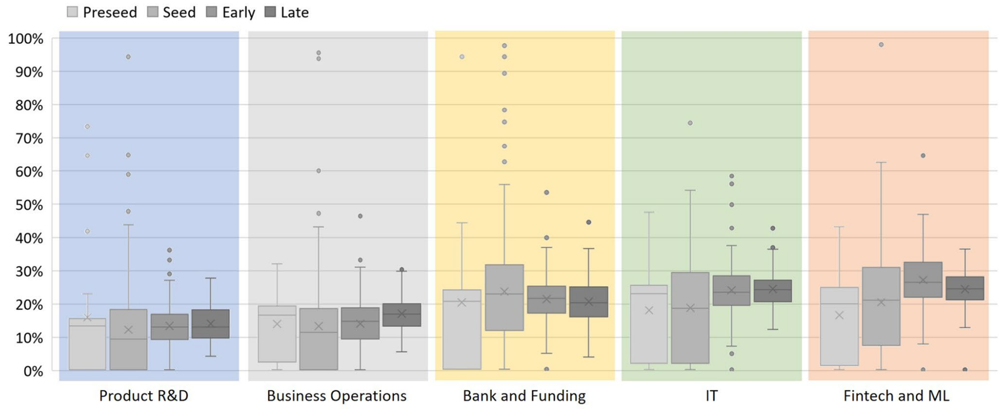


Fig. 15 Topics distribution over life cycle phases

constant presence of this topic demonstrates the importance of fundraising and financial matters for startups and supports the fact that funding rounds are a dimension that characterizes startups.

5 Implications

The primary goal of this study was to understand how Twitter contents of IT startups evolve over the company's growth. Literature shows that startups experience characteristic phases due to companies changing through their life cycle and adjusting their goals.

The first contribution of this study is the conceptualization of a life cycle model. This proposal is based on two dimensions previously described in the literature: maturity evolution and funding rounds. Maturity regards the development of the product or service the startup is selling, and the funding rounds regard capitalization through investors' financing. Our proposal unites those dimensions, creating a natural flow of business evolution: the funding and product evolution model (FPEM).

The second important implication of this study is categorizing IT startups' social media activity. Understanding the Twitter content was achieved through topic modeling, leading to a well-defined set of five topics describing the main subjects in the startups' tweets, which are "Fintech and ML," "IT," "Business Operations," "Product/Service R&D," and "Bank and Funding."

The third implication brings light to the question of how the startup's phases within its life cycle may affect social media usage. Our findings suggest that Twitter content produced by IT startups changes over the FPEM phases, while the startups scale up. The results outline that startups' initial posts are primarily related to product development and, in more advanced maturity phases, tweets became related to operations and business concerns. As expected, one of the topics found, "Bank and Funding," constantly emerge in tweets over the entire life cycle, denoting financial matters are a cornerstone for startups, as should be expected due to the particularities of these companies.

6 Conclusions and future work

This study proposes a new startup's life cycle model based on funding rounds and the companies' product maturity: the Funding and Product Evolution Model–FPEM. The validity of FPEM is illustrated using an SMI cycle-based methodology to extract the main topics from eight IT startups founded by Portuguese or headquartered in Portugal. The Twitter posts were subjected to an automatic information extraction of topics to understand if the tweets' contents change while startups are scaling up. The tweets posted between 2015 and 2020 were subjected to a topic model analysis for the IT startups chosen, adding up to 15 577 selected tweets. The results were combined with the FPEM life cycle model, creating a diachronic profile for each one of the startups. It was possible to perceive that the startups' key topics are: "Fintech and ML" and "IT," which regard the startups' core business; "Business Operations" and "Product/Service R&D" about enterprise subjects and product development; and "Bank and Funding" concerning startups' financing.

Nevertheless, results reveal that IT startups' Twitter topics change over time according to the company's current life cycle. The number of tweets published also varies according to the startup phase, showing that newer and more mature IT startups post more on Twitter when compared to companies in an intermediate phase. In terms of content, "Bank and Funding" is the only one of the five topics present throughout a startup's life cycle, demonstrating the great importance of financial investments and capital enabling the company's growth. On the other hand, another uncovered topic, "Product R&D," is predominant during the preseed phase, showing that startups begin as product-focused companies. In contrast, the topic "Business Operations" is prevalent in the late phase, revealing that business concerns take the place of the product development content with the startup's growth. Therefore, social media content evolves with the startups' evolution and scaling stages.

This study has several academic and practical contributions to how startups can employ social media in their growth process. Future research can map startups' maturity and scaling using this study's FPEM. The proposed life cycle model can guide researchers through the distinct phases. The results obtained in this study, namely the identified topics and their distribution through the startup life cycle, can be used by startups to create better marketing strategies. Regularly posting about "Bank and Funding" throughout the different phases seems to be a feasible approach. Lastly, investors can use the model proposed in this study to monitor startup's phases based on their social media activity and improve their investment decisions.

Like all studies, the study has limitations that should be considered in future research. Firstly, it focused only on IT startups based in Portugal. Future research should explore startups from other industries and countries to confirm whether the results are similar, regardless of the industry and region. Secondly, this study relies solely on publicly available Twitter data. Future studies should use data from other social media platforms, such as LinkedIn, to understand if posted contents vary for different platforms or if complementary topics emerge. Thirdly, the startups in this study were at different FPEM phases, which may have limited the possibility of a complete startup life cycle for some. Therefore, future research could focus on studying other startups at the same phase of the FPEM for more comprehensive results. Finally, we only validated the FPEM with the topics extracted from social media. Future work must use other data sources concerning startups to revalidate the model, like interviews with founders and venture capital experts.

Author contributions Conceptualization: All; methodology: All.; software: AP; validation: All; investigation: All; writing—original draft preparation: AP; writing, review and editing: All; supervision: AA e NA; All authors have read and agreed with the published version of the manuscript.

Funding Open access funding provided by FCT|FCCN (b-on). This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) namely by ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020; UIDB/04152/2020 (MagIC/NOVA IMS); and UIDB/50021/2020 (INESC-ID).

Declarations

Competing interests The authors declare no competing interests.

References

Alash HM, Al-Sultany GA (2020) Improve topic modeling algorithms based on twitter hashtags. J Phys Conf Ser 1660(1):012100. https://doi.org/10.1088/1742-6596/1660/1/012100

Alotaibi B et al (2020) Startup initiative response analysis (SIRA) framework for analyzing startup initiatives on twitter. IEEE Access 8:10718–10730

Azinhaes J, Batista F, Ferreira JC (2021a) EWOM for public institutions: application to the case of the Portuguese army. Soc Netw Anal Min. https://doi.org/10.1007/s13278-021-00837-w

Barry AE, Valdez D, Padon AA, Russell AM (2018) Alcohol advertising on twitter—a topic model. Am J Health Educ 49(4):256–263. https://doi.org/10.1080/19325037.2018.1473180

Bird, Steven., Ewan. Klein, and Edward. Loper. 2009. Natural language processing with Python Natural Language Processing with Python. O'Reilly. https://www.oreilly.com/library/view/natural-language-processing/9780596803346/ (January 16, 2023)

Blei DM, Ng AY, Jordan MT (2002) Latent Dirichlet allocation. Adv Neural Inf Process Syst 3:993–1022

Campos-Domínguez E (2017) Twitter y La comunicacíon política. In El Profesional De La Información. https://doi.org/10.1007/978-3-319-44700-1_23

Casero-Ripollés A (2018) Research on political information and social media: key points and challenges for the future. El Prof De La Inform 27(5):964 (https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2018.sep.01)

Castillero-Ostio E, Gil-Ramírez M, Castillo-Esparcia A (2021) Redes Sociales Como Espacios Comunicativos de Articulación de Movimientos Sociales: Revolución de Los Frijoleros (Guatemala). Chasqui Revista Latinoamericana De Comunicación 1(148):67–88 (https://revistachasqui.org/index.php/chasqui/article/view/4561)

Castillo-Esparcia A, Castillero-Ostio E, Castillo-Díaz A (2020) Los Think Tanks En España. Análisis de Sus Estrategias de Comunicación Digitales. Revista Latina 2020(77):253–273 (http://nuevaepoca.revistalatinacs.org/index.php/revista/article/view/386)

Chao Y, Margolin DB, Fownes JR, Eiseman DL, Chatrchyan AM, Allred SB (2021) Tweeting about climate: which politicians speak up and what do they speak up about? Social Media + Society 7(3):205630512110338. https://doi.org/10.1177/20563051211033815

Choi HJ, Park CH (2019) Emerging topic detection in twitter stream based on high utility pattern mining. Expert Syst Appl 115:27–36. https://doi.org/10.1016/j.eswa.2018.07.051

Choi J et al (2020) Social media analytics and business intelligence research: a systematic review. Inform Proc Manag 57(6):102279. https://doi.org/10.1016/j.ipm.2020.102279

Chu S-C, Kim Y (2011) Determinants of consumer engagement in electronic word-of-mouth (EWOM) in social networking sites. Int J Adv 30(1):47–75. https://doi.org/10.2501/IJA-30-1-047-075 (https://www.tandfonline.com/doi/full/10.2501/IJA-30-1-047-075)

Curiskis SA, Drake B, Osborn TR, Kennedy P (2020) An evaluation of document clustering and topic modelling in two online social networks: twitter and reddit. Inform Proc Manag 57(2):102034. https://doi.org/10.1016/j.ipm.2019.04.002

Curran K, O'Hara K, O'Brien S (2011) The role of twitter in the world of business. Int J Bus Data Commun Netw 7(3):1–15

