Departamento de Ciências e Tecnologias da Informação

Melhoria da Atratividade Internacional das Instituições de Ensino Superior através de Análise de Sentimentos

Carolina Leana Lopes dos Santos

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Sistemas Integrados de Apoio à Decisão

Orientador:
Doutor Paulo Miguel Rasquinho Ferreira Rita, Professor Catedrático,
ISCTE-IUL

Co-orientador:
Doutor João Ricardo Paulo Marques Guerreiro, Professor Auxiliar,
ISCTE-IUL

Setembro, 2016

“IT’S A TRAP!!”

Admiral Ackbar
Star Wars, Return of the Jedi (1983)

AGRADECIMENTOS

Se estamos hoje a ler esta página quer dizer eu consegui. Ao longo da escrita da dissertação passei por momentos de muito entusiasmo mas também por momentos em que quis realmente desistir. Ter chegado aqui deve-se indiscutivelmente a algumas pessoas que me acompanharam e, indubitavelmente merecem um agradecimento para a posteridade ☺.

Primeiramente gostaria de agradecer ao meu orientador e coorientador, respetivamente. Obrigada Professor Paulo Rita pela disponibilidade, partilha de conhecimento e todas as suas sugestões. Obrigada Professor João Guerreiro pelos conselhos, pelo incentivo, por me ter ajudado sempre em todas as fases deste trabalho. Mas, obrigada principalmente pela compreensão e por me ter mandado descansar quando percebeu que eu precisava.

Quero agradecer à professora Elsa Cardoso que desde o final da minha licenciatura representa sem dúvida uma peça chave no meu percurso académico. Obrigada não só pela partilha da sua experiência, mas também pela disponibilidade e muitas palavras de conforto e incentivo que ao longo destes anos recebi.

Obrigada a todos os meus colegas de mestrado com quem passei tão bons momentos. Obrigada Águeda, pois sei que compreendias bastante bem o meu caminho e tantas vezes me motivaste não só com palavras, mas com a garra com que levas a tua vida.

Obrigada Marta por toda a amizade desde sempre. Conhecemo-nos nas primeiras semanas de ISCTE-IUL e desde aí que estás sempre, sempre ao pé de mim. Com certeza fazes parte da minha família.

Obrigada André pela amizade verdadeira. Por toda a força, por teres me teres apoiado em tantos momentos de desespero e me teres proporcionado tantos momentos de parvoíce e diversão.

Obrigada Andrew por me guiares, por teres sido sempre paciente até mesmo nos momentos de impaciência. Obrigada por me conseguires ler tão profundamente e por veres tão bem através de mim.

Obrigada pai, mãe e irmã. A vocês tenho que agradecer tudo. Obrigada por me terem feito chegar até aqui. Obrigada pela paciência, pela compreensão, pelo apoio. Mas principalmente, obrigada pelo amor incondicional ao longo dos meus 25 anos. Obrigada por serem a base da pessoa que sou hoje. A vocês dedico o meu trabalho.

RESUMO

A Europa tem investido em várias políticas para aumentar a atratividade do ensino superior europeu devido aos benefícios que a mobilidade internacional de alunos traz, não só às instituições de ensino superior (IES) mas também aos países anfitriões. No entanto as experiências partilhadas na Internet por *ex*-alunos podem influenciar fortemente a reputação das instituições e consequentemente, a escolha do local de estudo de futuros alunos internacionais. Desse modo, analisar a informação partilhada no meio *online* é crítico para que as IES compreendam melhor as opiniões dos alunos gerindo a sua reputação e indo de encontro àquilo que realmente é atrativo para os mesmos.

Contudo, atualmente a quantidade de informação partilhada na Internet é enormíssima sendo impraticável ler todos os comentários no meio *online*. Assim, técnicas automatizadas capazes de lidar com grandes conjuntos de dados textuais são fundamentais para obtenção de conhecimento mais preciso. No entanto, não foram encontradas investigações que recorressem a estas técnicas para potenciar a atratividade das IES, neste contexto.

Na presente dissertação, foram analisadas experiências de alunos internacionais publicadas no meio *online,* com vista a compreender os principais aspetos que influenciavam a sua satisfação face a uma *business-school* da Europa. Para tal recorreu-se a técnicas de *text mining,* análise de sentimentos e *topic modelling.* Os resultados demonstraram que a satisfação dos alunos face a uma IES é influenciada por vários critérios relacionados com fatores financeiros, motivações pessoais, a imagem do país e a imagem da IES.

Palavras-chave: *Text mining, Análise de Sentimentos, Topic Modelling, Ensino Superior Europeu, Mobilidade internacional de Alunos, Sistemas de Apoio à Decisão*

ABSTRACT

Europe have been spending efforts in several policies to increase the attractiveness of the European higher education due to the benefits provided by international mobility of students provide to the higher education institutions (HEI) but also to the host countries. However, experiences written in the Internet by former students can strongly influence the institutions' reputation and, therefore, influence the choice of future students regarding the HEI they will attend. Thus, analysing the information shared online is critical for HEIS to better understand the students' opinions, leading HEI to better manage their reputation and meeting what is really attractive to international students.

Nevertheless, the amount of information on social media nowadays is so large that it is impractical to read all the reviews. Thereby, automated techniques able to handle large sets of textual data are essential for extracting accurate knowledge. Nonetheless, we haven't found any investigation that resorts to these techniques to enhance the attractiveness of HEI in this context.

On the present dissertation, online reviews of international students about their experience towards a business-school in Europe were studied in order to understand the main drivers influencing students' satisfaction. For that purpose text mining, sentiment analysis and topic modelling techniques were used. Results showed that students' satisfaction regarding a business-school is influenced by several criteria related to financial factors, personal motivations, image of the country and image of the HEI.

Keywords: Text mining, Sentiment Analysis, Topic Modelling, European Higher Education, international mobility of students, Decision Systems Support

ÍNDICE

1 Introdução 1
1.1 Enquadramento 1
1.2 Problema 3
1.3 Objetivos 4
1.4 Motivação 4
1.5 Âmbito 5
1.6 Metodologia 6
1.7 Estrutura 6
2 Estado da arte 7
2.1 Mobilidade internacional no ensino superior 7
2.1.1 Natureza do serviço de educação 8
2.1.2 Fatores que influenciam a escolha do local de estudo 9
2.2 Análise de sentimentos de opiniões de estudantes (*online reviews*) 15
2.3 Text Mining 16
2.3.1 Análise de Sentimentos 18
2.3.2 Abordagens de Análise de Sentimentos 20
2.3.3 Níveis de Análise de sentimentos 23
2.3.4 *Web services* 24
2.3.5 *Text mining* e análise de sentimentos no setor da educação 26
3 Descoberta de conhecimento em texto 28
3.1 Crisp-dm 28
3.2 Extração 29
3.3 Preparação dos dados 29
3.3.1 Tokenization 29
3.3.2 Normalização dos termos 30
3.3.3 Part-of-Speech 31
3.3.4 Tratamento da negação 31
3.3.5 Redução de dimensionalidade 32
3.3.6 Document-by-term matrix 32
3.4 Métodos não-supervisionados e supervisionados 33
3.4.1 *Métodos não-supervisionados* 33
3.4.2 *Métodos* Supervisionados 36
4 Metodologia 38
4.1 Compreensão do negócio 38
4.2 Compreensão dos dados 39
4.2.1 Extração dos dados 39
4.2.2 Análise da qualidade dos dados 40
4.2.3 Análise descritiva dos dados 42

4.3 Preparação dos dados 45
4.3.1 Processo 1 (P1) - Construção dos tópicos 48
4.3.2 Processo 2 (P2) - Construção do input com matriz frequências TF-IDF 52
4.3.3 Processo 3 - Construção de matriz com valores de sentimento (Semantria) 55
4.4 Modelação 56
4.5 Avaliação 58
4.5.1 Avaliação do *input 1* – Matriz TF-IDF 60
4.5.2 Avaliação do *input* 2 – Matriz Sentimentos 61
5 Resultados (Implementação) 63
5.1 Teste da Hipótese 3 65
5.2 Teste da Hipótese 2 69
5.3 Teste da Hipótese 1 73
6 Conclusões, Limitações e trabalho futuro 77
Bibliografia 81
Anexos 88
Anexo A - Conjunto de dados extraído 88
Anexo B - Resultados teste Dunn Bonferroni 94
Anexo C - Relação de monotonicidade 95

ÍNDICE DE FIGURAS

Figura 1 Fatores que influenciam a escolha de um local de estudo internacionalmente (sistematização própria) 14
Figura 2 Síntese das abordagens de Análise de Sentimentos (Fonte: adaptado de Medhat et al. (2014)) 20
Figura 3 Processo de funcionamento do Semantria (fonte: adaptado de Lawrence (2014)) 25
Figura 4 Fases do processo CRISP-DM (fonte: P.Chapman (2000)) 28
Figura 5 Exemplo de documento transformado em biterms (Fonte: Cheng et al. (2014)) 35
Figura 6 Probabilidade condicionada dos biterms (Fonte: Cheng et al., (2014)) 35
Figura 7 Esquema do funcionamento do LDA (a) e do BTM (b) (Fonte: adaptado de(Cheng et al., 2014)) 36
Figura 8 Hiperplano SVM (Fonte: Cortes e Vapnik (1995)) Figura 9 Margens SVM (Fonte: Pontil e Verri (1998)) 37
Figura 10 Excerto de uma review do iagora 38
Figura 11 Processo de extração dos dados 40
Figura 12 Exemplo de identificação de review 41
Figura 13 Distribuição de reviews por país de destino 43
Figura 14 Distribuição de frequências OV_Stars 44
Figura 15 Exemplos de substituições pelo QDA Miner 45
Figura 16 Esquema da preparação de dados 46
Figura 17 Processos da preparação de dados 47
Figura 18 Primeira DTM do P1 49
Figura 19 Segunda DTM do P1 49
Figura 20 Wordcloud termos mais frequentes do P1 50
Figura 21 Log-Likelihood 51
Figura 22 Perplexity 51
Figura 23 Tópicos detetados através do CTM 51
Figura 24 Tópicos detetados através do BTM 52
Figura 25 Tratamento da negação 53
Figura 26 Tratamento de pontuação 53
Figura 27 DTM depois stemming do P2 53
Figura 28 DTM final segundo processo do P2 54
Figura 29 Wordcloud top 65 termos do P2 54
Figura 30 Síntese dos conjuntos de dados 56
Figura 31 Distribuição da variável stars 57
Figura 32 Distribuição da variável starsbinaria 57
Figura 33 Distribuição variável stars - partição de treino balanceada 57
Figura 34 Distribuição variável starsbinaria - partição de treino balanceada 57
Figura 35 Testes de normalidade OV_Stars 70
Figura 36 One-way Anova - Teste Levene não paramétrico 70
Figura 37 Resultados do teste Kruskal Wallis 71

Figura 38 Mediana dos tópicos.....72
Figura 39 Wordcloud de temas detetados pelo Semantria.....72
Figura 40 Wordcloud de entidades detetadas pelo Semantria.....73
Figura 41 Síntese de correlações com os fatores.....77

ÍNDICE DE TABELAS

Tabela 1 Sumário das fatores que influenciam a escolha de um local de estudo internacionalmente (sistematização própria) 15
Tabela 2 Scores de sentimento do Semantria 26
Tabela 3 Term-by-document matrix 32
Tabela 4 Quantidade de atributos por tipo de estrutura 40
Tabela 5 Quantidade de comentários por atributo não estruturado 42
Tabela 6 Top 7 países de origem 42
Tabela 7 Top 7 de quantidade de reviews por universidade 43
Tabela 8 Quantidade de comentários por atributo não estruturado – conjunto de dados final 43
Tabela 9 Estatísticas descritivas do atributo Experiência 43
Tabela 10 Estatísticas descritivas das estrelas 44
Tabela 11 Estatísticas descritivas das 15 variáveis "overall" 45
Tabela 12 Exemplo de tratamento 48
Tabela 13 Excerto de comentário classificado pelo POS 48
Tabela 14 Correlações dos 6 termos mais frequentes do P1 50
Tabela 15 Correlações termos mais frequentes do P2 54
Tabela 16 Dataset input 1 55
Tabela 17 Dataset input 2 56
Tabela 18 Quantidade de variáveis por experiência 58
Tabela 19 Parametrização dos algoritmos 58
Tabela 20 Matriz de confusão 59
Tabela 21 Resultados input 1 60
Tabela 22 Resultados input 2 sem feature selection 61
Tabela 23 Resultados input 2 com feature selection 62
Tabela 24 Quantidade de reviews por tópico 65
Tabela 25 Síntese comparativa da revisão literária com os tópicos do BTM 67
Tabela 26 Quantidade de reviews por tópico sem o Student life 69
Tabela 27 Significâncias ajustadas do post hoc teste Dunn Bonferroni 71
Tabela 28 Resultados do teste Kendall-tau 74

ACRÓNIMOS

AACSB	*Association to Advance Collegiate Schools of Business*
AS	Análise de Sentimentos
BTM	*Biterm Topic Model*
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
CTM	*Correlated Topic Model*
DSR	*Design Science Research*
DTM	*Document-by-Term Matrix*
ECTS	Sistema Europeu de Transferência e Acumulação de Créditos
EDM	*Educational Data Mining*
EFMD	*European Foundation for Management Development*
EQUIS	*EFMD Quality Improvement System*
IDF	*Inverse Document Frequency*
IES	Instituição/Instituições de Ensino Superior
LBD	*Literature-Based Discovery*
LDA	*Latent Dirchlet Allocation*
MOOC	*Massive Open Online Course*
OCDE	Organização para a Cooperação e Desenvolvimento Económico
PCN	*Previous Current Next*
PNL	Processamento Natural da Língua
POS	Part-Of-Speech
SVM	*Support Vector Machines*
TF-IDF	*Term-Frequency-Inverse Document Frequency*
TM	*Text Mining*

1 INTRODUÇÃO

1.1 ENQUADRAMENTO

No final do século XX, as atividades económicas, políticas e sociais no mundo começaram a expandir-se internacionalmente cada vez mais, contribuindo para o fenómeno da globalização (Altbach & Knight, 2007). No setor da educação, o ensino terciário recebeu um grande investimento com o intuito de se criar uma "sociedade de conhecimento" pelo que as instituições de ensino superior (IES) acrescentaram a internacionalização como uma dimensão aos seus processos fazendo com que nas últimas três décadas as suas atividades internacionais aumentassem fortemente. Em 2014, a Organização para a Cooperação e Desenvolvimento Económico (OCDE) reportou que a quantidade de alunos do ensino superior a estudar num país estrangeiro aumentou drasticamente desde 1975 a 2012, sendo que em 1975 a quantidade de estudantes a estudar no estrangeiro foi de 0.8 milhões ao passo que em 2012 a quantidade foi de 4.5 milhões (OCDE, 2014).

A globalização contribuiu para uma imensa liberdade de escolha no sector da educação (Mazzarol, 1998; Šontaitė-Petkevičienė, 2013) como tal, os estudantes passaram a ter acesso a um vasto leque internacional de IES onde podem complementar os seus estudos, o que fez nascer o conceito de aluno internacional. A UNESCO *institute statistics* define como aluno internacional o "aluno que tenha atravessado uma fronteira nacional ou territorial para fins de educação e que encontra-se agora matriculado fora do seu país de origem"[1].

A mobilidade de alunos para realização de cursos fora do seu país de origem é o meio mais poderoso de internacionalização das IES, sendo em muitas delas uma das suas maiores fontes de rendimentos (Comissão Europeia, 2013). Em alguns países, chegam mesmo a ser exigidas propinas mais elevadas a alunos internacionais ao passo que, noutros opta-se por diminuir ou suprimir propinas com vista a aumentar a atratividade (Beine, Noël, & Ragot, 2014). Contudo, independentemente do pagamento ou não de propinas, os alunos internacionais proporcionam fortes benefícios económicos, não só nas instituições mas também nos próprios países de acolhimento (Comissão Europeia, 2013). Outro benefício diz respeito aos alunos que passam por uma experiência internacional e, muitas vezes, permanecem no país anfitrião após os seus estudos, aumentando fortemente a quantidade de trabalhadores altamente qualificados fomentando, assim, o crescimento do país (Beine et al., 2014; González, Mesanza, & Mariel, 2011). Adicionalmente, facultar estudos a alunos estrangeiros é um meio de difundir os valores culturais e políticos dos países anfitriões (Beine et al., 2014; González et al., 2011).

Na Europa, a internacionalização académica é desde há muito uma das principais preocupações existentes nas políticas do ensino superior. Em 1999 entrou em vigor a Declaração de Bolonha (1999) com o propósito de aumentar a competitividade e a atratividade do sistema de ensino superior europeu. O Processo de Bolonha sublinha a importância de se criar um sistema de ensino europeu que funcione de forma integrada dando enfase a vários aspetos,

[1] http://glossary.uis.unesco.org/glossary/en/home

entre eles a promoção da mobilidade no espaço europeu. Além do Processo de Bolonha, programas de mobilidade internacional como o Erasmus, Tempus, Marie Curie, entre outros e o Sistema Europeu de Transferência e Acumulação de Créditos (ECTS) contribuíram para uma forte internacionalização intraeuropeia (Comissão Europeia, 2013). Em 2001, foi reforçada a necessidade de tornar atrativo o ensino superior europeu bem como apontada a importância do envolvimento dos estudantes na organização interna das universidades (Comissao Europeia, 2001). A Estratégia de Lisboa (2000) foi outro veículo na promoção da internacionalização do ensino superior europeu, pelo que tinha como objetivo até 2010 tornar a Europa "a economia do conhecimento mais competitiva e mais dinâmica do mundo, capaz de um crescimento económico duradouro acompanhado de uma melhoria quantitativa e qualitativa do emprego e de maior coesão social". Em 2010, a Estratégia de Lisboa de 2000 foi atualizada para a atual Estratégia Europeia com objetivos para a década seguinte. A Estratégia Europeia promove novamente iniciativas com o intuito de aumentar a atratividade das instituições do ensino superior europeu até 2020.

Perante todos os esforços, a Europa tornou-se um dos destinos mais atrativos para alunos internacionais acolhendo atualmente 45% de alunos em mobilidade internacional (Comissão Europeia, 2013). No entanto, a concorrência no Médio Oriente, na Ásia e na América Latina encontra-se a crescer exponencialmente pelo que, os fluxos de mobilidade internacional estão a alterar-se rapidamente. A fim de conseguir vingar neste contexto de competitividade, a Europa necessita de aumentar ainda mais a sua atratividade promovendo eficazmente a mobilidade internacional (Comissão Europeia, 2013).

Até 2013, apenas meios tradicionais tais como feiras internacionais e *sites online* tinham sido utilizados para promover a atratividade da Europa no contexto do ensino superior (Comissão Europeia, 2013). Em 2013, a Comissão Europeia (2013) reconheceu a importância de apoiar o desenvolvimento de um instrumento mais transparente e orientado ao utilizador internacional que permitisse um maior ajustamento às necessidades dos mesmos e assim conseguisse atrair um maior número de alunos, o U-Multirank[2]. O U-Multirank nasce da importância da classificação das IES como suporte à tomada de decisão do aluno internacional no momento de escolha do seu local de estudo. Os métodos de classificação tradicionais das IES são os *rankings*, contudo estes são essencialmente focados na investigação pelo que, acabam por ter um forte impacto na reputação da instituição sem ter em conta o que é realmente atrativo para o aluno internacional (Comissão Europeia, 2013). Desse modo, o U-Multirank é um sistema baseado em informação proveniente das IES e questionários a alunos que estudaram nas mesmas, que permite explorar IES detalhadamente por pesquisa autónoma.

