INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Searching for associations between social media trending topics and organizations

João Pedro Sousa Henriques

Master in **Integrated Business Intelligence Systems**

Supervisor

Doctor João Carlos Amaro Ferreira, Assistant Professor

ISCTE-University Institute of Lisbon

Supervisor

Doctor Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor

ISCTE-University Institute of Lisbon

December, 2020

TECNOLOGIAS
E ARQUITETURA

Searching for associations between social media trending topics and organizations

João Pedro Sousa Henriques

Master in **Integrated Business Intelligence Systems**

Supervisor

Doctor João Carlos Amaro Ferreira, Assistant Professor

ISCTE-University Institute of Lisbon

Supervisor

Doctor Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor

ISCTE-University Institute of Lisbon

December, 2020

Resumo

Este trabalho foca-se na forma como as micro e pequenas empresas podem tirar partido dos *trending topics* para as suas campanhas de marketing. Os *trending topics* são os tópicos mais discutidos em cada momento nas redes sociais, particularmente no Twitter e no Facebook. Enquanto o acesso aos *trending topics* é gratuito e generalizado, os especialistas em marketing e o software especifico são dispendiosos, pelo que as pequenas empresas não têm o orçamento para suportar esses custos. O principal objetivo é procurar associações entre *trending topics* e empresas nas redes sociais e para isso foi criado um protótipo chamado HotRivers. Uma solução que pretende ser acessível, rápida e automatizada. Foram realizadas análises detalhadas para reduzir o tempo e maximizar os recursos disponíveis a preço baixo. O utilizador final recebe uma lista dos *trending topics* relacionados com a empresa alvo. O HotRivers foi testado com diferentes técnicas de pré-processamento de texto, um método para selecionar tweets chamado Estratégia Centroid e três modelos, uma abordagem de vectores embedding com o modelo Doc2Vec, um modelo probabilístico com *Alocação de Dirichlet Latente*, e uma abordagem de classificação com uma Rede Neural Convolucional, selecionada para a arquitetura final. A Estratégia Centroid é utilizada nos *trending topics* para evitar tweets indesejados. Nos resultados destacam-se o *trending topic "Nike"* que tem uma associação com a empresa Nike e *#WorldPatientSafetyDay* que tem uma associação com a Universidade dos Hospitais de Portsmouth. Embora o HotRivers não possa produzir uma campanha de marketing completa, pode apontar a direção para a campanha seguinte.

Palavras-chave: Trending topics, Similaridade de texto, Classificação de texto, Associações.

Abstract

This work focuses on how micro and small companies can take advantage of trending topics for marketing campaigns. Trending topics are the most discussed topics at the moment on social media platforms, particularly on Twitter and Facebook. While the access to trending topics is free and available to everyone, marketing specialists and specific software are more expensive, therefore small companies do not have the budget to support those costs. The main goal is to search for associations between trending topics and companies on social media platforms and HotRivers prototype is designed to accomplish this. A solution that aims to be inexpensive, fast, and automated. Detailed analyses were conducted to reduced the time and maximize the resources available at the lowest price. The final user receives a list of the trending topics related to the target company. For HotRivers were tested different pre-processing text techniques, a method to select tweets called Centroid Strategy and three models, an embedding vectors approach with Doc2Vec model, a probabilistic model with Latent Dirichlet Allocation, and a classification task approach with a Convolutional Neural Network used on the final architecture. The Centroid Strategy is used on trending topics to avoid unwanted tweets. In the results stand out that trending topic *Nike* has an association with the company Nike and *#WorldPatientSafetyDay* has an association with Portsmouth Hospitals University. HotRivers cannot produce a full marketing campaign but can point out to the direction to the next campaign.

Keywords: Trending Topics, Text Similarity, Text Classification, Associations.

Acknowledgements

I would like to thank my family that always believed in me. This work would not be possible without them and the endless support and love from my girlfriend. Also, to my friends João and Nuno that were always ready to help me with great inputs. Also, a special thanks to Professor João Ferreira for all corrections and time invested in this work. I would also like to thank Professor Elsa Cardoso for the support given. Additionally, a special thanks to Professor Ricardo Ribeiro for all the suggestions made, which contributed deeply to this work.

Contents

List of Figures

List of Tables

Abbreviations

API	**A**pplication **P**rogramming **I**nterface
Bi-LSTM	**Bi**directional - **L**ong **S**hort **T**erm **M**emory
CNN	**C**onvolutional **N**euronal **N**etwork
CS	**C**entroid **S**trategy
DL	**D**istributed **L**ag
Doc-p	**D**ocument-**P**ivot
FPM	**F**requent **P**attern **M**ining
GFeat-p	**G**raph-Based **Feat**ure-**P**ivot
GTP	**G**eneral **T**ext **P**rocessing
IDF	**I**nverse **D**ocument **F**requency
JSON	**J**ava **S**cript **O**bjet **N**otation
LDA	**L**atent **D**irichlet **A**llocation
LSTM	**L**ong **S**hort-**T**erm **M**emory
MNB	**M**ultinomial **N**aïve **B**ayes
OMECO	**O**ne **S**top **S**hop **M**arketing **E**cosystem
PHU	**P**ortsmouth **H**ospitals **U**niversity
RMCH	**R**oyal **M**anchester **C**hildren's **H**ospital
SFPM	**S**oft **F**requent **P**attern **M**ining
STP	**S**pecialized **T**ext **P**rocessing
SMM	**S**ocial **M**edia **M**arketing
SVM	**S**upport **V**ector **M**achine
TBG	**T**rend **B**ipartite **G**raph

TF	**T**erm **F**requency
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency

Chapter 1

Introduction

The number of internet users is growing exponentially, as well as the offer of new social media platforms over the course of the years, and the amount of users has also increased substantially [1, 2]. Social media has the incredible power of making information, opinions and complains accessible to everyone [3]. An example that illustrates this fact is the video that captured the death of George Floyd, which gained international attention after hitting an astonishing number of comments, likes and shares across multiple social media platforms. As a result of that, protests appeared across the world [4].

The following examples given by the authors Hoffman and Fodor [5], in 2010, illustrate the power that social media has to destroy a marketing campaign and products reputation. The case study about Motrin, a medicine that was so criticized that climbed to the top of trending topics on Twitter and gained such visibility that reached mainstream media. One impressive fact is that all this happened throughout a period of 24 hours during a weekend. The other case study is related to a milk-based product, Raging Cow and the failed attempt to create a good reputation around it. The campaign was not well-received by the blogosphere community and bloggers attacked and boycotted the marketing campaign, making the product disappear from the market.

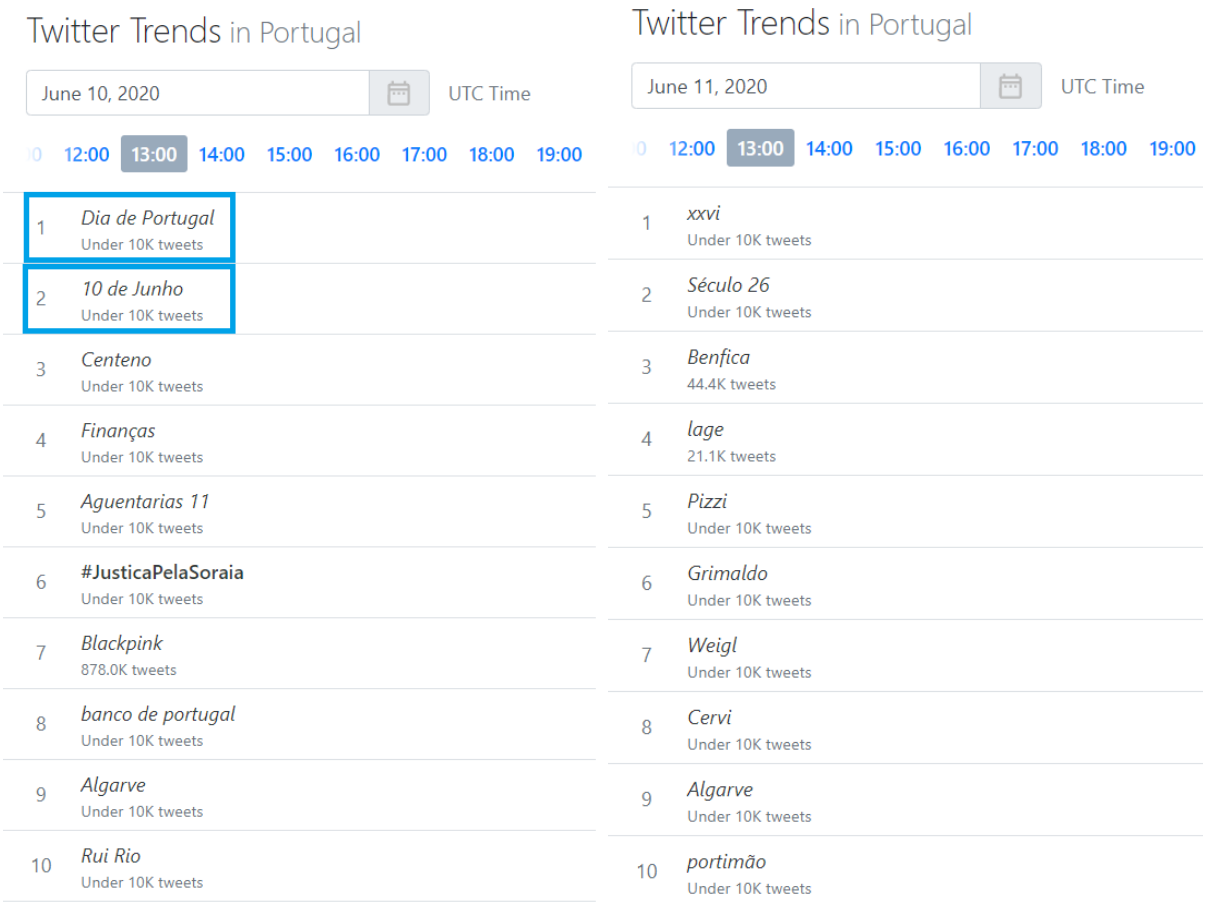
On the other hand, the next example is a mistake that could have damaged the image of Red Cross, but turned out to be a successful blood-donation campaign. In 2011, an employee from Red Cross made a tweet about drinking beer with an uncommon hashtag (#gettingslizzerd) from the company's account in the middle of the night. This event was noticed on Twitter getting attention from the users. In order to reverse this situation, Red Cross acknowledged the mistake and took action with humor [3]. Both companies Red Cross and Dogfish Head Brewery took advantage of the trending hashtag to their own benefit.

The given examples were triggered without intention and became a discussion topic across multiple platforms. Therefore, knowing the social media environment can be a very powerful tool to avoid harm or to improve social media metrics [3, 6].

1.1 Motivation

A trending topic is defined by Twitter as an emerging discussion topic that is popular in the present moment. To be considered trending, a topic, needs to be discussed more then what it usually is [7]. The authors Carrascosa et al. [8] defined trending topics as the official Twitter description of 2010 "the hottest emerging topics (or the "most breaking" breaking news), rather than the most popular ones".

Additionally, as the authors Zubiaga et al. [9] refereed in their work, in the year of 2011, trending topics became interesting to users, journalists, applications developers and social media researchers. Besides being new and relevant to people at that moment, the active time of a trending topic is limited [8]. Therefore, if it takes too long to evaluate trending topics, companies may not have time to do something effective [10]. As it can be observed in Figure 1.1 (a) the trending topics from 10 of June at 1 p.m are different from the ones of the following day at the same time (Figure 1.1 (b)).



(a) 10 of June top ten trending topics on Twitter adapted from [11]

(b) 11 of June top ten trending topics on Twitter adapted from [12]

FIGURE 1.1: Top 10 trending topics on Twitter on 10 and 11 of June

Trending topics are relevant to companies' marketing as the authors Carrascosa et al. [8] said "(...)Trending Topics present a comparable visibility to other traditional advertisement channels and thus they can be considered a useful tool in marketing and advertisement contexts.". Instead of companies spending more time and money to increase the visibility of their products or brand, they might take advantage of the already spoken topics on social media to reach their marketing goal.

There are companies already taking advantage of current events to communicate. On the 10 of June, Control a Portuguese brand, made a post on Instagram with the quote "it is day to raise the flag" (Figure 1.2 (a)) to take advantage of the national holiday of

Portugal. On the same day, the number one trending topic in Portugal on Twitter was Portugal's National day. (Figure 1.1 (a)).

On February 8, 2019, in the 21st round of the Portuguese first league, a match between *FC Porto* and *Vitória de Guimarães*, Marega, FC Porto player, was the target of racist chants and shouts by supporters of the *Vitória de Guimarães* team. This topic was very discussed in social and traditional media. Nine days after that event, Super Bock and Sagres, two Portuguese competitor beer brands, made a post together on Facebook with the quote "Against racism, there are no rivals"(Figure 1.2 (b)).

(a) Control Portugal post on Instagram adapted from [13]

(b) Super Bock and Sagres post on Facebook adapted from [14]

FIGURE 1.2: Control Portugal post on Instagram and Super Bock and Sagres post on Facebook

Finally, having an account on a social media platform is free and trending topics are easily available to everyone. Therefore, even organizations that may not have the budget to invest on specialized human resources and software to make social media marketing (SMM) can take advantage of trending topics. The key is to analyse which trending topics are relevant to each company in a timely manner.

1.2 Objectives

The goal of this work is to find text associations between the social media account of an organization and trending topics. Also, compare the results by using text similarity, a probabilistic and classification task approach.

Another objective is to create a solution that is inexpensive, fast in deliver results, automated, and with no need for specialized staff. Additionally, understand if companies working in the same market have associations with the same trending topics.

This proof of concept is accomplished by building a simple prototype, the artifact of this work. The prototype is called HotRivers and needs to be capable of:

1. **Collecting data from a social media platform:** By using the name of the company social media account (e.g. Adidas, Nike, Pull & Bear, and others) and desired location of the trending topic (e.g. United Kingdom, Lisbon, New York, and more);
2. **Preparing data:** Apply pre-processing techniques to clean and transform the data;
3. **Modeling:** Use different approaches to measure the similarity between data from trending topics and companies' social media accounts.

This idea came as an answer to a challenge from a company's Portugal2020 project called One Stop Shop Marketing Ecosystem (OMECO). In OMECO project was identified that micro and small enterprises have a lack of specialized resources and budget for SMM strategies, because the various sectors involved (e.g. design, communication) are mainly composed by small companies working in a discontinuous course with cost effect and minimal response to consumer needs. The goal of the project was to automate and integrate those services and deliver high quality and low cost service.

Unfortunately, due to COVID-19 the OMECO project has surfered several schedule delays. However, to guarantee that this work was delivered on time, it was decided to work in parallel with OMECO project schedule.

1.3 Dissertation Structure

In this section, is introduced the structure of this work. The design science research methodology model [15] was adopted as a guideline to the current work. This work was organized into six chapters.

In Chapter 1 were presented the problem and the motivations to perform this work. Also, the goals and solution were defined. In Chapter 2 is defined what is social media, the impacts of marketing and the presence of companies in social media platforms. Additionally, which social media should be used for this work and why. Furthermore, a research was conducted on works done with trending topics and similar prototypes and systems. In Chapter 3 is introduced the surrounding of HotRivers and user requirements. Then, the architecture is detailed from high-level to low-level. In Chapter 4 is described the implementation of HotRivers phase by phase and is presented the minimal requirements for HotRivers' work and in each section is pointed out the problems and solutions found. In Chapter 5, four experiments are conducted. The first intends to demonstrate the result of all phases and to discuss what went wrong and which aspects need more testing. The second experiment continues testing the techniques and models that were not abandoned in the first experiment. The third and fourth experiments confirm that the chosen model is suitable for HotRivers. In Chapater 5, the results of associations between trending topics and companies are also discussed. Finally, in Chapter 6 is presented the conclusions, future work and limitations of the current work.

Chapter 2

State Of The Art

In this chapter is reviewed the definition of social media and the impact of marketing and firms on social media platforms. Also, is discussed which social media platforms should be used for HotRivers prototype and why the top fifteen social media applications in terms of numbers of users were analyzed based on a set of conditions created specifically to this work. A systematic review was conducted on similar works to HotRivers and on related works on trending topics. The studies found on the systematic review of Section 2.5 can be divided in trending topics classification, spam and spammers detection, trending topics detection, summarization and exploratory analysis, and last trending topics forescasting. In the end is summary of the works analysed during the systematic review.

2.1 Social Media

In 2010, Kaplan and Haenlein [16] defined social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content.". According to Carr and Hayes [17], in the year of 2015, "Social media are Internet-based channels that allow

users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others.". The Cambridge English Dictionary [18] has defined social media as a format of media that allow individuals to use a computer or cell phone to share information. While researchers have distinct ways to formalize a definition for social media sites, they seem to agree that a social media platform is an online application that allows users to create and share information content with others.

In 2010, Hoffman and Fodor [5], classified social media applications as Blogs (e.g. WordPress), Microblogging (e.g. Twitter), Cocreation (e.g. NIKEiD), Social Bookmarking (e.g. StumbleUpon), Forums and Discussion Boards (e.g. Google Groups), Product Reviews (e.g. Amazon), Social Networks (e.g. Bebo, Facebook, Linkedin) and Video and Photo-sharing (e.g. Flickr, Youtube). In the same year, according to Kaplan and Haenlein [16], social media can also be classified as blogs, social networking sites (e.g. Facebook), virtual social worlds (e.g. Second Life), collaborative projects (e.g. Wikipedia), content communities (e.g. Youtube) and virtual game worlds(e.g. World of Warcraft).

2.2 Marketing in Social Media

In the past, according to Tiago and Veríssimo [19], in 2014, marketers disseminated information related to the company or it's products through e-mail blasts, direct marketing, telemarketing, informational websites, television, radio, and others. However, if costumers are on social media then firms should be as well. A survey conducted by Pew Research Center [20] between 2005 to 2019 found that 72% of U.S adults in 2019 use at least one social media platform. Of those adults the age group from 18 to 29 and the group from 30 to 49, 90% and 82% correspondingly, use at least one social media platform. And almost 70% of the adults aged between 50 and 64 use at least one social media site. While those

numbers are impressive, it is important to understand what is the position of companies in embracing those new channels and the benefit of using them on the current markets.

In the year of 2014, an online survey was conducted on the managers of the largest companies in Portugal. The authors concluded that 87% of the managers agrees that digital presence improves information gathering and feedback. Also, 85% of them acknowledged that digital presence increases knowledge and 82% admitted that digital presence promotes internal and external relationships [19].

In other study made in 2013, the authors Abeza et al. [21], made a case study where they conducted interviews with the staff of running events. They concluded that the gain of using social media in relationship marketing in sport was getting higher acknowledgement from consumers, improved communication client-organization, better customer engagement and more efficient use of resources.

An alternative research in the year of 2012, made by Erdoğmuş and Cicek [22], studied the impact of social media on brand loyalty. The data was collected through questionnaires. They found out that the following items were positively related to brand loyalty, campaigns in social media, relevancy of the content, popularity of the content among other users and friends, and variety of platforms and applications.

2.3 Why Twitter?

A set of conditions were selected to choose which social media should be used in this work. One condition was that the core of publications on that social media platform was text, for example, messages, posts, micro-blogging publications. Instead of video, photography or image, which are not text-based. Even though chat messages or messaging services are text-based, they are not suitable for this work.

For this work, another condition is that legal entities have to have some visibility in a social media application. Visibility, in this context, is not paid publicity, sponsorship or partnership. But an user account that represents a legal entity.

The language used must be English and it was required to be a worldwide application and not specific to some part of the globe. English was the language chosen to be used in this work, because many sophisticated linguistic models were developed for English.

Additionally, this work uses the topics of the day, so it is important to choose a social media platform where the hottest topics are already filtered, because that's the focus of this work. Last but not least, the appliance to access the data must be easy and fast.

Table 2.1 was constructed in order to compare the top fifteen most used social media platforms and to see which conditions each one check [23]. As said before, messaging service providers such as Whatsapp, Facebook Messenger, WeChat and QQ were disregarded. Then, there are social networks based on videos or images like Youtube, Instagram, TIKTOK, Kuaishou, Snapchat and Pinterest which were also disregarded. While Sina Weibo was possible to explore and a few companies were represented there, in QZone was not clear if legal entities played a relevant role. However, both platforms were made for Mandarin speakers and were therefore disregarded. Reddit was excluded, because it is a social news media aggregation, and no legal entity has presence on Reddit [24]. The most proper candidates were Facebook and Twitter. Facebook was excluded too, due to the difficulty to access data and the hottest topic aggregator, because it is not clear how it works. Twitter seemed to be the most accessible of all and fulfilled all requirements.

2.4 Twitter, Tweets and Trending topics

Twitter is a microblogging social media platform with 14 years [5, 25]. Twitter allows users to follow others and to receive their messages called tweets. A tweet is composed by a text message up to 280 characters and can include URLs', photos, GIFs or video [26]. Sharing

Top 15 most used social media platforms						
Social Media Platforms	Text Based	Legal Entities Visibility	Available in English	World Wide Platform	Appliance Conditions Difficulty[3]	Hottest Topics Filter
Facebook	Yes	Yes	Yes	Yes	Hard	Yes
Youtube	No	-	-	-	-	-
Whatsapp	Yes	N.A.[2]	-	-	-	-
Facebook Messenger	Yes	N.A.[2]	-	-	-	-
WeChat	Yes	N.A.[2]	-	-	-	-
Instagram	No	-	-	-	-	-
TIKTOK	No	-	-	-	-	-
QQ	Yes	N.A.[2]	-	-	-	-
QZone	Yes	N.K.[1]	-	-	-	-
Sina Weibo	Yes	Yes	No	-	-	-
Reddit	Yes	No	-	-	-	-
Kuaishou	No	-	-	-	-	-
Snapchat	No	-	-	-	-	-
Twiiter	Yes	Yes	Yes	Yes	Easy	Yes
Pinterest	No	-	-	-	-	-

[1] N.K.: Not Known.
[2] N.A.: Not Applicable.
[3] Difficulty: Easy, Medium, Hard.

TABLE 2.1: Table with which characteristic each social media has

other user tweets is called "retweet" and it is represented on the message by a *RT* mark (retweet) [27]. A reply is an answer to a tweet or to a retweet. Other well-known marks are *@mention* (e.g. *@adidas*) for mentioning users and *#hashtag* (e.g. *#MondayMotivation*) to sign that a message is part of a specific topic.

Twitter uses an algorithm that is customized to each user, based on followers, interests and location, or not personalized, based on a specific location. The algorithm analysis which topics are popular at the moment, instead of those who have been popular for a while or continuously. That's how the hottest topics of the day are found [7].

2.5 Related Works Systematic Review

In order to search for similar prototypes, systems and works, the main research database used to conduct a systematic review was Scopus. As a second source for papers it was used Google Scholar, which needs more selective search query due to the great amount of works. The query used was the following:

("trending topics" OR "trend topics" OR "hot topics") AND (association OR relationship OR relation OR connection OR link) AND (companies OR firms OR corporation OR institution OR organization)

The number of documents was enormous. In order to distinguish the relevant from the non-relevant, it was used the filtering conditions described in Table 2.2. On the left side of the table are the acceptance criteria and on the right side the elimination criteria.

Filtering Criteria	
Inclusion criteria	Exclusion Criteria
Written exclusively in English	Not written exclusively in English
Work developed to English language	Work developed to other languages
Publication after 2010	Publication before 2010
Free or inside ISCTE's scientific license	Paid documents
Papers in conferences or journals	Paper or journals published in non-trust sources
Title, abstract or keywords related to association between trending topics and companies	Non-applicability to association between trending topics and companies

Table 2.2: Filtering criteria for associations between trending topics and companies systematic review

The search for the selected query gave 7,729 results as present in Table 2.3. Scopus filtering features were used to search for title, abstract and keywords related to the query used which returned 330 documents. The most common words were "hot topic", while fewer instances of "trending topics" and "trends topics" appeared. Those documents were individually analysed and applying the filtering criteria, only five works remained to be fully analysed. The five studies found were about the following topics: spam, exploratory analysis and event classification. Unfortunately, it was not possible to identify similar studies to this work.

Filtering Steps	Number of works
Search for query	7,729
Title, abstract or keywords related to association between trending topics to companies	330
Applied inclusion and exclusion criteria	5
Full-work analyse	0

Table 2.3: Filtering steps of systematic review related to association between trending topics and company

As no related studies were found, the next sections are about works on trending topics, including the three topics identified previously. It was given preference to those studies that used Twitter, Facebook or other social media trending topics as data. Even though there are many techniques to extract topics, the focus is on studies that used trending topics given by social media algorithms. The search query on Scopus were the following:

("trending topics" OR "trend topics") AND (classification OR detection OR analyze)

The filtering criteria used was the same present in Table 2.2 except that were accepted studies that used other languages than English (e.g Mandarin) and works that in the title, abstract or keywords were related to trending topics. The results of Table 2.4 show a total of 1,895 works related to trending topics. The exclusion of the words "hot topic" from the previous search query was due to the fact that is common to use these words to refer to hot topics in areas of study (e.g. In biology the king-raccoon is the hot topic at the moment (...)), which conducted to a lot of results irrelevant to this work. After filtering for titles, abstracts and keywords that matched with the chosen search query were identified 302 works. A careful analysis was conducted on those works and were chosen twenty-one studies.