Doogan C, Buntine W, Linger H, Brunt S (2020) Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of twitter data. J Med Internet Res 22(9):e21419

Dutot V, Mosconi E (2016) Social media and business intelligence: defining and understanding social media intelligence. J Decis Syst 25(3):191–192

Emilia SL, Almansa-Martínez A (2021) Estudio de La Producción científica sobre social media. El caso de las revistas españolas de comunicación en JCR y SJR. Revista De Ciencias De La Comunicación e Información 26:15–38 (http://revistaccinformacion.net/index.php/rcci/article/view/124)

Feld B, Hathaway I (2020) The startup community way: evolving an entrepreneurial ecosystem. Wiley, Hoboken

Godoy-Martín (2022) Las agencias de comunicación ante las nuevas redes sociales. ¿Early adopters o incorporación tardía? Revista Internacional de Relaciones Públicas 12(23):225–244. https://doi.org/10.5783/RIRP-23-2022-12-225-244

Gulati R, Alicia DS (2016) "Startups that last." Harvard Business Review 2016(March). https://hbr.org/2016/03/startups-that-last (January 16, 2023)

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? J Interact Mark 18(1):38–52 (https://linkinghub.elsevier.com/retrieve/pii/S1094996804700961)

Hidayatullah AF et al. (2018). "twitter topic modeling on football news." In: 2018 3rd international conference on computer and communication systems, ICCCS 2018: 94–98

Jelodar H et al (2017) Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78:183–198 (http://arxiv.org/abs/1711.04305)

Kaila RP, Prasad AVK (2020) Informational flow on twitter-corona virus outbreak–topic. Int J Adv Res Eng Technol (IJARET) 11(3):128–134

Kapoor KK et al (2018) Advances in social media research: past, present and future. Inf Syst Front 20(3):531–558

Keller Kd (2007) Unleashing the power of word of mouth: creating brand advocacy to drive growth. J Adv Res 47(4):448–452

Landauer TK, McNamara DS, Dennis S, Kintsch W (2007) Handbook of latent semantic analysis. Psychology Press, Handbook of Latent Semantic Analysis

Lee Daniel D, Sebastian Seung H (2001) "Algorithms for non-negative matrix factorization." advances in neural information processing systems 13. https://proceedings.neurips.cc/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf (January 16, 2023)

Loria S (2020b) "TextBlob: simplified text processing—TextBlob 0.16.0 documentation." https://textblob.readthedocs.io/en/dev/ (January 16, 2023)

Lugović S, Wasim A (2015a) "An analysis of twitter usage among startups in Europe." In: 299–308. Lugovic, ahmed, an analysis of twitter usage among startups in EU.pdf. http://infoz.ffzg.hr/infuture/2015a/images/papers/8-02

Nguyen-Duc A, Seppänen P, Abrahamsson P (2015b) Hunter-gatherer cycle: a conceptual model of the evolution of software startups. ACM Int Conf Proc Ser 24(26):199–203

Olanrewaju AS, Temitope MA, Hossain NW, Mercieca P (2020) Social media and entrepreneurship research: a literature review. Int J Inf Manage 50(2018):90–110. https://doi.org/10.1016/j.ijinfomgt.2019.05.011

Paschen J (2017) Choose wisely: crowdfunding through the stages of the startup life cycle. Bus Horiz 60(2):179–188. https://doi.org/10.1016/j.bushor.2016.11.003

Pedregosa F et al. (2011) 12 Journal of machine learning research scikit-learn: machine learning in python (2023).http://scikit-learn.sourceforge.net

Rehurek R, Sojka P (2011) Gensim–python framework for vector space modelling. Faculty of Informatics, Masaryk University, Brno, Czech Republic, NLP Centre, p 3

Roesslein J (2020) "Tweepy: twitter for python!" https://github.com/tweepy/tweepy (January 16, 2023)

Ruggieri R et al (2018) The impact of digital platforms on business models: an empirical investigation on innovative startups. Manag Mark 13(4):1210–1225

Saravanakumar M, Suganthalakshmi T (2012) Social media marketing. Life Sci J 9(4):1097–8135 (http://www.lifesciencesite.comhttp//www.lifesciencesite.com.670)

Saura JR, Palos-Sanchez P, Grilo A (2019) Detecting indicators for startup business success: sentiment analysis using text data mining. Sustainability (switzerland) 11(3):1–14

Sha H, Hasan MA, Mohler G, Jeffrey Brantingham P (2020c) "Dynamic topic modeling of the COVID-19 twitter narrative among U.S. Gov Cabinet Exec 2:2–7 (http://arxiv.org/abs/2004.11692)

Skala A (2019) Digital startups in transition economies: challenges for management, entrepreneurship and education. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-01500-8

Virtanen P et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods 17(3):261–272 (http://www.nature.com/articles/s41592-019-0686-2)

Wang X et al (2016) Key challenges in software startups across life cycle stages. Lect Notes Bus Inform Proc 251:169–182

Wolny J, Mueller C (2013) Analysis of fashion consumers' motives to engage in electronic word-of-mouth communication through social media platforms. J Mark Manag 29(5–6):562–583

Xiong S, Wang K, Ji D, Wang B (2018) A short text sentiment-topic model for product reviews. Neurocomputing 297:94–102. https://doi.org/10.1016/j.neucom.2018.02.034

Yang S, Zhang H (2018) Text mining of twitter data using a latent Dirichlet allocation topic model and sentiment analysis. Int J Comput Inform Eng 12(7):525–529

Yu D, Dengwei Xu, Wang D, Ni Z (2019) Hierarchical topic modeling of twitter data for online analytical processing. IEEE Access 7:12373–12385

Zeng D, Chen H, Lusch R, Li SH (2010) Social media analytics and intelligence. IEEE Intell Syst 25(6):13–16

APPENDIX D

Comparison of community detection algorithms

Community detection algorithms may produce different results. Due to this limitation, one should apply the one that better suits the problem domain. In our scenario, we chose to apply an algorithm based on modularity optimization that has shown good performance in similar social media network problems. However, these algorithms usually suffer from the resolution limit problem. These cannot efficiently detect smaller communities even if they are well-defined in the network, performing better in the limit of a few large communities than many small communities. At the beginning of the community analysis, only eight large communities were expected to be detected since the data is from the *followers* and *following* users of eight Twitter accounts. With this in mind, we have selected a modularity-based algorithm due to its success in these social network scenarios with few large communities.

Nonetheless, we chose to apply another method to understand the stability of our results, and we selected the algorithm proposed by Traag et al. (2015). The algorithm uses asymptotical surprise, a metric that, like modularity, is employed to evaluate the quality of community detection in networks. This metric is a statistical approach that calculates the probability of observing at least a certain number of internal edges within the communities, given the total number of edges in the network. We choose to apply this algorithm because it is nearly unaffected by the resolution limit problem, the modularity optimization primary weakness.

The results of applying the new algorithm are identical to previous results, as shown in Figures 4.5 and D.1. Additionally, since the figures are not at scale, Table D.1 shows each algorithm's results' node counts.

Table D.1. The resultant communities size of the application of the different algorithms

Community	**Number of nodes (Leicht and Newman, 2008)**	**Number of nodes (Traag et al., 2015)**
@Unbabel	3916	3916
@codacy	4678	4678
@prodsmart	1595	1599
@attentivemobile	7003	6980
@Virtuleap	873	872
@Definedai	1964	1960
@feedzai	3579	3579
@Talkdesk	6957	6981

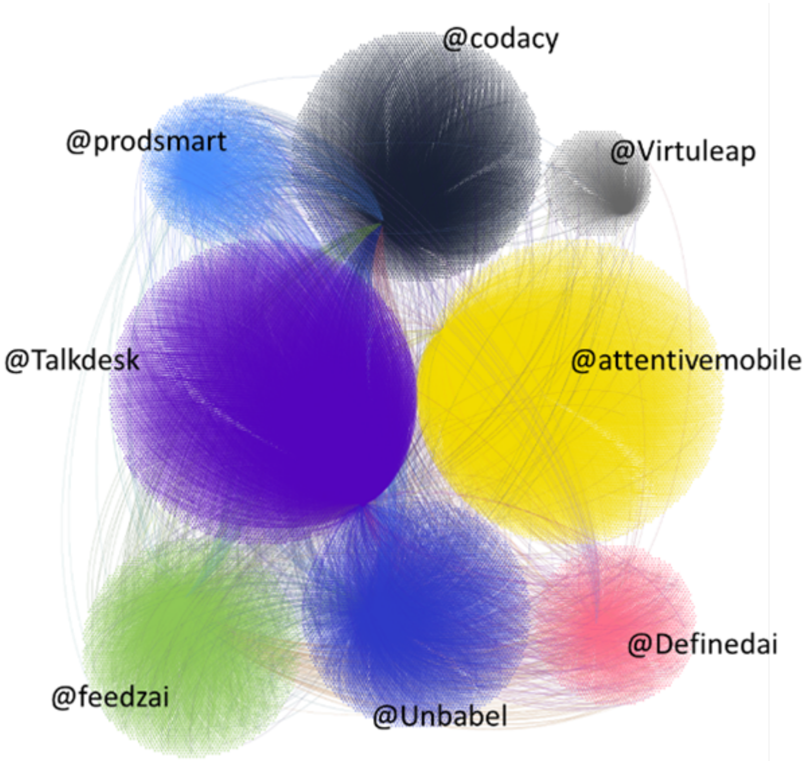


Figure D.1. Social digraph with community detection algorithm using asymptotical surprised proposed by Traag et al. (2015).