Num contexto de mobilidade internacional, a Internet tem um papel especialmente importante uma vez que os alunos não conseguem deslocar-se às instituições para as conhecerem mais profundamente (Gomes & Murphy, 2003). Neste sentido, o U-Multirank presta já um papel importante como auxilio ao aluno internacional. Contudo, os alunos internacionais recorrem frequentemente a outro tipo de fontes independentes *online* onde têm o intuito de obter

[2] http://www.umultirank.org/

informações escritas por ex-alunos pois acreditam que estas são mais confiáveis e detalhadas do que as das próprias instituições (Gomes & Murphy, 2003). Os meios *online* encontram-se repletos de informação sobre as experiências passadas pelos consumidores em relação a um serviço e, os alunos como consumidores do serviço de educação não são exceção. Os alunos caracterizam-se por ser cada vez mais verbais exprimindo frequentemente as suas opiniões face ao que pensam sobre o setor da educação (Binsardi & Ekwulugo, 2003). Assim, quando um aluno pesquisa sobre uma instituição educacional na Internet facilmente se depara com publicações escritas por ex-alunos partilhando a sua experiência.

No entanto, a partilha de experiencias *online,* quando não monitorizada cuidadosamente, representa uma ameaça para as instituições pois pode influenciar fortemente a reputação das mesmas (Bourke, 2000; Chun, 2005; Herbig & Milewicz, 1993). A partilha de experiências negativas na Internet tem um impacto muito grande para as organizações pois, o que é lido por potenciais consumidores afeta a impressão que se cria sobre a organização (Walsh, Mitchell, Jackson, & Beatty, 2009). No caso da tomada de decisão de alunos internacionais esta questão não é menos importante, podendo levar a que estes optem por selecionar uma IES ao invés de outra, levando consequentemente a uma diminuição de candidaturas desta última (Kotler & Fox, 1995). Neste sentido, os *social media* caracterizam-se por ser uma mina de informação para as organizações (Mostafa, 2013), neste caso para as IES, permitindo não só analisar o que pensam e sentem os alunos mas também, avaliar o seu desempenho interno percebendo como estão a corresponder às expectativas dos alunos. Assim, a obtenção de conhecimento sobre as opiniões, contribui para que as instituições consigam aprimorar as suas tomadas de decisão indo de encontro ao seu publico-alvo (Pang & Lee, 2008). Torna-se então crítico adaptar a gestão da reputação das IES de maneira a contemplar os ambientes *online* em que os alunos interagem com o propósito de melhorar a reputação (Jones, Temperley, & Lima, 2009) e, consequentemente, aumentar a atratividade.

1.2 PROBLEMA

A informação partilhada no meio *online* é imensamente rica contendo informação objetiva sobre a opinião dos consumidores em relação aos aspetos das experiências por que passaram. Adicionalmente, como mencionado anteriormente, a literatura constatou já a importância da informação disponível em fontes independentes *online* escritas por ex-alunos num contexto de mobilidade internacional (Gomes & Murphy, 2003). Contudo, ao longo da revisão literária, quase não foram encontrados estudos que analisassem a opinião dos alunos expressa no meio *online* e, nenhum dos estudos encontrados foi num âmbito de promover a atratividade internacional. Tal sugere que as IES estão a deixar de lado uma quantidade de informação bastante grande, que se encontra presente nos meios sociais *online,* passível de transmitir conhecimento útil e objetivo sobre o segmento dos alunos internacionais. Ao ignorar esta informação as IES podem estar a perder conhecimento sobre os aspetos que são realmente valorizados pelos alunos permitindo que a sua reputação seja fortemente influenciada por este meio e, consequentemente, a sua atratividade internacional seja prejudicada.

Posto isto, a problemática de investigação prende-se com as instituições não estarem a fazer tudo ao seu alcance para melhorar a sua atratividade internacional. Sendo que a questão que se coloca é a seguinte:

"Quais os principais termos e fatores a influenciar a atratividade de uma instituição de ensino superior, segundo as opiniões de alunos internacionais partilhadas no meio online?"

1.3 OBJETIVOS

A presente investigação pretende melhorar a compreensão sobre a mobilidade internacional para *business-schools* da Europa tendo como base a utilização de técnicas automatizadas de análise de conteúdo não estruturado (*text mining*) e experiências partilhadas por ex-alunos internacionais no meio *online.*

Para tal, numa primeira fase, pretende recorrer-se a comentários publicados no *website iagora.com*[3] e, através de técnicas de *text mining,* detetar-se os principais aspetos que os alunos referem sobre a sua experiência na *business-school* onde estudaram.

Numa fase subsequente, consoante os aspetos detetados, pretende analisar-se o sentimento de cada comentário face a cada um dos aspetos e, correlacionar os sentimentos associados a cada aspeto com a classificação dada pelo aluno à experiência na *business-school* onde estudou.

Pretende contribuir-se com maior conhecimento científico através da deteção dos principais determinantes que influenciam a satisfação dos alunos internacionais face a uma IES, bem como fornecer contribuições práticas aos gestores das IES através da sugestão de recomendações sobre em que é que devem investir esforços para promover a atratividade das *business-schools* no segmento da mobilidade internacional.

1.4 MOTIVAÇÃO

A Internet é atualmente uma plataforma com uma quantidade de partilha de opiniões abismal sendo que, quando se pretende perceber a opinião sobre algum serviço, produto ou entidade deparamo-nos com milhares de comentários. Face à quantidade de informação existente, perceber a opinião geral dos consumidores torna-se bastante moroso e, caso apenas sejam lidas algumas opiniões, muito possivelmente a conclusão que se retiraria seria incompleta e tendenciosa (Zhang, Ye, Zhang, & Li, 2011).

Posto isto, uma das motivações desta investigação diz respeito aos meios existentes para lidar com tamanha quantidade de dados. Na verdade, considera-se que atualmente é o momento ideal para se analisarem estes dados, uma vez que existem ferramentas capazes de lidar com altas quantidades de informação de forma muito mais automatizada e bastante mais eficiente. A análise de sentimentos (AS) é uma área concentrada na análise de opiniões de consumidores de forma automatizada com o propósito de extrair conhecimento das mesmas. As técnicas automatizadas são especialmente importantes por terem fortes vantagens quando comparadas às típicas abordagens de marketing (Yamanishi & Li, 2002). Estas últimas caracterizam-se por

[3] http://www.*iagora*.com/studies/

recorrerem normalmente a análises de questionários e por serem bastante morosas devido ao volume de informação (Yamanishi & Li, 2002). Adicionalmente, possuem o problema da análise feita por pessoas ser baseada na intuição o que faz com que seja altamente subjetiva e consequentemente pouco confiável (Yamanishi & Li, 2002; Zhang et al., 2011) pelo que, pretende tirar-se partido da investigação já existente na AS para o âmbito da presente dissertação.

Uma segunda motivação prende-se com a capacidade de demonstrar como é que as instituições educacionais podem aumentar conhecimento sobre os alunos internacionais de maneira a melhorar eficazmente a sua atratividade, num momento de alta competitividade e em que é dada tanta importância ao mercado internacional. A literatura defende que o marketing das instituições é um dos meios mais eficientes para atrair alunos internacionais (Bourke, 2000) e que a análise das redes sociais é extremamente valiosa para tomada de decisões mais acertadas (Mostafa, 2013). Desse modo, acredita-se que a aplicação de técnicas automatizadas para análise de comentários *online* de alunos internacionais é uma mais-valia para as IES otimizarem as suas tomadas de decisão e melhorarem o seu *marketing*.

1.5 ÂMBITO

Havendo a possibilidade de uma IES poder ser considerada mais atrativa por um aluno dependendo da área que estuda, o âmbito da dissertação é limitado a alunos internacionais que estudaram em *business-schools* da Europa. A fonte de onde foram extraídos os dados diz respeito ao *website iagora.com* onde alunos, que tenham feito algum tipo de mobilidade, escrevem partilhando a sua experiência relativamente a aspetos associados à IES onde decorreram os seus estudos. No *iagora*, cada instituição é avaliada qualitativa e quantitativamente tendo em conta a opinião do aluno segundo diversos fatores nomeadamente: habitação, vida estudantil, vertente académica, o idioma do país anfitrião bem como o idioma de instrução, despesas e considerações finais.

A fonte de dados contempla um vasto conjunto de faculdades distribuídas em vários países do mundo porém, uma vez que o âmbito é limitado a *business-schools* da Europa, foi necessário proceder-se a uma seleção de instituições a serem contempladas no conjunto de dados. Para a escolha das instituições de ensino, teve-se em conta a lista de acreditações da *Association to Advance Collegiate Schools of Business* (AACSB) uma vez que a AACSB nasceu em 1916 com o intuito de acreditar *business-schools* em programas de licenciatura, mestrados e doutoramentos caracterizando-se atualmente por ser uma das mais reconhecidas associações de acreditações nesta área. Existe também a EFMD Quality Improvement System (EQUIS), um sistema de acreditação europeia organizado pela European Foundation for Management Development (EFMD), no entanto devido à AACSB ser mais antiga e igualmente bastante divulgada na Europa considerou-se que seria uma boa opção.

Posto isto, o conjunto de dados final tem 1925 experiências internacionais de alunos distribuídas por 65 *business-schools* da Europa com acreditação AACSB até Maio de 2015.

1.6 METODOLOGIA

Com o propósito de guiar o desenvolvimento da investigação foi adotada a metodologia *Design Science Research* (DSR) proposta por Peffers et al. (2007). Esta metodologia é orientada para criação de artefactos na área dos sistemas de informação passando por seis etapas nomeadamente: a identificação do problema e da motivação; definição dos objetivos da solução; desenho e desenvolvimento; demonstração; avaliação; e, comunicação. Os autores defendem que o investigador é livre de poder executar o processo segundo uma ordem diferente sugerindo mesmo quatro possíveis pontos de entrada nomeadamente: centrado no problema; centrado nos objetivos; centrado no *design* e desenvolvimento; centrado num pedido de um cliente.

O ponto de entrada desta investigação é centrado no problema da melhoria da atratividade internacional de *business-schools* da Europa. Por sua vez, no que diz respeito às seis etapas, estas definem-se da seguinte forma:

- **Problema e motivação:** Apresentados nas secções 1.2 Problema e 1.4 Motivação.
- **Objetivos da solução:** Apresentados na secção 1.3 Objetivo;
- **Desenho e desenvolvimento:** Formulação de hipóteses de teste de acordo com revisão de literatura; Analisar conteúdo textual a partir da aplicação de modelos supervisionados e não supervisionados bem como análises estatísticas;
- **Demonstração:** Aplicação de técnicas de *text mining* e análise de sentimentos sobre comentários de alunos que passaram por uma experiência de mobilidade internacional numa *business-school* da Europa, extraídos do *website iagora*.com.
- **Avaliação:** Verificar se as hipóteses formuladas ao longo da literatura são ou não confirmadas tendo em conta os resultados.
- **Comunicação:** O artefacto e a sua importância será comunicado a partir da presente dissertação e de um artigo científico.

1.7 ESTRUTURA

A presente dissertação inicia-se com o enquadramento, identificação de problema, objetivos e motivações que levaram ao seu desenvolvimento. De seguida, dado que esta dissertação cruza a área tecnológica com uma área de negócio associada ao marketing, a secção 2 é composta por duas vertentes: a revisão literária referente à problemática de investigação acerca da atratividade da mobilidade internacional do ensino superior na Europa e, uma vertente associada ao *text mining* e análise de sentimentos. A secção 3, por sua vez, diz respeito aos métodos tradicionais utilizados para a descoberta de conhecimento em texto e servirá como base técnica para a secção seguinte. A secção 4 compreende toda a metodologia, isto é, toda a fase de desenvolvimento prática da investigação. Na secção 5 é feita a análise dos resultados obtidos relacionando com o conhecimento obtido ao longo da revisão literária. Por fim, na secção 6 são apresentadas as conclusões e formuladas algumas recomendações. Nesta última secção são ainda apresentadas as limitações e trabalho futuro da investigação.

2 ESTADO DA ARTE

2.1 MOBILIDADE INTERNACIONAL NO ENSINO SUPERIOR

Face ao ambiente de alta competitividade e diversidade de escolha, fortes estratégias de *marketing* são imperativas para que as instituições sejam atrativas internacionalmente (Bourke, 2000; Choudaha & Chang, 2012). Uma forte vantagem competitiva através do desenvolvimento de um forte posicionamento no mercado e uma imagem distinta é fundamental caso as instituições de ensino superior pretendam sobreviver e ter sucesso (Cubillo, Sánchez, & Cerviño, 2006; Kotler & Fox, 1995; Nicholls, Harris, Morgan, Clarke, & Sims, 1995)

De maneira a criar um forte posicionamento é necessário que, numa primeira fase, as instituições reconheçam a importância de segmentar os seus diferentes tipos de alunos tendo em conta que nem todos têm as mesmas necessidades e expectativas (Kotler & Fox, 1995; Nicholls et al., 1995). Por exemplo, as necessidades e expectativas de alunos locais possivelmente não serão as mesmas do que as de alunos internacionais. Ao compreender os diferentes segmentos, as instituições devem, numa fase subsequente, unir esforços de maneira a desenvolverem estratégias que lhes permitam criar uma forte posição dentro dos segmentos a que se quiserem dedicar. Num contexto de mobilidade internacional, devido ao crescente número de alunos que pretende estudar internacionalmente bem como a quantidade de países e instituições disponíveis para tal, as instituições de ensino superior necessitam de possuir um conhecimento profundo sobre o comportamento dos alunos como consumidores do serviço de educação numa perspetiva internacional, explorando os vários fatores que motivam os mesmos a escolher determinada escola (Choudaha & Chang, 2012; Cubillo et al., 2006).

A presente seção do estado da arte centra-se na tomada de decisão do aluno sobre o seu local de estudo, a partir do momento em que o aluno decide estudar internacionalmente. No âmbito do processo de tomada de decisão de uma instituição educacional, Kotler e Fox (1995) apresentam um modelo com cinco etapas: pré-pesquisa, pesquisa, candidatura, decisão de escolha e decisão de matrícula. Por sua vez, Maringe e Carter (2007, p.463) definem o processo de tomada de decisão de mobilidade internacional como um “processo complexo com várias fases realizado conscientemente e por vezes subconscientemente por um aluno que pretende ingressar no ensino superior pelo que o problema de escolha de destino de estudo e de curso é resolvido”. Não obstante a quantidade de fases envolvidas no processo de seleção de uma instituição, é consensual na literatura, que existe uma fase de pesquisa por parte do aluno num momento anterior à sua candidatura (Bourke, 2000; R. G. Chapman, 1986; Cubillo et al., 2006; Ivy, 2001; Kotler & Fox, 1995; Maringe & Carter, 2007; Mazzarol, 1998; Wilkins & Huisman, 2014). Desse modo, compreende-se que o aluno irá passar por várias fases até tomar a sua decisão. Porém, o presente estudo foca-se apenas na fase de pesquisa pretendendo explorar-se os fatores que levam os alunos a escolher determinada instituição educacional internacionalmente.

Durante o processo de tomada de decisão é frequentemente difícil conhecer aspetos detalhados sobre as instituições e os seus cursos pelo que, associado ao forte posicionamento surge a importância da imagem e da reputação como critérios poderosíssimos para a seleção do

local de estudo (Bourke, 2000; Cubillo et al., 2006; Nicholls et al., 1995). A definição dos conceitos imagem e reputação não é completamente consensual na literatura pelo que, será adotada uma definição de cada um destes conceitos de acordo com a literatura e com o que se pensa estar mais enquadrado com a presente investigação. Kotler e Fox (1995, p.59) definem imagem como "a soma das crenças, ideias e impressões que um indivíduo possui sobre um objeto" e, Bennett e Kottassz (2000) defendem que a reputação é o conjunto de opiniões formadas ao longo de um período de tempo sobre a organização. Segundo a literatura, entende-se que a imagem é algo mais pessoal que passa pelo que vai na mente de cada aluno sobre a instituição e, que pode rapidamente ser afetada tanto pela própria instituição, como pelas fontes de informação, como pela própria reputação da instituição (Cubillo et al., 2006; Gray & Balmer, 1998; Kotler & Fox, 1995). A reputação, por sua vez, tem um caráter mais coletivo, relacionando-se com a consistência dos serviços prestados pela organização e, como tal, é influenciada pelo agregado das imagens formadas ao longo do tempo pelos consumidores, neste caso, os alunos (Barnett, Jermier, & Lafferty, 2006; Bennett & Kottassz, 2000). Assim, a visão adotada de imagem tem uma conotação mais individual sobre as impressões que um aluno possui, num dado momento, sobre os aspetos da instituição tais como o corpo docente, as infraestruturas, os custos, entre outros. No que trata à reputação, defende-se que esta tem um sentido coletivo representando a opinião dos alunos a nível agregado sobre a instituição ao longo de um período de tempo.

Posto isto, nesta secção serão abordadas as principais características dos serviços, mais especificamente o serviço de educação e, serão explorados os principais fatores que influenciam a seleção de uma IES ao longo da fase de pesquisa, segundo a literatura existente.

2.1.1 Natureza do serviço de educação

De acordo com a literatura, os serviços implicam o desenvolvimento de estratégias de *marketing* específicas devido às características que apresentam nomeadamente a sua intangibilidade, inseparabilidade, heterogeneidade e perecibilidade (Edgett & Parkinson, 1993). Os serviços não podem ser tocados, provados, cheirados ou vistos pelo que, a intangibilidade caracteriza a qualidade do serviço difícil de ser percecionada sem uma experiência prévia. Durante a prestação do serviço, a produção e o consumo do serviço dão-se simultaneamente pelo que os serviços caracterizam-se também pela sua inseparabilidade. A heterogeneidade diz respeito à dificuldade de standardização de um serviço, o que traduz a dificuldade de controlo da qualidade. Por fim, a perecibilidade indica que os serviços não podem ser armazenados para consumir mais tarde, ou seja, têm de ser consumidos no momento em que são produzidos.

Dadas as suas particularidades, a qualidade da maioria dos atributos proporcionados por um serviço, não pode ser percebida sem que o mesmo seja consumido, levando a que a decisão de aquisição de um serviço seja associada a um nível de risco bastante elevado (Bourke, 2000; Cubillo et al., 2006). Devido ao nível de risco inerente à compra de um serviço sem qualquer experiência prévia, a reputação tem um papel primordial pois, devido ao seu caráter coletivo, é percebida como um selo de qualidade. Assim, os consumidores baseiam-se nesta de maneira a ponderarem sobre a aquisição (Herbig & Milewicz, 1993; Walsh et al., 2009).

A educação caracteriza-se por ser um serviço puro e, como tal, torna-se difícil inferir sobre a qualidade dos serviços prestados pelas instituições educacionais sem uma experiência prévia (Cubillo et al., 2006), principalmente quando estes dizem respeito a serviços prestados por um país diferente em que os alunos não se podem deslocar fisicamente às IES (Gomes & Murphy, 2003). Adicionalmente, num contexto internacional, a aquisição de um serviço de educação inclui, não só a aquisição do serviço primário académico, mas também de outros serviços secundários associados à instituição bem como ao país de destino, tais como infraestruturas do campus ou o custo de vida do país (Cubillo et al., 2006). Desse modo, compreende-se que o aluno precisa de ter em conta vários aspetos, o que amplifica a complexidade da tomada de decisão. Neste sentido, a imagem que o aluno forma sobre os possíveis locais de estudo ao longo da sua pesquisa influencia fortemente a sua decisão final (Kotler & Fox, 1995). Por sua vez, a reputação é crítica para formação da imagem do aluno pois, uma vez agindo como indicador de qualidade influencia fortemente a mesma (Bourke, 2000; Mazzarol & Soutar, 2002).