Filtering Steps	Number of works
Search for query	1,895
Title, abstract or keywords related to association between trending topics to companies	302
Applied inclusion and exclusion criteria	75
Full-work analyse	21

TABLE 2.4: Filtering steps of systematic review related to trending topics studies

Those twenty-one studies were grouped into five different categories: classification, spam detection, summarization and topic detection, exploratory analysis and trending topics forecasting. These categories are explored in the following sections.

2.5.1 Trending Topics Classification

Some authors present trending topics classification as away to assign trending topics into different categories such as sport, politics, technology, and more. However, trending topics classification can also be in the sphere of event classification (e.g. festival and commemorative days, news, memes, and more). There are other authors that analyse real-time events such as crisis and emergency events.

In the year 2011, according to Zubiaga et al. [9], it was used an Support Vector Machines (SVM) to classify trending topics into four types: *news, current events, memes,* and *commemorative events*. They used up to fifteen different features and tweets were written in twenty-eight different languages. The model was very successful in classifying *current events* and *memes*, with 82.9% and 73.1% accuracy, respectively. However, it was not so good at *commemorative events* with only 13.2% accuracy.

In the same year, Lee et al. [28], attempted a text classification and a network-based classification approach to classify eighteen classes. The categories are *Art & Design, Books, Business, Charity & Deals, Fashion, Food & Drink, Health, Holidays & Dates, Humor, Music, Politics, Religion, Science, Sports, Technology, TV & Movies, Other News* and *Other*. For text classification, they used a Bag-of-Words with term frequency-inverse document frequency (TF-IDF) weights and the classifier was a Multinomial Naïve Bayes (MNB). For the network-based classification, they used the similarity of topics based on the number of common influential users and the classifier used was a C5.0 decision tree learner. In a total of 768 trending topics, the network-based classification accomplishes 70% accuracy, while the text classification approach achieved only 65% accuracy.

Another work made by Zhu [29] in 2018, proposed a real-time information filtering approach with text classification. The method is a mix of the MNB classifier and short text aggregation. They classified the trending topics into four labels, *entertainment, news, others,* and *sports*. The accuracy of this approach was 73.33%, the precision was 73% and the recall was 73.3%. Besides, the author stresses that MNB excels in speed that is

a necessary requirement to implement a real-time system. The model only needed 1.5 seconds to build and classify thirty trending topics.

In the year 2019, Shalini et al. [30] contrary to all the previous works presented this was on Facebook. They collected trending topics and analyzed the author's comment stance sentiment, which could be in favor of or against the target topic. The data were transformed into a Bag of Tricks, word embeddings such as Word2Vec [31], and GloVe [32] and pre-trained embeddings. The models used for binary classification were a ConvolutionalNeural Network (CNN) [33], bidirectional long short term memory (Bi-LSTM) and a Bag of Tricks classifier. The proposed system extracted and cleaned the data from Facebook, then data was converted into a Bag of Tricks, word embeddings, and pre-trained embeddings. Finally, the models were trained and evaluated. The Bi-LSTM model had the worst performance in all six topics. The Bag of Tricks had slightly better results than the CNN model. It was not clear which input worked better.

The proposed system by Liu et al. [34], in 2019, called Oasis, were design to detect incivility and to classify sentiments. They used in their system a combo of CNN [33] and Long Short-Term Memory (LSTM). It is a quite simple network that takes as an input a sentence made of word embeddings [31, 32]. The CNN-LSTM model achieved an F1-score of 98% on classifying offensive and non-offensive tweets. Additionally, the model accomplishes an F1-score of 67.9% on classifying the sentiment *neutral, happy, sad, hate,* and *anger*, and only an F1-score of 66.1% on *worry, love*, and *neutral*.

2.5.2 Trending Topics Spam and Spammers Detection

Trending topics spam detection is an important topic due to the influence on users. Those works are related to malicious spam that may contain unsafe URLs, hate campaigns, factless information, provoking synthetic trending topics to gain visibility and more. Despite the negative actions presented by these studies, this work is intends to use trending topics as a vehicle for companies' campaigns with no harm to the blogosphere community.

In the year 2010, other study research by Benevenuto et al. [35], investigate the presence of spammers on trending topics. They studied three topics *#musicmonday*, *Jackson* and *Boyle*. They analyzed the number of daily tweets and found that *#musicmonday* has a pattern of peaks on Mondays, *Jackson* had a huge peak of almost 700.000 around Michael Jackson's death anniversary and *Boyle* had peaks on Susan Boyle's performances on TV. The model used to classify spammers was a SVM. They analyzed what were the most important attributes, they determined that *fraction of tweets with URLs, age of the user account, and an average of followers per followees* was the top three more crucial. They were able to classify correctly 70% of the spammers and 96% the non-spammers.

Another interesting work was conducted by Stafford and Yu [36], in 2013, they attempted to understand if trending topics were manipulated by spammers. They discovered that the most influential attributes were *URLs per word, URLs, number of hashtags, numeric characters, the rank of the topic,* and *whether the tweet was a reply.* Also, they observed the frequency of those attributes in tweets and concluded that spammers use more URLs, hashtags, and numeric characters than non-spammers. They used a Naïve Bayes model and classified it correctly 90% of the time. The authors found that the presence of spammers cannot change the rank of a trending topic and they do not necessarily target the top one rank trending topic. One of the most relevant conclusions was "Spammers don't drive topics in Twitter, but they do attempt to piggyback on their visibility".

A study conducted by Antonakaki et al. [37], in the year 2016, analyzed extensively how spammers are using certain Twitter features to increased the visibility of their campaign and to avoid Twitter's spam detection algorithm. They collected about 150 million tweets for four months. They found that there are trending topics that continue active for more than twenty days, however, 80% only stay active for around two days. While approximately 75% of the trending topics have linked 100 different URLs, 90% have related 1,000 distinct URLs. Another interesting fact is that out of 4,593,229 URLs encountered 250,967 lead to a spam domain. The dataset used were 8,2 million users, 7.2% of those users published at least one spam link. Additionally, they found that spammers use google

search result to mask their URLs. The authors train a decision tree regression classifier that was able to recognize correctly 73.5% of the spammers.

In 2017, Dang et al. [38] researched how to detect trending topics controlled by groups of spammers and how to detect who were the users that belong to those groups. The goal of those groups is to pretend to be normal users arguing about a topic, however, they are the ones who generate and manipulate those topics that normal users are commenting on. The authors proposed a topology-based method to detect spammer groups dispersed in numerous trending topics. Also, they proposed a new similarity measure based on sub-graph ranking for detecting anomalous topics and to cluster the users that join the trending topics to detect those who belong to a spammer group based on their authority on the topic. They analyzed the behavior of those groups and regarding one specific topic they found that however it took three hours to have 1,400 users in the first hour, 90% of the spammers entered the topic. The authors point out that the reason for this is to make the false appearance of a real trending topic, but in reality, it is a topic intentionally created to deceive normal users. Even though spammers are very good to masquerade their identity when examined the retweet relationship is observed their intention. The author claims this approach may be useful to prevent influencing political orientation and more.

2.5.3 Trending Topics Detection and Summarization

Another studies focus on finding sub-topics or events on trending topics by detecting pikes of unusual messages, analyze temporal metrics, examine the text, geographic information and trained clusters, and probabilistic models. Trending topics detection is more related to finding topics instead of using Twitter, Facebook or another social media trending topics.

One of those works was a summarization framework, made by Bian et al. [39], in 2013, for trending topics in Sina Weibo, one of the biggest microblogs. They proposed

a multimodal latent Dirichlet allocation model to automatically generate trending topics summaries. Their system collects text and image, pre-processes all the data and discover and summarize subtopics by using image and text information. The output of the framework is both textual and visual summaries of the trending topics.

In the same year, 2013, Aiello et al. [40] experimented six trending topic detection approaches on three real-world major events. The Twitter datasets were *FA Cup Final, Super Tuesday Primaries,* and *US Elections*. The data were collected using the official hashtags as well as popular hashtags for the three events (e.g.*#Elections2012*). The models used were latent Dirichlet allocation (LDA), Document-Pivot (Doc-p), Graph-Based Feature-Pivot (GFeat-p), Frequent Pattern Mining (FPM), Soft Frequent Pattern Mining (SFPM) and BNgram. Additionally, they analyzed how quality is affected by the type of input data such as timespan, topic breadth and more and pre-processing techniques. BNgram accomplishes the best topic recall in all datasets. FPM achieved the higher keyword precision in FA Cup and Super Tuesday datasets. While LDA had the higher keyword recall on FA Cup, SFPM had on Super Tuesday and the United States Elections. The most complete topic descriptions are obtained by SFPM and LDA, on the other hand, the most precise topic descriptions are produced by FPM. Relating to a pre-processing conclusion, while steaming lower the performance, tweets aggregation seems to improve topic recall. The performance on LDA can be very affected by "noisy" events.

Another study made by Melvin et al. [41] in 2017, proposed a phrase network model to detect and summarize events from tweets. They claimed that topic modeling techniques and keywords frequency have disadvantages to detect events on Twitter. The model starts by characterizing all topics as clustering of high-frequency phrases. Then, all trending topics are recognized based on temporal spikes of the phrase cluster frequencies. Therefore, their model filters events from other trending topics using three conditions, the number of peaks, peak intensity and the variation over time. The authors claim that the performance of their model is better than similar approaches. Their model achieved an F1-score of 54%, a recall of 84%, and a precision of 40%.

While non of the previous work used sentimental features to detect trending topics in real-time or in datasets, Peng et al. [42], in 2015, introduced a sentimental feature non-parametric supervised real-time model to detect trending topic. Additionally, they used multiple features such as tweet volume, temporal and social authority information, user volume and others. Their model was composed of an SVM classifier with unigram features to classify sentimental time series. To detect real-time trending topics was used a supervised model with sentimental time series. The performance of their model did not have the best efficacy but was the fastest to respond. However, by combining multiple features they were able to improve efficacy and reply time. The model achieved 73.3% of F1-score, 82.3% of recall and 65.2% of precision.

Instead of summarizing the trending topic in the study made in 2015 by Sharma et al. [43], intended to automatically generate a summarization of the sentiments of the trending topics. Their approach, the TopicDetect, was able to compare trending topics across four locations and observed where the same topics appear. This allowed them to compare how sentiments varied in different locations. Additionally, their method was capable to summarize the sentiments in trending topics. They claimed that their proposed approach was effective and extensive to cover important topics.

Contrasting with the previous approach for trending topic summarization, in the year 2019, Singh and Shashi [44] proposed a framework to identify the trending topics related to news articles and then catch opinion diversity from multi-document summarization of unreliable news articles based on the trending topics. There were experimented bag-of-word with TF-IDF weight, word embeddings and document embeddings with the cluster algorithm k-means. The framework was simple to find and extract trending topics' tweets and to search for tweets containing news. Then, detect the URLs present on tweets and webscrap and select the news based on content, social and temporal features. Then, transform the news articles into vectors and identify unreliable news articles by clustering them. Finally, generate multi-document summarization of the news. The word embeddings used were Word2Vec [45] and the document embeddings model applied were Doc2Vec [46]. The author said that TF-IDF produced better results, but offered fewer options than

Word2Vecand Doc2Vec. Even though, the purity score was 0.98 to TF-IDF, 0.89 to Word2Vec and 0.95 to Doc2Vec, which means that both Word2Vec and Doc2Vec were close to TF-IDF approach. Additionally, they found that with more data the Doc2Vec approach could deliver higher-quality results.

2.5.4 Trending Topics Exploratory Analysis

Many statistical, social, geographical, and other studies were conducted on trending topics. Those works intended to understand how long trending topics stay active, how they are correlated to traditional media, why the same trending topic appears in different parts of the globe and many other questions. Those studies are important to better understand how trending topics work.

In the year 2012, according to Wilkinson and Thelwall [47], trending topics included a wide range of categories such as festivals or religious events, media events, politics, human interest and sports. They studied a total of 178 topics across multiple countries: the United Kingdom, New Zealand, Australia, India, the United States (US) and the Republic of South Africa. They concluded that Twitter follows an identical pattern to media news. The authors found it normal because media news influence what people talk about.

In accordance with Asur et al. [48], in 2011, the rise of a trending topic is driven by a log-normal distribution. Also, they concluded that trending topics do not live long and those who live have a decay of a geometric distribution. Additionally, the authors found that the number of followers and tweet-rate of users does not provoke trends. The most important attribute is the retweet by other users because it is associated with the content being shared. Last, the authors said that most of the content shared is news from traditional media.

Another study made by Annamoradnejad and Habibi [49], in 2019, analyzed for one full year the following Twitter trending topic features tweets count, trending time, language and lexical analysis, trend reappearance and time to be on top of the ranking of trending topics. They found that the bigger trending topics were written with six words, but more than half were made by a single word. The number of characters presented in trending topics varied between two and thirty-one, however, the average was approximately thirteen characters and two words. Relating time to reach the top ranking positions was around 36.2 minutes to be at the top ten and 91.5 minutes to get to rank one. Although, not all trending topics get to the top, yet 977 got to rank one in less than ten minutes. Concerning reappearance on the list of trending topics, on average a trending topic only emerges 1.5 times, but there was a trending topic that appeared on the top ten ninety-five times in a year. The trending time was also examined, on average a trending topic only stayed at rank one for 99.3 minutes and in the top ten for 85.6 minutes. The longer duration of a trending top in 2018 was thirteen hours. The most predominant languages on the trending topics were English, Arabic and Korean.

2.5.5 Trending Topics Forecasting

Trending topic prediction is another field of study. The goal is to forecast what could be the next trending topics based on the information available.

In the year 2013, Liu et al. [50] proposed a framework cable of predicting if some posts could become a trending topic or not. Their framework started by collecting posts and user data and then filtering the data in quantity, quality and user specific features. Finally, create a feature vector to served as input to an SVM model. The results showed that this approach is better than Quantity-Centric approach, but narrowly. The accuracy of the authors' model is 87.8% and the Quantity-Centric approach is 84.8%. Additionally, it was found that their method offered more reliability than the Quantity-Centric approach because the false-positive ratio for the authors model is only 32.3% and for Quantity-Centric approach is 86.9%. Hence, they were able to predict trending topics with a low

figure of errors. Relating to the features used, the quantity feature didn't work due to the fact that the model classified almost all the test topics to be trending. While quality features have impacted in deciding a trending topic, user features have influence in avoid misclassification of a non-trending topic.

Althoff et al. [51], in 2013, made an analysis of Google Trends and News, Twitter and Wikipedia from US and Germany trending topics for a full year between September 2011 and September 2012. Additionally, they proposed a nearest neighbor forecast approach for the life cycle of a trending topic based on semantically similar topics. Their system first finds semantically similar topics, then searches for identical patterns and last predicts the behavior. They found that trending topics on Twitter and Wikipedia are faster to appear and to disappear than on Google. Also, they discovered that Google is more specialized in sports, celebrities, entertainment and politics whereas Twitter is more specialized in sports, celebrities, entertainment, products and holidays, last Wikipedia is more specialized in celebrities, entertainment and incidents. Regarding the lifetime of a trending topic on Twitter, 44% of them lived only one day and 24% two days, on Wikipedia, 50% of trending topics had a life cycle of one day and 16% of two days and on Google, 17% of the trending topics lived seven days,14% four days and 13% six days. The results of their forecast approach for a period of time up to fourteen days is nine to forty-eighth thousand views closer to the true value of views with an error of 19-45%.

In the same year, in 2013, Giummolè et al. [52] studied the capability of Twitter to predict and lead to a Google trend arise. They measured social and webtrends through lexical similarity and represented them in a Trend Bipartite Graph (TBG). Also, they used the TBG to analyze and compare the time series of Twitter and Google trending topics. The time series regression models used were Autoregressive model, Distributed Lag (DL) model, and Autoregressive Distributed Lag model. They discovered that DL model that used Google as the dependent variable and Twitter as the explanatory are significant 60% of the time and the model explained approximately 75% of the variance. Additionally, they found that 43% of the times Twitter trending topics caused an identical

Google trend. The best models were those who used both Twitter and Google information, they could explain 80% of the variance.

2.6 Major Findings of the Systematic Review

In a total of 9,624 works filtered in this systematic review, the title, abstract and keywords of 632 studies were examined. Of those 632, 80 were picked and after that 21 were selected to be fully analyzed.

One of the conclusions of the systematic review is that non of the works found tried to accomplish the goal of this work. As said before five categories of studies were found: classification, spam detection, trending topics detection and summarization, exploratory analysis and forecasting.

The first category examined was classification in Section 2.5.1. Different approach were used such as SVM, CNN, LSTM and others to solve problems of classify into classes, real-word events, offensive and sentiments. In Table 2.5 is a summary of classification works. In general it was found that the results were good with an accuracy higher than 70%.

Authors	Year	Approach used	Findings
Zubiaga et al. [9]	2011	SVM with 15 different features	Classified current events with an accuracy of 82.9% memes with 73.1%
Lee et al. [28]	2011	- MNB with bag-of-words TF-IDF - C5.0 decision tree learner	MNB accomplished 70% and C5.0 decision tree 65% accuracy
Zhu [29]	2018	MNB with short text aggregation	Model achieved 73.33% of accuracy and build and classifies in 1.5 seconds
Shalini et al. [30]	2019	- Bag of Tricks classifier - CNN - Bi-LSTM	The best were Bag of Tricks, then slightly worse CNN and last Bi-LSTM
Liu et al. [34]	2019	CNN-LSTM (A mix of a CNN and a LSTM)	Classification binary of offensive tweets attained 98% of F1-score and 67.9% of F1-score on classified sentiments

Table 2.5: Trending topics classification summary

Then the next category was Spam and Spammer Detection in Section 2.5.2. It become a relevant field due to the malicious intentions of some users. Their goal is influence and

deceive users on social media platforms. In Table 2.6 is a summary of Spam Detection works. It was found that spammer are very good on their job on deceiving users and Twitter algorithm. Also, the spammer use more URL and hashtags than a normal user. Additionally, this type of works are important to prevent hate propaganda, influence political orientation and more.

Authors	Year	Approach used	Findings
Benevenuto et al. [35]	2010	SVM	- The most important features are fraction of tweets with URLs, of the user account, and an average of followers per followees - Model classified correctly 70% of the spammers
Stafford and Yu [36]	2013	Naïve Bayes	- The model classified correctly 90% of the time - The most crucial features are URLs per word, URLs, number of hashtags, numeric characters, the rank of the topic, and whether the tweet was a reply
Antonakaki et al. [37]	2016	Decision Tree Regression Classifier	- Only 80% of the trending topics stayed active for around two days - The model recognized correctly 73.5% of the spammers
Dang et al. [38]	2017	Proposed topology-based framework	- Spammers are very good in masquerade their identity - This approach may be useful to prevent influential political orientation and more

TABLE 2.6: Trending topics spam and spammers detection summary

The following category is Trending Topics Detection and Summarization in Section 2.5.3. Hashtags maybe not be enough to understand what is the topic of discussion, since they are mainly composed by a single word [49]. So, this category has the practical use of produce informative reports about the trending topics. A few authors used trending topics detection as apart of their framework to generate summarization of documents. Trending Topics Detection and Summarization summary is in Table 2.7. Multiple approaches were used by the authors and highlighting that one work was able to produce a summary of textual and visual trending topics. Additionally, that Doc2Vec is capable of return high-quality results, steaming lower the performance and tweets aggregations have a tendency to improve and noisy events affects LDA.

The fourth category is Exploratory Analysis in Section 2.5.4. It is a relevant subject, because the authors study "laws" of the trending topics, in other words, discover what drives a trending topics, how and why they become trending, what are the key feature, and many other questions. Exploratory Analysis summary is in Table 2.8. The author found that the key attributes is the retweets and the users stats do not have influence. A

Authors	Year	Approach used	Findings
Bian et al. [39]	2013	Multimodal latent Dirichlet allocation	The framework output is textual and visual summaries of the trending topics
Aiello et al. [40]	2013	- LDA - Doc-p - GFeat-p - FPM - SFPM - BNgram	- The best topic recall: BNgram - The most complete topic description: SFTM and LDA - The most precise topic description: FPM - Steaming worsen the performance and tweets aggregation seems to improve topic recall - LDA is affected by noisy events
Peng et al. [42]	2015	SVM with unigram features	- Use of sentimental features - The model achieved the highest response time and accomplished 73.3% of F1-score
Sharma et al. [43]	2015	Proposed the algorithm TopicDetect	Approach effective and extensive to cover important topics
Melvin et al. [41]	2017	Phrase Network model	The model accomplishes an F1-score of 54%
Singh and Shashi [44]	2019	- Bag-of-words with TF-IDF - Word2Vec - Doc2Vec - k-means	- Bag-of-words with TF-IDF had a purity score of 0.98, Doc2Vec of 0.95 and Word2Vec of 0.89 - Bag-of-words with TF-IDF had slighly better performance, but offered less options than Word2Vec and Doc2Vec - Doc2Vec delivered the highest-quality results

Table 2.7: Trending topics detection and summarization summary

trending topic rise 1.5 times and on average a trending topic is composed by two words. Twitter act in accordance with the media news pattern.

Authors	Year	Findings
Asur et al. [48]	2011	- Trending topics are driven by a log-normal distribution - Trending topics have a decay of a geometric distribution - The most important attribute is the retweet by other users - The number of followers and tweet-rate of users does not provoke trends - The most content shared is news from traditional media
Wilkinson and Thelwall [47]	2012	Twitter follows identical pattern to media news
Annamoradnejad and Habibi [49]	2019	- Half of the trending topics were a single word - On average the trending topics had 30 characters and 2 words - Approximately a trending topic needed 36.2 minutes to get to top 10 and 91.5 minutes to be at top 1 - 977 trending topics in 1 year got to rank 1 in less than 10 minutes - A trending topic emerge 1.5 times - The longer duration of a trending topic was 30 hours

Table 2.8: Trending topics exploratory analyses summary

The last category is Trending topics forecasting in Section 2.5.5. This subject can be a powerful tool for any agent of marketing, due the fact of being able to predict what are the next trends. In Table 2.9 is a summary of forecast works. It was found that twitter cause Google trends 43% of the times. The duration of a trending topics on Twitter and Wikipedia approximately half of the times are one day, while the Google trends live much longer.

Authors	Year	Approach used	Findings
Liu et al. [50]	2013	- SVM - Quantity-Centric approach	- SVM achieved 87.8% of accuracy and Quantity-Centric approach 84.8% of accuracy - SVM had a lower false-positive ratio - Quality feature had influence in deceiving a trending topic - User feature had impact in avoiding misclassification of non-trending topics
Althoff et al. [51]	2013	- Nearest Neighnor - Forecaste approach	The duration of the lifetime of a trending topic: - on Twitter was 1 day in 44% and 2 days in 24% of the trending topics - on Wikipedia, 50% of trend stayed 1 day and 16% 2 days - on Google, 17% lived seven days, 14% four days and 13% six days - The model forecast 9,000-48,000 views up front to 14 days with an error of 19-45%
Giummolè et al. [52]	2013	- TBG - Autoregressive model - DL model - Autoregressive - Distributed Lag model	- The DL model explained approximately 75% of the variance - When Google was the dependent variable and Twitter the explanatory variable the DL model was significant 60% the time - Twitter trending topics caused an identical Google Trend 43% of the times

TABLE 2.9: Trending topics forecast summary

In conclusion, for this work the model Doc2Vec [44] seem suitable for the text similarity approach, the model LDA [40] for the probabilistic approach and the model CNN [30, 33, 34] for the classification task approach.

Chapter 3

HotRivers Prototype

In this chapter are presented the needs of HotRivers users, requirements needed for the prototype to work properly, the high-level architecture specification and finaly a more detailed explanation of each phase.

As discussed in Chapter 1, this simpler prototype was built to find an association between the social media account of an organization and trending topics. It was pointed that most of the trending topics have a life span of 24h to 48h, but they can survive longer [51]. It is important to clarify that only the primarily top 10 trending topics were used. That is the window of opportunity to act. In that time, HotRivers needs to find associations, marketing teams need to analyze the results, and then create engagement on social media platforms. All of this happens in a short period of time.

3.1 HotRivers Environment

The environment surrounding HotRivers is composed of Twitter users and companies. In Figure 3.1 is observed a company and Twitter users actors, a gray area that is part of Twitter authority, a yellow area that corresponds to an outside process of HotRivers sphere, HotRivers prototype and lasts the output.

Twitter users make tweets about a discussion topic marked with an *#hashtag*. The Twitter algorithm then classifies all those topics selecting some as trending, as explained in Section 2.4. Later, HotRivers collects the trending topics by specific locations.

On the other hand, companies use their social media Twitter account to fulfill their marketing agenda and to interact with the Twitter community. The inputs of HotRivers are social media accounts (e.g. Adidas) and trending topic locations (e.g. United Kingdom). Finally, the output of HotRivers goes to the company in the form of a list containing all the trending topics that matched.