APPENDIX E

Unlocking the power of Twitter communities for startups

Peixoto *et al. Applied Network Science* (2023) 8:66
https://doi.org/10.1007/s41109-023-00593-0

Applied Network Science

RESEARCH Open Access

Unlocking the power of Twitter communities for startups

Ana Rita Peixoto[1,2*], Ana de Almeida[1,2,6], Nuno António[3,7], Fernando Batista[1,4], Ricardo Ribeiro[1,4] and Elsa Cardoso[1,4,5]

*Correspondence:
rita_peixoto@iscte-iul.pt

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal
[3] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
[4] INESC-ID Lisboa, Lisbon, Portugal
[5] CIES-Iscte, Lisbon, Portugal
[6] CISUC, Coimbra, Portugal
[7] Desenvolvimento e Inovação em Turismo, CITUR, Centro de Investigação, Faro, Portugal

Abstract

Social media platforms offer cost-effective digital marketing opportunities to monitor the market, create user communities, and spread positive opinions. They allow companies with fewer budgets, like startups, to achieve their goals and grow. In fact, studies found that startups with active engagement on those platforms have a higher chance of succeeding and receiving funding from venture capitalists. Our study explores how startups utilize social media platforms to foster social communities. We also aim to characterize the individuals within these communities. The findings from this study underscore the importance of social media for startups. We used network analysis and visualization techniques to investigate the communities of Portuguese IT startups through their Twitter data. For that, a social digraph has been created, and its visualization shows that each startup created a community with a degree of intersecting followers and following users. We characterized those users using user node-level measures. The results indicate that users who are followed by or follow Portuguese IT startups are of these types: "Person", "Company," "Blog," "Venture Capital/Investor," "IT Event," "Incubators/Accelerators," "Startup," and "University." Furthermore, startups follow users who post high volumes of tweets and have high popularity levels, while those who follow them have low activity and are unpopular. The attained results reveal the power of Twitter communities and offer essential insights for startups to consider when building their social media strategies. Lastly, this study proposes a methodological process for social media community analysis on platforms like Twitter.

Keywords: Twitter data, Communities, Social media, Startups

Introduction

Startups are innovative and typically, when successful, present an above-average increase in customers and revenue (Lugović and Ahmed 2015). Nevertheless, they own limited resources having to reach customers on a restricted budget. In order to succeed, startups must raise funding. Social media platforms can serve as a gateway for various communities, allowing companies to gain customers, obtain funding and grow by achieving their goals (Rizvanović et al. 2023). Undoubtedly, social media has become a fundamental piece of the global information ecosystem, generating large amounts of data. Social media data can provide information about clients, products, and the overall market, helping to improve decision-making processes (Saura et al. 2021). However, social raw data must be structured, prepared, and interpreted to infer relevant information to support decisions. Understanding social media as a tool can help enhance the company's investment—ROI—while enabling better customer relationship management—CRM. Recent existing studies support this argument and consider social media data-driven projects a strategic business knowledge source (Saura et al. 2021).

Social media platforms offer cost-effective digital marketing opportunities that benefit startups (Ruggieri et al. 2018). Studies have shown that active engagement on social media platforms increases digital engagement and can lead to better funding from venture capitalists or significant success in crowdfunding projects (Zhang et al. 2017; Ko and Ko 2021; Hadley et al. 2018). Lugović and Ahmed (2015) found a positive correlation between the Twitter activity of European startups and the total investment in their country of origin. Additionally, by creating communities relating users and service providers, startups can monitor the market and take advantage of electronic word-of-mouth spread of positive opinions about their products (Ruggieri et al. 2018; Chu and Kim 2011). Social network analysis, also known as SNA, is commonly called the process of monitoring the market and allowing for data-driven marketing strategies based on social media data (Hansen et al. 2019). Several existing studies have explored digital data using this methodology. For example, Ruggieri et al. (2018) focus on finding startup success patterns based on their presence on digital platforms. Hingle et al. (2013) collected Twitter content to analyze dietary behavior, and the authors highlighted that data visualization allowed the identification of relationships between diet-related behavioral factors. Wu et al. (2016) show that visualization methods can help uncover social media analysis results and support data interpretation, leading to a network analysis and visualization process. The authors of Hansen et al. (2019) propose a methodology, Network Analysis and Visualization, or NAV, to act as a design process model for enabling meaningful network analysis and extracting relevant insights.

The primary aim of this study is to determine the degree to which startups use social media platforms to create communities, what distinguishes these communities, and if the individual startup communities intersect. Our findings might highlight the relevance of social networks and their online communities for startups. To the best of our knowledge, none of the existing literature aims to understand how startups create communities on Twitter or if these communities intersect and create a global ecosystem. Therefore, this study aims to contribute to filling the gap in the literature regarding the analysis of social media communities around startups. For this, we selected eight renowned Information Technology (IT) startups founded by Portuguese administrators or headquartered in Portugal from the *Sifted* 2020 Portugal startups list (Sifted 2020). The companies were chosen because they are currently at different stages in their life cycle and are considered active on Twitter. Portugal has created an excellent startup ecosystem by promoting initiatives like the Startup Portugal 200 M fund and several business incubators (Portugal Digital and Startup Portugal 2021). Additionally, Portugal is recognized for forging high-quality engineering talents and showing a very high English proficiency index (Education First 2022). Twitter was selected as the social media data source due to its simple API access. This social media platform is a short text source widely used in academic research to analyze online behavioral patterns and the structure of the formed social graphs (Antonakaki et al. 2021). Furthermore, Twitter differs from other social media platforms because it allows users to communicate with themselves publicly. It enables businesses to interact, create a community of users, and, instead of working solely as a marketing tool, allows for instant feedback, being crucial to companies that want to stay relevant and make immediate connections with their audience (Tanner 2023). Therefore, Twitter is an ideal platform for small businesses like startups, where they are now massively present (Ward 2020). Under these conditions, this case study intends to provide answers to the following research questions:

RQ1: Do IT Portuguese startups form their own social communities on Twitter?
RQ2: In the case of community formation, are these disjoint, presenting different (types of) users, or do they overlap?

Towards our goal, and to better perceive community creation, we have applied the NAV process to social media data. Specifically, a social digraph was built, representing the *followers* and *following* communities of the startups under investigation. We utilized a community detection algorithm to determine if the startups form communities on Twitter based on modularity. This algorithm enabled us to visualize the communities in the digraph structure using different colors. The resulting visualization revealed that each startup had indeed formed a community. Furthermore, we were able to identify links between nodes of different communities, indicating that there was some overlap between the communities. To address the second research question, we characterized the communities' users by analyzing their type, popularity, and activity level. This information enables the emergence of social media strategies that can be effective for startups to achieve their proposed goals, either for financial support or for product/service marketing actions.

This work highlights the benefits of using social media platforms for startups to establish user and service provider communities. Through a case study, we present a systematic process that enables the visualization of startup communities and allows for the detection of intersections between these groups, which can help startups monitor similar companies and select relevant users to follow or relevant public to attract. Additionally, we characterized the types of users in the overlap of Portuguese IT startup communities, including profiles related to the IT area or the startup ecosystem. Finally, the results indicate that startups should follow users showing high levels of activity and popularity and are relevant to their field to increase the effectiveness of their social media strategies.

The remainder of the article is organized as follows: In "Related work" section presents relevant related work, namely, the key role that informed visualizations perform for social media analysis. The following section describes the methodology employed for this research, which is based on the NAV process model (Hansen et al. 2019). After that, we present an analysis and discussion of the results. Lastly, the conclusions of this research are presented, and paths for future work are delineated.

Related work

Social media platforms are essential digital marketing tools for small businesses like startups. This section explains how we can analyze communities created by startups when using social media platforms. We present a literature review of the methods and tools available for mining social media data using visualization, including network analysis and community detection algorithms, focusing on their application to study social communities on Twitter. Furthermore, we describe the role of social media in facilitating online communities. Lastly, we explain some relationships between startups and their social communities.

Visual analytics in social media analysis

Social media can be a valuable data source for businesses to extract digital marketing knowledge. Additionally, it can serve as a means to interact with their clients and potentially facilitate funding. Social media platforms generate two types of data: content data and interaction data. The content is usually found in a non-structured format, such as text and images. It can be retrieved from tweets and users' comments. On the other hand, the interaction data can be represented by a network structure (a graph). An example of this interaction is the *following* relation, i.e., when a user follows another. Other examples are actions, such as: *likes*, *shares*, *replies*, and *mentions*.

Social media analysis is no more than the process of extracting information from social media data (Hansen et al. 2019). In Serrat (2017), the authors explain that such analysis can either focus on the social actors or their relationships. However, social media platforms generate high amounts of data, which can make it difficult to understand and analyze fully. Visual analytics is a promising approach for dealing with the challenges of understanding complex data (Keim et al. 2008). It aims to explore complex data through visualization using interactive visual interfaces. Visualization techniques can uncover social media patterns and trends and support data interpretation (Wu et al. 2016). Furthermore, it helps gather insights from larger datasets, combining visualization techniques with the human dimension for enhanced data analysis. The NAV methodology arises from the need to combine network analysis with visualization, and it can be applied to each type of social media data to attain different goals.

Recent studies employ social media analysis through visualization regarding the content data. Saura et al. (2023) study tweets and apply topic modeling and sentiment analysis intending to mine the opinion of Twitter users about open innovation. They used a graph-based visualization to unveil the relation between the topics. Hu et al. (2017) also analyzed tweets and performed topic modeling. However, this study's originality lies in the design of a particular technique for visualizing the content of unstructured social media text. Likewise, Smith et al. (2014) developed a new visualization to disclose the relationships between words and topics in topic models applied to unstructured social media data. The creation of novel visualization methods is key since it acts as an enabler to improve data understanding and unveil new insights. Hingle et al. (2013) use Twitter content to extract dietary behaviors and highlight that data visualization helped identify relationships between diet-related behavioral factors.