2.1.2 Fatores que influenciam a escolha do local de estudo

Vários estudos defendem que a decisão de estudar num país estrangeiro é explicada por uma combinação de fatores *push* e fatores *pull* (Gomes & Murphy, 2003; Maringe & Carter, 2007; Mazzarol & Soutar, 2002). Os fatores *push* dizem respeito a características do país de origem que fomentam o desejo de continuar os estudos num país estrangeiro. Por sua vez, os fatores *pull* dizem respeito a características, do país ou da instituição, atrativas para o aluno, influenciando positivamente a decisão de estudar nesse país.

Os fatores *push* têm uma influência especialmente significativa numa fase inicial, em que o aluno decide deixar o seu país de origem para estudar internacionalmente (Mazzarol & Soutar, 2002). Maringe e Carter (2007) apuraram que os principais fatores *push* são económicos, políticos e falta de capacidade no ensino superior do país de origem. A nível económico, a pobreza extrema de alguns países leva a que alguns alunos considerem inaceitável continuar a estudar no país de origem incentivando a que estes procurem outro país para estudar e viver. Outro fator económico diz respeito às oportunidades de trabalho que se caracterizam por ser escassas ou, quando existentes, são bastante mal remuneradas. Por outro lado, a instabilidade política bem como a rigidez de algumas forças políticas incentivam também a decisão de estudar noutro país. Além destes fatores, as instituições de ensino, em alguns países, não têm capacidade para dar lugar à quantidade de alunos que se candidata pelo que estes vêm a necessidade de procurar outros países para estudar. Mazzarol e Soutar (2002) realçam também como fatores *push* a não disponibilidade do curso desejado no país de origem bem como a qualidade de ensino no país de origem ser considerada de qualidade inferior quando comparada a outros países. Counsell (2011), verificou ainda que os alunos ambicionam estudar internacionalmente devido a motivos profissionais, uma vez que experiências internacionais bem como o domínio de uma segunda língua são bastante valorizados no mercado de trabalho do país de origem.

Através da revisão de literatura apuraram-se cinco grandes fatores *pull* capazes de influenciar a tomada de decisão de um aluno na escolha de uma instituição de ensino superior

aquando de um processo de mobilidade internacional e que serão explanados ao longo das próximas secções. O primeiro fator diz respeito à imagem que os alunos formam sobre o país. O segundo fator diz respeito à imagem da própria instituição de ensino superior. O terceiro fator relaciona-se com questões financeiras. O quarto fator diz respeito a motivações pessoais . Por último, o quinto fator diz respeito à influência exercida pela informação recolhida.

2.1.2.1 Imagem sobre o país anfitrião

Vários estudos defendem que, uma vez tomada a decisão de estudar internacionalmente, a decisão sobre o local de estudo passa tipicamente pela escolha do país e, apenas de seguida, é escolhida a instituição de ensino (Bourke, 2000; Mazzarol & Soutar, 2002), o que sugere a importância das características do país anfitrião no processo de escolha do local de estudo internacional.

Srikatanyoo e Gnoth (2002, p. 140) enfatizam a importância da imagem sobre país definindo-a como "crenças cognitivas do aluno sobre a industrialização do país, qualidade *standard* nacional, e outras informações associadas aos seus produtos e serviços". A definição apresentada por Srikatanyoo e Gnoth (2002) vai de encontro ao que Cubillo et al. (2006) defendem quando afirmam que o processo de mobilidade implica, além da aquisição do serviço primário académico, a aquisição de serviços secundários prestados não só pela instituição mas também pelo país.

A boa reputação é um dos fatores que mais influencia a imagem sobre o país (Beine et al., 2014; Bourke, 2000; Mazzarol & Soutar, 2002). Segundo Bourke (2000) a educação de ensino superior é uma característica de cada país sendo um reflexo da sua cultura pelo que, os alunos têm tendência a crer que países com boa reputação proporcionam serviços de educação de qualidade superior. O reconhecimento da qualidade de ensino do país é também importante uma vez que, muitos estudantes tomam a decisão de instruir-se em determinado país com o intuito de serem mais valorizados no mercado de trabalho do país de origem (Counsell, 2011; Maringe & Carter, 2007).

A reputação das instituições é influenciada pela reputação do país sendo que, a transmissão de uma imagem favorável sobre o país, por parte das instituições, poderá levar os alunos a formarem uma boa imagem sobre as mesmas (Srikatanyoo & Gnoth, 2002). Caso um país não seja fortemente reconhecido pelos serviços de educação prestados, as instituições devem apostar em estratégias de *marketing* agressivas de maneira a alterar a perceção que os alunos detêm (Srikatanyoo & Gnoth, 2002).

Não menos importante é a cidade. Esta, juntamente com a instituição e as suas instalações, representa o ambiente físico onde o aluno irá "consumir" o serviço de educação (Cubillo et al., 2006). Cubillo et al. (2006) destaca a importância da cidade dando como exemplo a cidade de Salamanca, em Espanha. O autor refere que Salamanca tem uma imagem fortemente associada à aprendizagem de espanhol e à sua cultura sendo, como tal, uma forte atração para alunos internacionais.

Por fim, de acordo com vários autores, a perceção de uma forte presença de estudantes internacionais num determinado país funciona também como forte atração para futuros alunos

pelo que se considera relevante a promoção de um ambiente internacional por parte das instituições (Beine et al., 2014; Cubillo et al., 2006; Maringe & Carter, 2007). Outros critérios pode levar o país a ser considerado mais atrativo são: a perceção sobre a segurança e discriminação, a facilidade do processo migração para alunos internacionais (Cubillo et al., 2006; Maringe & Carter, 2007; Mazzarol & Soutar, 2002), a distância/proximidade cultural e a distância/proximidade da língua entre o país de destino e origem (Beine et al., 2014; Cubillo et al., 2006; Maringe & Carter, 2007)

Tendo em conta a forte influência deste fator formulou-se a seguinte hipótese:

H1a. Com base na revisão literária, o sentimento dos termos relacionados com o país está positivamente correlacionado a satisfação dos alunos face a uma IES.

2.1.2.2 Imagem sobre a instituição de ensino superior

Srikatanyoo e Gnoth (2002, p.140) definem a imagem da instituição como "as perceções gerais sobre a qualidade da instituição, detidas pelo aluno". A imagem de uma instituição pode não corresponder à realidade da mesma pelo que, alunos que possuam uma imagem negativa sobre uma instituição vão evitá-la ou desprestigiá-la, apesar da mesma poder, na verdade, ter uma qualidade superior (Kotler & Fox, 1995).

A boa reputação académica é considerada um dos principais critérios a influenciar a escolha pois tem um impacto poderosíssimo na imagem que o aluno forma sobre a instituição (Bourke, 2000; Cubillo et al., 2006; Hemsley-Brown, 2012; Maringe & Carter, 2007; Mazzarol & Soutar, 2002). Garvin (1980, p.15) reforça ainda este aspeto defendendo que a "qualidade de uma instituição é menos importante do que a sua reputação pois, na realidade, é a sua reputação que comanda as decisões de potenciais alunos". De maneira a avaliar a reputação académica, os *rankings* exercem uma influência bastante significativa sendo que os alunos recorrem frequentemente aos mesmos a fim de conhecer as instituições com posições superiores nos países selecionados (Beine et al., 2014).

A disponibilização do curso procurado existir na escola é determinante na escolha (Bourke, 2000; Price, Matzdorf, Smith, & Agahi, 2003). A reputação e reconhecimento internacional do curso são também extremamente decisivos pois, os estudantes acreditam que terão maiores benefícios no mercado de trabalho quer a nível salarial quer a nível de novas oportunidades ou progressão de carreira (Bourke, 2000; Counsell, 2011).

Mazzarol & Soutar (2002) apuraram, através de questionários, que a reputação da qualidade e experiência do *staff*, especialmente do corpo docente, foi um dos principais motivos que levaram os alunos a selecionar uma instituição em favor de outras, o que reflete a importância da imagem que o aluno tem acerca do *staff*.

Price et al. (2003) salienta a relevância das instalações das instituições para a atração de novos alunos. Os autores, na sua investigação, realçam como primariamente importantes a existência de bibliotecas e disponibilização de material tecnológico. Além destes fatores, os autores realçam a importância da vida social na instituição bem como nos arredores da mesma. Por sua vez, Maringe & Carter (2007) alertam para a necessidade de as instituições terem o

cuidado de fornecer alojamento aos alunos ou, pelo menos, proporcionar informação sobre a disponibilização de acomodação por parte da instituição educacional.

Outros fatores que podem contribuir para a criação de uma imagem mais positiva da instituição dizem respeito à disponibilização de áreas de estudo e à simplicidade do procedimento de candidatura na mesma(Bourke, 2000; Maringe & Carter, 2007).

Dito isto, a literatura sugere que à medida que os alunos ficam mais satisfeitos com a instituição anfitriã, também as suas opiniões tendem a ser mais positivas o que leva à formulação da segunda hipótese:

H1b. Com base na revisão literária, o sentimento dos termos relacionados com a instituição está positivamente correlacionado com a satisfação dos alunos face a uma IES.

2.1.2.3 Fatores financeiros

Estudar internacionalmente é possivelmente uma das experiências mais dispendiosas no percurso académico de um estudante (Mazzarol, 1998). Nesse sentido, é consensual na literatura que os custos a suportar nesta experiência têm um peso bastante significativo no momento da escolha do local de estudo.

Através de uma análise por questionário, Binsardi e Ekwulugo (2003) verificaram, segundo respostas dadas por alunos internacionais, que a melhor maneira das instituições atraírem alunos internacionais seria a diminuição do valor das propinas bem como o aumento dos valores das bolsas de estudo em contexto internacional. Contudo, outros estudos alertam para a importância que os alunos dão, não só aos custos diretamente relacionados com o serviço de educação prestado, mas também aos custos indiretos associados a toda a experiência internacional, tal como o custo de vida (Beine et al., 2014; Mazzarol & Soutar, 2002) e os custos de viagens (Beine et al., 2014; Mazzarol & Soutar, 2002).

Beine et al. (2014) salientam a forte influência do custo de vida comparativamente ao custo das propinas no processo de tomada de decisão. A investigação dos autores defende que o alto valor das propinas é, por vezes, associado a uma qualidade superior da instituição pelo que, instituições que têm uma boa reputação podem dar-se ao luxo de praticar valores de propinas mais elevados pois a grande parte dos alunos predispõem-se a pagar tais quantias em troca de uma qualidade de ensino de topo. Além desta razão, os autores sugerem que frequentemente os alunos não beneficiam de ajudas financeiras que suportem despesas pessoais, sendo que aquelas apenas suportam despesas académicas tornando, desse modo, o custo de vida um fator extremamente importante.

Adicionalmente, alguns alunos consideram também a possibilidade de trabalhar em *part-time* no país anfitrião (Maringe & Carter, 2007; Mazzarol & Soutar, 2002). Os estudantes nem sempre têm possibilidades para pagar a totalidade dos anos requeridos, pelo que precisam de garantir que podem trabalhar no país de destino (Maringe & Carter, 2007). Vários países permitem já que alunos internacionais possam trabalhar em *part-time* através dos seus vistos de estudante, constituindo esta, por vezes, a única maneira dos mesmos poderem usufruir de uma experiência internacional (Mazzarol & Soutar, 2002).

À semelhança dos fatores anteriores, a literatura levou à formulação da seguinte hipótese:

H1c. Com base na revisão literária, o sentimento dos termos relacionados com questões financeiras está positivamente correlacionado com a satisfação dos alunos face a uma IES.

2.1.2.4 Motivações pessoais

A decisão de estudar internacionalmente é fortemente movida por aquilo que o aluno acredita que o irá beneficiar (Binsardi & Ekwulugo, 2003).

Counsell (2011) e Bourke (2000) identificaram, entre diversos fatores, que os principais motivos que levam alunos a decidir passar por uma experiência internacional passam pelos mesmos acreditarem que terão melhores oportunidades de carreira no futuro. Cubillo et al. (2006) aprofundam a ideia anterior apontando a importância dos alunos acreditarem que irão beneficiar de rendimentos superiores no mercado de trabalho.

O desejo de desenvolver uma segunda língua é muitas vezes uma motivação que leva alunos a ambicionar estudar internacionalmente (Bourke, 2000; Counsell, 2011). Alunos de vários países, principalmente alunos provenientes de países cuja língua não é a do país de destino, optam por estudar internacionalmente para que os seus estudos sejam lecionados em inglês (Bourke, 2000).

Por outro lado, vários alunos querem estudar noutro país simplesmente pela vontade que têm de passar por uma experiência internacional (Counsell, 2011; Maringe & Carter, 2007). Segundo Maringe e Carter (2007) uma experiência internacional é aquela em que os alunos têm o intuito primário de conhecer pessoas de culturas diferentes e trocar conhecimentos e vivências com as mesmas, sendo educados por equipas de especialistas internacionais.

Outras motivações apontadas pela literatura são a vontade sentida pelo aluno de ganhar independência (Bourke, 2000; Counsell, 2011), aumentar a rede de contactos internacional (Cubillo et al., 2006; Maringe & Carter, 2007), distanciar-se do seu país de origem e conhecer uma nova cultura (Bourke, 2000; Counsell, 2011; Maringe & Carter, 2007).

H1d. Com base na revisão literária, o sentimento dos termos relacionados com motivações pessoais está positivamente correlacionado com a satisfação dos alunos face a uma IES.

2.1.2.5 Informação

Bourke (2000) na sua investigação defende que são vários os fatores a influenciar a tomada de decisão de onde estudar internacionalmente. Porém, devido às características associadas à natureza dos serviços, a mais importante variável é a informação (Bourke, 2000). Na realidade, a informação obtida sobre o país, instituição e curso irá moldar fortemente a imagem criada pelo aluno ao longo do processo de seleção e, como tal, afetará a tomada de decisão do aluno independentemente da sua veracidade (Kotler & Fox, 1995). Kotler e Fox (1995) defendem mesmo que a informação constitui a base da formação da imagem. Tal, sugere que a informação é transversal aos restantes fatores identificados anteriormente, caracterizando-se por ser um fator base capaz de influenciar os restantes (Figura 1).

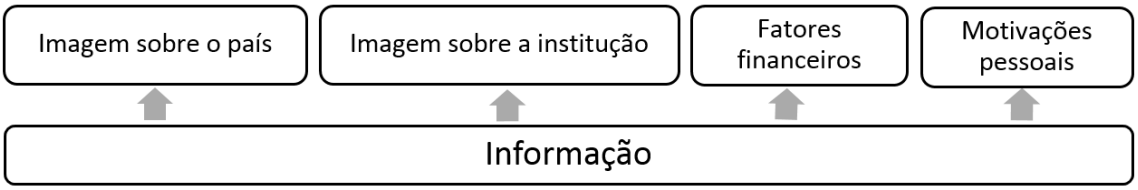


Figura 1 Fatores que influenciam a escolha de um local de estudo internacionalmente (sistematização própria)

A decisão sobre onde estudar internacionalmente pode ser afetada pela informação adquirida pelas próprias instituições. Contudo, a influência exercida por recomendações provenientes de ex-alunos, família, amigos, professores, redes sociais, entre outros tem um impacto especialmente grande (Maringe, 2006; Mazzarol & Soutar, 2002).

Como já mencionado, uma vez que não podem deslocar-se às IES, os alunos recorrem frequentemente à Internet em busca de recomendações (Gomes & Murphy, 2003). Gomes e Murphy (2003), através de um questionário a alunos internacionais, verificaram que aproximadamente 70% dos alunos recorreram à Internet, além do próprio *website* da instituição educacional, para procurar informações sobre as IES durante o processo de seleção da instituição, o que reflete a importância da opinião dos ex-alunos expressa neste meio para a tomada de decisão.

O *word-of-mouth* é considerado um dos meios mais fortes para a promoção de instituições num âmbito de educação internacional (Mazzarol & Soutar, 2002). Wilkins e Huisman (2014) alertam para a necessidade das instituições satisfazerem as necessidades dos alunos e encorajarem, ou até mesmo recompensarem, um *word-of-mouth* positivo pois, alunos que exprimem opiniões negativas sobre as instituições podem afetar fortemente a reputação das mesmas e, consequentemente levar a uma diminuição quer de candidaturas quer de recomendações. A imagem formada pelo aluno sobre as instituições educacionais é bastante influenciada pelo que é dito por terceiros (Wilkins & Huisman, 2014). Porém, é importante notar que a imagem partilhada por terceiros pode ela própria ser influenciada pela reputação existente da instituição (Wilkins & Huisman, 2014).

Binsardi e Ekwulugo (2003) verificaram que a maioria dos alunos considera que a melhor maneira de promover a instituição, para atrair mais alunos internacionais, é através de redes de alunos. Tal reflete, novamente, a relevância da influência das informações provenientes de ex-alunos sugerindo que as instituições devem apostar em fortes redes Alumni *online* como meio de se promoverem internacionalmente (Gomes & Murphy, 2003). Segundo a literatura, as redes Alumni provaram já ser eficazes fontes de *word-of-mouth* sobre as instituições que provocam um forte impacto na decisão do aluno ao longo do processo de seleção de uma instituição (Bourke, 2000; Cubillo et al., 2006; Mazzarol & Soutar, 2002; Wilkins & Huisman, 2014).

Contrariamente aos fatores identificados anteriormente, considera-se que este não está diretamente relacionado com características que possam ser vivenciadas ao longo da experiência, mas sim com a informação encontrada, na fase de pesquisa, acerca dos serviços prestados independentemente da sua veracidade. Assim, considera-se que a formulação de uma hipótese de estudo idêntica à dos fatores anteriores não faz sentido em relação a este.

2.2 ANÁLISE DE SENTIMENTOS DE OPINIÕES DE ESTUDANTES (*ONLINE REVIEWS*)

A Tabela 1 apresenta de forma sintetizada os cinco grandes fatores que influenciam a escolha do local de estudo segundo a literatura.

Tabela 1 Sumário das fatores que influenciam a escolha de um local de estudo internacionalmente (sistematização própria)

F1. IMAGEM SOBRE O PAÍS	F2. IMAGEM SOBRE A INSTITUIÇÃO	F3. FATORES FINANCEIROS	F4. MOTIVAÇÕES PESSOAIS	F5. INFORMAÇÃO
1.1 Reputação	2.1 Reputação e prestígio de ensino	3.1 Custo de vida	4.1 Experiência internacional	5.1 Social Media
1.2 Proximidade/Distancia de Idioma	2.2 Reputação e experiência do *staff*	3.2 Custos académicos	4.2 Ganhar independência	5.2 Redes de alunos
1.3 Proximidade/distância Cultural	2.3 Disponibilização e qualidade do curso	3.3 Custos de viagens	4.3 Conhecer nova cultura	5.3 Amigos
1.4 Facilidade de acesso à informação	2.4 Reconhecimento internacional do curso e da instituição	3.4 Possibilidade de trabalhar durante os estudos	4.5 Melhorar capacidades linguísticas	5.4 Família
1.5 Características pelas quais a cidade é reconhecida	2.5 Vida social dentro e fora da instituição		4.6 Aumentar redes de contactos	5.5 Professores
1.6 Ambiente internacional	2.6 Disponibilização de bibliotecas e áreas de estudo		4.7 Melhores oportunidades de carreira	5.6 Contacto direto com a instituição
1.7 Procedimentos de migração simples	2.7 Disponibilidade de material tecnológico		4.8 Rendimentos superiores	
1.8 Segurança e Discriminação	2.8 Disponibilização de alojamento			
	2.9 Procedimentos de ingresso simples			

Uma vez que os primeiros quatro fatores podem efetivamente ser experienciados e devido aos critérios entre eles serem bastante distintos entre si, considerou-se que faria sentido se a satisfação dos alunos face a uma IES após a experiência dependesse do fator com a qual estivesse mais correlacionada. Assim, colocou-se a seguinte hipótese de estudo para os primeiros quatro fatores:

H2. A satisfação dos alunos face a uma IES difere de forma significativa consoante o fator.