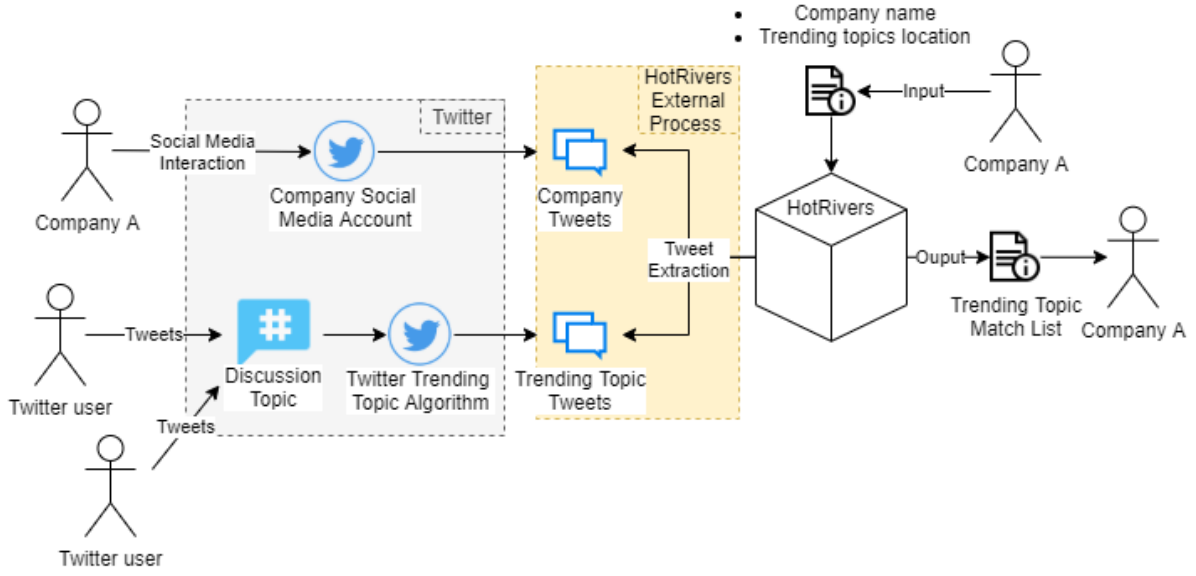


FIGURE 3.1: Environment of the actors and HotRivers prototype

3.2 HotRivers Users Requirements

Both actors have requirements that are meant to be satisfied as can be observed in Table 3.1. Those requirements were concluded as a mix of literature review and OMECO project needs.

General Data Protection Regulation laws came to protect users more. Also, Twitter want their data to be used with responsibility by applying a rule to give access to their data. So, it is important to guarantee that Twitter users' data are also being used with transparency and to fair use.

Regarding companies, their desire is having cheap, fast and useful information that can help them to take decisions. Thus, trending topics results must be ready in less than a day of work, for companies to have time to analyze and to decide what to do. Because the target of the OMECO project was micro and small companies this service must be low-cost. Plus, those organizations have a shortage of trained personnel to manage marketing platforms, meaning that more automated processes are better. Additionally, many studies have concluded that social media exposure and interaction increase sales, brand awareness and other metrics [3, 5, 8, 53]. Even though, measuring and increasing social interaction is a request from companies is out of the scope of this work.

Actor	Requirements
Company	Trending topics match information
	Information available in working time
	Low-cost solution
	More automated solution
	Increase social media interaction
Twitter users	Transparency and ethical use of the data

Table 3.1: Resume of actors needs. In gray color is the requirement that is out of the scope of this work

3.3 High-Level Architecture

The presented architecture was designed to be capable of collecting data from Twitter, then able to perform clean and transform processes and finally train models. The division of the architecture was made in three phases, two inputs and one output, as it is observed in the Figure 3.2. Although separating the inputs was a necessity imposed by Twitter Application Programming Interface (API) architecture, it also, allows a better organization, give more structure to the development work, and grant independence of the processes.

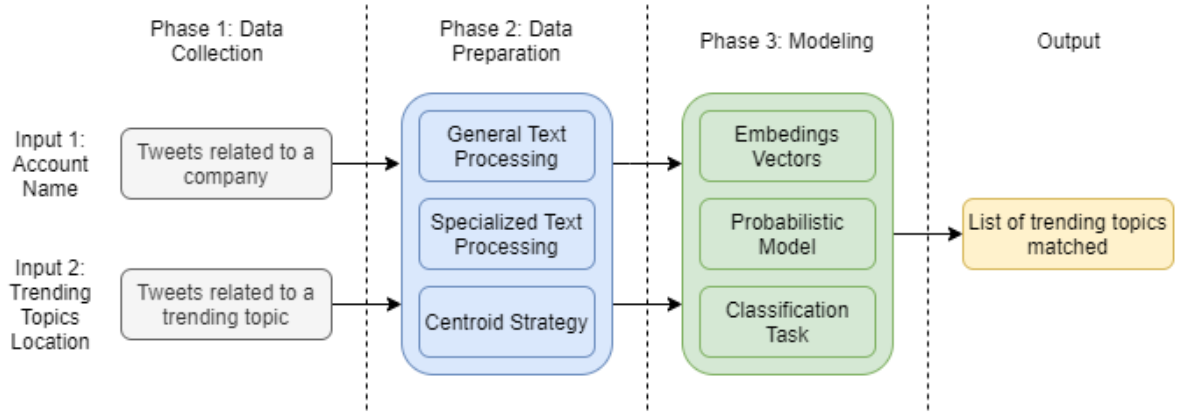


FIGURE 3.2: HotRivers main modules and its phases. Each phase have a different color. Along the this work the color scheme is kept. The modules of phase two and three were tested and in the end it was chosen the best modules for the prototype

The first phase is data collection where HotRivers collects data from Twitter. There are two components, one collects tweets from specific accounts related to a company and the other collects tweets from trending topics. Even though they seem similar, they are two distinct API methods and have separate collecting approaches.

Then, the second phase is data preparation, where the prototype cleans and transforms tweets into more suitable text objects for the next phase. In Section 3.5, it is explained in more detail the three different components.

The third phase is modeling. In this phase the models are trained using the data collected and prepared for each model. There are not any transformations, only a few preparations for training the models such as split the data. The modeling phase has three components and is where the models are retrained with new data.

The architecture has two inputs. One of them is the company's social media account name, and the other is the name of the desire location if available. It is described in the Section 3.4 with greater detail in from which locations are possible to extract data. In a scenario of a production environment situation, it is possible to keep both collecting

processes running independently. The prototype was designed in a way to improve performance and to mitigate the impact of API rate limits. It is explained with further detail in Section 4.2.

3.4 Data Collection

In this section, it is discussed in more detail the components of the data collection phase. The data collection was designed to be capable of using the full potential of Twitter API resources.

The key points of this phase are:

- **Collecting tweets from a company social media account:** By using the name of a company social media account (e.g. *@adidas*, *@Nike*, *@pull & bear*, *@shrinershosp* and more);
- **Collecting tweets from trending topics:** By using the desired location (e.g. United Kingdom, Lisbon, New York and others);
- **Filtering information:** By passing through multiple layers of filters to save the desired information (e.g. check language, retweet and other filters).

3.4.1 Pre-Data Collection

Before the data collection phase begins, two decisions need to be made. The first was to get all locations available on Twitter and to save it, this was performed only once and therefore is not demonstrated in Figure 3.3. The prototype user can choose what is more convenient to him if to extract all trend locations available or a specific one.

The second decision is to choose one company that has a Twitter account. It can be any company, but it is explained in subsection 4.3.2.2 with greater detail how to choose and the implications.

3.4.2 Data Collection: Components and Processes

As explained before, this phase has two components. In Figure 3.3, process one is related to extract companies tweets and process two is about to collect trending topic tweets. The green color represents the methods that affect only process one and the blue color is related to the process two methods. There is also a shared area in gray color, which both processes have to go by.

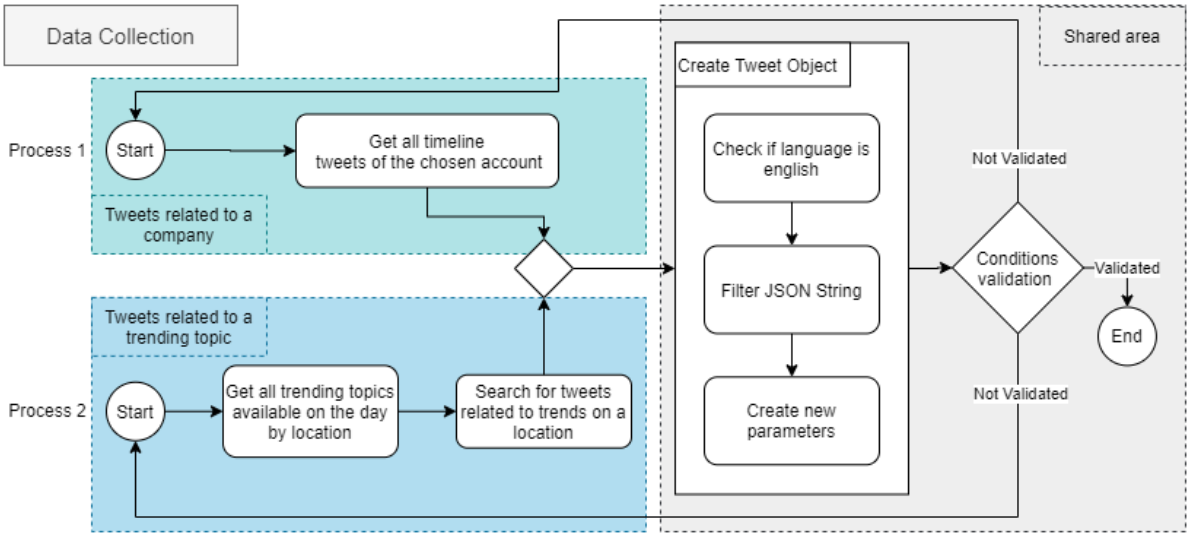


FIGURE 3.3: Scheme of data collection phase. The green color identifies the methods exclusive to process one. The blue color determines the methods limited to process two. Finally, the gray color is the common methods for both processes.

Process one starts by using the chosen company's account name to extract the target timeline. Then a new object is created, using only English tweets (labeled by Twitter), next there is a check for retweeted tweets, and last other parameters are saved like text, the author, if it is a retweet, tweet ID and more, in Subsection 5.2.1 there is an example of a tweet object saved. The JavaScript Object Notation (JSON) object given by API Twitter is discarded. Finally, it is necessary to check if the required conditions are satisfied before going to the next phase, as it is explained in Section 4.1. If affirmative, this phase is over, alternatively, process one or two should be repeated (e.g. with other companies or locations).

Process two task is to extract trending topics' tweets. The list of Twitter locations described in Subsection 3.4.1 is needed in this process. The locations can only be countries and respective cities with English as a native language. They are chosen based on the list of overseas countries as majority native English speakers by the United Kingdom government [54, 55]. For this work, only the country level was used. The trending topics have a query associated to them (e.g. "%23WhenCoronaViru-sIsOver", "ewan" or even more complex "%22Joe+Diffie%22"). This query is used to search for tweets related to that topic. The validation procedure is equal to process one.

Both processes are independent of each other because they use different Twitter API methods. This fact allows them to run at the same time, but, as it is explained in Section 4.2, the number of requests on Twitter API is limited and can run out very quickly if the process is duplicated.

3.5 Data Preparation

The data preparation phase is responsible for cleaning and transforming the data collected. There are three components, General Text Processing (GTP) in Figure 3.4, Specialized Text Processing (STP) in Figure 3.5 and Centroid Strategy (CS) in Figure 3.7. In Section 5 all those components arrested to find which is more suitable for HotRivers.

The important points of this phase are:

- **Data preparation processing:** Automated cleaning and transforming text processes;
- **Analyse data preparation:** Run analysis on data and minimum operating requirements verification;
- **Data preparation optional processing:** Apply more aggressive techniques to cleaning the data.

3.5.1 Data Preparation: General Text Processing

GTP component is straight forward as it is observed in Figure 3.4. There are three options for processing text: lemmatization, steaming and without treatment. All collected data need to go through this process.

The first task of GTP is to delete all types of emojis and hashtags. Additionally, the tweets' noise such as usernames, URLs, paragraphs, RT initials and others Unicodes are also removed. After this point, no more Unicodes should exist. If were detected an Unicode after this step, the noise treatment should be updated. The next task is numbers removal, then lowercase and punctuation removal. Before finalizing the process, stop words are deleted, tokenization and the process is over. Depending on the type of experiment the options are between no treatment, lemmatization, and stemming [56, 57].

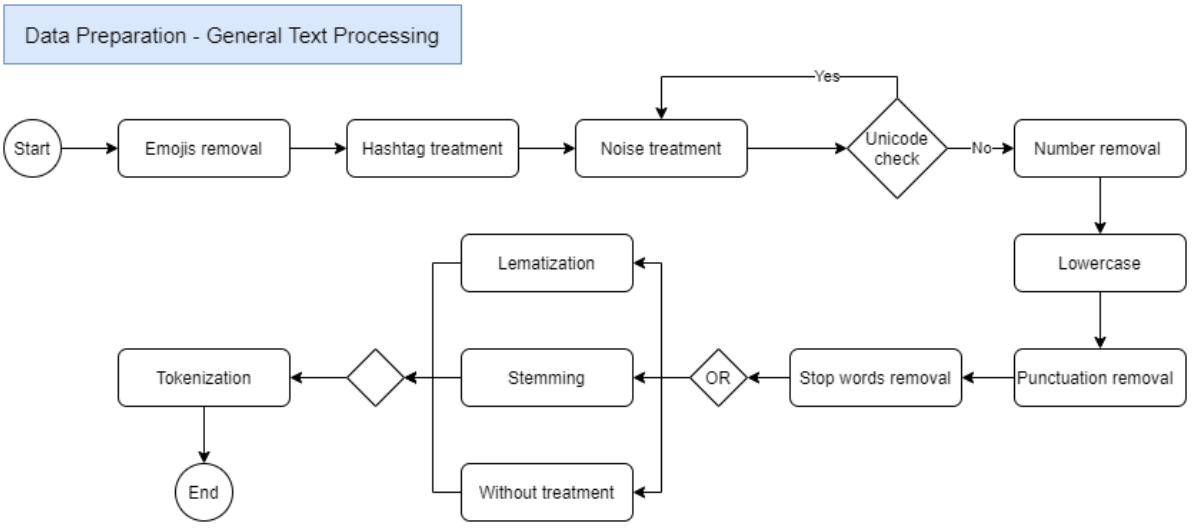


FIGURE 3.4: Scheme of data preparation GTP component. The methods such as Unicode check was created during implementation and validation of this phase. There are three components that were tested to identify which is more suitable for the task

3.5.2 Data Preparation: Specialized Text Processing

STP was created because of the necessity of evaluating the state of the data and to apply more cleaning techniques. This process was made to balance the problems of small and duplicated documents discussed in Subsection 4.4.2. This component was created to help investigate the state of the data and to improve the performance of the models.

This process is simpler than the previous process, as it can be observed in Figure 3.5. However, it is not automated. For STP to work properly, it needs a few statistical methods to evaluate the quality of the data. These parameters are the count of the length of each document, count of duplicated documents, count of the repetition of each word and total of documents. It is always recommended to observe the state of the data after each task.

First step is to apply statistical methods on data, then to choose one of the following options to start small words removal, small documents removal, or duplication documents removal. Then, to apply statistical methods on the data again and evaluate the necessity of any other treatment. When the desired results are achieved, this process is over.

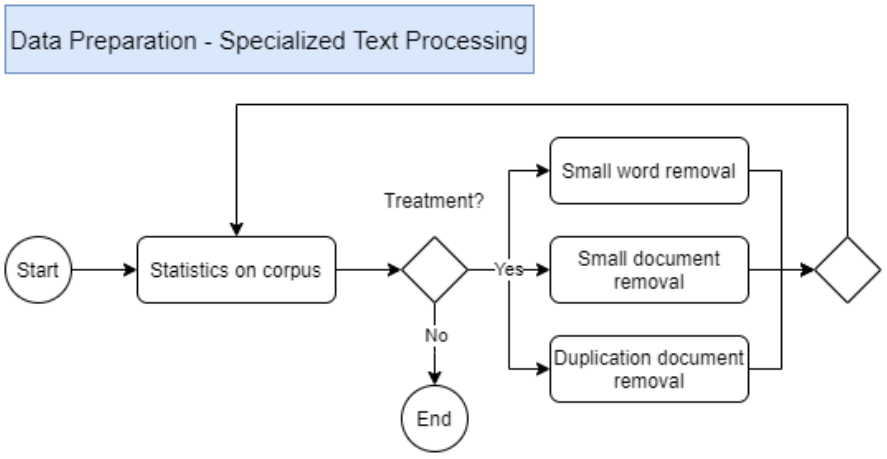


FIGURE 3.5: Scheme of data preparation STP component. This component is iterative and not all methods need to be used. Those methods can be quite aggressive and eliminate data that was not supposed to be deleted

3.5.3 Centroid Strategy

The trending topics are a set of tweets grouped by Twitter, in other words, trending topics could be seen as clusters of tweets. A centroid is the central point of a cluster, i.e. the tweet that best represents the trending topic [58].

CS came to solve the problem of having to extract large quantities of tweets, leading to significant periods of waiting time. This allows lowering the number of tweets collected and to reduce the long waiting time of extraction. Also, the CS is useful to filter tweets as shown in Figure 3.6, which do not say explicitly anything about the topic.

FIGURE 3.6: A tweet selected from the trending topic #SackWhitty

The CS component first needs to calculate all cosine similarity distances of all tweets versus all tweets. Then, to simplify the process, the values are saved on a matrix and the diagonal values are disregarded, which are the value of one tweet versus itself. The next step is to average the values by row or column, the output is the same because it was used always the same tweets. Finally, sort the values by descending order and return the top N similar tweets.

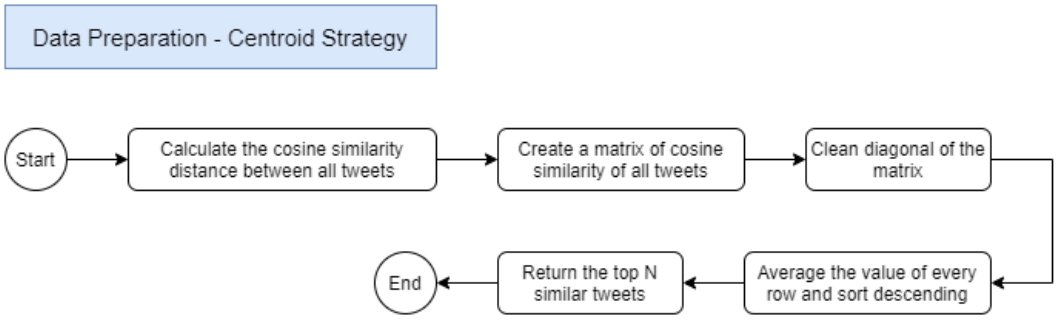


FIGURE 3.7: Scheme of data preparation CS component. This method is specially useful to select only the most similar tweets with positive cosine similarity distance value

3.6 Modeling

Modeling phase is to train and to evaluate models. In this phase there are three components equivalent to the numbers of approaches tried. The fundamental points of this phase are:

- **Embedding Vectors Modeling:** Train Doc2Vec model with companies topics and infer trending topics' tweets;
- **Probabilistic Model:** Train a latent Dirichlet allocation (LDA) model with companies' topics and infer trending topics' tweets;
- **Classification Approach Modeling:** Train a Convolutional Neuronal Network (CNN) and perform a binary classification;
- **Models Output:** The output is a list for the company of each trending topics that might be interesting for them.

The idea of document embedding vectors approach [44] is to use the model Doc2Vec [46] from Gensim [59], in other words, transform documents into embedding vectors. For the probabilistic model approach [40] is to use the LDA [60] model from Gensim [59]. Both algorithms have an identical scheme, which is to train the algorithm and infer the trending topics as unseen data and calculate the similarity value. The Doc2Vec uses the cosine similarity distance and the LDA uses the Hellinger distance. While the cosine similarity distance measures the angle of two vectors, the Hellinger distance quantifies the similarity between probability distributions.

The classification task approach uses a CNN model [34, 30]. The architecture used was proposed by Yoon Kim [33]. The CNN is made by an embedding layer and three convolutional layers. Then, it is applied max pooling on each convolution layer output and concatenate the layer. Finally, a dropout layer and softmax output. CNN is trained with the companies' tweets. This is a problem of binary classification. One of the labels is the company and the other label is called Others, it is composed by tweets from different companies. The output of each trending topic is a value between 0 and 1.

Chapter 4

HotRivers Implementation

In this chapter is discussed HotRivers prototype requirements, implementation, problems found and solutions developed.

4.1 HotRivers Minimum Operating Requirements to Work

This architecture has some requirements to work properly. These requirements were constructed during implementation and after testing the phases. It is worth stressing that they are created to insure quality in results. Non-compliance with these conditions could lead to poor quality and not trustworthy results. As an example, if the classifier is not properly trained, it could try to guess the data given randomly, which is not supposed to happen. A few requirements are for the data collection phase and others for the modeling phase. In Table 4.1 are summarized the minimum operating requirements to HotRivers work.

The first condition is to have a developer account to have access to Twitter API. Twitter makes it mandatory to have an account, either free with fewer resources or paid.

In Section 4.2 it is explained in more detail the impacts of choosing each type of developer accounts.

The second requirement is the minimal number of tweets for the modeling phase, which is 1,000 tweets. This value was set to guarantee that after cleaning and transforming the data it is possible to have enough tweets to properly train the models.

This leads to the third requisite, which is to choose only accounts that have more than 1,500 tweets published. It is recommended to choose an account with near 3,200 tweets posted, because during the data preparation phase, the number of tweets can be cut to half. Twitter allows the extraction of a maximum of 3,200 tweets and always the most recent. Depending on how much a company interacts on Twitter the range of tweets collected vary.

The fourth requirement was found during extraction. The accounts cannot be private, otherwise, it is impossible to collect the data. Unless the developer account follows the private account.

The fifth requirement is to have only words, abbreviations e.g. *"I love u"* (I love you) and different word spelling e.g. *"It's a raccooooooooooon!!!"* (It's a raccoon) after cleaning and transforming the data. In other words, usernames, emojis, hashtags, hyperlinks, or automatic text produced by Twitter needs to be deleted. A point worth of stress is that after filtering the data, empty tweets might appear because there are tweets only constituted by usernames, hyperlinks, hashtags, or emojis. This reinforces the need for the second requirement.

The sixth requirement is to have between two and five repetitions of each repeated tweet. Depending on the information repeated, it might make sense having more duplicates or less. The number of repetitions is low to avoid having the CS select the same sentences because they are all equal. This requirement is mandatory when the CS is used. This scenario is explained with greater detail in Subsection 4.4.4.

The seventh requirement is to extract only trend topic tweets from official English speaking countries. The countries are listed on the website of the University of Northampton and The University of Sheffield [55, 54].

The eighth requirement is to collect at least 100 tweets of each topic because Twitter already categorized those tweets as topics as it was explained in Section 2.4.

The ninth requirement is to limit the maximum time of extraction of trending topics to 4 hours. Since time is a crucial feature for this work, the extracting phase should not take longer than half a working day. A few points are worth of stress, the first is that Twitter has rate-limited on endpoint, and it can affect the performance when collecting the data, as it is described in the next Section 4.2. So, depending on the number of topics to be extracted, the number of tweets for extraction per topic may need to be adjusted.

	HotRivers Requirements
Collecting Phase	Own a Twitter developer account
	Minimum of 1,500 posted tweets in an account[1]
	Target account cannot be private
	Minimum of 100 tweets per topic
	Only hot topics' tweet from official English speakers countries
	Maximum extraction time 4 hours
Modeling Phase	Only words, abbreviations and different word spelling
	Number of tweets repeated between 2 to 5[2]
	Minimal of 1,000 tweets after cleaning and transformation

[1] It is recommend accounts with at least 3,200 tweets posted.

[2] Only when the CS is used.

TABLE 4.1: HotRivers minimum operating requirements to work

4.2 Developer Account and Twitter Rate Limited

Twitter API have two types of accounts, standard API developer account, which is free, and premium API developer account, which is paid. The type of developer account has a minor impact on this work. The most significant difference is in the search method,

because paid accounts have more data available. Since the scope of this work is to analyze the trending topics of the day, the free account is enough to extract the newest tweets. In Table 4.2, it is possible to observe in which requisites the type of account has influence. If the emphasis were on older topics, a premium developer account would be required.

Requisites	Twitter Standard API developer account	Twitter Premium API developer account
Own Twitter Developer	Free	Paid
Minimum of 100 tweets of a topic	only last 7-days query based search	30-days to full-archive query based search[1]

[1] It is not necessary for this work

TABLE 4.2: Affected requirements by developer accounts [61]

The type of developer account has no impact on extracting efficiency, because the rate limits are the same for both types of accounts. The default parameters allow the extraction of 10,800 tweets per hour, whereas maximized parameters allow 72,000, as it can be observed in Table 4.3. Both free and paid accounts allow maximized parameters.

These values were calculated based on Twitter API rate limits. The maximum number of calls possible to make to the API with the account used in 15 minutes window is 180. If the methods are maximized, in 1 hour is viable to extract four periods of a 15 minutes window. This makes a rate of tweets per hour of 72,000 and a total of 288,000 tweets by respecting the requirements. Both accounts allow maximized parameters.

A point worth of stress is in the next section those numbers are significant because only tweets are used. As an example, in 5,000 tweets of a topic, only 200 are tweets and the other are retweets. So, 4,800 tweets are discarded.

	Number of tweets returned per page	Number of tweets per hour	Number of tweets per 4 hour
Twitter default parameters	15	$180 \times 15 \times 4 = 10,800$	43,200
Maximized parameters	100	$180 \times 100 \times 4 = 72,000$	288,000

TABLE 4.3: Table of default parameters versus maximized parameters [61]

4.3 Data Collection Implementation

The data collection phase was a critical point because, without data, it is not possible to achieve the goal of this work. While the next phases were more dependent on the architecture decisions, this phase was limited by Twitter rules. In this implementation was used a free account and authentication OAuth 1.0a method [62]. It allowed a maximum of 180 requests per 15 minutes [61].