In the same way, the literature highlights visualization methods and intrinsic data aggregation as tools to understand and extract knowledge from social media interaction data. When a user follows another, this relation is represented by a directed link (edge), with the source being the follower and the sink on the followed one. The users connected through the *following* relation generate the so-called social graph (Gabielkov and Legout 2012). The visualization of the social graph can reveal network features that can help answer important research questions. As an example, Molla et al. (2014) performed sentiment analysis of user Twitter contents and applied the results to color the edges of the graph. This visualization highlighted where, in the social graph, negative, neutral, and positive opinions about a company existed. Abdelsadek et al. (2018) applied a community detection algorithm to a social graph, visually revealing the community's structure and related characteristics.

Based on the previous works, we can conclude that visualization is an essential tool in order to extract insights and knowledge from social media data. Additionally, we can improve the visualization by selecting and computing features that will be applied to the structure, such as color or format.

Social graphs and communities detection

As previously mentioned, the social media interaction data derived from the action of "user following another user" is encapsulated into a network structure called the social digraph (Gabielkov and Legout 2012). In the digraph, a node represents a user and an edge represents the user-following-user relation.

Studies regarding social graphs can be performed either at the node level or at the network level. Antonakaki et al. (2021) explains methods that can be applied to node-level studies of social graphs to measure users' activity, popularity, and influence. Activity means how frequently the user interacts. In the case of Twitter, activity is measured by the number of tweets and retweets the user performs. Popularity measures how well a user is recognized, which usually can be estimated by the number of *followers*. A simple popularity measure is the *Structural Advantage* (Cappelletti and Sastry 2012), a ratio between the number of *followers* and of *followings*. Lastly, influence estimates how a user's action influences (the actions of) other users, being the most used metric at the node level for studies involving this type of graph. Influential users are better disseminators of information through social platforms because they are more central in the graph. Consequently, graph centrality measures like *PageRank*, *betweenness centrality*, and closeness centrality are applied to evaluate the user's influence (Das et al. 2018). Furthermore, a recent work by Esposito et al. (2022) evaluated the relationship between network centrality measures and a firm's success. The work's results suggest success has a strong positive association with centrality measures of the firm and its large investors.

Regarding the network level, network metrics enable quantitative comparison between graphs and analysis of temporal evolution (Hansen et al. 2019). Between the metrics used, we can find counts of nodes and links, average counts, or the application of concepts such as *density* and *centrality*. Antonakaki et al. (2018) used the average node degree and the average of incident edges to measure the evolution of a Twitter social graph over time. Said et al. (2019) conclude that Twitter communities have unique attributes that may impact the social media usage of their users.

Another way to dissect and extract information from a network is to apply algorithms that output some relevant structure or characteristic in the data. An essential concept in networks is that of the group or community: a set of nodes more densely connected between themselves than to others. The methods that find those groups are called community detectors and work as cluster algorithms (Hansen et al. 2019). These social media communities are essential for business, enabling a fast way to cultivate online brand awareness (Zaglia 2013). The community detection algorithms commonly used in the literature are based on modularity optimization. Modularity measures the strength of the division of a graph. High values imply the graph has dense connections between the module's nodes and sparser connections between nodes of the different modules (Blondel et al. 2008). The modules represent the clusters and, in this case, the communities. The modularity optimization algorithms will explore every node if the modularity score increases when changing between modules. The specific steps and parameters depend on the algorithm used since, in the literature, many adaptations exist depending on the graph characteristics. Regarding social networks, Devi and Poovammal (2016) performed a complete review of the applicable options. One algorithm that stands out for social media platforms is found in the work of Leicht and Newman (2008), which considers the direction of the edges (connections). An example of an application of modularity optimization in Twitter communities can be found in Cruickshank and Carley (2020). The authors used multi-view modularity clustering to characterize and analyze hashtag COVID-19 pandemic Twitter communities.

In summary, social graphs represent social media relationships formed by users following each other. Analyzing these graphs at the node level provides insights into individual users' activity, popularity, and influence. At the network level, based on modularity, community detection algorithms can identify highly connected community nodes, acquiring information that is valuable for businesses looking to build brand awareness.

Startups and social media communities

Existing literature reveals investigations on the possible connections between startups and social media platforms. Lugović and Ahmed (2015) found a positive correlation between startups' Twitter usage and the total investment in their country of origin. Zhang et al. (2017) analyzed startups' Facebook and Twitter metrics, discovering that active engagement positively correlates with startup crowdfunding success. Ko and Ko (2021) conducted a social media analysis regarding fashion startups using Instagram as the data source. They conclude that the startups presenting a higher number of *followers* showed a higher probability of succeeding in crowdfunding projects, meaning their popularity on Instagram helps raise funds. Hadley et al. (2018) conducted a study regarding startups to analyze how their influence and popularity may affect their funding by combining US-based technology startups with venture capitalists and using Twitter as the data source. The authors found that the more central startups in the network, i.e., the most influential ones, received better funding and presented a more significant revenue. A similar study by Esposito et al. (2022) found that the network centrality of a firm and its large investors positively affects business success.

Ruggieri et al. (2018) aimed to identify trends in thriving startups' digital activity. The study indicates that startups predominantly use digital platforms because of their cost-effective functionality. In fact, social media platforms possess a widespread reach, are easy to access, and incur low operating costs, making them the ideal digital marketing gateway for startups to monitor the market. Furthermore, the study inferred that a community of clients and companies, as service providers, is crucial for business success.

Communities are fundamental for a positive impact on digital platforms on the startup, primarily social media communities. since they provide positive or negative opinions on both the products and the companies. Word-of-mouth is critical in everyday oral communication, creating an impression or idea about a specific subject (Keller

2007). In the realm of digital platforms, opinions are called electronic word-of-mouth (eWOM) (Hennig-Thurau et al. 2004), and social media are ideal tools for eWOM. Chu and Kim (2011) describe that eWOM enables the creation of a large community, which allows for increased digital engagement via social interactions, such as *comments*, *likes*, *shares*, and *followings*. The large quantity of those interactions might help raise a positive feeling in the social media profile (Wolny and Mueller 2013).

Following the literature presented, the relationship between startups and social media platforms provides startups with cost-effective digital marketing opportunities to monitor the market and create communities of users and service providers. These communities help spread positive opinions about the products and companies via electronic word-of-mouth, increasing digital engagement through social interactions. Several studies have found that startups with active engagement on social media platforms have a higher chance of succeeding in crowdfunding projects or receiving better funding from venture capitalists.

Methodology

To visualize each startup's Twitter community, we employed a methodology based on the NAV process model (Hansen et al. 2012). This methodology stresses the need for a heavy interactive process built around the following phases: (1) Define the visualization goal, (2) Collect and structure data, (3) Interpret data, and (4) Report results. Figure 1 illustrates the process pipeline, describing each one of the step phases.

As we aim to understand how the *following/follower* relations create communities, in the first step, we defined our goal as that of constructing an informed visualization of Portuguese IT startups social media communities in Twitter, in alignment with the previously proposed research questions.

The next step consisted of collecting and transforming the social media data extracted via the startups' Twitter accounts into structured data. This study case features the startups in the information technology domain that are active on Twitter, founded by Portuguese, or have headquarters in Portugal. The eight chosen startups are: *attentiveMobile*, *codacy*, *DefinedAi*, *feedzai*, *prodsmart*, *Talkdesk*, *Unbabel*, and *Virtuleap*. The names of the startups are presented using the Twitter account username. *attentiveMobile* is a B2B company that offers a personalized mobile messaging platform; *codacy* is an automated code review platform; *DefinedAi* is a company that develops artificial intelligence training data services and solutions; *feedzai* is an artificial intelligence startup, and its core business is finance risk management; *prodsmart* deals with transforming factories into digital and smart ones by employing automation software to control workflows and production; *Talkdesk* is a platform to support sales teams for customer satisfaction and cost savings; *Unbabel* enables companies to serve customers in their native language with a scalable translation across digital channels; Lastly, *Virtuleap* offers a virtual reality application that promotes brain health, supported by a library of games designed by neuroscientists.

Then, we extracted data from the Twitter accounts of users who follow the companies and users whom the startups follow. The extracted data format corresponds to the Twitter user object, from which the following features have been considered: *id*, *screen_name*, *followers_count*, and *friends_count*. Subsequently, the data was structured into a

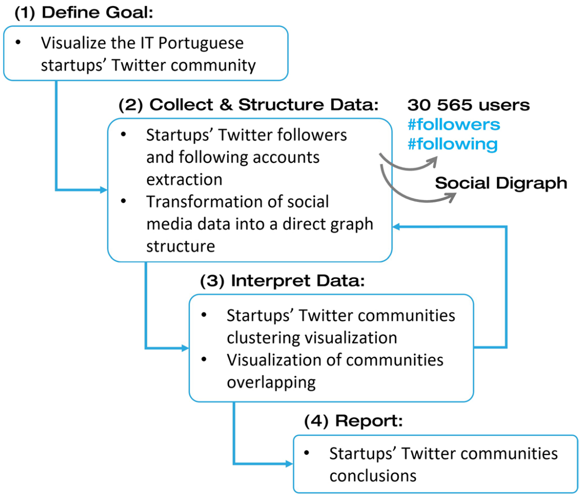

(1) Define Goal:
• Visualize the IT Portuguese startups' Twitter community
(2) Collect & Structure Data:
• Startups' Twitter followers and following accounts extraction
• Transformation of social media data into a direct graph structure
30 565 users
#followers
#following
Social Digraph
(3) Interpret Data:
• Startups' Twitter communities clustering visualization
• Visualization of communities overlapping
(4) Report:
• Startups' Twitter communities conclusions

Fig. 1 Current project's pipeline using the NAV process model (Hansen et al. 2019)

social digraph, with each node representing a user and each link denoting a following relationship, thus achieving a dataset with users that follow or are followed by startups.