Além dos quatro fatores, a literatura salientou como especialmente importante o impacto exercido pelas fontes de informação no processo de pesquisa, especialmente a Internet, devido às implicações associadas a um serviço de educação internacional. No entanto, verificou-se que as investigações encontradas são na sua totalidade apenas estudos teóricos ou estudos que recorreram a análises manuais de questionários a alunos internacionais.

Desse modo, consciente do impacto da Internet bem como do impacto que a reputação exerce sobre a formação da imagem, compreende-se facilmente que a análise das opiniões de ex-alunos presentes no meio *online* não pode ser desprezada, pois são estas opiniões que vão influenciar a reputação de uma IES e formar a imagem dos futuros alunos internacionais, consequentemente definindo a sua escolha.

A partir dos comentários deixados no meio *online* é possível extrair conhecimento sobre as experiências dos alunos internacionais e o sentimento em relação às mesmas. Por exemplo, um ex-aluno de uma instituição pode partilhar na Internet o seguinte comentário: “As instalações da faculdade são fantásticas mas os cursos não são lecionados em inglês e por isso é muito

difícil acompanhar a matéria!". O conhecimento sobre a opinião dos alunos em comentários como este permite que as instituições, não só tenham uma visão mais completa sobre as opiniões dos serviços que prestam mas também que, invistam em aspetos que são realmente importantes para o seu público-alvo (Feldman, 2013), neste caso os alunos internacionais.

Posto isto, uma vez que serão analisadas as experiências de alunos internacionais partilhadas no meio *online* e dada a diversidade de critérios apurados através da revisão literária, formulou-se a seguinte hipótese de estudo:

H3. Os critérios de escolha de uma IES apurados na revisão literária são consistentes com o que os alunos mais realçam nos comentários online após a experiência internacional.

No entanto, tendo em conta a quantidade de informação existente, perceber a opinião geral dos consumidores, neste caso dos alunos internacionais, é um processo bastante moroso e, analisar apenas algumas opiniões é limitador levando a que as conclusões retiradas sejam tendenciosas (Zhang et al., 2011). Como tal, pretende recorrer-se a técnicas de análise textual que permitam analisar grandes quantidades de dados e os seus sentimentos de forma automatizada. O recurso a estas técnicas possibilita analisar uma quantidade de dados bastante maior do que as comuns análises manuais e de maneira mais eficiente (Fan, Wallace, Rich, & Zhang, 2006). As técnicas automatizadas permitem também analisar opiniões expressas livremente pelo próprio consumidor, o aluno, livres de qualquer análise subjetiva e tendenciosa resultante de perceções humanas (Yamanishi & Li, 2002). Por último, este tipo de técnicas permite ainda fazer análises em "tempo real", ou seja, ao contrário das análises comuns por questionários que se caracterizam por serem estáticas no tempo, as técnicas automatizadas permitem analisar informação nova atualizada de forma contínua. Desse modo, nas próximas secções, as técnicas automatizadas de análise textual e análise de sentimentos serão elucidadas com maior detalhe dando a conhecer os benefícios da sua utilização.

2.3 TEXT MINING

À técnica que trata da descoberta de conhecimento automatizada a partir de dados não estruturados ou semiestruturados dá-se o nome de *text mining* (TM) (também conhecido por *text data mining* (Hearst, 1999) ou *text knowledge mining* (Sánchez, Martín-Bautista, Blanco, & Torre, 2008)). A descoberta de conhecimento diz respeito à obtenção de novas associações, hipóteses ou tendências não explícitas em documentos não estruturados ou com pouca estrutura, tais como documentos de negócio, questionários de campos abertos, comentários, páginas *web*, etc. (Delen & Crossland, 2008). Assim "*text mining* refere-se à descoberta de conhecimento não-trivial, previamente desconhecido e potencialmente útil, a partir de uma coleção de textos" (Sánchez et al., 2008).

O desenvolvimento tecnológico fez com que atualmente fosse possível capturar grandes quantidades de dados (Lee, Baker, Song, & Wetherbe, 2010). Parte destes dados caracteriza-se por ser estruturada, porém estima-se que a grande maioria dos dados existentes esteja em formato não estruturado, ou seja, textual (Sánchez et al., 2008). Na realidade, nos últimos anos a quantidade de dados textual tem crescido exponencialmente, sendo o melhor exemplo disso a

Internet. O recurso a ferramentas colaborativas tais como *blogs*, redes sociais, *wikis* e fóruns de discussão nos dias de hoje permite que seja possível aceder-se a um volume de informação muito maior do que em qualquer outra altura na história (Blake, 2011). No entanto, enquanto os dados estruturados podem ser analisados através de técnicas tradicionais de análise de dados estruturados (*data mining*), a informação textual requer análises específicas que consigam lidar com os desafios presentes no texto (Lee et al., 2010). O texto encontra-se repleto de obstáculos (tais como erros ortográficos, calões e significados implícitos) e, embora o ser humano tenha capacidade de interpretar o texto passando por vários desses obstáculos, a quantidade de informação é de tal forma grande que é impossível para um humano ter capacidade de a analisar eficientemente (Lee et al., 2010; Mostafa, 2013). É, para responder a esta necessidade que nasce o TM como um método automatizado capaz de analisar grandes quantidades de informação não estruturada ou semi-estruturada eficientemente.

Embora segundo a literatura, o que se entende por TM seja consensual, os investigadores que têm estudado e desenvolvido o TM ao longo dos anos têm contribuído com diferentes técnicas de geração de conhecimento a partir dos textos. Neste sentido, existem duas abordagens a ter em conta para este fim: indutiva – tipicamente através de algoritmos de *machine learning* – e a dedutiva – através de regras e associações (Blake, 2011; Sánchez et al., 2008). Em ambas as abordagens o intuito é o mesmo, ou seja, pretende descobrir-se novo conhecimento; no entanto, o que é distinta é a forma como se gera o novo conhecimento bem como a fonte de dados de onde se parte.

Na abordagem indutiva, recorre-se tipicamente à aplicação de técnicas de *machine learning* ou outras análises estatísticas (Blake, 2011; Sánchez et al., 2008). O processo de TM desta abordagem identifica-se com o processo de descoberta de conhecimento apresentado por Fayyad et al. (1996) residindo nos seguintes passos: seleção, pré-processamento, transformação, *data mining,* interpretação/avaliação (Blake, 2011). Neste processo, as técnicas de *data mining* são aplicadas após o tratamento e estruturação dos dados textuais sendo os padrões identificados através de modelos de *machine learning* (Fayyad et al., 1996). Este processo faz com que a geração de conhecimento seja efetuada através de inferência indutiva uma vez que os modelos aprendem com os casos que analisam, produzindo conhecimento consistente com a amostra presente e induzindo o conhecimento obtido para a restante população. Face a isto, alguns investigadores consideram que *o text mining* é uma extensão do *data mining* (Inniss, Lee, & Light, 2006; Tan, 1999). Tal facto, associado ao sucesso obtido através destas técnicas, fez com que a maioria dos estudos de TM recorressem a esta abordagem (Sánchez et al., 2008).

No entanto, Hearst (1999) e Sánchez et al. (2008) defendem que, o "verdadeiro *text mining"* não vem da análise de tendências e padrões detetados através dos dados estruturados da amostra pois tal restringe a expressividade da informação existente nos documentos expressos em linguagem natural. Desta forma, numa abordagem dedutiva parte-se de repositórios com mais conteúdo semântico e que contêm já premissas verdadeiras e factuais (Sánchez et al., 2008). Nesta abordagem o novo conhecimento é criado baseado em

conhecimento já existente através de regras dedutivas ou do recurso a outros procedimentos complexos que permitam obter conhecimento de forma dedutiva. A *literature-based discovery* (LBD) é um exemplo de abordagem dedutiva (Blake, 2011; Sánchez et al., 2008) que tem como objetivo deduzir uma relação, antes desconhecida, entre X e Z, a partir do conhecimento existente das associações entre X e Y bem como todas as associações conhecidas entre Y e Z (Hristovski, Peterlin, Mitchell, & Humphrey, 2005).

Apesar da geração de conhecimento ser diferente entre as duas abordagens, independentemente da abordagem utilizada é necessário que os dados passem por uma fase de tratamento e estruturação para que possam ser analisados (Sánchez et al., 2008). É neste sentido que a área de *Computational Linguistics*, também apelidada por processamento natural da língua (PNL) desempenha um papel fundamental (Blake, 2011). Os esforços despendidos por esta área não são tão concentrados a nível de descoberta de novo conhecimento mas sim a nível de melhorias na análise da linguagem (Hearst, 1999). Desta forma, os investigadores têm-se esforçado em analisar como é feita a estruturação da língua com o intuito de desenvolver ferramentas computacionais que consigam manipular e compreender os dados textuais (Chowdhury, 2003). O PNL tem em conta tarefas tais como: análise do contexto em que as palavras aparecem (Caro & Grella, 2013), a compreensão de regras sintáticas e semânticas, identificação do que está explícito ou implícito (Cambria, Schuller, Xia, & Havasi, 2013), entre outras.

O TM é transversal a várias áreas permitindo resolver bastantes desafios. Por exemplo, a área de Information Retrieval permite identificar mais eficientemente a informação, tal como a identificação de informação *online* em que são desenvolvidos motores de busca para encontrar informação mais facilmente (Liaw & Huang, 2006); A área de Information Extraction prende-se com a descoberta de informação precisa e relevante num documento particular e posterior transformação de dados não estruturados em dados estruturados permitindo que, por exemplo, facilmente se extraia a informação subjacente a um determinado conteúdo de forma resumida (Liaw & Huang, 2006); Por sua vez, a área de análise de sentimentos dedica-se à e compreensão do que sentem as pessoas sobre algum produto, entidade, assunto, etc. servindo como base para a tomada de decisões a nível de negócio (Cambria et al., 2013). A presente dissertação é centrada nesta última sendo que na próxima secção se explorará esta área em maior detalhe.

2.3.1 Análise de Sentimentos

A análise de sentimentos (AS) é uma aplicação na área de *text mining* que diz respeito ao processo automatizado de descoberta de padrões a partir da análise de opiniões presentes em documentos como por exemplo *blogs*, *tweets* ou *reviews* (Mostafa, 2013). Diversas designações têm sido associadas à análise de sentimentos tais como *opinion mining* (Pang & Lee, 2008; Rushdi Saleh, Martín-Valdivia, Montejo-Ráez, & Ureña-López, 2011), análise de subjetividade ou *sentiment orientation* (Rushdi Saleh et al., 2011) sendo que todas elas se referem ao mesmo campo de estudo (Cambria et al., 2013; Pang & Lee, 2008; Rushdi Saleh et al., 2011).

2.3.1.1 Conceitos

Liu & Zhang (2012) alertam e exploram um conjunto de conceitos importantes a ter em conta nesta área. Assim, de maneira a expor as várias noções subjacentes à análise de sentimentos recorrer-se-á a um exemplo de um comentário e serão expostas algumas noções de acordo com os autores.

" (1) Eu comprei um computador na semana passada. (2) O processador é muito rápido e tem uma ótima memória RAM. (3) O problema é que eu não contei à minha mãe. (4) Quando ela descobriu, ficou super chateada porque achou o computador muito caro."

Opinião "Uma opinião é simplesmente um sentimento, uma atitude, emoção ou avaliação positiva ou negativa sobre uma entidade ou um aspeto de uma entidade expresso por um detentor de opinião" (Liu & Zhang, 2012, p. 418).

Polaridade Também apelidada de orientação semântica ou *sentimental orientation,* refere-se ao sentimento associado a uma opinião, por exemplo positivo, negativo ou neutro. Neste caso, a frase (2) denota uma polaridade positiva ao passo que as frases (3) e (4) denotam uma polaridade negativa.

Detentor de opinião O detentor da opinião é aquele que detém e expressa a opinião. Na frase (2) o detentor da opinião é o próprio autor do comentário porém na frase (4) o detentor da opinião é a mãe. A deteção do detentor da opinião é normalmente mais importante ao analisar notícias pois podem existir vários sujeitos envolvidos que exprimem explicitamente a sua opinião. No caso de comentários sobre produtos, serviços ou blogs, tipicamente o detentor da opinião é o autor da publicação. Contudo, quando se pretende agregar a opinião geral dos autores, conhecer cada detentor de opinião não é necessariamente relevante.

Entidade O conceito entidade refere-se ao objeto que está a ser avaliado pelo detentor de opinião. Neste caso, a entidade diz respeito ao computador contudo, poderia ser qualquer outro produto, serviço, evento, tópico, entre outros.

Aspeto Os aspetos referem-se aos atributos das entidades. No exemplo apresentado pode percebe-se que dois dos aspetos são facilmente detetados, nomeadamente o processador e a memória. Contudo, embora implícito, o preço é também um aspeto contemplado no comentário.

Subjetividade Uma frase pode ser subjetiva ou objetiva. Trata-se de uma frase objetiva quando são apresentados factos sobre uma realidade por sua vez, trata-se de uma frase subjetiva quando são expressos sentimentos pessoais, desejos, pontos de vista, suspeitas ou crenças. Desse modo, a frase (1) é objetiva ao passo que as frases (2), (3) e (4) são subjetivas. Uma frase pode ser subjetiva e não exprimir uma opinião. Por exemplo, a frase "Eu quero uma mala de qualidade para guardar o meu computador" é subjetiva pois exprime um desejo mas não apresenta qualquer opinião positiva ou negativa. Por outro lado, uma frase pode ser objetiva e conter uma opinião implícita. Por exemplo a frase "Foram precisos cinco meses para as obras da sala ficarem acabadas" é objetiva no entanto está implícito um sentimento negativo.

Emoção As emoções vão além dos sentimentos serem positivos ou negativos estando associadas a um nível mais profundo tais como o amor, alegria, surpresa, raiva, tristeza e medo.

2.3.2 Abordagens de Análise de Sentimentos

Existem essencialmente duas grandes abordagens para detetar a polaridade dos sentimentos – *machine learning* ou léxicos de sentimentos (Caro & Grella, 2013; Medhat, Hassan, & Korashy, 2014). Sendo que a primeira se divide em métodos supervisionados ou não supervisionados e a segunda em abordagens *corpus-based* e *dictionay-based* (Figura 2). Existem ainda investigações híbridas, ou seja, que utilizam uma combinação de ambas as abordagens.

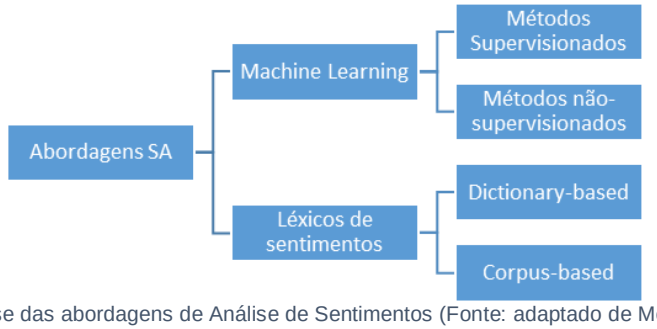


Figura 2 Síntese das abordagens de Análise de Sentimentos (Fonte: adaptado de Medhat et al. (2014))

2.3.2.1 *Machine Learning*

Nas análises através mecanismos de ***machine learning*** são aplicados algoritmos de aprendizagem automática que, após terem sido treinados num conjunto de dados representativo, ficam aptos a classificar novos documentos (Pang & Lee, 2004; Zhang et al., 2011). Como referido, dentro desta abordagem existem dois principais tipos de métodos – os métodos supervisionados e não-supervisionados. De seguida, serão explanados os métodos indicados e, na secção 3, estes métodos serão explorados a um nível mais técnico tendo em conta o que será utilizado no desenvolvimento prático da investigação.

Nos **métodos supervisionados** é definida uma quantidade finita de classes em que cada documento pode ser classificado (Feldman, 2013). Nas investigações de análise de sentimentos as classes correspondem frequentemente à polaridade do sentimento que se pretende determinar (por exemplo positivo, negativo e neutro). Após definidas as classes, é necessário construir-se um conjunto de dados de treino. Para tal, os documentos precisam de ser anotados com uma das classes possíveis. Em várias investigações, de maneira a anotar a polaridade, os documentos são associados a uma pontuação já associada ao comentário, ou seja, os comentários sobre um produto podem, por exemplo, ser associados às estrelas dadas pelos consumidores. Rushdi Saleh et al. (2011) no conjunto de dados que criaram, definiram que um comentário classificado com três, quatro ou cinco estrelas seria associado a um sentimento positivo ao passo que, um comentário classificado com uma ou duas estrelas seria considerado negativo.

Com o conjunto de treino criado, é posteriormente aplicado um algoritmo de *machine learning* que aprende os padrões de classificação dos documentos para cada classe. Finalmente, o modelo, após ter aprendido a classificar cada documento de acordo com a sua classe, encontra-se apto de categorizar novos documentos na classe respetiva.

Nos **métodos não-supervisionados** o propósito é juntar documentos semelhantes entre si no mesmo grupo, separando em grupos diferentes documentos distintos (Blake, 2011). Ao

contrário dos métodos supervisionados, neste método não é necessário saber-se à priori a que classe pertence cada documento uma vez que os documentos vão sendo agrupados com outros de acordo com a sua similaridade formando grupos (*clusters*). Após formados todos os *clusters*, os mesmos são rotulados de acordo com as suas características, definindo-se desta forma as classes. Finalmente, após as classes construídas, o modelo encontra-se pronto para classificar novos documentos nas classes a que mais se assemelharem. Turney (2002) utilizou um método não-supervisionado para classificar comentários de quatro domínios diferentes (filmes, destinos de viagem, automóveis e bancos). O algoritmo utilizado pelo autor calculava a polaridade de expressões escritas nos comentários que contivessem adjetivos ou advérbios. Caso a média das polaridades das expressões de cada documento atingisse um valor positivo, o comentário era associado à classe "recomendado". Caso a média da polaridade das expressões atingisse um valor negativo, o comentário era associado à classe "não recomendado".

2.3.2.2 Léxicos de sentimentos

Em comentários positivos estão habitualmente presentes palavras ou expressões que expressam sentimentos positivos assim como, em comentários negativos estão habitualmente presentes palavras ou expressões que expressam sentimentos negativos. Quando um conjunto de palavras ou expressões é classificado de acordo com a sua polaridade dá-se o nome de léxico de sentimentos. Assim, um léxico de sentimentos é no fundo uma lista de palavras em que cada uma encontra-se associada à sua polaridade. A criação de léxicos pode ser feita manualmente contudo, existem dois métodos primordiais para criar os léxicos de sentimentos: método baseado em dicionários *(dictionary-based)* ou método baseado num *corpus (corpus-based).* Devido à morosidade dos métodos manuais, a classificação manual tipicamente é combinada com as técnicas automáticas identificadas anteriormente (Liu & Zhang, 2012).