The object tweet given by Twitter API is an enormous and complex JSON object. In order to help to filter it, it was chosen a community library as it is explained in the next Subsection 4.3.1. It was important to choose in the moment of the extraction which parameters were required. If a missing parameter is necessary for the following phases, the data collection phase would need to be repeated.

The following subsection answers the requirements identified in Table 4.1 and other problems encountered during the development of the data collection phase.

4.3.1 Native API vs Community Libraries

Twitter offers a Rest-Full API entry point to anyone who has access to a developer account to create an app using Twitter data. The answer to the Twitter Rest-Full API is frequently a big and complex JSON object. One approach to deal with this problem is using your own methods to filter the information needed or to choose one of the many community libraries and use their already built functions to filter the information desired. The process for choosing one of the community libraries started by looking for community libraries recognized by Twitter [63], then looking for each of them on GitHub [64] and compare which had the most significant star rate and the higher number of users. The chosen library was the Tweepy [65] with approximately 6,700 ratings and roughly 14,800 users on GitHub. A few test trials were conducted to ensure that the information collected was correct. Random tweets were chosen and checked to verify, for example if the date,

number of likes, retweets and text were the same as in corresponding Twitter timeline account.

4.3.2 Pre-Data Collection Implementation

4.3.2.1 Which Company to Select?

As explained in section 3.4, choosing an organization is the first thing to do. For this work, it was decided that only companies were allowed. Although, sports leagues, charity associations, non-profitable organizations and more were a possibility. It could be any company that met the conditions detailed in section 4.1. It was used for more prominent companies instead of small business to be easier.

The focus of this work was on micro and small companies. Also, in Chapter 1 was mentioned that those firms have a lack of specialized marketing resources, this might represent a problem because of non-existent accounts or accounts with not enough tweets.

One solution to this issue is using a similar business market or a direct competitor. An example, the company *Raccoshoes* sells shoes and have a recent social media account. *Raccoshoes* has two options, either they can use *Adidas, Nike, Puma, Reebok* or other big companies in the shoe market, or on the other hand, they can use *Coyboots*, their direct competitor, which has a two years old active account and is located in the same area. The type of analysis to evaluate which option is better is out of the scope of this work.

During extraction, private accounts were found impossible to extract. The only option to solve this issue is by following them, but the owner needs to accept it. It might take time and reveals not viable.

4.3.2.2 Low number of tweets

Two possible scenarios that can result in a low number of tweets are: after collecting, the number of tweets extracted, they are not enough to fulfill the requirements, or after cleaning the same happens. So, one solution is to collect more tweets, if possible, to increase the chances of success, or to use more than one account. Nevertheless, these options might not be possible, so other methods were developed for data collection phase.

It was created a method to search for similar account names on the friend list of the target account. The method uses a regex expression to look for similar variants of the given name (e.g. with adidas was used before_adidas, adidas_after, and before_adidas_after). The reason for only looking on the friend-list is to ensure that those accounts are recognized by the owner of the account (e.g. @adidastennis is the account acknowledged by the company whereas @adidasseller, @makerplaceadidas are accounts made by other people).

This method is useful to increase the number of tweets. Although is out of the scope of this work to recommend which accounts shoud be used. This was one of many features created during HotRivers implementation to give more options to the prototype. As an example, Adidas has the following accounts *@adidasUK*, *@adidasUS* and *@adidasfootball*. *@adidasfootball* can be excluded because it is specific to football and Adidas is more than that. After a quick inspection through their feed, it can be observed that *@adidasUS* replies more tweets from the main account *@adidas* then it is pair *@adidasUk*. So, using *@adidasUS* might be a good option. However, this type of analysis is not the goal of this work.

Another option to increase the number of tweets is not using some treatments, for example not deleting so many repeated tweets, small documents or keeping small words. This solves one problem but could worsen the performance of the training.

4.3.3 Data Collection: Components and Processes Implementation

4.3.3.1 Faster Extraction

In the beginning, the collection phase was built using page iteration per item. It became a problem when the excessive timing for extracting was noticed. So, it was changed to page iteration per results. That makes it possible to achieve the performance of 72,000 tweets per hour, as presented in Table 4.3.

The equilibrium between extract efficiency and having enough tweets of a topic to this work is around 250. It takes about one hour to complete the process of extracting six locations with ten topics each. Collecting more than 500 might take too long on a few topics. As an example, in Table 4.4 there is a comparison between saving 200 or 500 tweets of the trending topic *#VMAs*. The reasons for the number of tweets extracted being so despair is explained in Subsection 4.3.3.3.

Number of tweets saved	Number of tweets extracted	Time necessary
200	3,700	a few seconds
500	67,100	$\approx$ 45 minutes

TABLE 4.4: Efficiency comparison between extracting 200 or 500 tweets of the trending topic *#VMAs* on Ireland on 1st of August

4.3.3.2 Data Quality and Quantity

The quality of the tweets is essential to this work. Twitter has a parameter that regulates the return type for search queries method the options are mixed, recent, and popular. It was tried with popular option only, but the results were scarce in numbers of tweets. Therefore, the mixed option, which includes both popular and most recent results in the response, was chosen. The greatest problem identified was tweets such as in Figure 3.6 that do not say anything about the topic and difficult the task of finding associations. This reinforces the tactic of extract a higher quantity of tweets and the use of the CS. Relating

companies' tweets depends on companies' communication strategy, in other words, it can be short tweets with less text, smiles, mentions, and hashtags or bigger tweets. For this work, it is preferable to have bigger tweets (at least in terms of text), because smiles, mentions and hashtags are deleted and more text means more information to be analyzed.

The quantity has a vital role in this work. Quantity is not a problem to trending topics, because a trending topic is a topic that has a considerable amount of tweets [35, 48]. Although, for companies, the situation is slightly different, due to Twitter API limitations of 3,200 tweets. One option is to extract from similar accounts as it is explained in Subsection 4.3.2.2 and the second option is to pick a different company. It is important to stress that combines low quality and low quantity might result in not fulfilling the minimum operating requirements, due to processing techniques deleting to many tweets causing models to not learn the data properly.

4.3.3.3 Retweets and Tweets

During the collection phase it was pondered if retweets should be part of training data. For this work, retweets were not used. When someone makes a tweet using an hashtag related to a subject is almost guaranteed that the user is talking about that subject. Even though that's not always true as showed in Figure 3.6. However, retweets are answers to tweets and are more challenging to guarantee that the discussion is still about the subject related to the hashtag. Also, in this project, it is assumed that Twitter algorithm has an high assurance gathering tweets about the same topic. Retweets are detected in the filter JSON string component as presented in Figure 3.3. If the key *retweeted_status* is found on the JSON object, then it is a retweet.

A few points about extracting only tweets are longer collection time, higher chances of quantity, and quality problems. The high figure of retweets make the collection time longer, in Table 4.4 is express this problem, and the difference of extract 200 or 500 tweets is 1,800%. In order to solve this problem, it was set as a requirement a minimum of extract only 100 tweets per topic and to ensure the quality of the tweet was used the

CS technique. Since, it allowed to select the similar tweets. Additionally, this solution allowed to give preference to more topics than more tweets.

4.4 Data Preparation Implementation

Data processing has a key role because it helps improving model learning by eliminating noise and unwanted words. In the following sections it is explained how GTP, STP and CS were implemented.

4.4.1 Data Preparation: General Text Processing Implementation

A tweet may contain emojis, hashtags, usernames, other symbols and noise, numbers, and punctuation. On JSON object tweet emojis are represented with a Unicode (e.g. *u2139*). A regex expression was created to be able to identify all emojis, every time a new emojis is not identify the expression needs an upgrade. For the following scenarios, more regex expressions were created. What makes a hashtag is the use of the symbol # and some text after (e.g. #Covid-19). A username always starts with a @ behind the account name (e.g. @adidas). Retweets have in the beginning the letters *RT*. Another noise needed to be detected was in paragraphs represented as $\backslash n$. It is important to only clean numbers after all Unicodes were deleted, otherwise, Unicodes are not properly handle. For lowercase it was used a native Python method. The last point on general processing is tokenization, there are many python packages for this task. For this work, it was used mainly *nltk*.

4.4.2 Data Preparation: Specialized Text Processing Implementation

One of the most useful components in STP is the statistical methods. One method counts the number of times each document is in the corpus. Another method uses the package *collections* to count the number of words and the length of the documents in the corpus. All information is printed and it is possible to analyze the length and frequency of the documents, also the length and frequency of the words and the variety of words.

The rest of the methods are for small words removal, small document removal and duplicated document removal. The implementations are relatively similar and simple. Small words task deletes all words smaller than a given length. The method small document removal removes all documents smaller than a given number of terms and duplicated document removal excludes smaller documents than a given size.

4.4.3 Centroid Strategy Implementation

The first step of the CS is to create a vector of embeddings for each document. This is made to transform each word into a word embedding and then average all embeddings into one vector. This also could be made by the sum or concatenate the word embeddings. The next step is to calculate the cosine distance of all documents versus all documents. The output is a matrix of distances. Thus, the diagonal values of the matrix are disregarded. Average is applied to catch column to find which column have higher values. For this work GloVe [32] Twitter embeddings were used, but Fasttext [66] were also a possibility to use.

This method is sensitive to duplicate documents. Because if the centroid strategy finds high similarity with duplicat documents, it puts them at the top. This might not tell the reality of the data. As an example, if a corpus has 200 documents and 50 documents have the following sentence "We love your photo, can we post it on our website?".

Those, 50 documents have the highest similarity between them, but probably those tweets do not have a higher similarity with the other 150 documents. It means that the overall similarity might be bad, if repetitions and extremes documents are removed. While extreme documents are removed with the centroid strategy, repetitions must clean before this method.

4.4.4 Small and Duplicated Documents

Tweets are small documents with just a few words, which it represents a problem, because the variety of words may be low when more information is expressed in fewer words and models usually tend to be better when there is more information available. One solution is to collect more tweets to try having more variety of words and tweets. However, due to Twitter API limitations, the maximum for companies is 3,200, which may not solve this problem.

Repeated tweets can worse the training of models by over-fitting to those cases. In Table 4.5 it can be observed all documents that have more than 20 repetitions on Adidas' tweets. The total of Adidas documents before the cleaning process was 3,082. Those tweets seem to be interactions between Adidas and other users, they are very similar between them and share almost the same words. Also, it is possible to observe a large number of tweets with only one or two words and with no words at all. That reinforces the need to extract a large number of tweets to ensure that deleting them do not lead to not fulfill the requirements.

4.5 Modeling Implementation

The following subsections describe the process and the configuration of the models. However, during the test phase were made slight improvements on initial configurations to

Document	Number of repetitions
[ ]	86
['love', 'share', 'photo', 'adidas', 'community', 'reply', 'agree']	73
['create']	55
['love', 'share', 'rest', 'reply', 'agree']	31
['nice', 'kick', 'love', 'share', 'strip', 'community', 'reply', 'agree']	26
['ready', 'create']	22
['creator']	20
Total	313

TABLE 4.5: Duplicated document higher then 20 times on Adidas' tweets

get better results. The configurations used for Chapter 5 are the same described in this section. The data for training the models is always companies' tweets.

4.5.1 Modeling: Document Embedding Vector Implementation

This model is the simplest since is only needed to train the model with the companies' data. The model was trained with a vector size of 200, minimum word frequency was 5, the number of epochs was 250, the algorithm was distributed bag-of-words with train word-vectors simultaneous with DBOW doc-vector training, an alpha of 0.025 and the default minimum alpha. In the sanity test of the model, the vectors were inferred with a configuration of alpha 0.025, default minimum alpha, and 275 epochs. The trending topics vectors were inferred with a configuration of alpha 0.025, default minimum alpha and 500 epochs.

4.5.2 Modeling: Latent Dirichlet Allocation Model Implementation

To implement LDA it is need first to convert the corpus into a dictionary, then transform the corpus into a bag-of-word representation and train the model. Before training, words that appear in less than five documents and more than 75% of the documents were filtered.

All the methods used were from the package *gensim* [59]. The LDA model was trained with 100 topics, a chunk size of 2000, the number of passes was 20, the value of iterations was 400, the alpha and eta values were set as auto and the default minimum probability. The sanity test of the model used the same data, but no additional configuration was needed. It was also applied to trending topics.

4.5.3 Modeling: Convolutional Neuronal Network Model Implementation

For this CNN the embedding layer used was the GloVe embeddings [32]. First, the embeddings were loaded into memory. Then, the data was split into 75% for training and 25% for testing. The second step is to update the vocabulary with companies' tweets. The maximum number of each word based on frequency was set to 15,000. Then, was performed label encoding to train and test target data. The maximum length set was 27. After that is padding tweets with zeros to ensure that all tweets have the same length. Then the next step, adding the existent embeddings, build an embedding matrix and add with zeros the non-existent embeddings, because some of the vocabularies may not exist in the previously loaded embedding. Last, create the model with every layer explained in Subsection 3.6. The embedding layer is set with 200 dimensions. The model was set with the number of filters 100, document maximum length 27, number of classes 2, batch size of 64, number of epochs of 350, and a dropout of 0,30. The activation function was softmax, the loss was binary cross-entropy and the optimizer was adam.

Chapter 5

Experiments and Results

In this chapter, HotRivers was tested phase by phase. The companies used for the test were chosen based on the number of tweets published, the number of followers and area of business (e.g. health care, sport and fashion, technology and more). Additionally, it was taken into consideration to include both bigger and smaller social media accounts because of the HotRivers company target. Pairs of companies were selected to analyze the similarity of the results, i.e, companies that work in the same area of business such as two hospitals, for example. At the end of this chapter it was decided which treatment from the data preparation phase to use and which model is more suitable to integrate HotRivers.

5.1 Data

The data used for the test of the prototype were only tweets from companies and trending topics. The countries collected were Australia, Canada, Ireland, New Zealand, the United Kingdom and the United States. Only the top ten trending topics were collected for each country. In total, 28 days between 31st of August and 9st of November were collected, in a total of 1,680 trending topics and approximately 450,000 tweets.

Due to the high volume of data gathered during the development of this work, only a sample was selected. From the six countries it was chosen to work only the United Kingdom (UK) and the United States (US). Eight days out of the twenty-eight days were randomly chosen. For UK days one, seventeen, twenty-two and twenty-four of September and for the US days two, three, four and eight of September. In total the sample comprises 160 trending topics and approximately 40,000 tweets.

The companies used were Adidas, Nike, Royal Manchester Children's Hospital (RMCH), and Portsmouth Hospitals University (PHU). Table 5.1 shows the number of tweets extracted from those companies' social media accounts. Those accounts were selected, because of the distinct areas of business and because the high number of followers on their social media account. Adidas and Nike are both in sport, fashion, tennis, clothes areas and both seem to be competitors in the same business market. RMCH and PHU are both hospitals, representing the medical area. RMCH and PHU have smaller accounts comparing with Adidas and Nike.

Name of the company	Number of tweets collected	Number of tweets published	Number of followers
Adidas	3,104	13,800	3,800,000
Nike	2,889	36,800	8,200,000
Royal Manchester Children's Hospital	1,512	6,352	6,300
Portsmouth Hospitals University	2,261	17,000	8,790

TABLE 5.1: Table with companies and total of tweets extracted

The training data were composed by the target company (e.g. Adidas) and by the following companies as the opposite label RMCH, PHU, Pull & Bear, Springfield, Sheffield Hospital, Tesco, Sainsbury's, CVS Pharmacy, Pharmacy Times. While experiments one and two did not have the companies Adidas, Nike and Puma as the opposite label, experiment three and four did not have RMCH, PHU, Sheffield Hospital, CVS Pharmacy, and Pharmacy Times as the opposite label. This was decided to help the model learning better and to decrease classification errors.

5.2 Experiment 1: Adidas

The first experiment is with Adidas. The first phase to be tested is Data collection, followed by Data Preparation and last Modeling. A random example was used to present how data was collected and processed. The same example was used to validate each phase. The same rule was applied for trending topics.

5.2.1 Experiment 1: Data Collection

In Twitter documentation there are many examples of the JSON object given by the Twitter API. As described before, that JSON object is filtered and a new object is created. Table 5.2 is an example of a random Twitter collected and the attributes filtered. Attributes such as "Reply to", "Reply", "Date", "Retweets Count", and "favorite Count" help navigate on the Adidas account in order to confirm that tweets were correctly collected and filtered. The most important attribute is the text.

Attributes	Tweet n°42
ID	1260590034215993351
User	adidas
Source	Twitter for iPhone
Date	2020-05-13 15:17:39
Text [1]	This week on The Huddle, @KAKA and @joaofelix70 discuss all things, their careers and life goals. Watch the full conversation on YouTube now: https://t.co/hHPEZwcGio #hometeam https://t.co/vHgkUur5uZ
Location	-
Retweet	False
Reply	True
Reply to	KAKA
Company replay	False
Original ID	1258132300719624192
Retweet Count	14
Favorite Count	142

[1] In the original tweet there was an emoji, unfortunately it was not possible to display in this table as text.

Table 5.2: Table with an example of extracted and filtered tweet

The trending topics have an identical structure to companies' tweets. So, attributes in Table 5.2 are equivalent to companies' tweets and trending topics' tweets. As explained before, only the methods for collecting tweets are different. The extraction of six locations, ten topics per location and 250 tweets per topic take approximately one hour.

5.2.2 Experiment 1: Data Preparation

After tweets are collected, they are cleaned. The techniques tested were the following no treatment, lemmatization, stemming and STP. The CS is tested in a different section. After cleaning stopwords, hashtags, URLs, user mentions, and using one of the tree techniques tweets are expected to look as the one present in Table A.1.

Relating to STP, depending on the number of words per tweet, the length of the words, the quantity of duplicate and the total of tweets, it is decided how aggressive cleaning should be. At the begging of STP, Adidas had 3,082 tweets in total. The biggest tweet had 29 words, and the top five word frequency was "create" with 267 occurrences, "us" with 253, "love" with 244, "share" with 162 and "agree" with 156. First it is necessary to remove small words, i.e. words smaller than three characters. The number of tweets remains the same. Now, the biggest tweet had 27 words, and the top five word frequency was "create", "love", "share", "agree" and "adidas". Then, all documents smaller than three words were removed. Adidas now has 1,131 tweets and the biggest tweet remain the same. The top five-word frequency was "love" with 193 occurrences, "share" with 162, "reply" with 149, "agree" with 149 and "adidas" with 140. Finally, after removing tweets with more than three duplicates, Adidas had a total of 988 tweets. The top five word frequency was "create" with 93 occurrences, "creativity" with 74, "like" with 73, "time" with 72 and "adidas" with 70.

This process, as explained previously can be aggressive, and depending on the company should be adjusted. At the begging of this process, the number of documents containing only two words was 796, three words were 598 and four words were 455. However,

in the end, the number of documents containing four words was 363, five words were 214 and six words were 107. In Table 5.3 it is possible to observe the statistics of the final STP configuration, document length bigger than three words, two repetitions per-document, and word length bigger than two.

	Results
Number of tweets	1,332
Top 5 words frequency	love: 195, share: 162, wed: 154, reply: 149, agree: 149
Biggest tweet	28
Document length frequency	4 words: 428, 5 words: 238, 6 words: 170, 8 words: 136, 7 words: 120,

TABLE 5.3: Table of Adidas dataset statistic after applied the STP configuration document length bigger than three, document repetition lower than two and word length bigger than two

5.2.3 Experiment 1: Modeling Embedding Vectors

In this subsection Embedding Vectors are the first approach to be tested. All techniques were tested, starting with no treatment, then Lemmatization and Stemming. After is performed the STP and last the CS. Finally, experiment on a trending topic and discussing the results.

5.2.3.1 Metrics for Embedding Vectors

The metrics used are average and standard deviation or average and median. Additionally, the disparity of the mean or median values being wide is desirable. Otherwise, it might indicate that the model is considering every tweet as similar to any tweet. The main range for acceptance of a trending topic is the highest average figure less one standard deviation, otherwise is the highest median figure less one standard deviation. The average is used when average and median have similar values. When average and median are different, the median is used.

The average and median tell how close to minus one, zero, or one is the corpus. Positive values mean that documents are similar to each other. Negatives values say documents are similar to the contrariwise. When the value is near zero it means that documents are not similar.

The other essential factor in evaluating how the algorithm is learning is the Ranks. The Rank is the similarity position of each tweet to itself. As an example, the algorithm calculates the similarity of tweet A versus all tweets. Then, the algorithm sorts it by similarity and find the Rank of tweet A. If tweet A is in position zero, then means that tweet A is the most similar to itself. However, if tweet A is in position one, then it says that tweet A is more similar to another tweet than itself, which is not supposed to occur. So, a higher frequency of rank 0 means that the model is learning correctly. In other words, the higher frequency on top Ranks, the better the model is learning.

5.2.3.2 Embedding Vectors, GTP and STP

The first test done is the sanity test. Sanity test is to check if the algorithm is learning how it is supposed. It utilizes the same data that was used to train the algorithm. That data is given to the algorithm as new unseen data. This is how the Ranks are calculated. It is not expected to have always the exact same similarity value, due to bulk-training and randomness of the train. So, re-training the algorithm gives slightly different values.

The results were not great. The sanity test of Adidas' tweets is observed in Table 5.4. The average and median values are close to each other, which means that the data is distributed around the average. Also, without STP the Rank is not better than with STP. With no treatment, only 60.8% of the tweets are similar to themselves, and for lemmatization and steaming are approximately 62.5%. This means that roughly 37% of the tweets are more similar to other tweets than to themselves.

The STP configuration was the same as presented in Subsection 5.2.2. The ranks are better with STP because there are more tweets with Rank zero. By using no treatment

92.1% of the tweets are similar to themselves, for lemmatization is 93.7% and for steaming is around 93.2%. That indicates that about 8% of the tweets are more similar to other tweets.

While STP improves data model understanding, GTP is not clear. However, of all techniques, lemmatization improves slightly the model.

	Techniques	Average	Median	Rank	Number of Tweets
No STP	No treatment	0.357	0.360	[0: 1875, 1: 348, 2: 181, 3: 108, 4: 75, ...]	3,081
	Lemmatization	0.358	0.359	[0: 1928, 1: 355, 2: 168, 3: 104, 4: 77, ...]	
	Steamming	0.359	0.361	[0: 1923, 1: 367, 2: 180, 3: 105, 4: 64, ...]	
With STP	No treatment	0.356	0.354	[0: 1082, 1: 69, 2: 14, 3: 2, 831: 1, ...]	1,175
	Lemmatization	0.358	0.349	[0: 1078, 1: 59, 2: 6, 3: 4, 4: 1, ...]	1,151
	Steamming	0.357	0.349	[0: 1082, 1: 61, 2: 13, 3: 1, 4: 2, ...]	1,161

TABLE 5.4: Table of the results of the Adidas sanity test with the GTP and STP techniques

5.2.3.3 Embedding Vectors and Centroid Strategy

In this experiment, the CS values tried were 50, 100, 250, 500, and 1,000 documents. As explained before, CS chooses the documents with higher similarity, on average. For the algorithm, it is expected that with fewer tweets the average is higher because the probability of having fewer tweets with opposite similarity (negative values) is higher. As an example, tweet number 45 may be the opposite of tweet 473, 580, 983, and a lot more, so the overall average in 1,000 tweets might be lower.

It is worth stressing that it was employed only tweets from Adidas. However, giving to the algorithm only 50 or 100 tweets might be a reduced number of data, at least they are the most similar documents. In Table 5.5 it is observed that the ranks have good values, for 50 documents, 86% of them have higher similarity with themselves, for 100 documents are 97%, for 250 is 94.8%, for 500 is 94%, and last for 1,000 is 93,1%. When given only 50 documents the algorithm did a good job understanding the data by presenting a high figure of similarity. The rest of the tests were poorer, which might indicate higher levels

of opposite similarity or non-similarity. Nevertheless, it is important to test with more companies and with trending topics to investigate the possibility of overfitting.

Tweets	Average Similarity	Median Similarity	Standard Deviation Similarity	Ranks
50	0.977	0.981	0.017	[0: 43, 1: 3, 2: 1, 3: 1, 6:1, 7:1]
100	0.506	0.509	0.201	[0: 97, 1: 3]
250	0.392	0.376	0.171	[0: 237, 1: 11, 3: 1, 2: 1]
500	0.354	0.343	0.132	[0: 470, 1: 28, 2: 2]
1,000	0.356	0.347	0.112	[0: 931, 1: 55, 2: 10, 3: 3, 630: 1]

TABLE 5.5: Table of Adidas sanity test with GTP, STP, CS techniques

5.2.3.4 Embedding Vectors Experiment on Trending Topics

The model used was Doc2vec and data preparation techniques used were GTP with lemmatization, for STP were used three minimum number of terms, two minimum number of repetitions per document and words longer than two characters, and finally for the SC the top 50 documents. The trending topics used were the ones from days two, three, four, and eight of September and for the US days one, twenty, twenty-two, twenty-four of September for the UK. It was used the CS technique to select the top ten most similar tweets of each trending topic. The reason to use the CS technique is to avoid tweets such as shown in Figure 3.6. The discussion of the results is in Subsection 5.3.3

5.2.4 Experiment 1: Modeling LDA

In this subsection the LDA model is tested with Adidas tweets. The techniques used were the same as in previous experiments.