After the organization of the data into a digraph structure, we recurred to using different visualizations to interpret the data and enable information extraction. In order to visualize the social graph, we employed *Gephi* (Bastian et al. 2009) and defined a layout and community clustering for data interpretation. As displayed in Fig. 1, steps (2), data structuration, and (3), interpretation, occur in an iterative fashion, where the visualization and respective interpretation may require a different organization of the data to explore emerging insights further. For this case study, this happened mainly when trying to visualize the communities' overlap. In the related literature, no specific visualization for the overlap between communities in a large social graph has been found, which led to a deeper exploration of possible visualization techniques, such as the ones presented in the coming sections, that in turn required different data organization.

Dataset

The Twitter API was used to extract relevant data, that is, data from the users that follow the eight chosen startups active on Twitter - *followers* - or users that the startups follow - *following*. The extraction occurred on May 31st, 2022, resulting in 30,565 accounts of Twitter users. Table 1 presents the number of *followers* and *following* users for each one of the companies, and Fig. 2 shows an illustration of the respective percentages in terms of the total number of links (edges) for each company using a stacked bar chart.

In terms of descriptive quantities, notably, all startups present a higher number of *followers* than of *followings*, meaning that their communities are mostly composed of Twitter users who follow their accounts. The only exception is *prodsmart*, for which the distribution of *followers* and *following*, although being approximately identical, shows

Table 1 Comparison of the startups counts of *followers* and *following*

Startups	*Following*	*Followers*
@attentivemobile	387	6842
@codacy	283	5048
@Definedai	170	1908
@feedzai	920	3132
@prodsmart	1037	905
@Talkdesk	685	7252
@Unbabel	1116	3627
@virtuleap	4	877

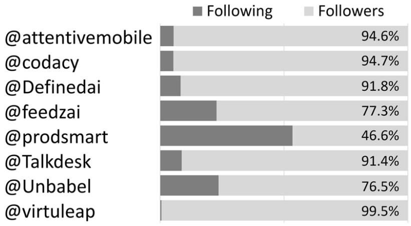


Fig. 2 Distribution of *followers* and *following* users by startup

that this company mostly follows others. Interestingly, while *Virtuleap* is the startup presenting the smaller community, it also presents an expressively higher percentage of *followers* than of *following*, being the startup showing the highest rate of *followers*, 99.5%, closely followed by *attentivemobile* and *codacy*, with 94.6% and 94.7%, respectively. *Talkdesk* is the startup showing the highest number of *followers* (7252), followed by *attentivemobile* (6842), and *codacy* (5048). *Virtuleap* stands out as the startup with the smallest number of links, likely attributed to its relatively recent foundation year: 2018.

Social digraph creation

After extraction, data has been structured into a social digraph, that is, a directed graph, where a node represents a user and a directed edge represents the user-following-user relation. This action resulted in a graph consisting of 30,565 nodes and 34,184 directed links/edges. The graph's density, 7.32×10^{-5}, indicates that it is a very sparse graph, meaning that it presents very few edges compared to the maximum possible number of edges for this number of nodes. This sparsity was expected since the graph nodes represent mostly users who follow the startups, while information about the other nodes those users may follow or about their followings was not extracted. No weights were used since we have not extracted any quantitative information towards this end.

To enable community visualization, a community detection algorithm was used. As previously mentioned in the related work section, based on the description of the analysis performed by Devi and Poovammal (2016), we chose to use the modularity algorithm for social digraphs created by Leicht and Newman (2008). Modularity measures the density of the connections within a graph's structure and groups it into modules. As expected, the results showed eight modules, one for each startup community, which has been validated by discovering the company at the center of each founded community. To evaluate the results, we measured the modularity score, ranging between 0 and 1, with higher values indicating a stronger community structure (McDiarmid and Skerman 2020). Our case study graph achieved a modularity value of 0.768, suggesting a robust community structure. We used the Python library *CDLib* (Rossetti et al. 2019) for the algorithm and evaluation.

For our analysis, we have selected a modularity-based algorithm due to its success in these social network scenarios showing few large communities, even though they usually suffer from the resolution limit problem. However, community detection algorithms are known to be quite unstable, with different algorithms sometimes producing different results. To gauge the stability of our findings, we have also tested an alternate method proposed by Traag et al. (2015) to evaluate the eventual differences that may arise. This alternate algorithm uses asymptotical surprise, a metric that, like modularity, is employed to evaluate the quality of community detection in networks. This metric is a statistical approach that calculates the probability of observing at least a certain number of internal edges within the communities, given the total number of edges in the network. We choose to apply this algorithm because it is nearly unaffected by the resolution limit problem, the modularity optimization primary weakness. The results with the new algorithm are mostly identical to the previous results and with few large communities. We decided to carry on using the results obtained by the algorithm proposed by Leicht and Newman (2008) for the analysis.

Figure 3 shows the number of nodes—users—for the found communities, comparing it with the respective number of *followers* plus *following* users of the startup, i.e., the total count of links of each of the companies. We have numbered each modularity class from 1 to 8, representing each one of the startups' communities.

Bearing in mind the quantities presented in Fig. 3, *Talkdesk* and *attentivemobile* present the larger communities and *Virtuleap* the smallest, as expected. Interestingly, the number of nodes for each of the communities is lower than the sum of each startup's number of *followers* and *followings*. This fact indicates that some of the users are shared between communities, meaning that the users who follow or are followed by the startups may intersect. *Virtuleap* appears as an exception, presenting an identical number of linked users and community nodes: 881 linked users and 873 nodes in the community. This means that only eight users are shared with different companies.

Data visualization and interpretation

This section presents visualizations of the social Twitter communities built around the different startups, focusing on the extent and characterization of the overlap between them. Overlap in this context means that a user follows more than one of the startups or is followed by more than one startup. To extract knowledge that may inform the creation of social media marketing strategies, We examined information at the user level for the ones found in an overlap situation to understand their type profiles and characterize the general communities of *followers* and *following*.

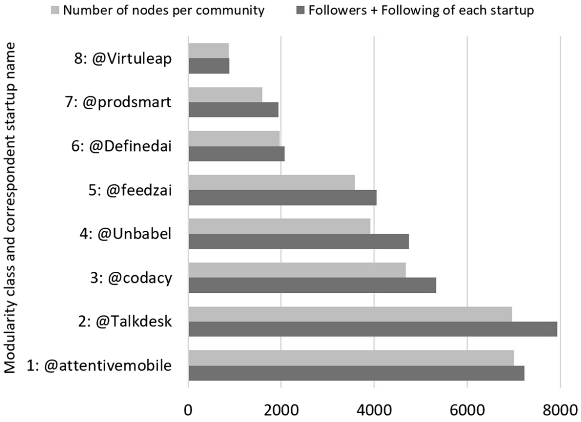


Fig. 3 Communities size distribution

Social digraph visualization

In a first step towards informative visualizations, the *circle pack layout* (Groeninger 2015) was applied to the social digraph to perceive what and how the communities are outputted by the modularity algorithm, as described in the previous section. This layout organizes the network into circles using a selected set of network features. In this case, the only selected feature has been the modularity class, which allowed us to create a visualization where each circle represents one of the communities. To help interpret the graph, the nodes and links were colored using different colors. Since the central (centroid) of each formed cluster was found to be each of the startups, the coloring rule used a color based on the logotype of each company to color the nodes corresponding to their modularity class value. However, the links were colored using a mix of the colors from the source and the target nodes. Figure 4 displays the visualization thus obtained for this social digraph, where each circle represents one community. The name of the correspondent startup community central point, the startup, is also shown.

As expected, the graph shows that the communities varying sizes, with the *Talkdesk* community showing its larger number of members and the *Virtuleap* community the smallest, as seen in Fig. 3. Furthermore, as anticipated, the graph shows a significant overlap between the several communities, with nodes (members) connected to more than one community in the social network graph. However, the exact degree of overlapping cannot be determined from this particular visualization alone, and additional analysis is needed to assess the implications of this overlap for the startups involved. It is essential to understand the degree of overlapping and how it may or may not differ in terms of *followers* and of *followings* since this distinction may have meaningful implications regarding actions in a startup's social media strategy.

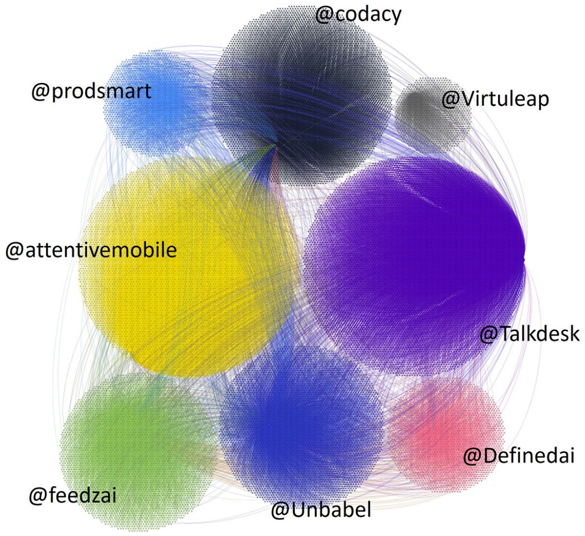


Fig. 4 Social graph visualization: circle pack layout using the modularity class

Overlap of communities

Understanding the communities' common points may help startups expand their networks and gain a competitive advantage in their respective markets. The following section aims to answer this question and grasp how the communities overlap. The overlap between communities in the context of social media refers to the situation where some users in the community around one startup are also part of other communities established around different startups. In other words, as depicted by the edges traversing different communities in Fig. 4, these users follow more than one startup or are themselves followed by more than one startup.