2.3.2.3 *Dictionary-based*

A técnica *dictionary-based* inicia-se pela construção manual de uma pequena lista de palavras associadas ao seu sentimento (*seed list*) a partir do conjunto de dados a analisar (Feldman, 2013). Seguidamente, recorre-se a um dicionário com o propósito de expandir a *seed list* através dos sinónimos e antónimos detetados pelo mesmo. O WordNet[4] é um exemplo de dicionário que contempla um grande léxico de palavras em inglês sendo capaz de detetar os vários conjuntos de sinónimos de uma palavra segundo a sua classe gramatical.

Existem várias experiências que recorrem a dicionários de maneira a detetar a orientação do sentimento. Uma das práticas foi explorada por Hu & Liu (2004). Os autores partiram do pressuposto de que os adjetivos têm a mesma polaridade dos seus sinónimos. Baseando-se nesta ideia, o processo utilizado pelos autores consistiu em criar manualmente uma pequena *seed list* de adjetivos e, recorrendo ao dicionário WordNet foram detetados os sinónimos dos adjetivos anotados. Por fim, os sinónimos detetados foram acrescentados à *seed list* inicial.

A desvantagem da utilização de dicionários para a deteção do sentimento é que este método não tem em conta o domínio em estudo (Liu & Zhang, 2012). Nesse sentido, observe-se

[4] https://wordnet.princeton.edu/

uma frase no domínio de máquinas de sumos “esta máquina tem um motor lento que permite preservar as enzimas das frutas e legumes” e, uma frase no domínio de operadores de Internet “este operador fornece um serviço de Internet lento”. No domínio da primeira frase, a palavra “lento” está a associada a um sentimento positivo ao passo que, no domínio da segunda está associada a um sentimento negativo.

2.3.2.4 *Corpus-based*

O método *corpus-based* é utilizado quando é pretendido utilizar um léxico de palavras dedicado a um domínio específico (Feldman, 2013). Neste método, os léxicos são frequentemente criados a partir de regras sintáticas, padrões de ocorrência e pequenas *seed lists* (Caro & Grella, 2013; Liu & Zhang, 2012). Para a construção de léxicos dependentes do domínio são muitas vezes tidas em conta técnicas de processamento natural da língua, as quais permitem que a análise textual seja feita a um nível mais profundo contemplando tanto o contexto dos documentos como as suas polaridades (Caro & Grella, 2013).

Qiu, Liu, Bu, & Chen (2011) utilizaram uma abordagem de *bootstraping* através da construção de regras de propagação para criar um léxico de palavras sobre avaliações de produtos. Para o desenvolvimento do algoritmo, os autores apenas precisaram previamente de uma pequena *seed list* e um analisador sintático (*parser*). O *parser* foi utilizado para analisar as relações sintáticas das frases com o objetivo primário de suportar a construção das regras de propagação. Por exemplo, na frase “o computador tem um bom processador” o *parser* conseguia identificar que o adjetivo “bom” modificava o substantivo “processador” e, desse modo, os autores conseguiram criar regras tais como: “se um substantivo é modificado através de um adjetivo então o substantivo corresponde a um aspeto”. A *seed list* foi utilizada para encontrar as expressões sentimentais e os aspetos a que estas se referiam no conjunto de dados a analisar. De seguida, através das regras construídas, o algoritmo detetava: novos aspetos associados às expressões sentimentais da *seed list,* novos aspetos associados aos aspetos extraídos, novas expressões sentimentais associadas aos novos aspetos e, novas expressões sentimentais associadas às da *seed list* e às novas expressões sentimentais detetadas. À medida que as expressões sentimentais e os aspetos eram detetados eram acrescentados à *seed list*.

Existem vários léxicos disponíveis publicamente para uso, por exemplo Mostafa (2013) utilizou o léxico criado por Hu & Liu (2004) e aplicou-o no seu caso de estudo pois verificou que outros autores, com investigações semelhantes à sua, tinham obtido bons resultados através da utilização do mesmo léxico. No entanto, é muito difícil construir e manter um léxico universal que contemple os vários domínios que possam existir pois, como já explanado, uma palavra pode ser positiva num domínio e negativa noutro (Qiu et al., 2011). Adicionalmente, a adoção de léxicos disponibilizados para classificação de sentimentos por si só é limitadora pois, por muito que os léxicos até possam contemplar expressões relacionadas com o domínio em causa, dificilmente contemplam todas as expressões existentes (Liu & Zhang, 2012) e, nesse sentido é importante expandir os léxicos já existentes específicos de cada domínio (Qiu et al., 2011).

2.3.3 Níveis de Análise de sentimentos

A AS pode ainda ser efetuada a vários níveis ou seja, pode ser aplicada a documentos completos (nível do documento), aplicada a cada uma das frases do documento (nível da frase) ou aos vários aspetos mencionados ao longo do documento (nível do aspeto) (Feldman, 2013). Assim cada um destes será explanado seguidamente.

2.3.3.1 Nível do documento

A análise a nível do documento classifica a polaridade de um documento como um todo, por exemplo a polaridade de uma *review* sobre um produto ou serviço (Liu & Zhang, 2012). Quando é feita uma análise a nível de documento assume-se que a opinião é expressa apenas por um detentor de opinião e sobre uma única entidade (Feldman, 2013; Liu & Zhang, 2012). Tipicamente ao serem analisadas *reviews* sobre produtos ou serviços, o autor dá a sua opinião pessoal sobre um produto em concreto. Contudo, por exemplo ao serem analisadas notícias podem, num único documento, estar envolvidos vários detentores de opiniões ou, no caso de *blogs* o autor pode estar a escrever uma *review* comparativa sobre dois produtos diferentes.

A nível do documento, Zhang et al. (2011) classificaram *reviews* em cantonês sobre restaurantes. Os autores utilizaram uma abordagem de *machine learning* e, para contruírem o conjunto de dados, pediram a dois nativos para etiquetar manualmente as *reviews* de acordo com a sua polaridade considerando como unidade básica toda a *review*.

2.3.3.2 Nível da frase

Quando um único documento possui várias opiniões sobre várias entidades é importante ir a um nível mais detalhado na análise (Feldman, 2013). Assim, é importante descer de uma análise a nível de documento para uma análise a nível frásico. Algumas investigações a nível frásico examinam duas problemáticas nomeadamente: classificação de subjetividade e a classificação de sentimentos das frases subjetivas (Feldman, 2013; Liu & Zhang, 2012). Investigações que recorrem a esta abordagem, após detetarem a subjetividade das frases, descartam as frases objetivas e concentram-se apenas na deteção da polaridade das frases subjetivas. Existem ainda trabalhos que apenas se focam numa das duas tarefas.

Pang e Lee (2004) e Turney (2002) nas suas investigações detetaram inicialmente a subjetividade das frases e apenas analisaram a polaridade de frases subjetivas. Por sua vez, Caro e Grella (2013) analisaram apenas a polaridade de opiniões no serviço da restauração propondo um algoritmo com regras de propagação em que, a cada termo frásico era associado um *score* de polaridade que era propagado segundo a estrutura sintática da frase. Os autores criaram manualmente um pequeno léxico a partir de cem *reviews* considerando nomes, verbos e modificadores de sentido. De seguida o modelo assumia que, segundo diversas regras, cada um dos componentes da frase influenciava os valores dos componentes seguintes. Por exemplo, na frase “o restaurante tem uma má atmosfera”, o elemento “atmosfera” tem associado um valor de sentimento que é influenciado pelo modificador “má”, neste caso a palavra “atmosfera” é influenciada por um valor negativo proveniente do modificador de sentido “má”. Desta forma, o valor de sentimento de cada frase era calculado a partir da propagação dos valores de cada um

dos componentes da mesma bem como da influência que cada um destes exercia no seguinte, isto permitia que toda a sequência da frase fosse tida em conta em vez de classificar os *reviews* apenas como palavras soltas.

2.3.3.3 Nível do aspeto

Análises a nível da frase, embora sejam a um nível mais detalhado do que as análises a nível do documento, podem por vezes ser incompletas (Liu & Zhang, 2012). As análises a nível frásico têm geralmente um bom desempenho quando a frase apenas expressa a opinião sobre um aspeto. Porém, muitas vezes, as frases têm um nível de complexidade maior sendo que os autores dão opiniões sobre mais que um aspeto na mesma frase. Por exemplo, na frase "o meu telemóvel tem uma câmara espetacular e um desempenho fantástico mas, as colunas têm uma qualidade de som péssima", embora a maioria do comentário até possa exprimir um sentimento positivo, o autor refere-se a aspetos positivos e aspetos negativos. A deteção de sentimento a nível frásico é bastante útil porém, ignorar a polaridade dos aspetos leva a uma perda de informação enormíssima (Feldman, 2013). Assim, de maneira a detetar informação mais detalhada sobre os vários atributos é necessário analisar-se o sentimento ao nível dos aspetos. Para tal, é importante detetar inicialmente os vários aspetos que estão a ser avaliados pelo detentor de opinião e seguidamente avaliar a polaridade sobre cada uma das avaliações (Liu & Zhang, 2012).

Cruz, Troyano, Enríquez, & Vallejo (2013) criaram um sistema de extração de opiniões a nível do aspeto recorrendo a uma abordagem taxonómica. Foram explorados comentários sobre três domínios diferentes (*headphones,* hotéis e carros) e, para cada um destes foi criada uma taxonomia de aspetos. Cada taxonomia tinha na sua raiz o próprio produto (por exemplo o *headphone)* e, cada nível da taxonomia correspondia aos aspetos sobre o mesmo (por exemplo a qualidade de som, o aspeto, entre outros). O objetivo era que o sistema conseguisse detetar os vários aspetos existentes em todas as frases de todos os comentários dos diferentes domínios e, por fim, mapeasse as várias opiniões nas taxonomias de aspetos criadas.

2.3.4 *Web services*

Além dos métodos tradicionais supracitados, atualmente existem já vários *Web services* para análise de sentimentos tais como o AlchemyAPI[6], o Lymbix[7], o Repustate[8] e, o Semantria[9]. Segundo alguns estudos o AlchemyAPI e o Semantria são considerados dos *Web Services* mais estáveis e com melhores resultados (Gao, Hao, & Fu, 2015; Peisenieks & Skadiņš, 2014; Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015).

No desenvolvimento desta dissertação será também utilizado um *Web service* para a análise automatizada de sentimentos. O AlchemyAPI embora tenha apresentado muito bons resultados em investigações anteriores, a sua versão de teste gratuita, impõe uma análise

[6] http://www.alchemyapi.com/
[7] http://www.lymbix.com/
[8] https://www.repustate.com/
[9] https://www.lexalytics.com/

máxima de 1000 documentos diários. Assim, o *Web service* a utilizar será o Semantria pois permite a análise de 15000 documentos por conta de utilizador.

O Semantria é um *Web service,* criado pela Lexalytics, que recorre a técnicas de PNL e algoritmos de *machine learning* para classificar sentimentos positivos, negativos e neutros. Para tal, o Semantria tem um *plug-in* para o Excel que permite analisar múltiplas linhas com vários documentos de texto. Ao longo do processo de classificação de sentimentos é detetado o sentimento de várias instâncias tais como: expressões de sentimento, entidades, temas e, categorias (Lexalytics, 2015a, 2015b, 2015c). As expressões de sentimento são normalmente combinações de nomes-adjetivos por exemplo "*solid account"* ou "*public safety*"; As entidades dizem respeito a nomes próprios por exemplo uma pessoa ("Steve Jobs"), um local ("*Denmark*"), um produto ("*iPhone*"), etc., segundo a Lexalytics (2015d) as entidades representam "sobre quem se está a falar"; Por sua vez, os temas, representam conceitos mais genéricos, através de nomes que não sejam próprios por exemplo "*international student*". Segundo a Lexalytics (2015b) os temas identificam "sobre o que é que se está a falar"; As categorias representam conceitos mais alargados do que os temas, estando associadas a várias palavras-chave. Por exemplo, uma das categorias do Semantria chama-se "*Aviation*" e tem associado as palavras "*airplane", "flying"* e "*aviation*".

O processo de classificação de sentimentos por documento passa pelas seguintes fases (Lawrence, 2014; Lexalytics, 2015c; Williamson & Ruming, 2015): (1) segmentação de termos e classificação da sua classe gramatical, (2) Identificação de expressões de sentimentos, (3) Ponderação do *score* de sentimento de cada expressão de sentimento através de uma escala logarítmica entre -10 e 10, (4) Atribuição do sentimento global do documento através da ponderação dos *scores* das várias expressões de sentimento dos documento.

No que diz respeito à análise de sentimentos das entidades, temas e categorias são também efetuadas ponderações consoante a expressão de sentimento associada (Lexalytics, 2013, 2015c). O Semantria tem ainda em conta a análise de negações, *emoticons*, emojis, acrónimos (por exemplo "OMG" e "LOL") e, *hashtags*.

No que diz respeito às transformações de dados e aos modelos utilizados, o Semantria funciona como uma *black box* (Figura 3) (Lawrence, 2014; Lexalytics, 2015e). Ou seja, apenas é possível visualizar os documentos de *input* e os resultados obtidos no *output* sendo que, sobre o funcionamento interno quer das transformações aos dados quer dos modelos do *plug-in* não existe informação. No entanto, sabe-se que o Semantria recorre a perto de 40 algoritmos distintos e a grandes léxicos de palavras (Lexalytics, 2015e).

Figura 3 Processo de funcionamento do Semantria (fonte: adaptado de Lawrence (2014))

Após a classificação das várias instâncias detetadas pelo Semantria, a atribuição de sentimentos positivos, negativos e neutros correspondem aos intervalos de *scores* apresentados na Tabela 2.

Tabela 2 Scores de sentimento do Semantria

	Negativo	Neutro	Positivo
Documentos e Expressões	[-2,-0.05[	[-0.05,0.22]	]0.22,2]
Entidades, Temas e Categorias	[-10, -0.45[	[-0.45,0.5]	]-0.45,10]

2.3.5 *Text mining* e análise de sentimentos no setor da educação

No setor da educação foi já bastante reconhecida a importância da utilização de técnicas automatizadas para análise de padrões em grandes quantidades de dados. Vários projetos de *data mining* começaram a ser desenvolvidos na educação tendo mesmo surgido uma área dedicada exclusivamente a este efeito apelidada de *Educational Data Mining* (EDM) (He, 2013; C. Romero & Ventura, 2007).

A EDM parte da análise de dados provenientes de ambientes educacionais (aulas tradicionais presenciais, plataformas de cursos online, entre outros) com o objetivo de extrair conhecimento que permita melhorar os ambientes em que os alunos estudam bem como os processos de aprendizagem dos mesmos (Cristobal Romero & Ventura, 2013).

Embora grande parte do trabalho existente no âmbito da área EDM sejam aplicações de *data mining,* existem já bastantes aplicações em *text mining.* He (2013), por exemplo, analisou dois conjuntos de dados provenientes de um sistema educacional *online* em *streaming* de vídeo, ou seja, provenientes de uma plataforma onde as aulas eram dadas em tempo real por transmissão de vídeo *online.* O propósito da sua investigação consistia em fazer uma análise exploratória sobre o comportamento da participação dos alunos e correlacioná-lo com as notas finais dos mesmos. Os resultados da investigação demonstraram uma correlação positiva entre a quantidade de questões postas pelos alunos e as suas notas sugerindo que os alunos que fazem mais questões prestam provavelmente mais atenção às aulas levando-os a ter melhores notas.

Com o propósito de analisar a performance dos estudantes Goda, Hirokawa e Mine (2013) propuseram o método *Previous Current Next* (PCN). O método PCN implica que os alunos façam comentários sobre as suas atitudes, comportamentos e aprendizagem no final de cada aula. Os comentários deste método abordavam as atividades efetuadas para preparação da aula (P), os conhecimentos adquiridos ao longo da aula (C) e, os comentários sobre os objetivos a cumprir até a próxima aula (N). Numa primeira fase, os autores utilizaram uma análise de regressão múltipla para calcular *scores* dos vários comentários de maneira a perceber se os mesmos se enquadravam devidamente em cada um dos pontos P,C e N. Numa segunda fase, os comentários C foram utilizados para fazer uma previsão das notas finais dos alunos através de *support vector machines*. Sorour, Mine, Goda, & Hirokawa (2015) fizeram uma extensão da investigação feita pelos autores Goda et al. (2013) explorando o mesmo conjunto de dados através da técnica *Latent Semantic Analysis* (LSA) com intuito de explorar os motivos que melhor ajudam a prever determinada nota.

A análise de sentimentos na educação foi também já explorada e, embora se tenha notado uma menor quantidade de investigações neste âmbito, a maioria dos trabalhos encontrados são relativamente recentes sugerindo que a análise de sentimentos na educação encontra-se

atualmente a despertar maior interesse por parte dos investigadores. Wen, Yang, e Rosé (2014) investigaram como é que as opiniões dos alunos podiam estar relacionadas com desistência dos mesmos ao longo de cursos *Massive Open Online Course* (MOOC). Os MOOC são cursos gratuitos *online* em que qualquer pessoa pode inscrever-se para aumentar os seus conhecimentos. O conjunto de dados utilizado para a investigação compreendia comentários publicados nos fóruns de três MOOCS do *site* Coursera.org. Os autores inicialmente procuraram perceber a tendência dos sentimentos ao longo da duração dos cursos tendo verificado que rácios superiores de sentimento estavam associados a uma quantidade menor de desistências. Leong, Lee, e Mak (2012) criaram uma plataforma que recebe mensagens com as opiniões dos alunos sobre palestras, seminários, conferências e desenvolveram um método automatizado para analisar as opiniões dos alunos e assim otimizar a entrega do *feedback.* Na investigação os autores exploraram várias técnicas de *text mining* tais como *Part-of-speech tagging* (POS), *Stemming,* um modelo de correção de erros gramaticais, entre outras.

A grande maioria das investigações de análises de dados na educação foram no âmbito da EDM (Minami & Ohura, 2013). Os dados utilizados nos trabalhos desta área são provenientes de ambientes educacionais, tal como mencionado anteriormente. Existem ainda algumas investigações no âmbito da educação que recorrem a dados externos a sistemas educacionais. Isik, Öztaysi, e Fenerci (2012) recolheram comentários dos *social media* com o objetivo de perceber as opiniões sobre a Universidade Técnica de Istambul. Os autores inicialmente aplicaram técnicas de *clustering* de maneira a perceber o que era mais abordado sobre a universidade nos *social media* tendo detetado três *clusters* descritos como: “qualidade de educação”, “campus” e “reputação corporativa”. Seguidamente, através de um método estatístico e um modelo baseado em regras utilizando o *software Sentiment Analysis Studio of SAS*, os autores analisaram a polaridade das opiniões em cada um dos *clusters.*

Uma vez explorada a literatura subjacente ao estado da arte do *text mining* e da análise de sentimentos, na próxima secção serão exploradas as principais técnicas utilizadas para a descoberta de conhecimento em texto que servirão de base para o desenvolvimento prático no capítulo 4.

3 DESCOBERTA DE CONHECIMENTO EM TEXTO

Um processo de descoberta de conhecimento em texto passa pelo "refinamento do texto de forma a obter uma forma intermediária computacionalmente manejável representativa do texto, um procedimento de mineração para obter novo conhecimento e, uma última fase para avaliar o conhecimento obtido" (Sánchez et al., 2008). Contudo, o recurso a uma metodologia com um processo bem definido pode ajudar a obter resultados com uma qualidade superior (Miner et al., 2012, p. 73). Desse modo, antes de se abordar as técnicas elucidar-se-á a metodologia utilizada ao longo da restante investigação.