5.2.4.1 Metrics for LDA

The LDA model gives the probability distribution for each topic of each document. So, cosine similarity distance is not the most suitable metric and instead it was used Hellinger

distance. The range of Hellinger distance is between zero and one. When the value is close to zero, it means a smaller distance and, therefore higher similarity. But when close to one, it means the opposite, higher distance and lower similarity. The other important metric that allows the analysis of how the model is learning is by observing the comparison of similarity between the same document to itself. This metric is the diagonal average, similar to the Rank as discussed in Subsection 5.2.3.1. In a matrix of documents versus itself, the values on the diagonal are the result of one document versus itself.

5.2.4.2 LDA and GTP

The results of the experiment using LDA and GTP were poor. The first signal that this model would not work was on the number of unique tokens. Even though the number of documents was a significant figure, this did not turn into a sizable number of unique words. In Table 5.6 it is observed that the number of unique tokens is not higher than 513, which is a low figure. The overall Hellinger distance average indicates low similarity. A low standard deviation and median value close to the average, validate the poor result. The Diagonal average shows a relatively good value, which shows that the model knows the data.

Technique	Number of unique tokens	Number of documents	Average Similarity	Standard Deviation Similarity	Median Similarity	Diagonal Average Similarity
No treatment	513	3,082	0.800	0.092	0.816	0.001
Lemmatization	502		0.814	0.089	0.831	0.001
Steaming	507		0.811	0.087	0.820	0.001

Table 5.6: Table of Adidas sanity test with GTP techniques

It does not make sense to use the CS technique, because it reduces the number of documents, hence decreases the number of unique tokens. Thus, use any technique that cleans more data also reduces the number of unique tokens. The unique words metric is really important to this model because LDA is a generative statistical model, that word frequency is taken into account, but word order is ignored. Which makes so import the number of unique tokens. The only way to fix this issue was by adding more data, but

unfortunately, it was not possible. This makes the use of this model not viable for this work. For a total of approximately 3,000 tweets, this low number of unique tokens was not expected.

Unfortunately, for Adidas' case, this model did not work, however it can not be assumed that it does not work for other cases. Since, HotRivers should be able to accept any kind of company that can fulfill the minimum operating requirements and present good results. This approach was disregarded.

5.2.5 Experiment 1: Modeling Classification Task Approach

In this subsection, the approach tested was the classification task approach with a CNN. Similar to the previous subsection, the first tests were with no treatment, Lemmatization, or Steaming. Then with STP and with the CS. Finally, testing the trending topics and discussing the results.

There are only two labels, Adidas and Others. While label Adidas only comprises Adidas' tweets, the Others aggregates a set of tweets from other companies (RMCH, PHU, Pull & Bear, Springfield, Sheffield Hospital, Tesco, Sainsbury's, CVS Pharmacy, Pharmacy Times). There are many ways of balance both classes. The option chosen were using a percentage of each company, the more tweets a company has more percentage has. It was needed a method to create Others label dataset and this way seems the better. As an example, in total, Adidas had 1,000 tweets, RMCH 1,000, PHU 750 tweets, Springfield 2,000, and Tesco 1,500. The training data is composed by the number of Adidas' tweets and a percentage of the number of tweets of each company. The percentage is calculated by the total of Adidas' tweets divided by the sum of the total of each company tweets. It would be around 19%. So, from RMCH is 190, PHU is 143, Springfield is 380 and Tesco 285. It makes a total of 997 tweets. So, the training dataset is composed by 1,000 Adidas' tweets and 997 others' tweets. It worth to point that if collected a total of 50,000 tweets from multiple companies and Adidas had 1,000 tweets the percentage would be 2%, i.e. in

long term the percentage of tweets of each company should be low enough to the number of each company be homogeneous.

5.2.5.1 Metrics for Classification Task Approach

In the classification task approach, the metrics collected on training and testing were accuracy, recall, precision and F1-score. F1-score and accuracy were the metrics to evaluate the model. Similar to the previous approaches, to find which trending topics are selected, all trending topics were picked in a range of the trending topic with higher similarity less one standard deviation. This approach is a task of supervised classification, however evaluate the trending topics per se is complicated, because without anyone from those companies is not possible to ensure that the trending topics picked are totally correct.

5.2.5.2 Classification Task Approach, GTP and STP

The results with the training dataset were satisfactory. The model was tested with GTP techniques and STP. In Table 5.7, is observed that the average and F1-score present values higher than 90%. The STP results were always worse than without the STP, even though different STP configurations were tested. The best configuration for STP was any document length and word length and seven repetitions per document. It seems that CNN learns better with more data and also duplication did not seem to have a negative impact. Lemmatization had in both techniques, with and without STP, better results than the rest of the techniques.

Techniques		Loss	Accuracy	Recall	Precision	F-1 score	Quantity of adidas tweets
No STP	No treatment	0.421	0.918	0.920	0.920	0.920	3,077
	Lemmatization	0.466	0.920	0.922	0.922	0.922	
	Steaming	0.543	0.910	0.912	0.912	0.912	
With STP	No treatment	0.479	0.901	0.903	0.903	0.903	2,886
	Lemmatization	0.447	0.903	0.903	0.903	0.903	2,874
	Steaming	0.543	0.900	0.901	0.901	0.901	2,886

TABLE 5.7: Table of Adidas CNN model results with GTP, STP techniques

In Table 5.8 is observed the confusion matrix of the CNN with lemmatization and no STP. The model has more difficulty in classifying the label Others than the label Adidas. In 759 Adidas tweets, 71 were misclassified and in 780 Others tweets, 52 were wrongly classified. It seems that the model has learned the data correctly. Analyzing the model loss graph in Figure A.1, there is not overfit on learning curves. Thus, on the model accuracy, precision and recall graph Figure A.2, Figure A.3, and Figure A.4, respectively, there is a gap between lines, which does not represent a problem. Both train and validation are in the same direction, so there is not a clear indication of overfitting.

Confusion Matrix		Predicted Label	
		Adidas	Others
True Label	Adidas	728	52
	Others	71	688

Table 5.8: Confusion matrix of Adidas

5.2.5.3 Classification Task Approach and Centroid Strategy

The CS values tested were the minimal number of tweets (1,000), 50% of the dataset (1,500), and 75% of the dataset (2,250). The cleaning techniques used were only GTP with lemmatization. Even though CNN does not use cosine similarity, it was still able to take some advantage of the most representative tweets and achieve interesting results. In Table 5.9 is observed that using CS with 1,000 tweets it was possible to accomplish close values of accuracy and f1-score with those in Table 5.7. Unfortunately, the results were not improved and the CS was disregarded.

Quantity of adidas tweets	Loss	Accuracy	Recall	Precision	F-1 score
1,000	0.664	0.910	0.909	0.909	0.909
1,500	0.573	0.905	0.905	0.905	0.905
2,250	0.569	0.874	0.877	0.877	0.877

Table 5.9: Table of Adidas model results using the CS with different number of tweets

5.2.5.4 Classification Task Approach Experiment on Trending Topics

The data preparation techniques used were GTP with lemmatization and the model was a CNN, and equally to the previous approach, the CS was used only for trending topics. The analysis and discussion of why those trending topics were selected is in Subsection 5.3.3.

5.3 Adidas Results Discussion

5.3.1 Techniques and Models Discussion on Adidas

The results of using GTP with lemmatization were consistently better than steaming and no treatment in both Doc2Vec and CNN. Relating to STP techniques, on the Doc2Vec model it helped to improve the model learning, but did not make the performance better. On the other hand, the CNN model did not help to increase the model learning and did not boost the performance. The CS truly improve the results on Doc2Vec but did not do the same for the CNN model.

The results, on both models, using trending topics were satisfactory. The CNN model was trained with a few health companies and the results reflect that decision, which indicates that with more data from other companies from different areas the results could be slightly different. On the other side, training the Doc2Vec with only 50 tweets seems too little to be able to generalize, even though the results were interesting, it might be a dangerous assumption to say that the model is good. It is worth testing with a few more companies and see how it behaves.

Until now the results showed that in the long term Doc2Vec has an disadvantage, which is work only with 50 tweets, which mean a lot of tweets would needed to be disregarded. However, the CNN model can be retrained with new tweets over the time, which is an advantage.

5.3.2 Trending Topics Analysis on Document Embedding Vector

The results are shown in Table A.2 for the US and in Table A.3 for the UK. The green color represents the selected trending topic and the yellow color the value of the standard deviation of the highest value of similarity average. The results were interesting with a good disparity of values. In the US, the disparity value range between the higher and lower average was between 0,674 and 0,330. In the UK, the disparity value was between 0,474 and 0,385. It suggests trending topics with similarity and non-similarity. Otherwise, lower disparity could mean that model sees every tweet as equal. The average and median had close values. Regarding the number of selected topics, the maximum was three topics per day, which did not seem to be excessive. Fifteen topics were picked by the model in a total of eighty topics. The lowest value of similarity was 0.276 for the trending topic *DYNAMITE CELEBRATION* and the highest was 0.970 for the trending topic *#NHSCOVID19app*. The relations between trending topics and Adidas on the Doc2vec model are harder to justify, which is a negative point of using this model.

Relating to the US trending topics in Table A.2 is observed that *steven adams* and *Lakers* are related to basketball, but both Nike sponsored. The trending topic *Hyrule Warriors* is a video game, the *#TheMandalorian* is a movie, *Sarah Sanders* is an American politician and the *#appleevent* is a technological event from the company Apple. The trending topic *Big Sean* is the only one directly related to Adidas. Big Sean made one pair of snickers in partnership with Adidas.

Concerning the UK trending topics, in Table A.3, trending topic *Nike* had a higher value, which does not surprise because both companies fight for sponsor players, teams and more. While the trending topic *Starmer* is a UK politician, *#GBBO* is related to a cooking show and the *#ps5preorder* is about a gamming console, the connection with Adidas is harder to understand. The trending topic *Thiago* has an easier connection because Thiago is a football player, which played on Bayern Munique, a team sponsor by Adidas.

5.3.3 Trending Topics Analysis on Classification Task

The US results of the trending topics of days two, three, four, and eight of September are observed in Table A.4. The green color means the selected trending topics and the yellow color is the standard deviation of the highest value of similarity of the day. Nine topics on the US were selected out of eighty. The trending topic with the highest average was *Justin Bieber* and with the lowest average was *Novichok*. The overall disparity of the average values was satisfactory, but on the three of September, on label Adidas, the average value was almost high in all topics. The day with more trending topics was the three of September.

The results were interesting. The high figure of trending topic *Justin Bieber* might be explained by the fact that he made many commercials for Adidas, especially with the shoe line NEO Adidas. Additionally, the trending topic *Big Sean* has a high average value too because he had designed a few Adidas shoes. Unfortunately, that trending topic was not selected by hotRivers prototype. Another interesting trending topic is *Odell*, who is an American football player that was sponsored by Adidas and disputed by Nike. The trending topic *DYNAMITECELEBRATION* refers to a song that hit an astonishing number of views, *FadedFtLOEY* is also a trending topic related to a song. However, there is not a clear connection, there might be indirect connection with singers, for example. Again, for the trending topics *David Blaine*, it was not found evident connections. However, the trending topic *steven adams* and *bryson tiller* are about celebrities that wear Adidas products, nothing solid was found beyond that. The last trending topic is *#JusticeForDeonthat* and it is related to the movement Black Live and Adidas was one of many companies that made public statements against racism.

The UK results on trending topics of days one, twenty, twenty-two, and twenty-four of September are described in Table A.5. The color scheme is equal to the one presented before. Only two out of eighty topics were selected by the HotRivers prototype. The selected topic with the highest similarity was *Ed Sheeran* and the lowest similarity was

#WorldPatientSafetyDay. The majority of the topics are classified as Others. The disparity of values and the number of trending topics selected is satisfactory. The number of trending topics labeled as Others does not matter to Adidas. It indicates that those topics do not have any associations with the company.

The trending topic *Ed Sheeran* refers to the singer Ed Sheeran, who, like other celebrities wear Adidas products, this fact might explain the high average. The model was not clear of which label for the trending topic *Nike*. There is two points worth of stress, first was good that the prototype did not classify it as Adidas, because they are different companies, but on the other hand, having a higher average would not be a surprise, since both companies sell similar products and seem to be direct competitors. Additionally, the trending topic *#PepsiMaxTasteOneStopwere* was classified as Others. This happened with other companies and products such as *Amazon UK*, *argos*, Xbox, Playstation 5, and *iOS 13*. It is important to point out that Amazon, Pepsi, Microsoft, Apple, or Sony were not introduced to the training dataset. In other words, the model was enough well trained to label those topics as Others. The trending topics *#WorldPatientSafetyDay*, *#Covid_19* and *#NHSCOVID19app* were labeled as Others with a high average, due to medical and pharmaceutic companies used on the training dataset. In other words, hospitals and pharmacies made multiple publications about the Covid-19 subject. Which made the model learn that trending topics related to Covid-19 are not related to Adidas. Regarding the trending topic *Leighton Buzzard*, it was not found obvious connections.

5.4 Experiment 2: Nike

In the last Subsection it was not conclusive which algorithm to use, the document embedding vector, or the classification task approach. In experiment two the company tested was Nike. All of these tests helped to find the final architecture for HotRivers.

The data collection procedure was the same to every company and trending topic. All GTP and STP techniques were used, even though, lemmatization was consistently the best cleaning technique in the last experiment.

5.4.1 Experiment 2: Modeling Embedding Vectors, GTP, STP and CS

Applying the same conditions as in experiment one, the results of the Nike sanity test with GTP and STP were not as satisfactory as compared to the previous experiments like it is possible to observe in Table A.6. In this experiment, not using any treatment was better than lemmatization and steaming with and without STP. With Nike dataset without STP and no treatment, 88.4% of the tweets were the most similar to themselves, with lemmatization 88.9%, and with steaming 89.2%. With STP and any of the techniques, approximately 96% of the tweets were the most similar to themselves.

The average similarity with the CS showed, in Table A.7, was not close to the values of the previous experiment. The average of the sanity test was only 0.529, the median was 0.512 and the ranks were worse, with only 54% of the documents showing the most similarity to themselves. The results were not even close to the first experiment, therefore this model is not as suitable for HotRivers. The model to be accepted should be able to adapt to any company that respects the minimal operating requirements and the consistency of good results is important to HotRivers. Hence, this model needs more technical knowledge to find a configuration that optimizes the model. For all those reasons this approach was disregarded.

5.4.2 Experiment 2: Modeling Classification Task Approach, GTP, STP and CS

Again, the condition was the same as in experiment one. The use of the STP did not improve the performance of the model, as can be observed in Table 5.10. Lemmatization continues to be the best technique, followed by no treatment and last steaming. In the next experiments, only lemmatization is used. As it can be observed in Table A.8, the CS did not improve the results. In the next experiments, the STP and CS are not used, expect the CS for trending topics.

	Technique	Loss	Accuracy	Recall	Precision	F-1 score	Quantity of tweets
No STP	No treatment	0.539	0.901	0.901	0.901	0.901	2,880
	Lemmatization	0.535	0.906	0.906	0.906	0.906	
	Steaming	0.654	0.899	0.899	0.899	0.899	
With STP	No treatment	0.547	0.898	0.898	0.898	0.898	2,853
	Lemmatization	0.654	0.899	0.899	0.899	0.899	2,844
	Steaming	0.607	0.882	0.882	0.882	0.882	2,854

TABLE 5.10: Table of GTP, STP and Nike train CNN results

The confusion matrix of the model with GTP and with lemmatization, illustrated in Table 5.11, is acceptable. The model committed approximately the same number of mistakes as in the previous experiment. It missed 70 Nike tweets ,out of 720, and 65 Others tweets, out of 720, which does not indicate that the model has not learned properly. The learning curve in Figure A.5 looks satisfactory and stabilized around 300 and 350 epochs. The metrics accuracy, recall, and precision showed in Figure A.6, in Figure A.8 and in Figure A.7, respectively, do not seem to overfit and are stabilized between 200 and 350 epochs, but the validation curve presents many spikes.

Confusion Matrix		Predicted Label	
		Nike	Others
True Label	Nike	650	70
	Others	65	655

TABLE 5.11: Confusion matrix of Nike Hospital

5.5 Nike Results Discussion

5.5.1 Techniques and Models Discussion with Nike

This experiment with Nike confirms that Doc2Vec was not a suitable model for HotRivers. The lack of good performance with other companies is a factor that matters. Additionally, Doc2vec needs more knowledge and more experiments in order to be optimized for a company, which make the automation objective more complicated to accomplish. Instead, CNN consistently showed good results with GTP with lemmatization and with both companies tests so far. Nevertheless, more experiments are necessary to ensure that the CNN maintains consistent results.

5.5.2 Trending Topics Analysis with Nike

Regarding the UK trending topics, presented in Table A.9, the results were interesting, particularly to the trending topic *Nike*, which gave almost the maximum average. The lowest trending topic was *#AutumnEquinox*. The model was able to recognize Nike very well. The color scheme is equal to the previous experiment, color green for the picked trending topics and color yellow for the standard deviation of the highest value of average. Two trending topics were selected out of eighty. The disparity of the values seems satisfactory and in two days there were not any associations with Nike.

The dataset for training the model was the same used in the last experiment, except for the targeted company. Again, the trending topic *#WorldPatientSafetyDay* were classified with one of the highest average values for the label Others. The trending topics *#Covid_19* and *#NHSCOVID19app* were also labeled as Others. Additionally, the trending topic *#AutumnEquinox* was classified as Others with a high average too. The main reason seems to be the data used in the training. Trending topics such as *iOS 13, #PepsiMaxTasteOneStop, argos*, and *Amazon UK* were also classified as Others and no

association was found to the trending topic *Harold Evans*. The high standard deviation and the low average made a lot of topics being picked.

Regarding the US trending topics, in Table A.10, the results were acceptable. Six trending topics were picked out of eighty. One day did not have any association and two days had only one selected topic. On day four of September, due to how rules to choose the trending topics were made, all topics were in the range of acceptance. Four of them had a connection with Nike, but a few with an average value of approximately 0.50, which was not expected. It means that the algorithm considers half of the tweets as being Nike and the other half as being Others. Nevertheless, the rest of the days showed satisfactory results. The trending topic with the highest average was *#JusticeForDeon* and the lowest was *#firstdayofschool*. On day two of September, the trending topic *David Blaine* is related to a commercial made by Nike. Still, on the same day, the trending topic *Novichok* is about a nerve agent, and it seems correctly labeled as Others because it appears to not have anything in common with Nike. On the three of September, the trending topics *Lakers* and *Rockets* are about NBA basketball teams sponsored by Nike, on the same day, the trending topic *steven adams* is about a basketball player of the Oklahoma City Thunder NBA team, which is also sponsored by Nike. Furthermore, the trending topic *bryson tiller* refers to a songwriter who helped to inspire the *Bryson Tiller Nike Air Force 1* shoes. Unfortunately, non of those trending topics were in the range due to how selection calculations were made. The trending topic *#JusticeForDeon* might have such a high average because Nike supports the *#BlackLivesMatter* movement. The trending topic *#JusticeForDeon* was an event related to the murder of a black person during an anti-racism protest. The four of September have peculiar results because the highest value of average also has a high standard deviation. On the four of September the only trending topics that may have direct connection with Nike were *Big Sean* and *Justin Bieber*. The trending topics with an average of around 0.50 mean that the algorithm is confused with the right label. The last day did not have any connection, apart from the trending topic *Odell* that could be classified with Nike. *Odell* is about an American football player that is sponsored by Nike.

5.6 Experiment 3: Royal Manchester Children's Hospital

In this experiment, the conditions are equivalent to the last experiments. RMCH is different from previous companies. RMCH has fewer tweets collected and the medical area is distinct from fashion, sport, and shoes, which influences the way tweets are written. This experiment and the next, test the consistency of the results and how a different business area is interpreted by HotRivers.

The data preparation method used was GTP with lemmatization. The trending topics tweets selected with the CS approach were ten. The validation results of the RCMH model are observed in Table 5.12. They were less satisfactory than the results obtain for previous companies, but the training dataset is also different, since no medical or pharmaceutic companies were on the training dataset. A few points worth of discussion are the 84.9% of F1-score and the 85.1% of accuracy, which are good results since the number of RCMH tweets are half of the ones from Adidas.

Loss	Accuracy	Recall	Precision	F1-score	Quantity of tweets
0.657	0.851	0.849	0.849	0.849	1,503

Table 5.12: Table of results of RMCH model test

The confusion matrix, in Table 5.13, show that the model misclassified less tweets. However this fact cannot be interpreted as better results since the number of tweets tested were fewer than Adidas and Nike datasets. The model misclassified 55, out of 376, RMCH tweets and 57, out of 376, Others tweets. The Loss curve showed in Figure A.9 seems to stabilize around 300 and 350 epochs. The accuracy curve showed in Figure A.10, the precision in Figure A.11, and the recall in Figure A.12, all have many spikes, but the curves move in the same direction and there is no clear gap, which suggests no overfitting. Again, the curves seem to stabilize between 300 and 350 epochs.

Confusion Matrix		Predicted Label	
		RMCH	Others
True Label	RMCH	321	55
	Others	57	319

TABLE 5.13: Confusion matrix of RMCH

5.7 Royal Manchester Children's Hospital Results Discussion

The CNN has presented a good performance until now as well as the data preparation phase techniques. Further tests are conducted with these conditions.

The US trending topics results are shown in Table A.11, the disparity of the average values was acceptable. The scheme of color is equal to the previous experiments, the color green is the picked trending topic and the yellow color is the standard deviation of the highest average of the day. The trending topic with the highest value of average was *#FadedFtLOEY* and the lowest was *#ThursdayThoughts*. Eight topics were selected in a total of eighty. The overall values were not that high as in the other experiments and this led to that on the two of September, due to how the selection method was made, eight topics were picked. Additionally, the overall values of average were low and a higher value of standard deviation led to many trending topics being picked. Unfortunately, a few topics did not have a high average value as for example *#firstdayofschool*.

On day two of September, the trending topic selected *Carole Baskin* is related to saving animal lives. The trending topic *#BillAndTed3Sweeps* was about a movie that launched a face mask as merchandising. The trending topics *#FadedFtLOEY* and *DYNAMITE CELEBRATION* were related to songs, therefore the connection is harder to find. The associations with trending topics *Big Sean, David Blaine*, and *Justin Bieber*, might be related to public thanks or events to help health professionals. The model also classifies as Other the trending topic *Sarah Sanders* and *John McCain*, these topics regard American politicians and so it does not seem to exist any relation to RMCH. The trending topic *John*

Boyega is related to an actor who is a supporter of anti-racist causes, and the training dataset has some companies that become publicly against racism. Finally, the following trending topic seems to be correctly classified as Others, *woojin* is a south-Korean singer, *Xbox* is a gaming product and *Novichok* is related to chemical weapons subject.

In Table A.12 is observed the UK trending topics results. The same color scheme was used, as the previous table. The results were peculiar, no associations were found with UK trending topics. The trending topic with a higher average was *Gigi* and the lowest was *#NHSCOVID19app*.

The training dataset may be the cause of non-associations with RMCH. The trending topics such as *Nike* and *argos* had companies included in the dataset, *#PepsiMaxTasteOneStop* did not have Pepsi, so the model labeled it correctly. Again, The trending topics *#ps5preorder*, *#XboxSeriesX*, and *iOS 13*, related to gaming and technological products, so not related to RMCH and correctly labeled. Unfortunately, the trending topic *#WorldPatientSafetyDay* did not have the expected association. Also, the trending topics associated with covid-19 and the pandemic such as *#SackWhitty, #northeastlockdown, #Covid_19* and #NHSCOVID19app were classified as Others. Possible reasons are the training dataset, the GTP process, and the Tweet decision discussed in Subsubsection 4.3.3.3. RMCH made many retweets about Covid-19 and not many tweets related to Covid-19, unfortunately, only tweets were used. Additionally, the hashtags were deleted and many RMCH publications had the *#Covid19, #Covid-19, #Covid_19*, or other variants. Another point contributing to that situation is that companies such as Nike, Adidas, Tesco and the others, that were in training dataset label as Others, made publications about their Covid-19 policies and giving to support to everyone suffering from the pandemic.

5.8 Experiment 4.: Portsmouth Hospitals University

As mentions before, the data collection phase is equal to all experiments, the data preparation phase is the same as the previous experiments, GTP with lemmatization, the CS only for trending topics, and the modeling phase is the classification task approach. CNN had a satisfactory performance until now, therefore this approach continued to be used. The next sections are equal to the previous sections, first the training results and then the trending topics. This is the last experiment to confirm HotRivers final archictecture.

The same dataset from the last experiment was used on the training dataset, except for the target company the PHU. The results of the validation test are observed in Table 5.14. The overall results were better than RMCH and worse than Nike and Adidas model, even though 87.2% of the F1-score and 87.2% of accuracy is a satisfatory result.

Loss	Accuracy	Recall	Precision	F1-score	Quantity of tweets
0.638	0.872	0.872	0.872	0.872	2,259

Table 5.14: Table of results of PHU model test

The confusion matrix in Table 5.15 shows more difficulty in classifying correctly the PHU tweets. As observed in Figure A.13, the loss curve stabilized between 300 and 350 epochs, like in the previous experiments. Regarding the metrics accuracy, precision, and recall, illustrated in Figures A.14, A.15 and A.16, respectively, showed a little overfited on validation test curve, but stabilized between 250 and 350 epochs. Nevertheless, the results suggest to be a competent model.