To distinguish between the two different situations—when a startup follows a user or when a user follows a startup—we divided the network in two: one representing only the user-*follow*-startup relation and the other representing the startup-*following*-user relation. The resultant graphs' dimensions show the difference in number between these two features that have been previously noticed in the analysis of Fig. 2: while the graph with the startups' *followers* has 27,682 nodes and 23,270 edges, the one for the users followed by the startups consists of 4306 nodes and 4225 edges.

However, for a better understanding of what is the overlap between startups, the visualizations must be more comprehensive. Data grouping is required since we discovered that some of the users in the overlap are shared between two or more communities. To accomplish this, we created two matrices: one for the *followers* of the startups and the other for the users the startups are *following*. The users in common for each combination of startups were counted, resulting in an overlap with 1289 *followers* (Fig. 5) and one with 249 *following* users (Fig. 6). The Python library *upsetplot* (Lex et al. 2014) was used to visualize the matrices, and the resultant plots represent the overlap in both of the

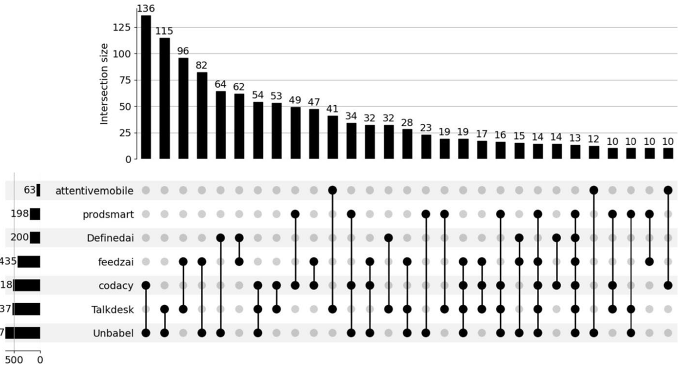


Fig. 5 Followers overlap visualization

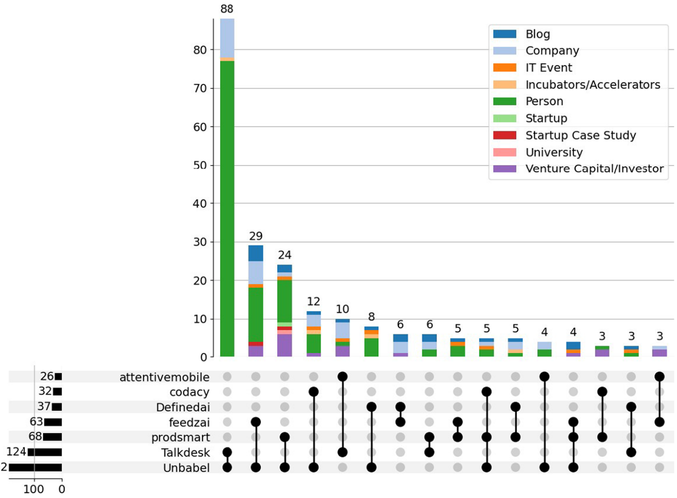


Fig. 6 Following overlap visualization, including user categories

domains: Fig. 5 plots the different combinations of startups having *followers* in common and how many and Fig. 6 regards combination and counts for startups *following* shared users.

In fact, Fig. 5 is a subplot of the total visualization of the overlap with *followers*. Since a considerable number of users was found to be in overlap, for a more effective visualization, the data were filtered by applying a threshold to show values only when the number of users in common was above nine. As seen in the previous section, *Virtuleap* displays only eight users in an overlap situation, and thus *Virtuleap* appears not to be sharing *followers* with other startups in this scenario where the threshold was applied. Notably, the startup sharing most *followers* with other startups is *Unbabel*, which shares a total of 667 *followers* with all the other six startups, even if under different combinations. The one sharing the least is *attentivemobile*, with a total of 63 shared *followers*. Furthermore, *attentivemobile* only overlaps with the startups showing the three highest numbers for overlapping. Interestingly, all these shares are duets that is these *followers* follow only two companies: *Talkdesk* (41 shared *followers*), *Unbabel* (12 shared *followers*) and *codacy* (10 shared *followers*). Nevertheless, this is an unexpected result since *attentivemobile* displays the biggest counts, both in its community dimension and for the total number of linked users. The pair of startups with more *followers* in common is *codacy* and *Unbabel*, showing an overlap of 136, followed by Talkdesk and *Unbabel* with 115, and *Talkdesk* and *feedzai* with 96. The visualization presents eight trios, with the one sharing more *followers* composed by *Unbabel*, *Talkdesk*, and *Codacy*, with a total of 54 users in common. Additionally, we can observe two quartets (with 19 and 16 users), one quintet (with 14 users), and one sextet (with 13 users).

Next, the shared *following* users have been analyzed. Figure 6 displays the overlap of the *following* by the startups, which may express coincident digital marketing options between them. Knowing the users in or not in the overlap can help to direct a digital marketing strategy. Therefore, we decided to manually categorize the 249 users in the overlap. Since the categorization was manual, we chose a set of global and vague categories of the startup ecosystem to facilitate manual categorization. The annotation procedure looked at the user's profile and bio description (usually stating the type of profile) and searched in Google for verification if needed. This annotation resulted in categories and in the colored visualization shown in Fig. 6. Similarly to what has been done with the *followers*' scenario, we applied a threshold and considered values only above two shared *following* relations. Again, *Virtuleap followings* do not appear in an overlap situation since this startup only follows 4 different users.

Both the startups showing the higher and the lowest levels of overlap are still the same: *Unbabel* shares more *following* users (192 users shared in total) and *attentivemobile* shares the least (26, in total). Notice that the latter has exchanged its behavior: while it shared many *followers* with other startups in the previous analysis, it now differentiates by following different users. By the analysis of Fig. 6, we can also conclude that *Talkdesk* and *Unbabel* are the pair presenting the highest number of *followings* in common, 88 users. Interestingly, most of the overlap occurs among Twitter users that are from the category "Person", mostly experts in the core business field of these two startups: applications encompassing natural language processing. The next profile of common *followings* are "Company" and, naturally, "Incubators/Accelerators." The next pairings and groups of startups show much less following in common, as can be noted by the abrupt decrease shown from the second column of Fig. 6. We can see two trios: one consisting of five shared users and the other shares four. The trio sharing more *following* relations—*codacy*, *prodsmart* and *Unbable*—consists of: the CEO of *codacy* (a "Person"), a Portuguese journalist (a "Person"), the Lisbon Investment Summit (an "IT Event"), *beta-i* (a "Company"), and a Portuguese blog (a "Blog").

Overlapping users characterization

In this section, we study the users that were found in overlapping communities. Understanding who startups follow and who follows them is important to better characterize the overlap.

Regarding the *followers* overlap, we found it appealing to understand what users follow most of the startups. These users are defined as the ones that follow four or more companies, comprising a total of 68 users. Next, node-level metrics were applied to evaluate their activity and popularity levels. These metrics were used both for the 68 users selected in the *followers* and for the 249 users selected in the *following*. Concerning user activity, we retrieved the total number of tweets and retweets in their Twitter profiles. We used a version of the *Structural Advantage* for popularity, the FF ratio (Cappelletti and Sastry 2012), that involves the number of *followers* and of *followings*:

$$\text{FF Ratio} = \frac{\#followers}{\#followers + \#following}$$

This ratio indicates how popular a user is on the social media platform, with higher ratios indicating that a user has more *followers* than is *following* others. Values between 0 and 0.5 indicate that the user is not particularly popular, following more users than being followed.

Concerning the *followers*' overlap, Table 2 shows the percentages for each type of Twitter user that has been encountered in this set of shared users and also the corresponding average values for the FF ratio and activity of each of the types.

Clearly, more than half of the users are classified as "Person," accounting for 56.9% of the total, followed by "Blog" (10.8%), "Incubators/Accelerators" (10.8%), "Company" and "Venture capital/Investor," both showing 6.2%. The categories showing the least are "IT event" and "Startup," with 4.6%. The type "Incubators/Accelerators" displays the highest average FF ratio (0.65), followed by "IT event" (0.60), "Venture capital/Investor" (0.54), all above-average level of popularity. The remaining types show less favorable FF ratios, especially the type "Person", which shows an average FF ratio of 0.27.

Regarding the activity levels, "Incubators/Accelerators" shows the highest average levels (3988), indicating that this type of user is more engaged with their *followers* than the remaining user types. Next, we see "Blog", with 2450, "Person" (2245), and "Company" (2216), all presenting similar levels of activity. Finally, we have "IT Event" (1184), "Venture capital/Investor" (989), and "Startup" (287).