3.1 CRISP-DM

Muitos dos trabalhos de TM são desenvolvidos a partir de tentativa-erro e orientados segundo experiências pessoais e preferências (Miner et al., 2012, p. 73) e, embora existam algumas propostas de metodologias para descoberta de conhecimento em texto, estas ainda não foram comumente adotadas essencialmente devido à subjetividade inerente ao que envolve o *text mining* e, às características dos dados em causa (Miner et al., 2012, p. 74).

O *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) é um processo desenvolvido para projetos de *data mining.* Contudo, nos dias de hoje, é um dos processos *standard* mais utilizados para desenvolvimento de projetos de descoberta de conhecimento, incluindo conhecimento em texto (Marbán, Mariscal, & Segovia, 2009, p. 5; Miner et al., 2012, p. 74). Desse modo, considerou-se que seria adequada a sua orientação ao longo do desenvolvimento.

O CRISP-DM apresenta seis fases sendo que, dependendo dos resultados obtidos por fase deve ser decidida a próxima fase ou tarefa a ser feita (P. Chapman et al., 2000). Neste sentido, é sempre possível voltar atrás no processo de maneira a otimizar os resultados obtidos em cada fase (Figura 4).

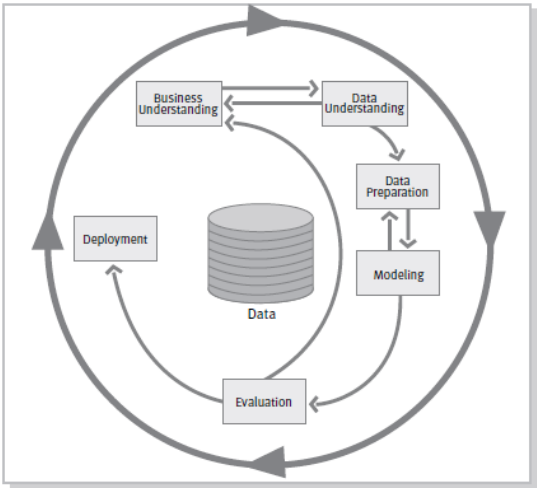


Figura 4 Fases do processo CRISP-DM (fonte: P.Chapman (2000))

A fase ***Business Understanding*** (compreensão do negócio) trata da compreensão do contexto de negócio e da tradução de objetivos num problema, neste caso, num problema de *text mining.*

Por sua vez, o ***Data Understanding*** (compreensão dos dados) é o momento em que é feito o acesso às fontes onde a informação está presente. Espera-se que, nesta etapa, seja feita uma exploração inicial dos dados que permita posteriormente um tratamento adequado aos mesmos

A fase ***Data Preparation*** (preparação dos dados) é uma das fases mais complexas e exigentes (Hotho, Nürnberger, & Paaß, 2005). Esta é a fase onde existe a principal distinção entre um processo de *data mining* tradicional e um

processo de TM. Num processo de TM espera-se que esta fase compreenda a aplicação de técnicas de processamento natural da língua que permitam estruturar o conteúdo textual de maneira a ser possível aplicar os tradicionais algoritmos de *machine learning* na fase de modelação (Hotho et al., 2005).

Na fase de ***Modeling*** (modelação) prefigura a seleção e aplicação de técnicas de modelagem apropriadas. Segundo P. Chapman et al. (2000) é natural retornar-se várias vezes à fase de preparação de dados para que sejam feitas alterações com o propósito de melhorar os resultados alcançados.

A fase ***Evaluation*** (avaliação) dedica-se à avaliação da qualidade do modelo (ou modelos) obtidos. Aqui é verificado se o modelo atingiu os objetivos definidos.

Por fim, a fase ***Deployment*** (implementação) diz respeito à apresentação dos resultados obtidos através de um relatório, apresentação ou até mesmo através da implementação dos modelos na própria organização (P. Chapman et al., 2000).

3.2 EXTRAÇÃO

Conforme as fases do CRISP-DM, depois da análise do contexto de negócio, dá-se início à fase de compreensão dos dados, a qual compreende o momento de extração dos mesmos. Para esta etapa, é fundamental ter uma estrutura que permita manusear documentos textuais e facilitar o uso de formatos heterogéneos (Sánchez et al., 2008). Assim, para que seja viável usar-se dados textuais a fim de os analisar, é criada uma base de dados onde está presente a coleção de documentos, a qual se dá o nome de *corpus* (Feinerer, Hornik, & Meyer, 2008). Os vários documentos podem, num momento inicial, estar em diferentes formatos desde ficheiros textuais, ficheiros XML, páginas *web,* etc. pelo que, precisam de ser transformados num formato único, por exemplo um ficheiro ASCII (Delen & Crossland, 2008). Assim, o *corpus* representa o conjunto da totalidade dos documentos a examinar que, após estar organizado num único documento, permite que seja possível proceder-se à preparação dos dados.

3.3 PREPARAÇÃO DOS DADOS

Após a fase de análise dos dados, segundo o CRISP-DM dá-se entrada na fase de preparação dos mesmos. Como já explicado, ao longo do processo de preparação dos dados, pretende-se partir de uma base de dados com informação não estruturada (o *corpus*) e atingir o momento em que essa informação se encontra numa base de dados estruturada (Delen & Crossland, 2008). Os documentos presentes no *corpus* no seu formato natural têm uma qualidade de conteúdo muito fraca como tal, é importante que estes passem inicialmente por uma fase de tratamento (Feinerer et al., 2008).

3.3.1 Tokenization

A primeira tarefa na fase de preparação de dados de um processo tradicional de *text mining* é, normalmente, a *tokenization* (Aranha & Passos, 2008; Webster & Kit, 1992). A *tokenization* tem em vista a segmentação do conteúdo textual em unidades de texto mais elementares (*tokens*) (Hassler & Fliedl, 2006). Um *token* (n-grama) é no fundo uma entidade (ou

termo) que pode ser representado como uma palavra, expressão ou frase e não pode ser decomposta em fragmentos mais pequenos (Miner et al., 2012, p. 13; Webster & Kit, 1992).

É importante notar que um *token* não necessita de ser exclusivamente um único termo pois, por vezes poderá fazer sentido que não sejam feitas determinadas separações (Hassler & Fliedl, 2006; Kaplan, 2005). Por exemplo a seguinte frase: "ontem estive a trabalhar perto das mesas de voto" poderia ser segmentada pelos vários unigramas existentes mas, seria também viável considerar os vários *tokens* como seis unigramas – "ontem", "estive", "a", "trabalhar", "perto", "das" – e um trigrama – "mesas de voto".

Após a *tokenization* a quantidade e variedade de termos obtida é imensa sendo indispensável uniformizar o texto e aplicar técnicas com vista à redução da dimensionalidade do *corpus* de maneira a reduzir ruído dos dados e o tempo de processamento das análises (Blake, 2011; Delen & Crossland, 2008; Feinerer et al., 2008)

3.3.2 Normalização dos termos

Muitos dos termos têm o mesmo significado embora estejam escritos de formas ligeiramente diferentes sendo, portanto, necessário uniformizá-los (Manning, Raghavan, & Schutze, 2008). Para tal, é importante ter em conta técnicas que lidem a normalização de acentuação e outros diacríticos, pontuação e maiúsculas.

O processo utilizado tipicamente em relação à acentuação passa por retirar os acentos das palavras contudo, é importante ter em conta a língua a analisar (Manning et al., 2008). Ao serem retirados os acentos de palavras inglesas, uma vez que as palavras inglesas com a acentuação tendem a ser estrangeirismos, estas permanecem com o mesmo significado. Porém, noutra língua que não seja a inglesa, o retirar de acentos pode resultar em palavras com significados diferentes.

No que diz respeito ao tratamento de pontuação e de outros diacríticos além da acentuação – por exemplo cedilhas, apóstrofos, hífens –, o processo consiste também em retirar os mesmos de todas as palavras, de forma a criar uma visão única sobre a palavra, por exemplo é considerado boa-prática a criação de uma visão única sobre palavras tais como "EUA" e "E.U.A.

Habitualmente a abordagem aplicada às letras maiúsculas passa por transformá-las em letras minúsculas (Manning et al., 2008), por exemplo em vez de existirem os termos "Good" e "good" passa apenas a existir um único termo "good". Apesar desta prática auxiliar na diminuição da variedade de termos no *corpus*, por vezes o significado das palavras pode alterar-se. Atente-se neste exemplo de língua inglesa, o acrónimo "C.A.T", após o tratamento de diacríticos passaria para "CAT" e, quando fosse feito o tratamento de maiúsculas a palavra passaria a ser "cat" que corresponde à palavra "gato", o que representaria uma alteração completa de sentido. Alguns trabalhos defendem que, na área de análise de sentimentos, pode ser interessante manter-se palavras que estejam completamente escritas em maiúsculas pois podem indicar uma maior intensidade de sentimento (Brooke, Tofiloski, & Taboada, 2009).

3.3.3 Part-of-Speech

O *Part-of-speech* (POS) é uma técnica que tem como fim etiquetar uma classe sintática a uma palavra (Jackson & Moulinier, 2007). No POS, cada palavra é avaliada, através de métodos estatísticos ou à base de regras, e é-lhe atribuída a sua categoria gramatical (Brill, 1992).

O POS pode ser bastante importante para selecionar o tipo de atributos que se quer analisar, ou seja, apenas fazer-se análises tendo em conta diferentes combinações de classes gramaticais tais como adjetivos; nomes; ou adjetivos, nomes e advérbios. Na área de análise de sentimentos trabalhos anteriores provaram que os adjetivos são bons indicadores da presença de opiniões e que auxiliam na deteção de sentimentos (Hatzivassiloglou & Wiebe, 2000; Wiebe, Bruce, Bell, Martin, & Wilson, 2001). Contudo, os adjetivos não devem ser as únicas categorias sintáticas a ter em conta (Pang & Lee, 2008), uma vez que existem também estudos que comprovaram com sucesso que a utilização de outras classes fortalecem a deteção do sentimento tais como: Turney (2002) que utilizou diversos padrões que associavam adjetivos com nomes, adjetivos com advérbios e advérbios com verbos de forma a classificar *reviews* de quatro domínios diferentes; Hu & Li (2004b) ao analisar sentimentos detetaram inicialmente os substantivos (nomes) com o propósito de perceber as opiniões sobre os vários aspetos abordados nos comentários.

3.3.4 Tratamento da negação

Na base de dados estruturada que se pretende atingir, no final da fase de preparação de dados, o conteúdo textual é representado através dos vários *tokens* e da sua frequência. No entanto, os *tokens* ao serem estruturados, ficam completamente independentes uns dos outros sendo apenas um conjunto de várias palavras cuja ordem é irrelevante (*bag-of-words*) (Munzert, Rubba, Meißner, & Nyhuis, 2015). Face a esta questão, a existência de negações nos documentos pode representar um obstáculo para a classificação da polaridade dos sentimentos.

Considere-se a seguinte frase "*This university isn't a good place to study*". Os vários *tokens* desta frase são por exemplo: "This", "university", "isn't", "a", "good", "place", "to", "study". Uma vez que os *tokens* se tornam independentes e com uma ordem irrelevante, apenas o termo isn't fica identificado como uma entidade negativa ao passo que os restantes termos da frase ficam identificados como uma entidade positiva (ou pelo menos neutra). De modo a evitar que isto aconteça, a literatura sugere que, quando surjam nos documentos termos que sugiram uma negação, seja aplicado um prefixo "NOT_" após o primeiro termo negado até ao sinal de pontuação seguinte (Das & Chen, 2007; Pang, Lee, & Vaithyanathan, 2002). Com este tratamento, os *tokens* da frase anterior ficaram da seguinte forma: "This", "university", "isn't", "NOT_a", "NOT_good", "NOT_place", "NOT_to", "NOT_study". Assim, apesar da independência dos termos, a partir do termo *isn't,* todos os restantes sofrem o impacto da negação. Adicionalmente, devem também ser tidas em conta conjunções adversativas como "but", "however" pois estas, à semelhança dos sinais de pontuação, podem invalidar novamente a negação caso não exista nenhuma pontuação anteriormente (Liu & Zhang, 2012).

3.3.5 Redução de dimensionalidade

Além das técnicas mencionadas na normalização de termos, outras das mais utilizadas na literatura para a redução de ruído e dimensionalidade passam pela remoção de *sites html* e numeração (Das & Chen, 2007; Guerreiro, Rita, & Trigueiros, 2015; Lin, He, Everson, & Rüger, 2012). Adicionalmente a remoção de *stopwords* assim como o *stemming* são também essenciais para minimizar este problema (Feinerer et al., 2008).

As ***stopwords*** caracterizam-se por ser termos vulgares que aparecem frequentemente no texto mas que o valor de informação fornecido pelas mesmas é muito baixo (Feinerer et al., 2008). São exemplos de *stopwords* preposições, determinantes, pronomes tais como: "de", "a", "o", entre outros. Adicionalmente existem palavras com pouco poder diferenciador próprias do domínio de estudo. Os termos que não trazem qualquer valor para a análise devem ser removidos do *corpus* pois, caso contrário, criarão ruído no restante conjunto de dados (Delen & Crossland, 2008). As *stopwords* podem ser detetadas através da análise dos termos mais frequentes no conjunto de dados ou recorrendo a listas específicas de *stopwords* do domínio em estudo (Blake, 2011).

A técnica ***stemming*** (Porter, 1980) é um método heurístico que tem como propósito identificar e reduzir a palavra ao seu *stem* através da remoção do seu sufixo. Quando o conteúdo não-estruturado a analisar está escrito em inglês, o algoritmo de Porter (1980) é um dos mais utilizados para executar este processo. Esta técnica identifica palavras que, apesar de não terminarem da mesma maneira, o seu significado é similar. No entanto, o *stem* não necessita de ser uma palavra que exista realmente e que faça sentido (Jackson & Moulinier, 2007). Por exemplo, ao ter em conta as palavras: "maravilhosamente", "maravilhoso", "maravilhas" verifica-se que todas elas têm como *stem* a palavra "maravilh". Esta técnica permite que a complexidade da análise seja reduzida uma vez que é possível generalizar sobre um determinado termo quando a raiz é a mesma sem que haja uma perda de informação severa (Feinerer et al., 2008; Ruch et al., 2006).

3.3.6 Document-by-term matrix

A *document-by-term matrix* (DTM) representa a base de dados estruturada que se pretende atingir ao longo da fase de preparação de dados. Nela está representado todo o *corpus* num formato estruturado, onde cada coluna representa um *token,* cada linha representa um documento e, cada célula de cruzamento indica a frequência do *token* no documento a que se refere (Tabela 3). A DTM é provavelmente a forma mais comum de *input* para análises de classificação textual (Feinerer et al., 2008; Munzert et al., 2015).

Tabela 3 Term-by-document matrix

	student	*teach*	*intern*	*good*
Doc 1	2	3	1	1
Doc 2	1	0	4	2
Doc 3	0	5	2	4

A DTM representa o *corpus* como um *bag-of-words*, ou seja, os termos são um conjunto de palavras soltas onde a sua ordem é irrelevante pelo que é impossível, a partir da DTM, saber-se a ordem de cada termo na frase (Feinerer et al., 2008; Munzert et al., 2015).

Frequentemente a maioria dos cruzamentos de termos por documento é igual a zero o que leva a matriz a ser altamente esparsa. De modo a tentar reduzir a esparsidade da DTM podem ser definidas frequências mínimas para que um *token* seja considerado válido (Munzert et al., 2015, p. 306) e ainda um número mínimo de caracteres (Grün & Hornik, 2011).

A esparsidade da matriz pode ainda ser minimizada através da alteração do peso de cada termo, utilizando por exemplo a *term-frequency-inverse document frequency* (TF-IDF), e consequente remoção dos termos com uma TF-IDF baixa (Grün & Hornik, 2011). A matriz apresentada na Tabela 3, tem os seus índices calculados a partir do cálculo das frequências dos termos por documento (*term frequency*) levando a que termos mais frequentes sejam considerados mais importantes. No entanto, palavras como “and” e “the” são muito frequentes mas pouco importantes dada a sua frequência. Nesse sentido, é importante uma medida que avalie a raridade dos termos *inverse document frequency* (IDF). A IDF avalia a palavra analisando quão raramente esta surge na totalidade dos documentos (Delen & Crossland, 2008). Posto isto, uma matriz que pondere os pesos dos termos através da TF-IDF irá valorizar termos mais raros. A TF-IDF faz uma ponderação entre a TF e a IDF para que termos que são muito frequentes num documento mas tenham uma baixa frequência na coleção total de documentos sejam associados a um peso mais alto com maior importância (Feinerer et al., 2008). Para tal, a TF-IDF utiliza a seguinte fórmula para cada palavra *i* e o documento *j* (Delen & Crossland, 2008):

$$idf(i,j) = \begin{cases} 0 \; se \; wfij = 0 \\ (1 + \log(wfij)) \log \frac{N}{dfi} \; se \; wfij \geq 1 \end{cases} \quad (1)$$

wf: Frequência de palavras
df: Frequência de documentos
N: total de documentos

3.4 MÉTODOS NÃO-SUPERVISIONADOS E SUPERVISIONADOS

Uma vez construída a DTM, as primeiras análises podem ser efetuadas recorrendo a métodos não-supervisionados e/ou supervisionados.

3.4.1 *Métodos não-supervisionados*

Tal como explanado anteriormente, os métodos não-supervisionados são aplicados em dados que não se encontram categorizados por nenhuma classe à partida, assim, o seu propósito recai sobre a deteção de padrões que são posteriormente etiquetados segundo as suas características (Blei, Griffiths, & Jordan, 2010). A investigação sobre métodos não-supervisionados desta dissertação recai essencialmente sobre a modelação de tópicos uma vez que pretende recorrer-se a esta técnica no desenvolvimento prático.

A área da modelação de tópicos surgiu da necessidade de tentar reduzir a diversidade de pesquisa dos leitores sendo que, tem como propósito analisar as palavras dos documentos e, através de métodos estatísticos, descobrir os vários assuntos/tópicos subjacentes (Chaney & Blei, 2012). Um dos modelos de tópicos mais utilizados pela literatura com este objetivo é o

Latent Dirchlet Allocation (LDA) e o *Correlated Topic Model* (CTM), técnica subsequente ao LDA (Cheng, Yan, Lan, & Guo, 2014; Guerreiro et al., 2015).

O racional do LDA e do CTM (Blei & Lafferty, 2006), parte do pressuposto de que todos os documentos contêm múltiplos tópicos, ou seja, um documento pode por exemplo ser sobre "análise de sentimentos na educação" e, como tal, contemplará termos relacionados com educação, com *text mining* e com análise de sentimentos. Assim, os termos do documento deverão sugerir a existência de pelo menos três tópicos nomeadamente a "educação", *"text mining"* e "análise de sentimentos".

Nestes modelos, um tópico corresponde a uma distribuição fixa do vocabulário existente nos documentos e, cada um dos termos do vocabulário encontra-se associado a uma probabilidade. Por sua vez, cada documento tem uma distribuição de tópicos. Os tópicos são transversais à totalidade dos documentos pois, todos os tópicos contêm todas as palavras presentes na coleção de textos. O que difere de documento para documento são as proporções de cada tópico e, o que difere de tópico para tópico é probabilidade associada a cada um dos termos.