Confusion Matrix		Predicted Label	
		PHU	Others
True Label	PHU	499	57
	Others	88	486

Table 5.15: Confusion matrix of PHU

5.9 Portsmouth Hospitals University Results Discussion

This last experiment validates the data preparation technique GTP with lemmatization and the CS for trending topics, and the modeling phase, the classification task approach with a CNN as the model. This is the final architecture of HotRivers as it can be observed in Figure 6.1.

Concerning the UK trending topics results showed, in Table A.13, only day seventeen of September had associations with PHU. In a total of eighty trending topics, only three were picked. The trending topic with the lowest value of average was *Ed Sheeran* and the highest was *#WorldPatientSafetyDay*. The disparity of values is satisfactory and the chosen standard deviations were close to zero, which means that the model classified the trending topics with confidence.

The trending topic *#WorldPatientSafetyDay* is related to the hospital affairs, but *#SackWhitty* and *#northeastlockdown* are associate to Covid19 and political issues, apart of Covid-19 subject, the political matter may not be the interest of the hospital. On the opposite side classified as Others, the trending topic *Ed Sheeran* was correctly classified since in PHU training dataset Adidas was labeled as Others. Again, it is worth to point that other companies trending tropics *#PepsiMaxTasteOneStop* and *Nike* had no association with PHU. As a trending topic *argos*, which is a subsidiary of Sainsbury's, was another company labeled as Others on the training dataset. On day twenty-two of September, the trending topic *#Covid_19* did not have enough value of average, because *Leighton Buzzard* had a high association with label Others. The lack of an association with *#Covid_19* and *#NHSCOVID19app* might be explained by the reasons as presented on Section 5.7.

Concerning the US trending topics results showed in Table A.14, there are no associations found. The trending topic with the lowest average was the *Matt Watson* and one

with the highest average was *John McCain*. The disparity of the values is not as satisfactory, but it is not bad enough to say that the model was unacceptable. The geographic location of the company and the trending topic, as well as, the training companies used may have played an important role in the average values.

On day two of September, most of the higher values of averages were related to North American public personalities, it is not surprising that they were not related. Again, on days three and four of September, the trending topics had a high value of association with companies labeled as Others, for example Adidas and Nike. On the eight of September was peculiar because the five selected topics, all had high values of average, but highly related to other companies on the training dataset. Nevertheless, this does not seem wrong, since are no associations between these topics and the PHU.

Chapter 6

Conclusion

To complete the objectives of this work, HotRivers prototype was designed and implemented. The final scheme of HotRivers, as observed in Figure 6.1, consists in data collection phase as described in Section 3.4, data preparation phase with GTP with lemmatization and CS, only for trending topics, and modeling phase with the classification task approach. This scheme complies with all minimal operating requirements and goals of this work.

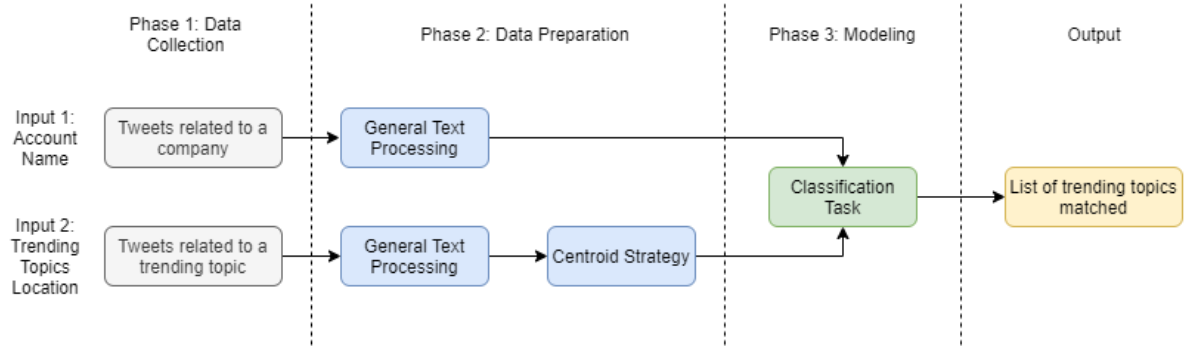


FIGURE 6.1: Final Scheme of HotRivers, for company tweets GTP with lemmatization, and for trending topics GTP with lemmatization and th CS and the model is the CNN

Relating to the models tested in the four experiments, it was concluded that tweets made by companies to marketing purposes are written in a way that results in low number

of unique words and short tweets, as illustrated in Table 5.6. Also, in Subsection 5.2.2 is observed that on Adidas' tweets the most frequent length of tweets is between two and four words per tweet, contrary to journalistic texts, reviews or books [43, 45], which are extensive and detailed. This is one of the reasons to justify the mediocre performance of models such as Doc2Vec and LDA. Tweet aggregation methods could have helped [29, 40] the models learning, however the variety of words would not change with tweet aggregation. On the other hand, the model CNN was able to learn the data well and classify the tweets correctly. A few points worth of stressing is that repetitions did not seem to have impact on learning, also when training the model with lower number of tweets, such as the experiment of RMCH, the model showed more difficulty in learning the data. Therefore, the CNN was the selected model for HotRivers prototype.

Relating to the techniques used in the four experiments, it was concluded that steaming technique did not improve the models [40], on the other hand lemmatization was consistently better in all the experiments. The STP did not have the desired impact on models and did not improve them. Again, the CS did not increase the results, except for trending topics to avoid unwanted tweets, as shown in Figure 3.6, however, it was an useful clustering technique [38, 41, 44] to select the most representative tweets and extract less tweets from trending topics.

One of the main goals of this work was to search for associations between companies and trending topics. The most evident result was the high association between the company Nike and the trending topic *Nike*, which had a value of 0.986, almost the maximum possible value. Another outcome worth of reference was that with the company Adidas, all trending topics selected had an association value higher than 0.809, which shows high confidence from the model. Also, between the hospital PHU and *#WorldPatientSafetyDay* and *#northeastlockdown*, there was high association, which was expected since both the company and trending topics are related to medical and health affairs. Unfortunately, the trending topics *#Covid_19* and *#NHSCOVID19app* were not picked by the prototype as associated with PHU, however both trending topics had an average value of approximately 0.68. The RMCH had the worst results of all and also had the lowest

number of tweets, which may have influenced the results. Unfortunately, the trending topics *#Covid_19, #NHSCOVID19app, #northeastlockdown* and *#WorldPatientSafetyDay* were not select for RMCH. Even though, it was possible to conclude that the connection between companies and trending topics exists.

The decision to exclude retweets might have conditioned the results, because by evaluating the RMCH Twitter account feed most of the tweets related to Covid-19 were retweets. This might be one of the reasons why PHU and RMHC did not have higher values of association with Covid-19 topics, for example.

Relating to the other objectives of this work, one of the them was creating a low cost solution, in Section 4.3 was study the impact of Twitter accounts, and in Section 4.2 the available resources. It was found that a free account was enough and all solutions presented were achievable with it.

Another objective of this work was fast delivered results, it was set on the minimum operating requirements that the list of trending topics associations should be delivered in less than four hours. In order to accomplish this goal, HotRivers' parameters were optimized, as explain in Section 4.2. Additionally, the implementation of the CS, in Subsection 4.4.3, made possible to decrease the extraction time of trending topics from one day to approximately one hour, as explain in Subsection 4.3.3.

Another goal of this work was to offer a solution that did not required specialized staff. Since, the HotRivers prototype is fully automated, it only needs two inputs to be operated, the target account and the location of trending topics, and the output of HotRiver is a list of trending topics associations, which are easily searched on twitter to find out more information about them. The complexity to use and to understand the results is low.

It was not conclusive if companies in the same market have similar associations, further investigation with more companies and topics is needed to confirm that. In the experiments done for Adidas and Nike, the same seven trending topics were picked, but

only the trending topics *David Blaine, #JusticeForDeon* and *Justin Bieber* were classified as associated to Adidas and Nike, the rest were classified as Others. In the case of RMCH and PHU, eight equal trending topics were selected out of eighty. Curiously, most of the trending topics classified as RMCH were classified as Others for PHU and vice versa. Even though they are both hospitals, they may have different targets, which may explain the polarity of the results.

The final appreciation is that HotRivers can be a powerful tool to point out the direction of the marketing campaign. It was possible to observe that the HotRivers prototype was capable of finding associations between companies and trending topics. Thus, the HotRivers prototype showed a lot of potentials to be used in a larger and complex marketing platform. In Table 6.1 is illustrated a summary of the objectives, conclusions and contributions of this work.

Objectives	Conclusion	Contributions
Finding association between companies and trending topics	The CNN were the model that was better adapted to the four companies	- A prototype cable of finding associations between trending topics for an economical price, easy to operate and fast delivering results - Deep analyses on Twitter API and tweets problems, and explanation on how to solve them - Three different approaches were tested on four different companies - First work that tried to search for associations between companies and trending topics
Building a solution that is: - Inexpensive - Fast - Automated - Understandable for non-specialized staff	- Free accounts are available - Results in less than one hour - Only requires two inputs from the user - Fully automated and the output is easy to understand	
Understanding if companies in the same area of business shared similar trending topics association	Not conclusive and more work needs to be done	

TABLE 6.1: Summary of objectives, conclusions and contributions

The present work has been submitted to the journal Multimedia Tools and Applications (https://www.springer.com/journal/11042)

6.1 Future Work

For future work, it is necessary to measure the impact of this prototype with more companies and more focus on the type of companies used. The companies used on the training data set for the label Others should be improved and further research is needed to understand better combinations and to better train the model. This would also reduce the influence of accompanies labeled as Other on the results. Also, more research on how to improve the techniques implemented and to generate more value for the user. The hashtags, mentions, and emojis need to be properly handled to be used to improve model learning. Also, more studies need to be done around the companies used on the dataset. While in this work a binary approach was used, a multi-class approach could solve the training dataset problem. Additionally, the use of other feature than text should be considered such as timeline of the events, geography, sentiments and others.

As future work it would be important to test HotRivers prototype in a real-life scenario with a company to measure the user interaction, number of followers and likes. Since HotRivers is a solo piece of software it could be integrated into a bigger solution that provides more information about how to structure the campaign. Additionally, a few trending topics appear on the top ten more than one time, this is something that companies could benefit by predicting when they will come back.

6.2 Limitations

The pandemic Covid-19 limited the OMECO project by delaying it several times. It affected the development of this work and the HotRivers prototype. Even though a different schedule was used, it difficulted the task.

Regarding confirming the associations between companies and trending topics, it was made by resorting the news and queries on google and by analyzing the content of tweets from some companies and trending topics. Even though it is possible to state there are

reasonable associations, more work needs to be done. It is important to stress that the data used for training had a big influence on the decision of the model and the fact that the model needed to chose between two classes, which may induce sometimes randomness of the output, instead of the probability of each class independently.

Another limitation is that those associations in trending topics that apparently did not have a clear connection in these work may have when analysed by a marketing specialist or the employees of the targeted company. Also, trending topics with not a solid connection might still have potential for marketing purposes. That is why judging the model decision veracity is a complicated task and further studies with companies are required. Nevertheless, it presents interesting results.

Appendices

Appendix A

HotRivers Experiments Results

Techniques	Tweet n°42
No treatment	['week', 'huddle', 'discuss', 'things', 'careers', 'life', 'goals', 'watch', 'full', 'conversation', 'youtube']
Lemmatization	['week', 'huddle', 'discuss', 'thing', 'career', 'life', 'goal', 'watch', 'full', 'conversation', 'youtube']
Steaming	['week', 'huddl', 'discuss', 'thing', 'career', 'life', 'goal', 'watch', 'full', 'convers', 'youtub']

TABLE A.1: Table of a tweet chosen randomly cleaned with various GTP techniques

Day	Trending Topic Name	Average Similarity	Standard Devitation Similarity	Median Similarity
02-09-2020	Kirk Cousins	0.751	0.018	0.752
	Carole Baskin	0.470	0.016	0.474
	#WednesdayWisdom	0.943	0.028	0.951
	#TheMandalorian	0.948	0.027	0.956
	David Blaine	0.677	0.015	0.681

	John Boyega	0.844	0.028	0.848
	#BillAndTed3Sweeps	0.392	0.014	0.394
	Sarah Sanders	0.946	0.021	0.950
	Keanu Reeves	0.855	0.021	0.859
	Novichok	0.842	0.027	0.847
03-09-2020	Lakers	0.948	0.027	0.953
	#ThursdayThoughts	0.871	0.017	0.874
	steven adams	0.957	0.023	0.964
	Rockets	0.823	0.026	0.827
	Dort	0.656	0.022	0.662
	#DokkanSquad	0.461	0.016	0.464
	#FadedFtLOEY	0.584	0.017	0.587
	#SexIsGreatButHaveYou	0.844	0.023	0.846
	bryson tiller	0.754	0.023	0.758
	#JusticeForDeon	0.601	0.012	0.605
04-09-2020	#DETROIT2	0.842	0.026	0.847
	Big Sean	0.967	0.015	0.972
	Matt Watson	0.784	0.020	0.788
	John McCain	0.779	0.034	0.779
	#TrumpHatesOurMilitary	0.837	0.028	0.840
	#FridayMotivation	0.939	0.025	0.946
	Wonho	0.725	0.025	0.728
	#BeyDay	0.637	0.026	0.640
	#Mulan	0.943	0.029	0.951
	Justin Bieber	0.687	0.013	0.690
08-09-2020	Odell	0.847	0.025	0.853
	Hyrule Warriors	0.950	0.024	0.956
	#StarTrekDay	0.851	0.021	0.856
	#appleevent	0.942	0.028	0.951

	#sendkeeblermagic	0.756	0.023	0.761
	#TuesdayThoughts	0.950	0.025	0.958
	#firstdayofschool	0.915	0.032	0.918
	Xbox	0.872	0.017	0.876
	DYNAMITE CELEBRATION	0.276	0.011	0.277
	woojin	0.901	0.034	0.905

TABLE A.2: Table of Adidas Doc2Vec model of days two, three, four and eight of September US trending topics similarity

Day	Trending Topic Name	Average Similarity	Standard Devitation Similarity	Median Similarity
01/09/20202	#September1st	0.916	0.033	0.922
	#BackToSchool	0.827	0.030	0.835
	Ritchie	0.746	0.024	0.752
	Marcus Rashford	0.919	0.029	0.923
	Ed Sheeran	0.575	0.020	0.576
	#TuesdayMorning	0.837	0.027	0.844
	#ThisMorning	0.763	0.019	0.768
	#PepsiMaxTasteOneStop	0.566	0.012	0.569
	Jim Davidson	0.861	0.020	0.865
	Nike	0.951	0.025	0.958
17/09/2020	Thiago	0.956	0.023	0.963
	Alex Scott	0.660	0.022	0.665
	#ThursdayThoughts	0.831	0.027	0.838
	#WorldPatientSafetyDay	0.735	0.022	0.738
	#SackWhitty	0.848	0.025	0.854
	#ps5preorder	0.961	0.021	0.967
	#northeastlockdown	0.856	0.021	0.862

	Amazon UK	0.487	0.014	0.490
	Neil Warnock	0.738	0.028	0.740
	argos	0.931	0.028	0.937
22/09/2020	#XboxSeriesX	0.927	0.031	0.935
	Starmer	0.942	0.026	0.947
	#TuesdayThoughts	0.848	0.024	0.854
	#AutumnEquinox	0.754	0.020	0.760
	Britain First	0.894	0.037	0.894
	Michael Gove	0.922	0.030	0.925
	#GBBO	0.957	0.019	0.961
	#Covid_19	0.665	0.019	0.670
	Leighton Buzzard	0.508	0.014	0.512
	Bake Off	0.865	0.017	0.869
24/09/2020	#NHSCOVID19app	0.970	0.021	0.978
	#ThursdayThoughts	0.583	0.018	0.586
	Harold Evans	0.563	0.015	0.568
	#Magic	0.928	0.031	0.934
	iOS 13	0.843	0.027	0.851
	Kent	0.767	0.020	0.772
	Gigi	0.505	0.013	0.508
	#thursdayvibes	0.917	0.033	0.923
	Downloaded	0.735	0.027	0.740
	#bbcbreakfast	0.921	0.032	0.929

TABLE A.3: Table of Adidas Doc2Vec mode of days two, three, four and eight of September UK trending topics similarity

FIGURE A.1: Model loss graph on train and validation on Adidas training dataset

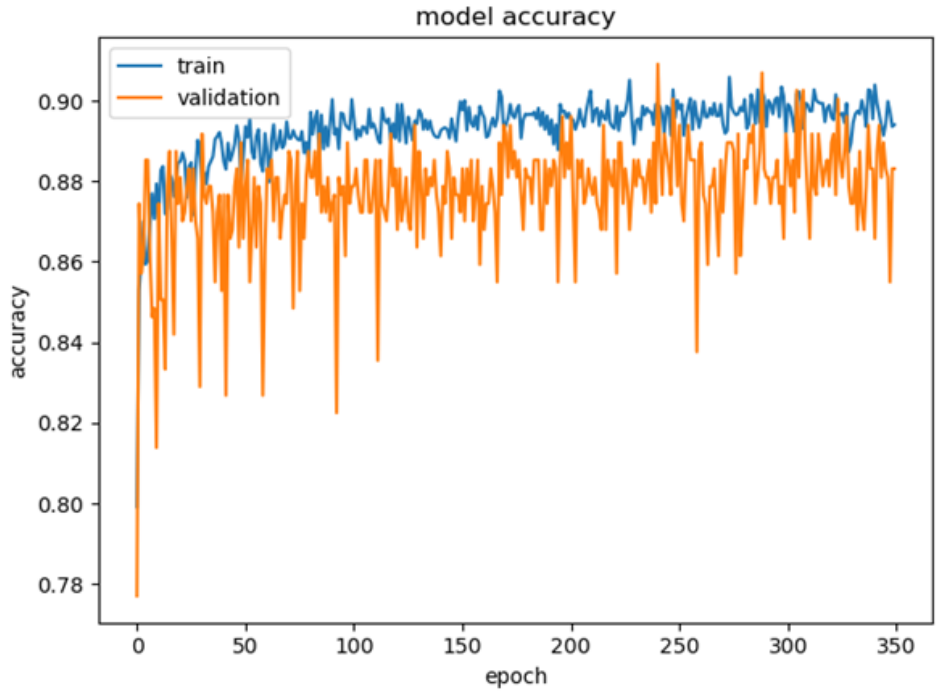


FIGURE A.2: Model accuracy graph on train and validation on Adidas training dataset

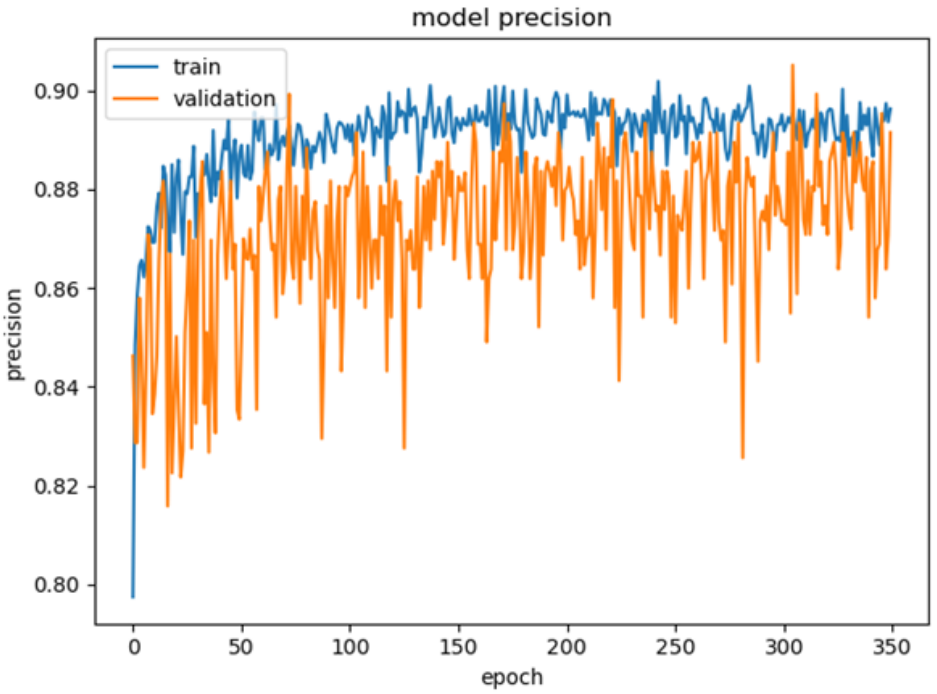


FIGURE A.3: Model precision graph on train and validation on Adidas training dataset

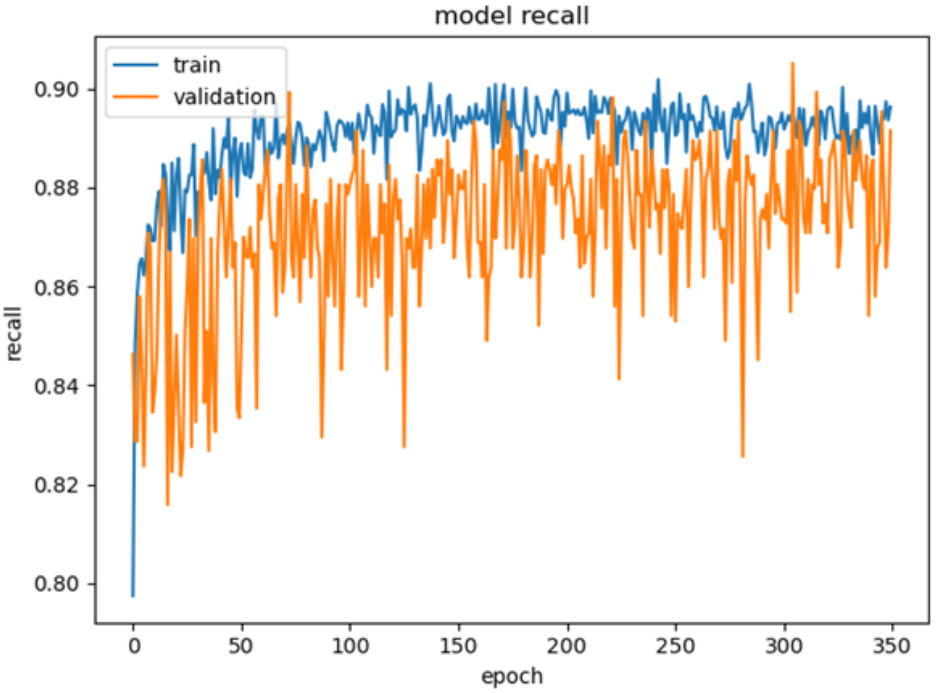


FIGURE A.4: Model recall graph on train and validation on Adidas training dataset

Day	Trending Topic Name	Average - Adidas	Average - Others	Standard Devitation - Adidas	Standard Devitation - Others
02/09/2020	Kirk Cousins	0.695	0.305	0.174	0.174
	Carole Baskin	0.745	0.255	0.100	0.100
	#WednesdayWisdom	0.466	0.534	0.311	0.311
	#TheMandalorian	0.606	0.394	0.130	0.130
	David Blaine	0.856	0.144	0.105	0.105
	John Boyega	0.466	0.534	0.222	0.222
	#BillAndTed3Sweeps	0.573	0.427	0.209	0.209
	Sarah Sanders	0.422	0.578	0.215	0.215
	Keanu Reeves	0.706	0.294	0.178	0.178
	Novichok	0.225	0.775	0.186	0.186
03/09/2020	Lakers	0.598	0.402	0.169	0.169
	#ThursdayThoughts	0.375	0.625	0.078	0.078
	steven adams	0.852	0.148	0.035	0.035
	Rockets	0.752	0.248	0.129	0.129
	Dort	0.787	0.213	0.139	0.139
	#DokkanSquad	0.705	0.295	0.162	0.162
	#FadedFtLOEY	0.885	0.115	0.079	0.079
	#SexIsGreatButHaveYou	0.608	0.392	0.233	0.233
	bryson tiller	0.809	0.191	0.100	0.100
	#JusticeForDeon	0.831	0.169	0.061	0.061
04/09/2020	#DETROIT2	0.499	0.501	0.184	0.184
	Big Sean	0.818	0.182	0.090	0.090
	Matt Watson	0.747	0.253	0.168	0.168
	John McCain	0.668	0.332	0.235	0.235
	#TrumpHatesOurMilitary	0.447	0.553	0.231	0.231
	#FridayMotivation	0.464	0.536	0.243	0.243

	Wonho	0.289	0.711	0.138	0.138
	#BeyDay	0.456	0.544	0.222	0.222
	#Mulan	0.374	0.626	0.196	0.196
	Justin Bieber	0.901	0.099	0.017	0.017
08/09/2020	Odell	0.855	0.145	0.089	0.089
	Hyrule Warriors	0.676	0.324	0.233	0.233
	#StarTrekDay	0.417	0.583	0.164	0.164
	#appleevent	0.522	0.478	0.289	0.289
	#sendkeeblermagic	0.570	0.430	0.275	0.275
	#TuesdayThoughts	0.584	0.416	0.179	0.179
	#firstdayofschool	0.438	0.562	0.153	0.153
	Xbox	0.461	0.539	0.157	0.157
	DYNAMITE CELEBRATION	0.812	0.188	0.109	0.109
	woojin	0.409	0.591	0.194	0.194

TABLE A.4: Table of Adidas CNN model of days two, three, four and eight of September US trending topics results