Table 2 *Followers'* overlap: percentage of profiles encountered and node-level measures

User type	n	%	FF ratio (average)	Activity (average)
Person	37	56.9	0.27	2245
Blog	7	10.8	0.44	2450
Incubators/accelerators	7	10.8	0.65	3988
Company	4	6.2	0.38	2216
Venture capital/investor	4	6.2	0.54	989
IT event	3	4.6	0.60	1184
Startup	3	4.6	0.34	287

In terms of the overlap that exists in the *following* relations, that is, the users that startups follow, Table 3 presents the percentage of each type of user in this overlap, as well as the average FF ratio and the activity counts. The majority of the users followed by the startups are "Person", accounting for 53% of the total, followed by "Company" (18.10%), "Blog" (10.50%), "Venture capital/Investor" (8.50%), "IT Event" (4.40%), "Incubators/Accelerators" (2.80%), and "Startup" (1.20%). In this profiling, we can also encounter the type "University" (0.40%), albeit showing the least number of followings. The specific university is Instituto Superior Técnico, Lisbon, Portugal, from which many of the case study startups are either spin-offs or from where their founders obtained their degrees.

Comparing these users' activity levels, we see that the type "Blog" shows the highest average activity level (107,924 tweets/retweets), which is consistent with this type's main function. A "Blog" engages with its audience by regularly sharing meaningful content, which also explains why this is one of the most popular user types found in this overlap. In fact, when comparing this set of blog users with one of the *followers* previously discussed, the averages now are considerably higher than before, which entails that the blogs followed by startups are respected and credited blogs in this ecosystem. The next ranking position in terms of activity is occupied by the type "Person" that, with a 28,999 average count of tweets/retweets, positions itself rather distant from "Blog". "Company", showing an average of 26,49 showing an average of 26,494, follows closely. All the remaining types show considerably less activity when compared with any of the previous ones. It should be noticed, however, that all these user types display higher activity levels on average than those found in the *followers* set. Namely, the lowest activity count for the *followings*—3216 average tweets/retweets (Table 3)—is still higher than the highest count for the *followers* (Table 2).

In terms of popularity, we can see that the user types with the highest FF ratios are "Blog" and "University" (both attaining 0.95), followed by the startups of this case study (0.88), whose popularity levels are close to "Company" (0.87), again close to "Venture capital/Investors" (0.84) and "Startup" (0.84) popularity levels. On the other hand, user types "IT Event", "Incubators/Accelerators," and "Person" show the lowest FF ratios, but still, all of these users can be classified as popular since any of them shows to have more *followers* than *following* others.

Notably, our case study startups mostly share the action of following persons and companies relevant of the ecosystem. Other *following* types are "Venture Capital/Investor"

Table 3 *Followings'* overlap: percentage of profiles encountered and node-level measures

User type	n	%	FF ratio (average)	Activity (average)
Person	132	53	0.71	28,999
Company	45	18.1	0.87	26,494
Blog	26	10.5	0.95	107,924
Venture capital/investor	21	8.5	0.84	13,563
IT event	11	4.4	0.74	18,759
Incubators/accelerators	7	2.8	0.74	3216
Startup	3	1.2	0.84	8459
Startup case study	3	1.2	0.88	7752
University	1	0.4	0.95	6664

and “Blog.” While both show a relatively high FF ratio, indicating popularity, the average activity levels are quite different. “Venture Capital/Investor” activity is not expressive, which may indicate that these users look upon Twitter more as an observational or promotional tool than with the intent of engaging with other users. “IT Event” and “Incubators/Accelerators” show the lowest average FF ratios and activity levels, suggesting these users albeit being both quite popular, are not as active on this Twitter ecosystem as the remaining types.

Significantly, the categories of users found in the studied overlaps all belong to the startup universe, highlighting the interconnectedness of this ecosystem. Upon examination of the categorization of users in both overlaps, it becomes evident that most of the users who follow startups and those whom startups follow are individuals, and the second most common user type that startups follow is companies. Furthermore, startups follow users showing higher activity levels, with an average of 19,669 tweet/retweet count, which strikingly compares with those users that follow startups, showing expressively lower activity levels. In terms of popularity, startups follow users that are more popular than those who follow them, with an average FF ratio of 0.46 among *followers*, compared to an average of 0.84 among those whom startups follow.

Conclusions

Startups, known for their innovation and limited resources, must raise funding and reach customers on a restricted budget. Social media platforms offer a cost-effective gateway to various communities, enabling startups to achieve their goals and expand their business. Active engagement on social media can lead to better funding and create communities of users and service providers. This work investigates how startups fare on social media platforms, namely on Twitter, and if they create their own communities. Startups can benefit from data-driven projects using social media data, like the one in this study, as a strategic digital marketing knowledge source and unlock the power of social networks.

Our primary research goal was to understand if the follower/following relations on Twitter’s social graph create social communities around startups. Using Portuguese IT startups as a case study, we proceed with collecting and treating the needed Twitter data to create meaningful visualizations, enabling us to extract relevant knowledge about the communities. The case study data, using eight IT startups having some type of connection with Portugal, was organized into a social digraph, representing the users and links between the different users found in the data, resulting in a graph with over 30,000 nodes. Applying a community detection algorithm enabled the identification of communities in the data. Notably, the results showed that the communities were built around our eight chosen startups. By encoding the color of the social graph, the created visualization highlighted each one of the startup’s community of users. Thus, we showed that IT Portuguese startups form their own social communities on Twitter and that these communities heavily relate to the fact that these companies are startups and Portuguese-related. Next, we used other types of visualizations, paired with manual user categorization and node-level metrics, that enabled us to characterize the found communities and find out that, as expected, these communities show an interesting degree of overlap between them, either from the perspective of the startups' *followers*, as from the perspective of whom the startups are *following*.

We discussed the concept of overlap between communities on social media, which occurs when users belong to multiple communities established by different startups. We presented two graphs, one for *followers* and one for *following*, and analyzed the overlap between the startups. The resultant plots provide a comprehensive understanding of the social digraphs of startups and show the overlap between different startups.

As understanding communities is essential, we did two visualizations representing the overlap, one showing the startups' *followers* and the other showing whom they follow, to analyze the overlap between startups. The resulting graphs fully represent the startups' communities overlap. Then, we manually categorized the users in the overlaps. We discovered that all categories of users belong to the startup universe, highlighting the interconnectedness of this ecosystem. Examining user categories in both overlaps revealed that most users who follow startups and those whom startups follow are persons. Companies represent the second most prevalent user category that startups choose to follow, whereas blogs and incubators/accelerators are the second categories that most follow startups. In addition, startups tend to follow users who post high volumes of tweets and have high popularity levels. On the other hand, those who follow startups have low activity levels and are not popular.

Theorectial and managerial implications

As stated in the related literature, social media platforms offer startups affordable digital marketing opportunities to monitor the market and establish user and service provider communities. This study proposes a methodological process for social media community analysis on platforms like Twitter based on Network Analysis and Visualization (Hansen et al. 2019). The specific process displayed in this study involves five steps. First, select the users intended to be analyzed as the center of the communities. In this case, the center users were the eight chosen startups. Second, collect information on the *follower* and *following* users. Third, perform data transformation, including: user classification and user profile description, using popularity and activity metrics; data structuring into a digraph; creating two distinct tables, one for the *followers* and another for the *followings*. Fourth, visualize the data: use clustering algorithms and color to illustrate the communities formed. Additionally, use visualization tools like *upsetplot* to gain insights into the overlaps between these communities. Finally, the fifth step consists of concluding and data-driven supported strategies gained by analyzing the communities created by the selected central users. This process can be applied in future social media community analysis studies.

This study's first managerial contribution consists of a viable process for the visualization of a startup's community and when in an ecosystem, understanding the existence of shared *followers* or *followings*. Startups can use this method to monitor similar companies, select users to follow, study which users others follow, and perceive their community. They can build marketing campaigns for extending it as needed, facilitating the creation of a wider social media community that might benefit them or benefit from the ecosystem.

The second implication relies on the type of users found in the IT startup's overlapping communities. After manually categorizing 317 Twitter users, we found that the user's profiles in the overlap of the Portuguese IT startup communities are: "Person," "Company," "Blog," "Venture Capital/Investor," "IT Event," "Incubators/Accelerators," "Startup," and, exceptionally, "University." The first three categories found were all profiles of the IT area or of the startups' core business. Furthermore, the last five categories relate to the startup's ecosystem; thus, the ecosystem is mostly considered a closed environment. Examining the overlap and the type of users comprising it enables us to perceive the communities' common points, which may help startups expand their networks and gain a competitive advantage in their respective markets.

The third relevant managerial implication relies on the activity and popularity of the users in the overlap of the communities. Startups seem to follow users who post high volumes of tweets and have high popularity levels, while those who follow them generally show low average activity levels and also, on average, are not considered as popular. Accordingly, the study recommends that startups keep studying their Twitter users' activity and popularity profiles to stay relevant in their field and reach a wider number of other users.

Limitations and future work

Like all studies, the study has limitations that should be addressed in future research. Firstly, this study used eight Portuguese IT startups to develop a proof of concept. Future work should explore wider communities of startups, study startups working in different areas or industries, and be based in different countries to confirm and expand our results. Secondly, we focused only on social media communities created by startups on Twitter. Future research should compare community creation and overlap within other social media platforms. Thirdly, we manually performed the user categorization, enabling us to understand the types of community users overlap. However, if larger in scale, future studies should automatize the categorization process to allow the scalability of the process. Fourthly, this research focuses on the social graph created by the action of following on Twitter. Therefore, future works should consider extending the graph with more social variables, for example, using the number of interactions (likes, replies, and mentions) between two users as edges' weights. Lastly, future research should consider the life cycle phases of startups as described in a related work (Peixoto et al. 2023) and explore the potential influence between those phases and the startups' Twitter activity.

Author contributions

Conceptualization: All; methodology: All; software: AP; validation: All; investigation: All; writing original draft preparation: AP; writing, review, and editing: All; supervision: AA e NA; All authors have read and agreed with the published version of the manuscript.