No seu funcionamento, o LDA recorre à distribuição *Dirichlet,* no entanto, devido às características da mesma, os termos dos tópicos são tratados como independentes e sem qualquer correlação estatística entre si (Blei & Lafferty, 2007). Neste sentido, o CTM (Blei & Lafferty, 2006) colmata a lacuna supracitada operando do mesmo modo porém, em vez de recorrer à distribuição *Dirichlet,* recorre à distribuição logística normal o que permite detetar correlações entre os tópicos.

Os algoritmos de modelação de tópicos tradicionais (como o LDA e o CTM) baseiam-se na deteção dos padrões que geram as palavras de cada documento. O funcionamento destes algoritmos tem demonstrado bons resultados quando são feitas análises a documentos grandes com bastante conteúdo, tais como artigos de notícias ou científicos (Guerreiro et al., 2015). Contudo, em textos mais pequenos e, principalmente, textos pequenos na *Web*, a deteção de padrões por documento torna-se um problema devido à co-ocorrência de termos ser muito escassa (Yan, Guo, Lan, & Cheng, 2013). Ou seja, a co-ocorrência de termos em documentos grandes permite que sejam mais facilmente percebidas as relações latentes no entanto, em pequenos documentos, uma vez que existem poucos termos, esta análise torna-se um obstáculo mais desafiante.

Assim, de maneira a minimizar o efeito da escassez dos termos, o *Biterm Topic Model* (BTM) (Yan et al., 2013) traz vantagens na análise de texto com conteúdos menores comparativamente com o LDA ou o CTM. O BTM deteta os tópicos a partir da modelação direta da co-ocorrência das palavras tendo em conta a totalidade dos documentos. As palavras dos documentos são todas agregadas o que permite criar padrões mais consistentes. Ou seja, métodos como os tradicionais esforçam-se por modelar a geração de documentos ao passo que o BTM preocupa-se em gerar a modelação de palavras. Na verdade, o BTM modela a correlação de pares de palavras, os chamados *biterms.*

Um *biterm* corresponde a "um par de palavras desordenado que co-ocorre num contexto pequeno" (Yan et al., 2013, p. 1447). Por exemplo, na frase "*I loved the university campus*", os *biterms* extraídos seriam "*loved university*", "*loved campus*", "*university campus*", excluindo as *stopwords* "*I*" e "*the*". Assim, um documento e os seus *biterms* seriam representados, por exemplo, através da seguinte forma em que cada *w* corresponde a um termo:

$$(w_1, w_2, w_3) \Rightarrow \{(w_1, w_2), (w_2, w_3), (w_1, w_3)\}$$

Figura 5 Exemplo de documento transformado em biterms (Fonte: Cheng et al. (2014))

A ideia é que duas palavras que ocorrem juntas frequentemente estão correlacionadas e como tal, em princípio, devem pertencer ao mesmo tópico (Yan et al., 2013). Assim, inicialmente todos os *biterms* são extraídos por documento e, o *corpus* passa a ser apenas um conjunto de *biterms*.

Para a explicação do funcionamento do BTM, considere-se N_b *biterms, K* tópicos constituídos por *W* palavras únicas e, Z como uma variável indicadora do valor de K ou seja, z ∈ [1, K]. θ é uma distribuição *K*-dimensional multinomial que representa a predominância de cada tópico sobre a totalidade da coleção. Por sua vez, $\emptyset_k$ é uma linha de uma matriz Φ de *K* × *W* com uma distribuição *W*-dimensional multinomial, que representa a predominância das palavras por tópico. Adicionalmente, à semelhança do LDA, o BTM recorre à distribuição Dirichlet para θ e Φ utilizando α e β como parâmetros iniciais.

Posto isto, os *biterms* são gerados a partir do seguinte processo generativo (Cheng et al., 2014; Yan et al., 2013):

1. É definida uma distribuição de tópicos $\theta \sim$ Dirichlet (α) para toda a coleção.
2. Para cada tópico $k \in [1, K]$ é definida uma distribuição de palavras $\emptyset_k \sim$ Dirichlet (β).
3. Para cada *biterm* b_i do conjunto de dados B.
    a. É extraído um $z_i \sim$ Multinomial (θ)
    b. São extraídas duas palavras $w_{i,1}, w_{j,2} \sim$ Multinomial ($\emptyset_{zi}$)

Assim, a probabilidade conjunta de um *biterm* é calculada tendo em conta o somatório dos $P(b_i)$ de todos os tópico:

$$\begin{aligned} P(b_i|\boldsymbol{\theta}, \boldsymbol{\Phi}) &= \sum_{k=1}^{K} P(w_{i,1}, w_{i,2}, z_i = k|\boldsymbol{\theta}, \boldsymbol{\Phi}). \\ &= \sum_{k=1}^{K} P(z_i = k|\theta_k) P(w_{i,1}|z_i = k, \phi_{k,w_{i,1}}) \cdot \\ &\quad P(w_{i,2}|z_i = k, \phi_{k,w_{i,2}}) \\ &= \sum_{k=1}^{K} \theta_k \phi_{k,w_{i,1}} \phi_{k,w_{i,2}}. \end{aligned} \tag{2}$$

Figura 6 Probabilidade condicionada dos biterms (Fonte: Cheng et al., (2014))

Para que seja mais claro, a Figura 7 apresenta o funcionamento do LDA (a) e o funcionamento do BTM (b). No LDA, primeiro, é escolhida uma distribuição de tópicos θ_d. De seguida, para cada palavra escolhe-se aleatoriamente um tópico *z* da distribuição de tópicos do documento e, por fim, extrai-se uma palavra *w* da distribuição de termos do tópico *z*. O problema é que a distribuição de tópicos do LDA é feita por documento, o que leva a que o tópico *z* dependa das restantes palavras presentes nesse documento (Yan et al., 2013). Num contexto grande, este método pode ser eficaz, contudo num contexto pequeno, este método não é o ideal porque a quantidade de palavras é pouca. No BTM, este problema não acontece uma vez que, não é feita uma distribuição de tópicos por documento θ_d, mas sim uma distribuição de tópicos global θ sobre todo o conjunto de biterms.

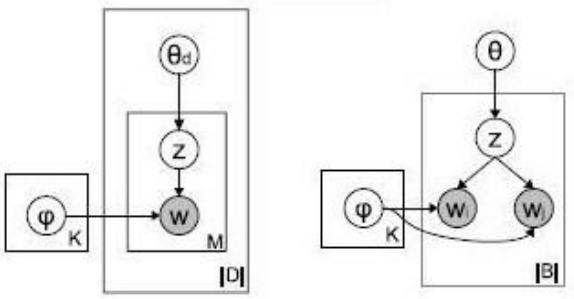


Figura 7 Esquema do funcionamento do LDA (a) e do BTM (b) (Fonte: adaptado de(Cheng et al., 2014))

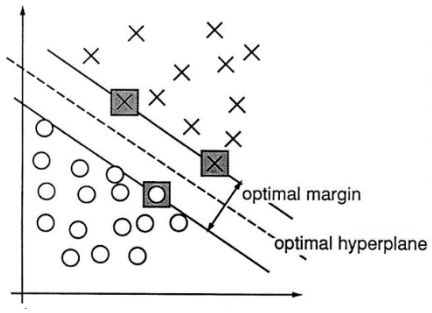
3.4.2 *Métodos* Supervisionados

Alguns dos algoritmos supervisionados mais frequentemente utilizados são *Support Vector Machines* e as *Naive Bayes* (Rushdi Saleh et al., 2011; Zhang et al., 2011). Neste sentido, uma vez que serão utilizados na secção seguinte, cada um dos algoritmos será sucintamente explanado de seguida.

3.4.2.1 *Support Vector Machines*

As *Support Vector Machines* (SVM) são utilizadas com o propósito de fazer classificações entre duas classes (Cortes & Vapnik, 1995). Para tal, as SVM ambicionam encontrar o melhor hiperplano que separa ambas as classes. Entende-se por melhor hiperplano (hiperplano ótimo) aquele que maximiza a margem entre as classes sendo esta limitada por casos particulares do conjunto de treino, aos quais se dá o nome de vetores de suporte (Cortes & Vapnik, 1995). A Figura 8 apresenta o hiperplano ótimo, a margem que separa as duas classes bem como os vetores de suporte de ambas as classes.

No entanto, tipicamente as classes não são inteiramente separáveis pelo hiperplano, o que faz com que seja necessário atenuar a rigidez da margem que as separa. É, então, necessário que haja um balanço entre a quantidade de erro permitida e a largura da margem que separa as classes, mesmo que isso implique uma perda de precisão no modelo (Pontil & Verri, 1998). A Figura 9 apresenta uma situação em que foi necessário aumentar-se a margem e permitido que fosse incorporado mais erro no modelo.



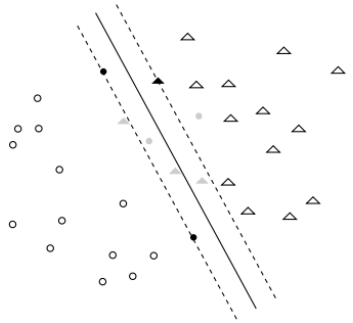


Figura 8 Hiperplano SVM (Fonte: Cortes e Vapnik (1995)) *Figura 9 Margens SVM (Fonte: Pontil e Verri (1998))*

3.4.2.2 *Naive Bayes*

As Naive Bayes (Lewis, 1998) são um algoritmo de classificação baseado no teorema de Bayes associado ao pressuposto da independência da ocorrência de atributos.

$$P(Ck|\vec{X}) = \frac{P(Ck)\times P(\vec{X}|Ck)}{P(\vec{X})} \quad (3)$$

A variável $\vec{X}$ representa um vetor de frequências de termos de cada documento *(d). C_k* representa uma das possíveis classes em que o documento pode ser classificado e, em que a ocorrência de cada X_j é independente da ocorrência de outro X_j dado um documento classificado como C_k (Lewis, 1998). A probabilidade condicionada $P(C_k|\vec{X})$ pretende identificar qual a classe (C_k) de um determinado documento dado um vetor ($\vec{X}$) (Lewis, 1998). Para tal, inicialmente calcula-se $P(C_k)$ indicativo da probabilidade de uma certa classe se verificar (Mitchell, 1997). Por exemplo, ao avaliar a opinião de diversos documentos e, considerando duas classes possíveis "Positiva" ou "Negativa" é calculada a $P(C_k$=Positiva) bem como a $P(C_k$=Negativa). De seguida, estima-se a probabilidade condicionada $P(\vec{X}|C_k)$. Contudo, devido à dificuldade de computar esta função assume-se que $P(\vec{X}|C_k)=\prod_{j=1}^{d} P(Xj|Ck)$, ou seja, é estimado o produto das probabilidades condicionadas de cada uma palavra presente num documento (d) (Mitchell, 1997).

4 METODOLOGIA

4.1 COMPREENSÃO DO NEGÓCIO

A partir da revisão de literatura compreendeu-se a importância das IES analisarem os comentários partilhados no meio *online* para fazerem tomadas de decisão mais acertadas de acordo com o público-alvo, neste caso os alunos internacionais, bem como conseguirem uma melhor gestão da reputação. Desse modo, serão ao longo desta secção analisados, através de técnicas de *text mining* e análise de sentimentos, os comentários de uma plataforma *online* – *iagora*[11] – onde alunos que participaram num processo de mobilidade internacional escrevem para partilhar a sua experiência.

O *iagora* funciona como uma espécie de um questionário *online* onde é pedido aos alunos que comentem a sua experiência internacional segundo seis aspetos gerais nomeadamente: *Housing* (habitação), *Student Life* (vida estudantil), *Academic* (aspetos académicos), *Learning Language* (idioma de aprendizagem), *Expenses* (despesas), *Overall* (experiência geral) e *Final Comments* (considerações finais). Adicionalmente é também pedido aos alunos que classifiquem qualitativa e quantitativamente, de acordo com escalas de *Likert,* certas características mais específicas associadas a cada um dos seis aspetos gerais tendo em conta a sua satisfação.

A Figura 10 apresenta um excerto de uma *review* do *iagora,* nomeadamente do aspeto *Overall.* No aspeto *Overall* é pedido ao aluno que escreva livremente (retângulos azuis) sobre o que desejaria ter sabido antes de fazer a experiência (retângulo 2) e recomendações pessoais (retângulo 4). Nos retângulos verdes é pedido ao aluno que dê classificações segundo uma escala de *Likert* quer ao campo *overall* em estrelas (retângulo 1) quer a vários sub-aspetos (retângulo 5). O retângulo 3, a amarelo, embora tenha conteúdo textual, não é escrito pelo aluno fazendo parte de um conjunto de possíveis respostas que o aluno escolhe como sendo a mais concordante com a sua opinião.

Figura 10 Excerto de uma review do iagora

Todos os aspetos gerais, excetuando o *Final Comments,* têm uma estrutura similar ao aspeto *Overall* possuindo campos de texto aberto, campos com resposta fechada à escolha e

[11] Http://www.*iagora*.com

alguns sub-aspetos para classificar de acordo uma escala de *Likert.* O aspeto *Final Comments* tem apenas um campo de texto aberto em que os alunos escrevem livremente algumas considerações finais que tenham ainda a dizer, não tendo este aspeto qualquer tipo de classificação.

É importante notar que as classificações dadas em estrelas (retângulo 1) quer ao aspeto *Overall* quer a cada um dos demais aspetos gerais, são calculadas a partir de uma média ponderada entre a classificação dada em estrelas (retângulo 1) e as classificações dadas aos vários sub-aspetos (retângulo 5).

Pretende ao longo da investigação extrair-se os dados presentes nesta fonte com o intuito primário de explorar o seu conteúdo não estruturado e correlaciona-lo com a satisfação dos alunos gerando, assim, conhecimento que permita perceber quais os termos e fatores que determinam o que é mais valorizado pelos estudantes de mobilidade internacional aquando da sua experiência.

4.2 COMPREENSÃO DOS DADOS

4.2.1 Extração dos dados

O processo de extração de dados para a construção do *dataset* implicou o desenvolvimento de três diferentes *scripts* programados em *python* com recurso à biblioteca *beautiful soup* (Figura 11). A execução dos *scripts* para a extração ocorreu no dia 31/05/2015.

O **primeiro script** percorreu todas as páginas das 65 *business-school*s escolhidas extraindo os URL's de cada *review* individual para um csv. Uma vez que cada *review,* no seu URL, é identificada por um identificador que não é sequencial por instituição, foi necessário em primeira instância, aceder ao URL de cada uma das *business-schools* de forma a ter acesso a todas as *reviews* associadas. Tal foi implementado partindo do racional de que todas páginas das instituições continham a mesma estrutura[12] de URL. Desse modo, ao aceder à página de cada *business-school* e recorrendo aos links *"Read full review"* e *"next"* da mesma, o *script* identificou e extraiu todos os URL's das *reviews* de cada instituição armazenando-os diretamente num ficheiro csv.

Após a extração dos URL's das *reviews,* um **segundo *script*** acedeu aos mesmos guardando localmente o HTML de todas as *reviews.*

O **terceiro *script*** leu cada ficheiro HTML e processou o seu conteúdo através da biblioteca *beautiful soup*. A partir da inspeção do código detetaram-se as *tags* html que continham os dados que interessavam extrair. Desta forma, o script foi estruturado de maneira a respeitar a sequência pretendida para o ficheiro csv bem como de maneira a tratar de alguns casos especiais (por exemplo, algumas páginas estruturadas de forma diferente). Adicionalmente, foi feita uma primeira limpeza de caracteres especiais como "aspas" e mudanças de linha. Por fim, os dados foram armazenados de forma ordenada numa *string* separada por vírgulas, tendo sido esta

[12] Http://www.*iagora*.com/studies/uni/Nome_da_Instituição_de_Ensino

escrita no ficheiro csv. O processo repetiu-se iterativamente passando por todas as páginas das *reviews*.

Posto isto, foram extraídas 1938 *reviews* sendo posteriormente efetuada uma verificação manual de 10% (194) das *reviews* de modo a confirmar a coerência dos dados extraídos.

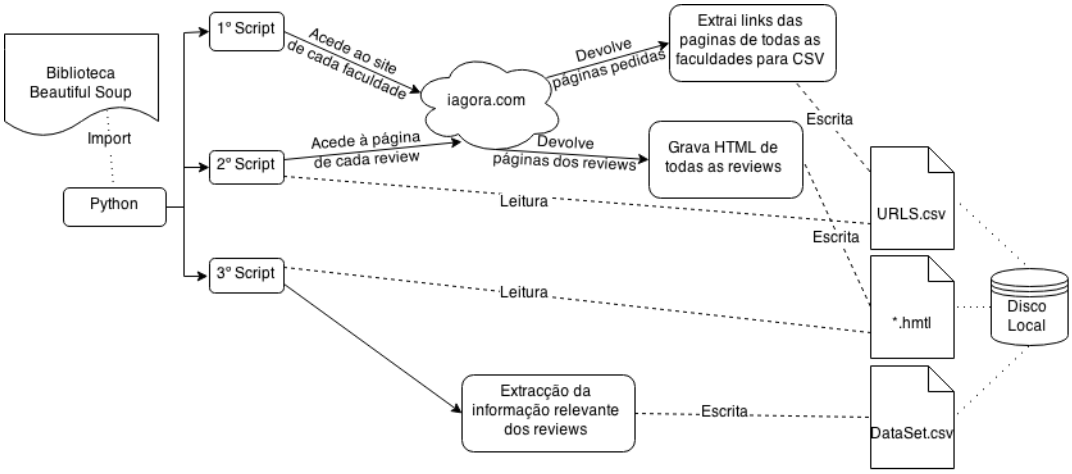


Figura 11 Processo de extração dos dados

4.2.2 Análise da qualidade dos dados

O *corpus* extraído caracteriza-se por ter um total de 92 atributos (colunas) e 1938 registos (linhas).

Ao analisar detalhadamente o *corpus* verificou-se que maioria dos atributos existentes tinha um tipo de formato estruturado (Tabela 4). Tal acontece por, tal como já referido, em alguns casos as questões expostas no *iagora* terem respostas previamente definidas à escolha do aluno e, noutros casos ser necessário avaliar alguns aspetos qualitativa ou quantitativamente de acordo com uma escala de *Likert*. Três dos atributos apresentam registos quer em formato estruturado quer em formato não estruturado uma vez que alguns campos da *review* permitiam, além das respostas predefinidas, escolher uma opção "outra/o" em que o aluno podia escrever outros motivos que justificassem a sua opinião, alheios aos sugeridos pelo *iagora*.

Tabela 4 Quantidade de atributos por tipo de estrutura

Tipo de estrutura de dados	Quantidade de atributos
Estruturado	78 (84.78%)
Estruturado e Não estruturado	3 (3.26%)
Não estruturado	11 (11.26%)

De seguida, observou-se que existiam problemas em alguns dos registos que poderiam comprometer a qualidade da amostra pelo que foi necessário fazer-se uma pré-preparação dos dados.

4.2.2.1 *Primeiro problema – uniformização de registos*

Verificou-se que, nos atributo pais_destino e pais_origem, alguns países estavam escritos de formas distintas contudo diziam respeito ao mesmo, por exemplo o país "*The Netherlands*" e "*Netherlands*". O mesmo verificou-se no atributo cidade de origem surgindo casos como "*Bogotá*"

e “*Bogota*”. Noutros casos, por vezes, em vez de estar escrito o país de destino no atributo pais_destino, estava escrita a cidade de destino existindo, por exemplo, o país “*Denmark*” e o país “*Copenhagen*”, pelo que foi também necessário corrigir-se este erro substituindo a cidade pelo país correspondente.