Day	Trending Topic Name	Average - Adidas	Average - Others	Standard Devitation - Adidas	Standard Devitation - Others
01/09/2020	#September1st	0.212	0.788	0.159	0.159
	#BackToSchool	0.269	0.731	0.154	0.154
	Ritchie	0.367	0.633	0.194	0.194
	Marcus Rashford	0.506	0.494	0.248	0.248
	Ed Sheeran	0.838	0.162	0.049	0.049
	#TuesdayMorning	0.416	0.584	0.197	0.197
	#ThisMorning	0.402	0.598	0.180	0.180
	#PepsiMaxTasteOneStop	0.263	0.737	0.216	0.216
	Jim Davidson	0.629	0.371	0.209	0.209

	Nike	0.560	0.440	0.329	0.329
17/09/2020	Thiago	0.532	0.468	0.251	0.251
	Alex Scott	0.308	0.692	0.168	0.168
	#ThursdayThoughts	0.332	0.668	0.137	0.137
	#WorldPatientSafetyDay	0.013	0.987	0.020	0.020
	#SackWhitty	0.130	0.870	0.073	0.073
	#ps5preorder	0.222	0.778	0.101	0.101
	#northeastlockdown	0.202	0.798	0.163	0.163
	Amazon UK	0.326	0.674	0.183	0.183
	Neil Warnock	0.414	0.586	0.288	0.288
	argos	0.382	0.618	0.232	0.232
22/09/2020	#XboxSeriesX	0.251	0.749	0.156	0.156
	Starmer	0.214	0.786	0.153	0.153
	#TuesdayThoughts	0.298	0.702	0.214	0.214
	#AutumnEquinox	0.191	0.809	0.069	0.069
	Britain First	0.304	0.696	0.164	0.164
	Michael Gove	0.440	0.560	0.259	0.259
	#GBBO	0.265	0.735	0.104	0.104
	#Covid_19	0.121	0.879	0.166	0.166
	Leighton Buzzard	0.826	0.174	0.022	0.022
	Bake Off	0.345	0.655	0.202	0.202
24/09/2020	#NHSCOVID19app	0.054	0.946	0.065	0.065
	#ThursdayThoughts	0.323	0.677	0.205	0.205
	Harold Evans	0.673	0.327	0.296	0.296
	#Magic	0.334	0.666	0.161	0.161
	iOS 13	0.130	0.870	0.070	0.070
	Kent	0.207	0.793	0.193	0.193
	Gigi	0.677	0.323	0.185	0.185
	#thursdayvibes	0.408	0.592	0.193	0.193

	Downloaded	0.355	0.645	0.145	0.145
	#bbcbreakfast	0.228	0.772	0.115	0.115

TABLE A.5: Table of Adidas CNN model of days one, seventeen, twenty-two and twenty-four of September UK trending topics results

	Techniques	Average	Median	Ranks	Number of Tweets
No STP	No treatment	0.307	0.305	[0: 2550, 1: 134, 2: 45, 3: 18, 4: 16, ...]	2,883
	Lemmatization	0.300	0.297	[0: 2565, 1: 130, 2: 43, 3: 25, 4: 13, ...]	
	Steamming	0.299	0.296	[0: 2573, 1: 125, 2: 47, 3: 28, 4: 14, ...]	
With STP	No treatment	0.299	0.293	[0: 2194, 1: 71, 2: 8, 916:1, 2130: 1, ...]	2,276
	Lemmatization	0.289	0.282	[0: 2188, 1: 76, 2: 9]	2,273
	Steamming	0.289	0.281	[0: 2199, 1: 66, 2: 8, 1021: 1]	2,274

TABLE A.6: Table of Nike sanity test results

Average Similarity	Median Similarity	Standard Deviation Similarity	Tweets	Ranks
0.560	0.575	0.241	50	[0: 27, 1: 7, 2: 5, 3: 4, 4:3, 5:1, ...]
0.333	0.312	0.196	100	[0: 99, 1: 1]
0.303	0.291	0.136	250	[0: 245, 1: 5]
0.290	0.278	0.118	500	[0: 489, 1: 10, 2: 1]
0.297	0.277	0.105	1,000	[0: 966, 1: 29, 2: 5]

TABLE A.7: Table of Nike sanity test with GTP, STP, CS techniques

Quantity of adidas tweets	Loss	Accuracy	Recall	Precision	F-1 score
1,000	0.544	0.874	0.876	0.876	0.876
1,440	0.615	0.879	0.879	0.879	0.879
2,160	0.616	0.899	0.899	0.899	0.899

TABLE A.8: Table of results of the CS with different quantities of tweets on Nike model train

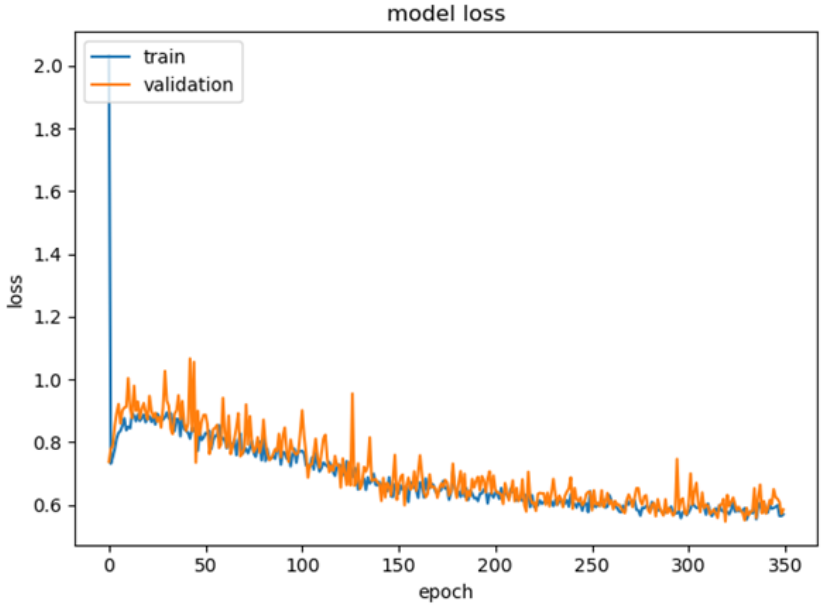


FIGURE A.5: Model loss graph on train and validation on Nike training dataset

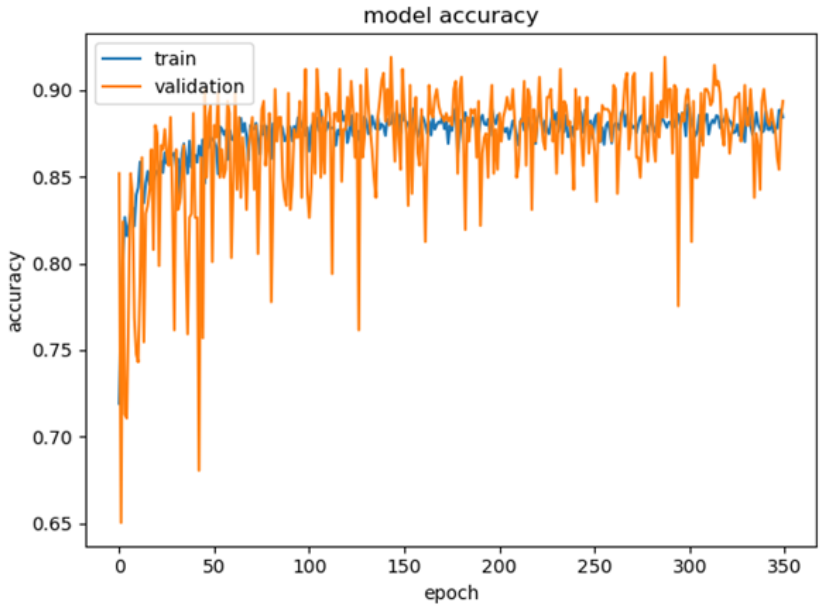


FIGURE A.6: Model accuracy graph on train and validation on Nike training dataset

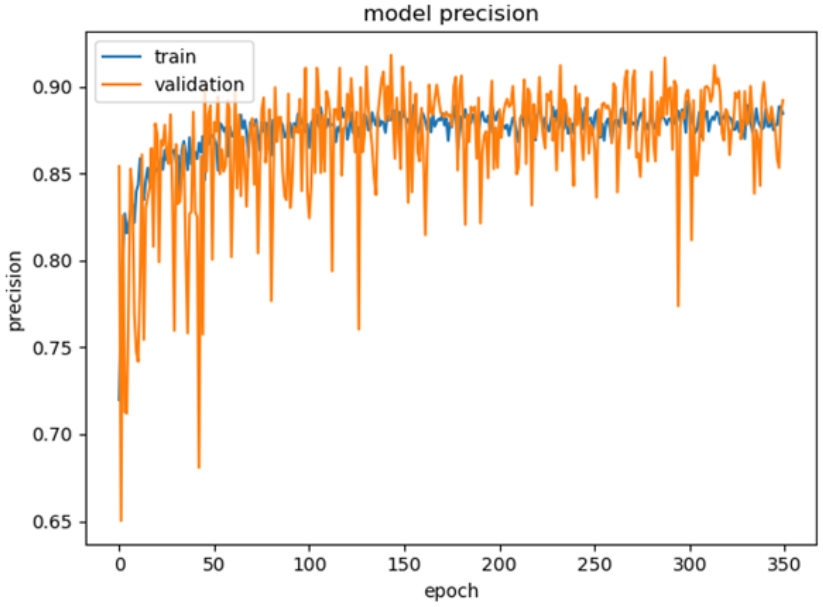


FIGURE A.7: Model precision graph on train and validation on Nike training dataset

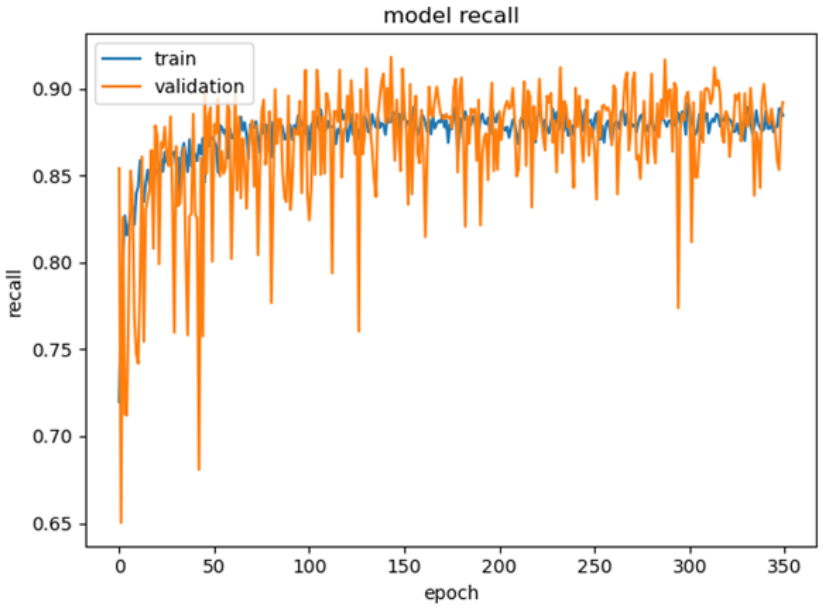


FIGURE A.8: Model recall graph on train and validation on Nike training dataset

Day	Trending Topic Name	Average - Nike	Average - Others	Standard Devitation - Nike	Standard Devitation - Others
01/09/2020	#September1st	0.220	0.780	0.199	0.199
	#BackToSchool	0.251	0.749	0.144	0.144
	Ritchie	0.525	0.475	0.135	0.135
	Marcus Rashford	0.500	0.500	0.163	0.163
	Ed Sheeran	0.352	0.648	0.134	0.134
	#TuesdayMorning	0.371	0.629	0.151	0.151
	#ThisMorning	0.362	0.638	0.186	0.186
	#PepsiMaxTasteOneStop	0.323	0.677	0.170	0.170
	Jim Davidson	0.435	0.565	0.204	0.204
	Nike	0.986	0.014	0.017	0.017
	Thiago	0.579	0.421	0.211	0.211
	Alex Scott	0.265	0.735	0.133	0.133
	#ThursdayThoughts	0.462	0.538	0.180	0.180

	#WorldPatientSafetyDay	0.036	0.964	0.024	0.024
	#SackWhitty	0.215	0.785	0.149	0.149
	#ps5preorder	0.566	0.434	0.173	0.173
	#northeastlockdown	0.351	0.649	0.191	0.191
	Amazon UK	0.400	0.600	0.236	0.236
	Neil Warnock	0.491	0.509	0.171	0.171
	argos	0.514	0.486	0.206	0.206
22/09/2020	#XboxSeriesX	0.603	0.397	0.210	0.210
	Starmer	0.268	0.732	0.202	0.202
	#TuesdayThoughts	0.345	0.655	0.234	0.234
	#AutumnEquinox	0.077	0.923	0.032	0.032
	Britain First	0.428	0.572	0.195	0.195
	Michael Gove	0.393	0.607	0.238	0.238
	#GBBO	0.229	0.771	0.132	0.132
	#Covid_19	0.237	0.763	0.201	0.201
	Leighton Buzzard	0.735	0.265	0.048	0.048
	Bake Off	0.286	0.714	0.189	0.189
24/09/2020	#NHSCOVID19app	0.200	0.800	0.191	0.191
	#ThursdayThoughts	0.459	0.541	0.230	0.230
	Harold Evans	0.616	0.384	0.229	0.229
	#Magic	0.303	0.697	0.191	0.191
	iOS 13	0.284	0.716	0.245	0.245
	Kent	0.328	0.672	0.231	0.231
	Gigi	0.378	0.622	0.077	0.077
	#thursdayvibes	0.510	0.490	0.200	0.200
	Downloaded	0.451	0.549	0.232	0.232
	#bbcbreakfast	0.457	0.543	0.258	0.258

TABLE A.9: Table of Nike model results of days one, seventeen, twenty-two and twenty-four of September UK trending topics

Day	Trending Topic Name	Average - Nike	Average - Others	Standard Deviation - Nike	Standard Deviation - Others
02/09/2020	Kirk Cousins	0.589	0.411	0.167	0.167
	Carole Baskin	0.426	0.574	0.172	0.172
	#WednesdayWisdom	0.326	0.674	0.204	0.204
	#TheMandalorian	0.555	0.445	0.145	0.145
	David Blaine	0.696	0.304	0.102	0.102
	John Boyega	0.337	0.663	0.151	0.151
	#BillAndTed3Sweeps	0.481	0.519	0.228	0.228
	Sarah Sanders	0.316	0.684	0.148	0.148
	Keanu Reeves	0.595	0.405	0.148	0.148
	Novichok	0.337	0.663	0.170	0.170
03/09/2020	Lakers	0.764	0.236	0.113	0.113
	#ThursdayThoughts	0.376	0.624	0.190	0.190
	steven adams	0.768	0.232	0.078	0.078
	Rockets	0.826	0.174	0.087	0.087
	Dort	0.697	0.303	0.123	0.123
	#DokkanSquad	0.706	0.294	0.122	0.122
	#FadedFtLOEY	0.782	0.218	0.147	0.147
	#SexIsGreatButHaveYou	0.585	0.415	0.173	0.173
	bryson tiller	0.760	0.240	0.076	0.076
	#JusticeForDeon	0.922	0.078	0.000	0.000
04/09/2020	#DETROIT2	0.539	0.461	0.151	0.151
	Big Sean	0.667	0.333	0.091	0.091
	Matt Watson	0.442	0.558	0.208	0.208
	John McCain	0.496	0.504	0.150	0.150
	#TrumpHatesOurMilitary	0.351	0.649	0.259	0.259

	#FridayMotivation	0.292	0.708	0.210	0.210
	Wonho	0.326	0.674	0.152	0.152
	#BeyDay	0.376	0.624	0.207	0.207
	#Mulan	0.552	0.448	0.264	0.264
	Justin Bieber	0.670	0.330	0.104	0.104
08/09/2020	Odell	0.611	0.389	0.203	0.203
	Hyrule Warriors	0.572	0.428	0.114	0.114
	#StarTrekDay	0.394	0.606	0.250	0.250
	#appleevent	0.467	0.533	0.222	0.222
	#sendkeeblermagic	0.445	0.555	0.197	0.197
	#TuesdayThoughts	0.467	0.533	0.224	0.224
	#firstdayofschool	0.226	0.774	0.113	0.113
	Xbox	0.556	0.444	0.208	0.208
	DYNAMITE CELEBRATION	0.639	0.361	0.103	0.103
	woojin	0.368	0.632	0.182	0.182

TABLE A.10: Table of Nike model results of days two, three, four and eight of September US trending topics

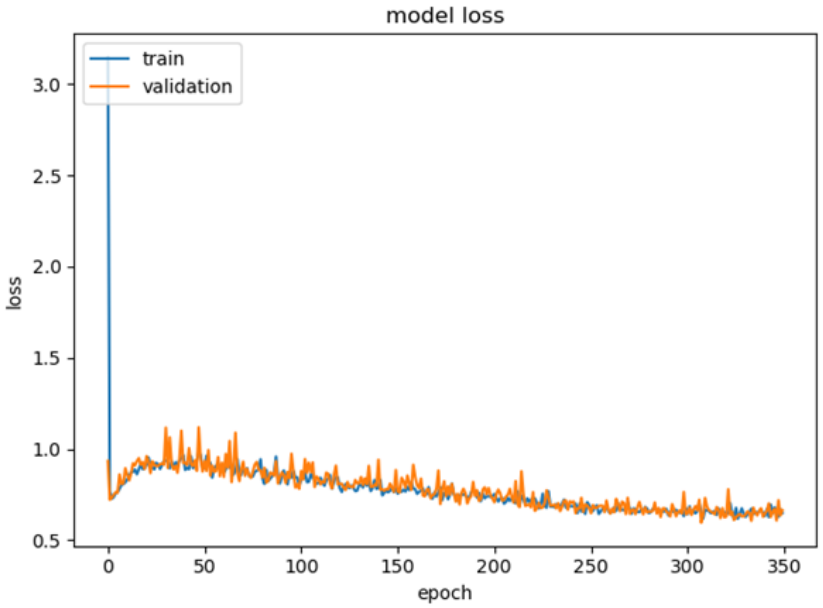


FIGURE A.9: Model loss graph on train and validation on RMC Hospital training dataset

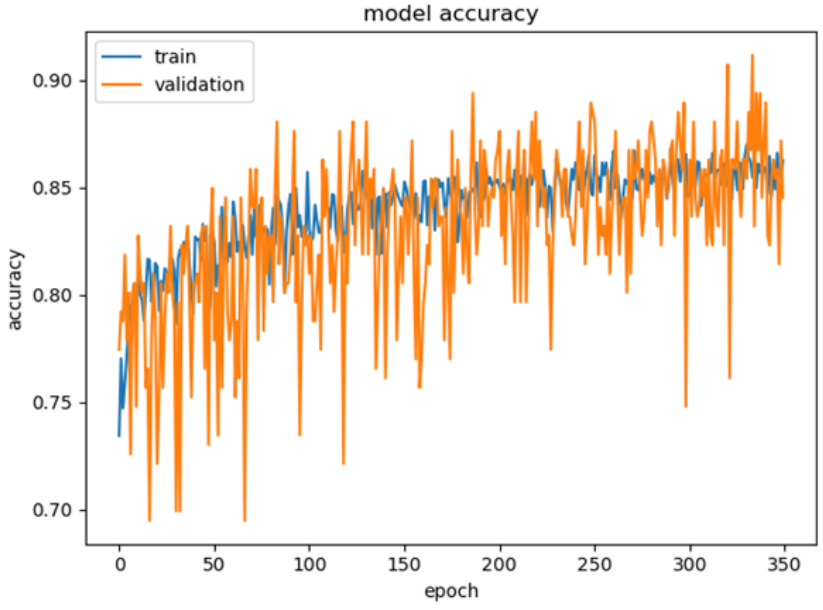


FIGURE A.10: Model accuracy graph on train and validation on RMC Hospital training dataset

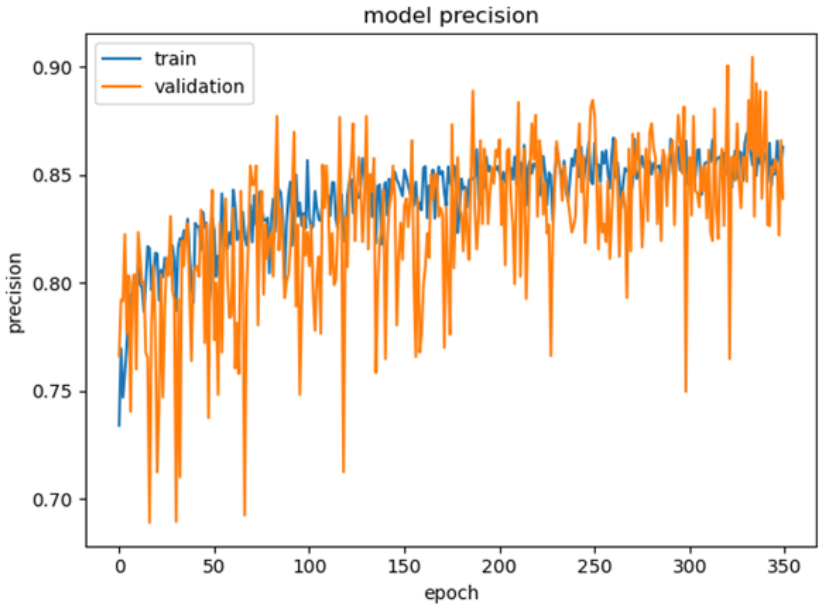


FIGURE A.11: Model precision graph on train and validation on RMC Hospital training dataset

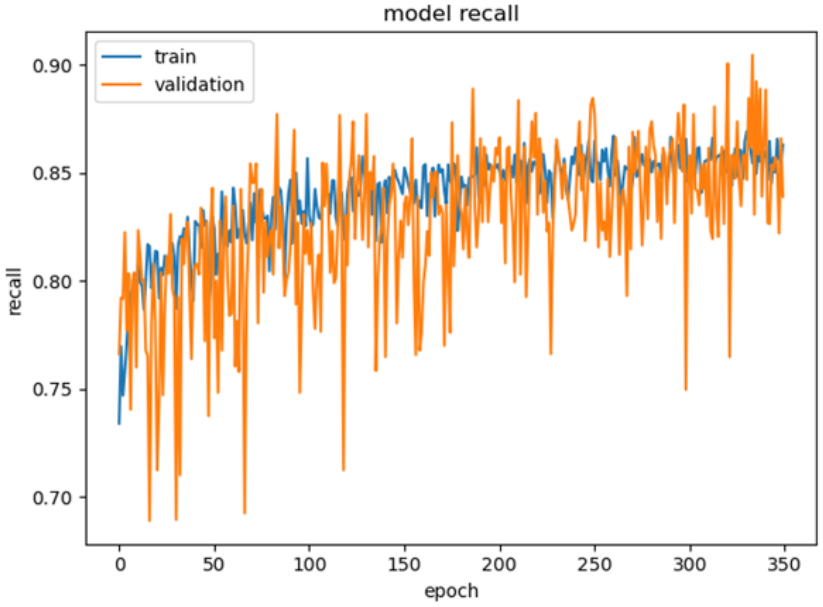


FIGURE A.12: Model recall graph on train and validation on RMC Hospital training dataset

Day	Trending Topic Name	Average - RMCH	Average - Others	Standard Deviation - RMCH	Standard Deviation - Others
02/09/2020	Kirk Cousins	0.411	0.589	0.148	0.148
	Carole Baskin	0.643	0.357	0.111	0.111
	#WednesdayWisdom	0.290	0.710	0.174	0.174
	#TheMandalorian	0.455	0.545	0.128	0.128
	David Blaine	0.624	0.376	0.073	0.073
	John Boyega	0.314	0.686	0.169	0.169
	#BillAndTed3Sweeps	0.622	0.378	0.155	0.155
	Sarah Sanders	0.353	0.647	0.107	0.107
	Keanu Reeves	0.566	0.434	0.138	0.138
	Novichok	0.271	0.729	0.134	0.134
03/09/2020	Lakers	0.430	0.570	0.215	0.215
	#ThursdayThoughts	0.154	0.846	0.096	0.096
	steven adams	0.609	0.391	0.094	0.094
	Rockets	0.546	0.454	0.129	0.129
	Dort	0.506	0.494	0.227	0.227
	#DokkanSquad	0.622	0.378	0.217	0.217
	#FadedFtLOEY	0.759	0.241	0.029	0.029
	#SexIsGreatButHaveYou	0.438	0.562	0.152	0.152
	bryson tiller	0.497	0.503	0.102	0.102
	#JusticeForDeon	0.496	0.504	0.000	0.596
04/09/2020	#DETROIT2	0.426	0.574	0.164	0.164
	Big Sean	0.662	0.338	0.164	0.164
	Matt Watson	0.477	0.523	0.193	0.193
	John McCain	0.336	0.664	0.066	0.066
	#TrumpHatesOurMilitary	0.444	0.556	0.143	0.143