Funding

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) namely by UIDB/04466/2020 and UIDP/04466/2020 (ISTAR_Iscte); UIDB/04152/2020 (MagIC/NOVA IMS); UIDB/50021/2020 (INESC-ID); and UIDB/03126/2020 (CIES_Iscte).

Availability of data and materials

The code used to extract the Twitter user accounts and to create the social digraph is available at https://github.com/arhpo-iscteiulpt/TwitterSocialDigraph.

Declarations

Ethical approval and consent to participate

This declaration is not applicable.

Competing interests

The authors declare no competing interests.

Received: 14 June 2023 Accepted: 13 September 2023
Published online: 20 September 2023

References

Abdelsadek Y, Chelghoum K, Herrmann F, Kacem I, Otjacques B (2018) Community extraction and visualization in social networks applied to Twitter. Inf Sci 424:204–223. https://doi.org/10.1016/j.ins.2017.09.022

Antonakaki D, Ioannidis S, Fragopoulou P (2018) Utilizing the average node degree to assess the temporal growth rate of Twitter. Soc Netw Anal Min. https://doi.org/10.1007/s13278-018-0490-5

Antonakaki D, Fragopoulou P, Ioannidis S (2021) A survey of Twitter research: data model, graph structure, sentiment analysis and attacks. Expert Syst Appl 164:114006. https://doi.org/10.1016/j.eswa.2020.114006

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp. https://doi.org/10.1088/1742-5468/2008/10/P10008

Cappelletti R, Sastry N (2012) Iarank: ranking users on twitter in near realtime, based on their information amplification potential. In: 2012 international conference on social informatics, pp 70–77

Chu S-C, Kim Y (2011) Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. Int J Advert 30(1):47–75. https://doi.org/10.2501/IJA-30-1-047-075

Cruickshank IJ, Carley KM (2020) Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering. Appl Netw Sci. https://doi.org/10.1007/s41109-020-00317-8

Das K, Samanta S, Pal M (2018) Study on centrality measures in social networks: a survey, vol 8, no 1

Devi JC, Poovammal E (2016) An analysis of overlapping community detection algorithms in social networks. Procedia Comput Sci 89:349–358

Education First (2022) Ef epi 2022 – ef english proficiency index. https://www.ef.com/wwen/epi/. Accessed 13 June 2023.

Esposito C, Gortan M, Testa L, Chiaromonte F, Fagiolo G, Mina A, Rossetti G (2022) Venture capital investments through the lens of network and functional data analysis. Appl Netw Sci 7(1):42. https://doi.org/10.1007/s41109-022-00482-y

Gabielkov M, Legout A (2012) The complete picture of the Twitter social graph. In: CoNEXT Student 2012—proceedings of the ACM conference on the 2012 CoNEXT student workshop, pp 19–20. https://doi.org/10.1145/2413247.2413260

Groeninger M (2015) Gephi - circular layout. https://github.com/mgroeninger/gephiplugins/ tree/master/modules/CircularLayout. Accessed 13 June 2023

Hadley B, Gloor PA, Woerner SL, Zhou Y (2018) Analyzing VC influence on startup success: a people-centric network theory approach. Stud Entrep Struct Change Ind Dyn. https://doi.org/10.1007/978-3-319-74295-3_1

Hansen DL, Rotman D, Bonsignore E, Milic-Frayling N, Rodrigues EM, Smith M, Shneiderman B (2012) Do you know the way to SNA?: A process model for analyzing and visualizing social media network data. In: 2012 international conference on social informatics, pp 304–313

Hansen D, Shneiderman B, Smith MA (2019) Analyzing social media networks with nodexl: insights from a connected world. Morgan Kaufmann, Burlington

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? J Interact Mark 18(1):38–52. https://doi.org/10.1002/dir.10073

Hingle M, Yoon D, Fowler J, Kobourov S, Schneider ML, Falk D, Burd R (2013) Collection and visualization of dietary behavior and reasons for eating using twitter. J Med Internet Res 15(6):1–16. https://doi.org/10.2196/jmir.2613

Hu M, Wongsuphasawat K, Stasko J (2017) Visualizing social media content with sententree. IEEE Trans Vis Comput Graph 23(1):621–630. https://doi.org/10.1109/TVCG.2016.2598590

Keim D, Andrienko G, Fekete J-D, Görg C, Kohlhammer J, Melançon G (2008) Visual analytics: definition, process, and challenges. Information visualization, vol 4950 LNCS. Springer, Berlin, Heidelberg, pp 154–175

Keller E (2007) Unleashing the power of word of mouth: creating brand advocacy to drive growth. J Advert Res 47(4):448–452. https://doi.org/10.2501/S0021849907070468

Ko J, Ko E (2021) What fashion startups should know before launching Crowdfunding projects: focusing on Wadiz reward Crowdfunding. J Glob Fash Mark 12(2):176–191. https://doi.org/10.1080/20932685.2020.1870521

Leicht EA, Newman ME (2008) Community structure in directed networks. Phys Rev Lett 100(11):118703

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H (2014) Upset: visualization of intersecting sets. IEEE Trans Vis Comput Graph 20(12):1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Lugović S, Ahmed W (2015) An analysis of twitter usage among startups in Europe. In: Infuture2015: e-institutions-openness, accessibility, and preservationproceedings. INFuture, pp 299–308

McDiarmid C, Skerman F (2020) Modularity of erdős-rényi random graphs. Random Struct Algorithms 57(1):211–243

Molla A, Biadgie Y, Sohn KA (2014) Network-based visualization of opinion mining and sentiment analysis on twitter. In: 2014 international conference on IT convergence and security, ICITCS 2014, 2012–2015. https://doi.org/10.1109/ICITCS.2014.7021790

Peixoto AR, de Almeida A, António N, Batista F, Ribeiro R (2023) Diachronic profile of startup companies through social media. Soc Netw Anal Min 13(1):52. https://doi.org/10.1007/s13278-023-01055-2

Portugal Digital & Startup Portugal (2021) Portugal, the best place to startup. https://portugaldigital.gov.pt/wpcontent/uploads/2022/02/Portugalthebestplacetostartup.pdf. Accessed 13 June 2023

Rizvanović B, Zutshi A, Grilo A, Nodehi T (2023) Linking the potentials of extended digital marketing impact and start-up growth: developing a macrodynamic framework of start-up growth drivers supported by digital marketing. Technol Forecast Soc Change. https://doi.org/10.1016/j.techfore.2022.122128

Rossetti G, Milli L, Cazabet R (2019) Cdlib: a python library to extract, compare and evaluate communities from complex networks. Appl Netw Sci 4(1):1–26

Ruggieri R, Savastano M, Scalingi A, Bala D, D'Ascenzo F (2018) The impact of digital platforms on business models: an empirical investigation on innovative start-ups. Manag Mark 13(4):1210–1225. https://doi.org/10.2478/mmcks-2018-0032

Said A, Bowman TD, Abbasi RA, Aljohani NR, Hassan SU, Nawaz R (2019) Mining network-level properties of Twitter altmetrics data. Scientometrics 120(1):217–235. https://doi.org/10.1007/s11192-019-03112-0

Saura JR, Ribeiro-Soriano D, Palacios-Marqués D (2021) Setting B2B digital marketing in artificial intelligence-based CRMs: a review and directions for future research. Ind Mark Manag 98(August):161–178. https://doi.org/10.1016/j.indmarman.2021.08.006

Saura JR, Palacios-Marqus D, Ribeiro-Soriano D (2023) Exploring the boundaries of open innovation: evidence from social media mining. Technovation. https://doi.org/10.1016/j.technovation.2021.102447

Serrat O (2017) Knowledge solutions: tools, methods, and approaches to drive organizational performance. https://doi.org/10.1007/978-981-10-0983-9

Sifted (2020) Top Portuguese startups to follow in 2020. https://sifted.eu/portugal-startups-top-rankings/. Accessed 24 July 2023

Smith A, Chuang J, Hu Y, Boyd-Graber J, Findlater L (2014) Concurrent visualization of relationships between words and topics in topic models. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces, pp 79–82

Tanner J (2023) The ultimate guide to twitter marketing for small businesses and startups. https://www.socialmediacollege.com/blog/the-ultimate-guide-to-twitter-forstartups-small-businesses. Accessed 24 July 2023

Traag VA, Aldecoa R, Delvenne JC (2015) Detecting communities using asymptotical surprise. Phys Rev E Stati Nonlinear Soft Matter Phys. https://doi.org/10.1103/PhysRevE.92.022816

Ward S (2020) Why your business should use twitter. https://www.liveabout.com/top-reasons-why-your-small-businessshould-use-twitter-2948523. Accessed 24 July 2023

Wolny J, Mueller C (2013) Analysis of fashion consumers' motives to engage in electronic word-of-mouth communication through social media platforms. J Mark Manag 29(5–6):562–583. https://doi.org/10.1080/0267257X.2013.778324

Wu Y, Cao N, Gotz D, Tan YP, Keim DA (2016) A survey on visual analytics of social media data. IEEE Trans Multimed 18(11):2135–2148. https://doi.org/10.1109/TMM.2016.2614220

Zaglia ME (2013) Brand communities embedded in social networks. J Bus Res 66(2):216–223. https://doi.org/10.1016/j.jbusres.2012.07.015

Zhang Q, Ye T, Essaidi M, Agarwal S, Liu V, Loo BT (2017) Predicting startup crowd funding success through longitudinal social engagement analysis. Proc Int Conf Inf Knowl Manag Part F 1318:1937–1946. https://doi.org/10.1145/3132847.3132908

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.