	#FridayMotivation	0.259	0.741	0.154	0.154
	Wonho	0.534	0.466	0.124	0.124
	#BeyDay	0.418	0.582	0.207	0.207
	#Mulan	0.444	0.556	0.148	0.148
	Justin Bieber	0.658	0.342	0.146	0.146
08/09/2020	Odell	0.532	0.468	0.134	0.134
	Hyrule Warriors	0.469	0.531	0.215	0.215
	#StarTrekDay	0.591	0.409	0.145	0.145
	#appleevent	0.429	0.571	0.200	0.200
	#sendkeeblermagic	0.497	0.503	0.254	0.254
	#TuesdayThoughts	0.340	0.660	0.178	0.178
	#firstdayofschool	0.608	0.392	0.167	0.167
	Xbox	0.291	0.709	0.184	0.184
	DYNAMITE CELEBRATION	0.673	0.327	0.127	0.127
	woojin	0.289	0.711	0.134	0.134

TABLE A.11: Table of RMCH model results of days two, three, four and eight of September US trending topics

Day	Trending Topic Name	Average - RMCH	Average - Others	Standard Deviation - RMCH	Standard Deviation - Others
01/09/2020	#September1st	0.360	0.640	0.164	0.164
	#BackToSchool	0.425	0.575	0.261	0.261
	Ritchie	0.414	0.586	0.145	0.145
	Marcus Rashford	0.441	0.559	0.180	0.180
	Ed Sheeran	0.671	0.329	0.080	0.080
	#TuesdayMorning	0.416	0.584	0.165	0.165
	#ThisMorning	0.487	0.513	0.186	0.186
	#PepsiMaxTasteOneStop	0.179	0.821	0.076	0.076

	Jim Davidson	0.456	0.544	0.175	0.175
	Nike	0.121	0.879	0.107	0.107
17/09/2020	Thiago	0.440	0.560	0.218	0.218
	Alex Scott	0.359	0.641	0.179	0.179
	#ThursdayThoughts	0.262	0.738	0.219	0.219
	#WorldPatientSafetyDay	0.334	0.666	0.211	0.211
	#SackWhitty	0.228	0.772	0.125	0.125
	#ps5preorder	0.256	0.744	0.164	0.164
	#northeastlockdown	0.325	0.675	0.130	0.130
	Amazon UK	0.407	0.593	0.194	0.194
	Neil Warnock	0.498	0.502	0.168	0.168
	argos	0.302	0.698	0.198	0.198
22/09/2020	#XboxSeriesX	0.324	0.676	0.177	0.177
	Starmer	0.322	0.678	0.059	0.059
	#TuesdayThoughts	0.362	0.638	0.153	0.153
	#AutumnEquinox	0.479	0.521	0.220	0.220
	Britain First	0.345	0.655	0.112	0.112
	Michael Gove	0.414	0.586	0.165	0.165
	#GBBO	0.518	0.482	0.222	0.222
	#Covid_19	0.223	0.777	0.116	0.116
	Leighton Buzzard	0.616	0.384	0.085	0.085
	Bake Off	0.563	0.437	0.184	0.184
24/09/2020	#NHSCOVID19app	0.116	0.884	0.084	0.084
	#ThursdayThoughts	0.311	0.689	0.206	0.206
	Harold Evans	0.422	0.578	0.120	0.120
	#Magic	0.400	0.600	0.180	0.180
	iOS 13	0.201	0.799	0.055	0.055
	Kent	0.396	0.604	0.194	0.194
	Gigi	0.704	0.296	0.120	0.120

	#thursdayvibes	0.353	0.647	0.133	0.133
	Downloaded	0.211	0.789	0.125	0.125
	#bbcbreakfast	0.221	0.779	0.204	0.204

TABLE A.12: Table of RMCH model results of days one, seventeen, twenty-two and twenty-four of September UK trending topics

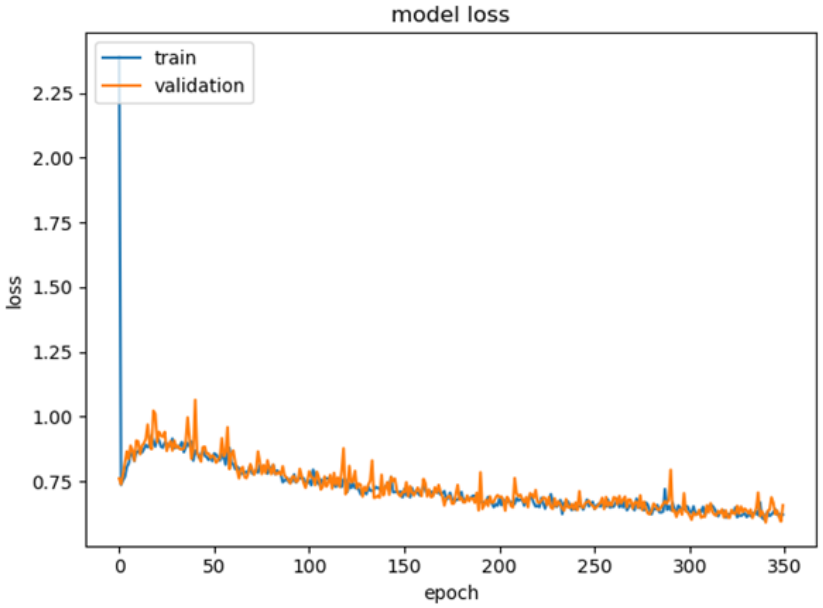


FIGURE A.13: Model loss graph on train and validation on PHU training dataset

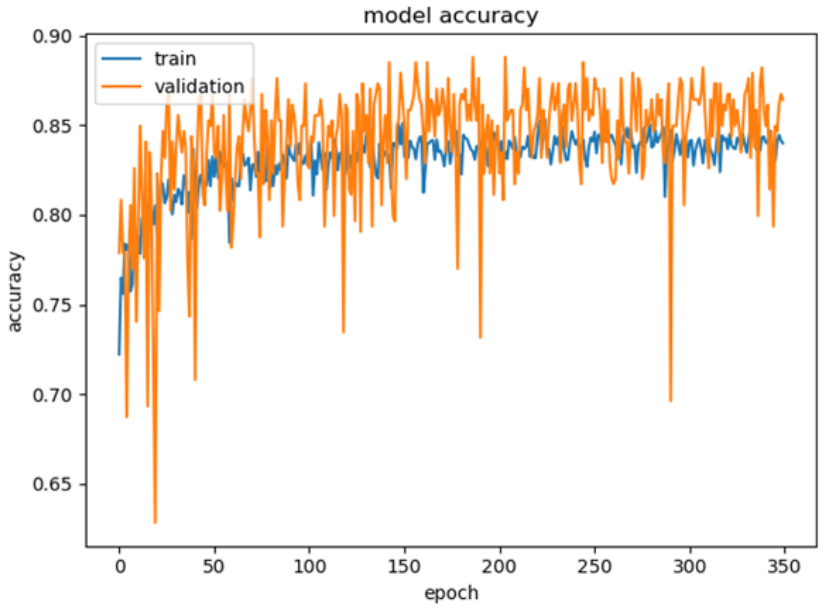


FIGURE A.14: Model accuracy graph on train and validation on PHU training dataset

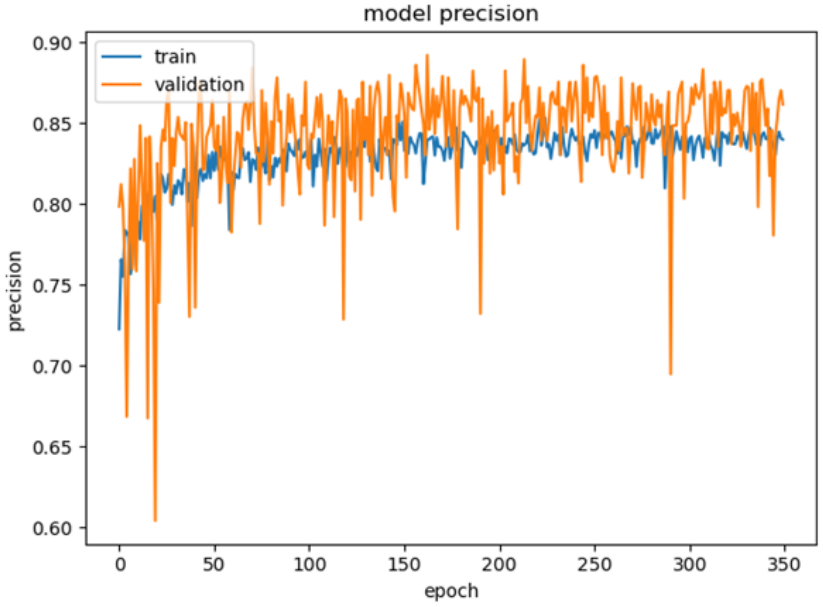


FIGURE A.15: Model precision graph on train and validation on PHU training dataset

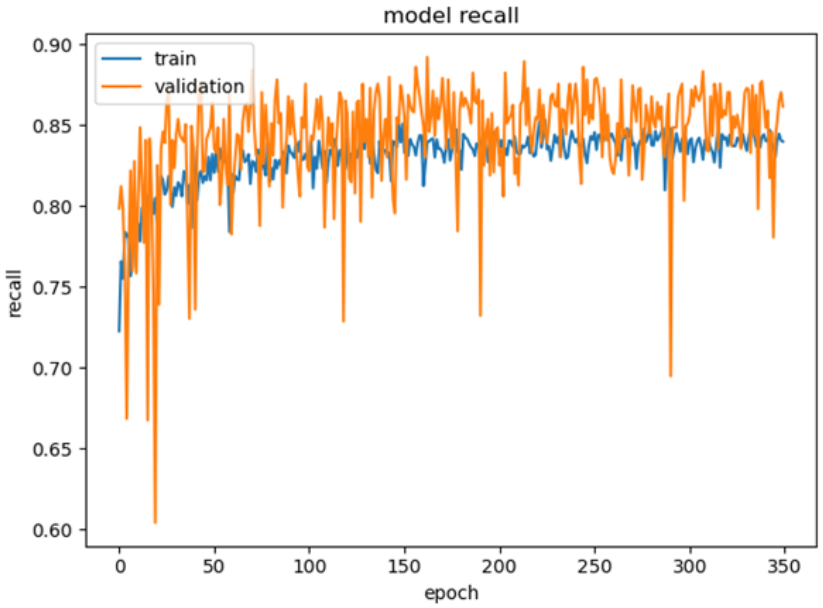


FIGURE A.16: Model recall graph on train and validation on PHU training dataset

Day	Trending Topic Name	Average - PHU	Average - Others	Standard Deviation - PHU	Standard Deviation - Others
01/09/2020	#September1st	0,486	0,514	0,152	0,152
	#BackToSchool	0.441	0.559	0.167	0.167
	Ritchie	0.457	0.543	0.184	0.184
	Marcus Rashford	0.506	0.494	0.238	0.238
	Ed Sheeran	0.123	0.877	0.020	0,020
	#TuesdayMorning	0.303	0.697	0.110	0.110
	#ThisMorning	0.375	0.625	0.129	0.129
	#PepsiMaxTasteOneStop	0.234	0.766	0.074	0.074
	Jim Davidson	0.201	0.799	0.105	0.105
	Nike	0.382	0.618	0.213	0.213
	Thiago	0.422	0.578	0.225	0.225
	Alex Scott	0.432	0.568	0.138	0.138
	#ThursdayThoughts	0.462	0.538	0.086	0.086

	#WorldPatientSafetyDay	0.816	0.184	0.168	0.168
	#SackWhitty	0.733	0.267	0.170	0.170
	#ps5preorder	0.373	0.627	0.143	0.143
	#northeastlockdown	0.722	0.278	0.139	0.139
	Amazon UK	0.369	0.631	0.169	0.169
	Neil Warnock	0.372	0.628	0.168	0.168
	argos	0.297	0.703	0.111	0.111
22/09/2020	#XboxSeriesX	0.392	0.608	0.099	0.099
	Starmer	0.610	0.390	0.189	0.189
	#TuesdayThoughts	0.458	0.542	0.132	0.132
	#AutumnEquinox	0.405	0.595	0.119	0.119
	Britain First	0.387	0.613	0.131	0.131
	Michael Gove	0.356	0.644	0.252	0.252
	#GBBO	0.484	0.516	0.154	0.154
	#Covid_19	0.689	0.311	0.193	0.193
	Leighton Buzzard	0.186	0.814	0.028	0.028
	Bake Off	0.481	0.519	0.193	0.193
24/09/2020	#NHSCOVID19app	0.683	0.317	0.154	0.154
	#ThursdayThoughts	0.390	0.610	0.173	0.173
	Harold Evans	0.180	0.820	0.102	0.102
	#Magic	0.401	0.599	0.191	0.191
	iOS 13	0.526	0.474	0.199	0.199
	Kent	0.523	0.477	0.148	0.148
	Gigi	0.137	0.863	0.036	0.036
	#thursdayvibes	0.402	0.598	0.152	0.152
	Downloaded	0.390	0.610	0.189	0.189
	#bbcbreakfast	0.513	0.487	0.157	0.157

TABLE A.13: Table of PHU model results of days one, seventeen, twenty-two and twenty-four of September UK trending topics

Day	Trending Topic Name	Average - PHU	Average - Others	Standard Deviation - PHU	Standard Deviation - Others
02/09/2020	Kirk Cousins	0.384	0.616	0.217	0.217
	Carole Baskin	0.187	0.813	0.078	0.078
	#WednesdayWisdom	0.360	0.640	0.195	0.195
	#TheMandalorian	0.327	0.673	0.184	0.184
	David Blaine	0.166	0.834	0.076	0.076
	John Boyega	0.372	0.628	0.262	0.262
	#BillAndTed3Sweeps	0.319	0.681	0.187	0.187
	Sarah Sanders	0.299	0.701	0.171	0.171
	Keanu Reeves	0.205	0.795	0.070	0.070
	Novichok	0.423	0.577	0.146	0.146
03/09/2020	Lakers	0.431	0.569	0.156	0.156
	#ThursdayThoughts	0.462	0.538	0.185	0.185
	steven adams	0.280	0.720	0.194	0.194
	Rockets	0.260	0.740	0.126	0.126
	Dort	0.292	0.708	0.128	0.128
	#DokkanSquad	0.229	0.771	0.111	0.111
	#FadedFtLOEY	0.155	0.845	0.002	0.002
	#SexIsGreatButHaveYou	0.286	0.714	0.122	0.122
	bryson tiller	0.155	0.845	0.027	0.027
	#JusticeForDeon	0.195	0.805	0.000	0.000
04/09/2020	#DETROIT2	0.456	0.544	0.190	0.190
	Big Sean	0.222	0.778	0.126	0.126
	Matt Watson	0.105	0.895	0.025	0.025
	John McCain	0.524	0.476	0.179	0.179
	#TrumpHatesOurMilitary	0.458	0.542	0.201	0.201

	#FridayMotivation	0.356	0.644	0.136	0.136
	Wonho	0.441	0.559	0.165	0.165
	#BeyDay	0.425	0.575	0.174	0.174
	#Mulan	0.346	0.654	0.176	0.176
	Justin Bieber	0.148	0.852	0.016	0.016
08/09/2020	Odell	0.223	0.777	0.074	0.074
	Hyrule Warriors	0.191	0.809	0.094	0.094
	#StarTrekDay	0.287	0.713	0.080	0.080
	#appleevent	0.380	0.620	0.169	0.169
	#sendkeeblermagic	0.284	0.716	0.151	0.151
	#TuesdayThoughts	0.280	0.720	0.103	0.103
	#firstdayofschool	0.423	0.577	0.066	0.066
	Xbox	0.375	0.625	0.135	0.135
	DYNAMITE CELEBRATION	0.203	0.797	0.067	0.067
	woojin	0.425	0.575	0.192	0.192

TABLE A.14: Table of PHU model results of days two, three, four and eight of September US trending topics

References

[1] E. Ortiz-Ospina, "The rise of social media," Available at https://ourworldindata.org/rise-of-social-media (2020/05/28).

[2] M. Roser, H. Ritchie, and E. Ortiz-Ospina, "Internet," Available at https://ourworldindata.org/internet (2020/05/28).

[3] W. Fan and M. D. Gordon, "The power of social media analytics," *Communications of the ACM*, vol. 57, no. 6, pp. 74–81, 2014.

[4] S. Smith, J. Wu, and J. Murphy, "Map: George floyd protests around the world," Available at https://www.nbcnews.com/news/world/map-george-floyd-protests-countries-worldwide-n1228391 (2020/06/18).

[5] D. L. Hoffman and M. Fodor, "Can you measure the roi of your social media marketing?" *MIT Sloan Management Review*, vol. 52, no. 1, p. 41, 2010.

[6] N. Deepa and S. Deshmukh, "Social media marketing: The next generation of business engagement," *International Journal of Management Research and Reviews*, vol. 3, no. 2, p. 2461, 2013.

[7] Twitter, Inc., "Twitter trends faqs," Available at https://help.twitter.com/en/using-twitter/twitter-trending-faqs (2020/07/30).

[8] J. M. Carrascosa, R. González, R. Cuevas, and A. Azcorra, "Are trending topics useful for marketing? visibility of trending topics vs traditional advertisement," in *Proceedings of the first ACM conference on Online social networks*, 2013, pp. 165–176.

[9] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez, "Classifying trending topics: A typology of conversation triggers on twitter," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 2461–2464. [Online]. Available: https://doi.org/10.1145/2063576.2063992

[10] R. C. Mosley Jr, "Social media analytics: Data mining applied to insurance twitter posts," in *Casualty Actuarial Society E-Forum*, vol. 2. Citeseer, 2012, p. 1.

[11] Available at https://getdaytrends.com/portugal/2020-06-10/13/ (2020/07/03).

[12] Available at https://getdaytrends.com/portugal/2020-06-11/13/ (2020/07/03).

[13] "Control portugal social media publication," Available at https://www.instagram.com/controlportugal/?hl=pt (2020/07/03).

[14] "Super book and sagres social media publication," Available at https://observador.pt/2020/02/18/contra-ao-racismo-nao-ha-rivais-sagres-e-super-bock-unem-se-contra-ao-racismo/ (2020/09/11).

[15] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. [Online]. Available: https://doi.org/10.2753/MIS0742-1222240302

[16] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[17] C. T. Carr and R. A. Hayes, "Social Media: Defining, Developing, and Divining," *Atlantic Journal of Communication*, vol. 23, no. 1, pp. 46–65, 2015.

[18] C. Dictionary, "Meaning of social media in english," Available at https://https://dictionary.cambridge.org/dictionary/english/social-media (2020/09/16).

[19] M. T. P. M. B. Tiago and J. M. C. Veríssimo, "Digital marketing and social media: Why bother?" *Business horizons*, vol. 57, no. 6, pp. 703–708, 2014.

[20] P. R. Center, "Social media fact sheet," Available at https://www.pewresearch.org/internet/fact-sheet/social-media/ (2020/09/16).

[21] G. Abeza, N. O'Reilly, and I. Reid, "Relationship marketing and social media in sport," *International Journal of Sport Communication*, vol. 6, no. 2, pp. 120–142, 2013. [Online]. Available: https://journals.humankinetics.com/view/journals/ijsc/6/2/article-p120.xml

[22] İ. E. Erdoğmuş and M. Cicek, "The impact of social media marketing on brand loyalty," *Procedia-Social and Behavioral Sciences*, vol. 58, pp. 1353–1360, 2012.

[23] Hootsuite and We Are Social, "Global social media overview," Available at https://datareportal.com/social-media-users (2020/09/21).

[24] A. Leavitt and J. J. Robinson, "The role of information visibility in network gatekeeping: Information aggregation on reddit during crisis events," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1246–1261. [Online]. Available: https://doi.org/10.1145/2998181.2998299

[25] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 591–600.

[26] Twitter, Inc., "How to tweet," Available at https://help.twitter.com/en/using-twitter/how-to-tweet (2020/09/18).

[27] Twitter, Inc., "How to retweet," Available at https://help.twitter.com/en/using-twitter/how-to-retweet (2020/09/19).

[28] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 251–258.

[29] Q. Zhu, "Classification of trending topics in twitter," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 274–277.

[30] K. Shalini, M. A. Kumar, and K. Soman, "Deep-learning-based stance detection for indian social media text," in *Emerging Research in Electronics, Computer Science and Technology*. Springer, 2019, pp. 57–67. [Online]. Available: https://doi.org/10.1007/978-981-13-5802-9_6

[31] Google, "word2vec," Available at https://code.google.com/archive/p/word2vec/ (2020/11/05).

[32] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[33] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[34] L. Liu, X. Huang, J. Xu, and Y. Song, "Oasis: Online analytic system for incivility detection and sentiment classification," in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 1098–1101.

[35] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.

[36] G. Stafford and L. L. Yu, "An evaluation of the effect of spam on twitter trending topics," in *2013 International Conference on Social Computing*, 2013, pp. 373–378.

[37] D. Antonakaki, I. Polakis, E. Athanasopoulos, S. Ioannidis, and P. Fragopoulou, "Exploiting abused trending topics to identify spam campaigns in twitter," *Social Network Analysis and Mining*, vol. 6, no. 1, p. 48, 2016. [Online]. Available: https://doi.org/10.1007/s13278-016-0354-9

[38] Q. Dang, Y. Zhou, F. Gao, and Q. Sun, "Detecting cooperative and organized spammer groups in micro-blogging community," *Data mining and knowledge discovery*, vol. 31, no. 3, pp. 573–605, 2017. [Online]. Available: https://doi.org/10.1007/s10618-016-0479-5

[39] J. Bian, Y. Yang, and T.-S. Chua, "Multimedia summarization for trending topics in microblogs," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, ser. CIKM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 1807–1812. [Online]. Available: https://doi.org/10.1145/2505515.2505652

[40] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268–1282, 2013.

[41] S. Melvin, W. Yu, P. Ju, S. Young, and W. Wang, "Event detection and summarization using phrase network," in *Machine Learning and Knowledge Discovery in Databases*, Y. Altun, K. Das, T. Mielikäinen, D. Malerba, J. Stefanowski, J. Read, M. Žitnik, M. Ceci, and S. Džeroski, Eds. Cham: Springer International Publishing, 2017, pp. 89–101.

[42] B. Peng, J. Li, J. Chen, X. Han, R. Xu, and K.-F. Wong, "Trending sentiment-topic detection on twitter," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Cham: Springer International Publishing, 2015, pp. 66–77.

[43] S. Sharma, K. Aggarwal, P. Papneja, and S. Singh, "Extraction, summariz ation and sentiment analysis of trending topics on twitter," in *2015 Eighth International Conference on Contemporary Computing (IC3)*, 2015, pp. 295–301.

[44] A. K. Singh and M. Shashi, "Vectorization of text documents for identifying unifiable news articles," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100742

[45] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[46] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014. [Online]. Available: http://arxiv.org/abs/1405.4053

[47] D. Wilkinson and M. Thelwall, "Trending twitter topics in english: An international comparison," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 8, pp. 1631–1646, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22713

[48] S. Asur, B. Huberman, G. Szabó, and C. Wang, "Trends in social media : Persistence and decay," *5th International AAAI Conference on Weblogs and Social Media*, 02 2011.

[49] I. Annamoradnejad and J. Habibi, "A comprehensive analysis of twitter trending topics," in *2019 5th International Conference on Web Research (ICWR)*, 2019, pp. 22–27.

[50] Y. Liu, W. Han, Y. Tian, X. Que, and W. Wang, "Trending topic prediction on social network," in *2013 5th IEEE International Conference on Broadband Network Multimedia Technology*, 2013, pp. 149–154.

[51] T. Althoff, D. Borth, J. Hees, and A. Dengel, "Analysis and forecasting of trending topics in online media streams," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 907–916.

[52] F. Giummolè, S. Orlando, and G. Tolomei, "Trending topics on twitter improve the prediction of google hot queries," in *2013 International Conference on Social Computing*, 2013, pp. 39–44.

[53] R. Shay and M. V. D. Horst, "Using brand equity to model roi for social media marketing," *International Journal on Media Management*, vol. 21, no. 1, pp. 24–44, 2019. [Online]. Available: https://doi.org/10.1080/14241277.2019.1590838

[54] The University of Sheffield, "List of majority native english speaking countries," Available at https://www.sheffield.ac.uk/international/english-speaking-countries (2020/03/29).

[55] University of Northampton, "Majority native english speaking countries," Available at https://www.northampton.ac.uk/international/english-language-requirements/majority-native-english-speaking-countries/ (2020/02/12).

[56] A. Madani, O. Boussaid, and D. E. Zegour, "Real-time trending topics detection and description from twitter content," *Social Network Analysis and Mining*, vol. 5, no. 1, p. 59, 2015.

[57] A. C. Arulselvi, S. Sendhilkumar, and S. Mahalakshmi, "Classification of tweets for sentiment and trend analysis," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2017, pp. 566–573.

[58] S. Giorgis, A. Rousas, J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "aueb.twitter.sentiment at SemEval-2016 task 4: A weighted ensemble of SVMs for Twitter sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 96–99. [Online]. Available: https://www.aclweb.org/anthology/S16-1012

[59] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[60] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[61] Twitter, Inc., "Twitter api v1.1," Available at https://developer.twitter.com/en/docs/twitter-api/v1 (2020/10/02).

[62] Twitter, Inc., "Authentication - oauth 1.0a," Available at https://developer.twitter.com/en/docs/authentication/oauth-1-0a (2020/08/31), 2020.

[63] Twitter, Inc., "Tools and libraries," Available at https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries (2020/09/21).

[64] GitHub, Inc., "Github," Available at https://github.com (2020/05/08).

[65] Tweepy, "Tweepy: An easy-to-use python library for accessing the twitter api." Available at https://www.tweepy.org/ (2020/03/12).

[66] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.