4.2.2.2 *Segundo problema – país de origem igual ao país de destino*

Detetou-se que 28 das *reviews* apresentavam o país de origem igual ao país de destino. Uma vez que o presente estudo diz apenas respeito a mobilidade internacional, ou seja a países de origem diferentes dos países de destino, *reviews* com o país de origem igual ao país de destino poderiam influenciar a coerência dos resultados. Ainda assim, antes de se proceder à remoção dos registos foram ler-se todos os ficheiros HTML das *reviews* correspondentes de maneira a garantir que os dados estavam corretos.

Verificou-se que algumas das *reviews* não estavam corretamente identificadas na fonte pois, por vezes, o nome do país de destino estava escrito no local onde deveria estar escrito o nome do país de origem no entanto, a cidade de origem presente na *review* do aluno correspondia ao real país de origem do aluno. Por exemplo na Figura 12, a utilizadora Silvia participou num programa de Erasmus para o curso de economia da EMLYON *Business-School* em França sendo que, a sua universidade de origem era a LUISS Guido Carli sediada em Roma no país de Itália e não de França como sugere a identificação.

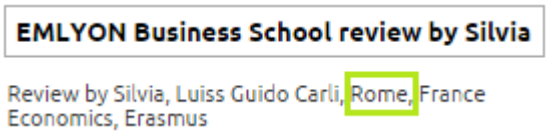


Figura 12 Exemplo de identificação de review

Posto isto, ao verificar-se as 28 *reviews* apenas duas diziam respeito a processos de mobilidade dentro do próprio país, sendo que estas foram removidas da amostra. Nas restantes 26 *reviews,* o atributo país_origem foi corrigido de acordo com a cidade de origem apresentada apresentada.

4.2.2.3 *Terceiro problema – obrigatoriedade de resposta*

A existência de campos específicos a preencher assim como a obrigatoriedade de algum deles levou a que bastantes alunos escrevessem comentários como “*No comments*”, “N/a”, “None”, etc. sem qualquer conteúdo. Considerou-se que levar este género de conteúdo para análise textual traria ruído ao *dataset* final pelo que foi feita uma pré-limpeza a estes comentários.

A pré-limpeza foi feita para os 11 atributos de conteúdo não estruturado, utilizando a ferramenta Excel. Para cada um dos atributos foi contado o número de caracteres por *review* e selecionados os comentários com um resultado inferior a 20. De seguida, verificou-se o conteúdo textual desses comentários tendo sido apurada uma lista de comentários a remover. Qualquer comentário que apenas tivesse este tipo de conteúdo foi removido.

4.2.2.4 *Quarto problema – campos vazios*

Após a pré-limpeza anterior, verificou-se a quantidade de comentários por cada um dos 11 atributos não estruturados, como se pode verificar pela Tabela 5. A partir dos resultados

obtidos, percebeu-se que alguns deles tinham a maioria dos campos vazios, ou seja, campos em que os alunos não escreveram qualquer comentário, como é o caso do AC_pComments com 73% de *missing values*.

Tabela 5 Quantidade de comentários por atributo não estruturado

1936 *reviews*	HO_why	HO_pComments	SL_pComments	LL_PComments	AC_courseRecommend	AC_pComments	EX_PersonalSpendingHabits	EX_PComments	OV_PRecomendation	OV_WishHadKnown	FinalComments
Quantidade de comentários	1567	1352	1316	1875	819	530	1192	634	1817	1343	1739
Missing values	369	584	620	61	1117	1406	744	1302	119	593	197
% Missing values	19%	30%	32%	3%	58%	73%	38%	67%	6%	31%	10%

Devido à falta de informação existente em alguns dos atributos, considerou-se então que seria mais interessante agregar todos os comentários de cada *review* num só atributo, ao invés de serem feitas análises por cada aspeto. Desse modo, foi criado um novo atributo – "Experiência" – através da concatenação do conteúdo dos 11 atributos não estruturados por *review*, representando este novo atributo uma *review* completa, ou seja, uma experiência internacional completa por aluno. O atributo "Experiência" será aquele onde irão recair todas as explorações textuais pois nele consta todo o conteúdo não estruturado escrito pelos alunos.

Por fim, analisando o atributo experiência observou-se que 11 *reviews* não tinham qualquer comentário textual, ou seja, os alunos apenas tinham avaliado as instituições quantitativamente e, nos campos de texto aberto, escreveram apenas comentários como "*none"*, "*nothing"* ou "n/a". Sendo o intuito principal desta dissertação fazer-se uma análise aos dados não estruturados, as 11 *reviews* foram removidas do conjunto de dados.

4.2.3 Análise descritiva dos dados

Finalizada a análise de qualidade, passaram a existir menos 13 documentos no conjunto de dados comparativamente à amostra anterior, contabilizando um total de 1925 *reviews.*

Foram então feitas algumas análises descritivas sobre o presente conjunto de dados. No que respeita à distribuição por país de destino, a Figura 13 mostra que a maioria das *reviews* são de alunos que viajaram para França ou para o Reino Unido sendo que, a soma de *reviews* destes dois países representa mais de 50% do conjunto de dados total. Por sua vez, os alunos provenientes de Espanha e da Áustria são os que têm mais opiniões no conjunto de dados (Tabela 6). Curiosamente, observa-se também que, embora a maioria dos países de origem seja europeia, os Estados Unidos bem como a China representam também uma das percentagens mais altas de *reviews* no conjunto de dados.

Tabela 6 Top 7 países de origem

País de origem	Qtd de reviews	%
Spain	241	12.5%
Austria	230	11.9%
Italy	169	8.8%
France	161	8.4%
United States	112	5.8%
China	88	4.6%
United Kingdom	77	4.0%

Figura 13 Distribuição de reviews por país de destino

Em relação à distribuição por instituição, a *Copenhagen Business School* na Dinamarca é aquela que apresenta maior quantidade de *reviews* seguida do ESADE *Business School* em Espanha (Tabela 7).

Tabela 7 Top 7 de quantidade de reviews por universidade

Universidade	Qtd Reviews	%
Copenhagen Business School	126	6.5%
ESADE Business School	112	5.8%
Erasmus Universiteit Rotterdam	93	4.8%
Universiteit Maastricht	71	3.7%
HEC Paris	70	3.6%
KEDGE Business School	68	3.5%
Rijksuniversiteit Groningen	58	3.0%

No que diz respeito à distribuição de comentários por atributo não estruturado a Tabela 8 mostra que a quantidade de *missing values* é bastante similar à anteriormente apresentada na Tabela 5 sendo que apenas o atributo AC_courseRecommend e o OV_WishHadKnown sofreram uma redução de 1%.

Tabela 8 Quantidade de comentários por atributo não estruturado – conjunto de dados final

1925 *reviews*	HO_why	HO_pComments	SL_pComments	LL_PComments	AC_courseRecommend	AC_pComments	EX_PersonalSpendingHabits	EX_PComments	OV_PRecomendation	OV_WishHadKnown	FinalComments	Experiência
Quantidade de comentários	1568	1351	1316	1875	819	529	1191	634	1816	1342	1738	1925
Missing values	357	574	609	50	1106	1396	734	1291	109	583	187	0
% *Missing values*	19%	30%	32%	3%	57%	73%	38%	67%	6%	30%	10%	0%

Em relação ao atributo Experiência, através da Tabela 9, observa-se que embora existam *reviews* com bastante conteúdo, a média de palavras é bastante mais baixa, encontrando-se nas 144 palavras por *review,* chegando mesmo a existir *reviews* apenas com um termo.

Tabela 9 Estatísticas descritivas do atributo Experiência

	N	Minimo	Máximo	Média	Desvio-Padrão
Qtd palavras do atributo "Experiência"	1925	1	2049	144,49	145,975
Valid N (listwise)	1925				

Relativamente às estrelas dadas pelos alunos aos vários aspetos gerais pedidos pelo *iagora*, as estatísticas descritivas são apresentadas na Tabela 10. Todos os aspetos receberam pontuações entre um e cinco. As médias das estrelas dadas ao aspeto sobre a habitação (HO_Stars) e principalmente às despesas (EX_Stars) são as mais baixas sendo que este último não atinge sequer uma média de 3 estrelas. Por sua vez, além da classificação dada à experiência no geral (OV_Stars), as estrelas dadas à experiência do estudante na cidade e aos seus momentos de lazer (SL_Stars) assim como os aspetos relacionados com a língua desenvolvida (LL_Stars) são as que têm médias de estrelas mais elevadas atingindo um valor médio próximo de quatro estrelas. A classificação média dada à experiência no geral (OV_Stars) é claramente a que tem um valor mais alto sendo o único aspeto que chega a atingir uma média de quatro estrelas. Os desvios-padrão não são muito elevados sedo que o OV_Stars é o atributo com o desvio-padrão mais baixo, o que reforça a consistência das opiniões neste atributo.

Os resultados apresentados sugerem que embora os alunos possam considerar certos aspetos na sua experiência como menos positivos, principalmente as despesas, no final consideram que a sua experiência foi bastante positiva existindo uma tendência a dar uma pontuação mais alta.

Tabela 10 Estatísticas descritivas das estrelas

	N	Mínimo	Máximo	Média	Desvio-Padrão
HO_Stars	1925	1,000	5,000	3,45143	,917379
SL_Stars	1925	1,000	5,000	3,93766	,924240
AC_Stars	1925	1,000	5,000	3,54701	,854089
LL_Stars	1925	1,000	5,000	3,89974	,880356
EX_Stars	1925	1,000	5,000	2,96286	,814721
OV_Stars	1925	1,000	5,000	4,06208	,616320
Valid N (listwise)	1925				

Uma vez que o atributo Experiência se refere a toda a experiência do aluno, considerou-se que a variável *target* mais adequada para as explorações textuais seria a "OV_Stars" pois esta representa a satisfação sentida pelo aluno de um modo geral. A distribuição da OV_Stars é enviesada à esquerda atingindo o seu pico de concentração nas 4.5 estrelas, sendo que existem apenas 50 *reviews* a dar uma classificação inferior a três (Figura 14).

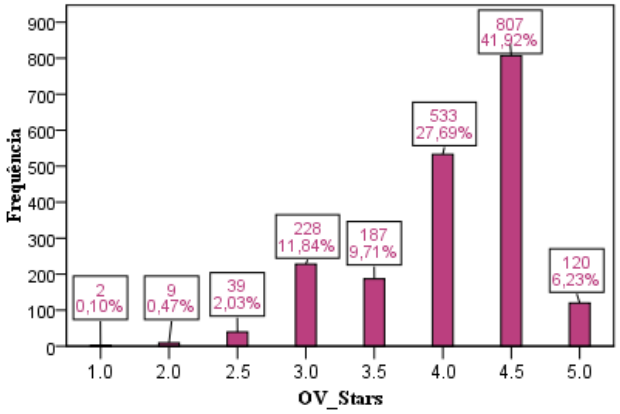


Figura 14 Distribuição de frequências OV_Stars

No que diz respeito às 15 variáveis quantitativas, o valor mínimo e máximo varia entre um e quatro respetivamente (Tabela 11), distintamente das variáveis *Stars* gerais que tinham um valor mínimo e máximo entre um e cinco (Tabela 10).

Tabela 11 Estatísticas descritivas das 15 variáveis "overall"

	N	Mínimo	Máximo	Média	Desvio-Padrão
OV_ICU_AcademicReasons	1925	1	4	3,44	,769
OV_ICU_Culture	1925	1	4	3,21	,792
OV_ICU_Costs	1925	1	4	2,33	,890
OV_ICU_Activities	1925	1	4	2,75	,833
OV_ICU_CampusLife	1925	1	4	2,66	,865
OV_ICU_PartyPeople	1925	1	4	2,99	,860
OV_ICU_WeatherLocation	1925	1	4	2,49	1,024
OV_DMEA_BecameFamiliarWithAnotherCulture	1925	1	4	3,44	,664
OV_DMEA_Traveled	1925	1	4	3,38	,713
OV_DMEA_ImprovedLanguageSkills	1925	1	4	3,39	,770
OV_DMEA_MetPeopleFromOtherCountries	1925	1	4	3,75	,505
OV_DMEA_BecameMoreIndependent	1925	1	4	3,59	,637
OV_DMEA_PartiedALot	1925	1	4	3,60	,645
OV_DMEA_ExperiencedChangeInLife	1925	1	4	3,61	,644
OV_DMEA_AdvancedStudiesCareer	1925	1	4	3,45	,733
Valid N (listwise)	1925				

As primeiras sete variáveis (OV_ICU) dizem respeito ao motivo que levou à escolha da universidade antes do processo de mobilidade (Tabela 11). Os resultados obtidos sugerem que os principais motivos que levaram à escolha da universidade foram em primeiro lugar razões académicas e de seguida a cultura do país. Por sua vez, os custos assim como o clima aparentam ter sido critérios com menor peso na escolha.

As restantes oito variáveis (OV_DMEA) dizem respeito a opiniões sentidas ao longo da experiência no estrangeiro. Todas as variáveis apresentam médias similares durante a experiência no estrangeiro sendo que a variável que mais se destaca positivamente é a experiência de conhecer pessoas de várias culturas diferentes.

É importante notar que a OV_Stars é uma classificação construída a partir de uma média entre a classificação em estrelas que o aluno dá ao campo OV_Stars com a classificação que o aluno dá às restantes OV_DMEA variáveis.

4.3 PREPARAÇÃO DOS DADOS

A preparação do conjunto de dados foi feita recorrendo à ferramenta R[13] no entanto, antes da importação do conjunto de dados no R, utilizou-se a ferramenta QDA Miner[14] apenas para corrigir e uniformizar alguns termos. O QDA Miner possui uma funcionalidade que permite detetar palavras desconhecidas de algum idioma a partir de um dicionário da própria língua, neste caso do inglês. O QDA Miner detetou vários termos desconhecidos com

Figura 15 Exemplos de substituições pelo QDA Miner

[13] https://www.r-project.org/
[14] http://provalisresearch.com/

frequências entre 1 e 374. Os termos foram observados e parte deles corrigidos. No total, foram efetuadas 112 substituições cuja frequência termos errados variou entre 10 e 374. Alguns exemplos das substituições efetuadas são apresentados na Figura 15.

Antes da exposição detalhada do restante desenvolvimento da preparação de dados, serão apresentados as várias etapas passadas ao longo da preparação dos dados até à fase de modelação recorrendo a um esquema (Figura 16).

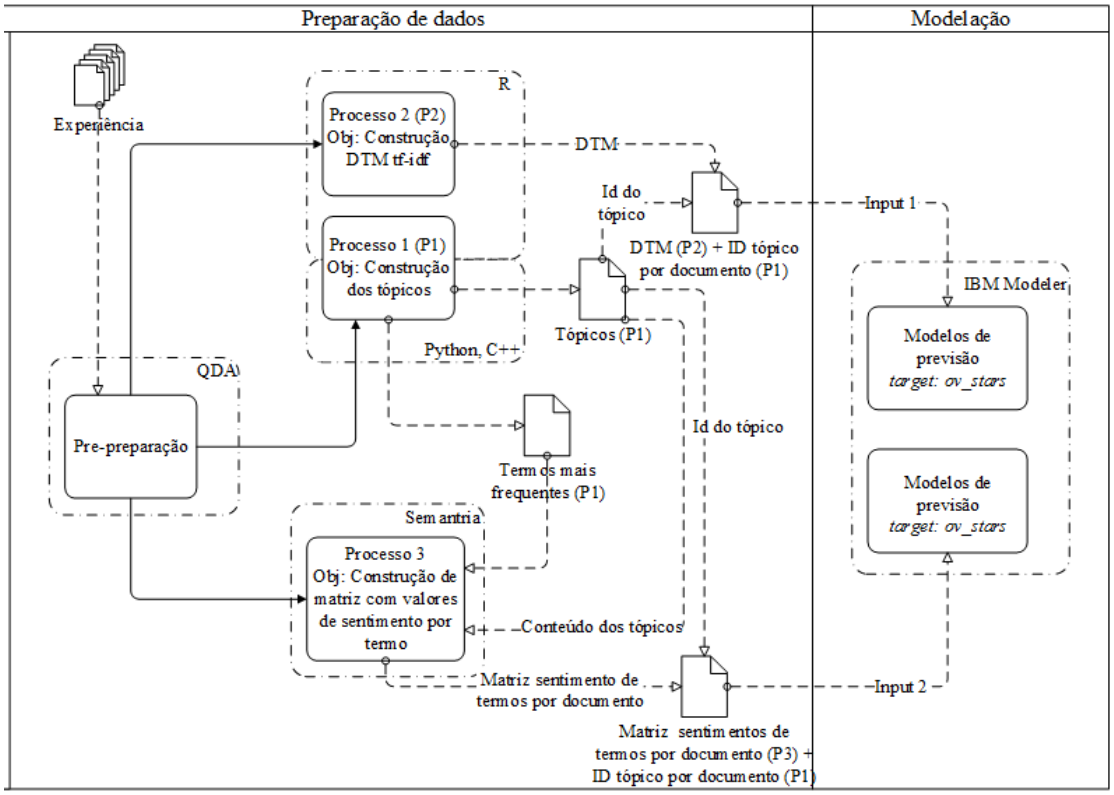


Figura 16 Esquema da preparação de dados

Após uma pré-preparação através do QDA Miner, foram desenvolvidos três processos distintos através dos quais foram produzidos dois *inputs* (*input* 1 e 2) para a fase de modelação (Figura 16). Cada um dos três processos encontra-se sistematizado na Figura 17.

No Processo 1 (P1), o principal intuito foi aplicar técnicas de modelação de tópicos de forma a ter uma contextualização mais aprofundada sobre o que abordavam as *reviews,* através da descoberta dos vários assuntos/tópicos escritos nas mesmas, independentemente dos seus sentimentos. Deste processo, foram gerados dois *outputs*: os tópicos latentes e, uma lista com os termos mais frequentes nas *reviews*. Para o desenvolvimento do P1, recorreu-se maioritariamente à ferramenta R no entanto, para a conceção dos tópicos foi necessário recorrer-se a outros métodos que serão explanados aquando da elucidação mais detalhada do P1.

O processo 2 (P2), deu origem a um dos *inputs* (*input* 1 – construção da DTM com TF-IDF com os tópicos) sobre a qual irão ser aplicados modelos de classificação, na fase de modelação, para prever a variável *ov_stars,* representativa da satisfação dos alunos. O input consiste numa DTM à qual foi adicionada uma nova coluna que identificava o tópico mais correlacionado com cada documento de acordo com a informação extraída dos tópicos criados no P1.

No processo 3 (P3), recorreu-se a uma alternativa à habitual da DTM. Pretendia utilizar-se os valores dos sentimentos associado a cada termo, em vez de ter como base os pesos dos termos associados às suas frequências (como acontece no P2). Neste sentido foi utilizado o Semantria. O P3 recebe como *input,* proveniente do P1, os termos mais frequentes bem como o conteúdo dos tópicos detetados, sendo que será calculado pelo *plug-in* o sentimento associado a cada um destes (Figura 17). O *output* do P3 (*input* 2) consiste numa matriz que cruza o sentimento dos termos por documentos e, à semelhança do P2, tem também uma coluna que identifica o tópico mais correlacionado com cada documento.A diferença entre o input 1 e 2 é que o primeiro é uma matriz de frequências ao passo que o segundo é uma matriz de sentimentos. E, embora os documentos sejam os mesmos nos dois inputs, os termos são distintos devido às técnicas de seleção de termos do Semantria serem diferentes das aplicadas no R.

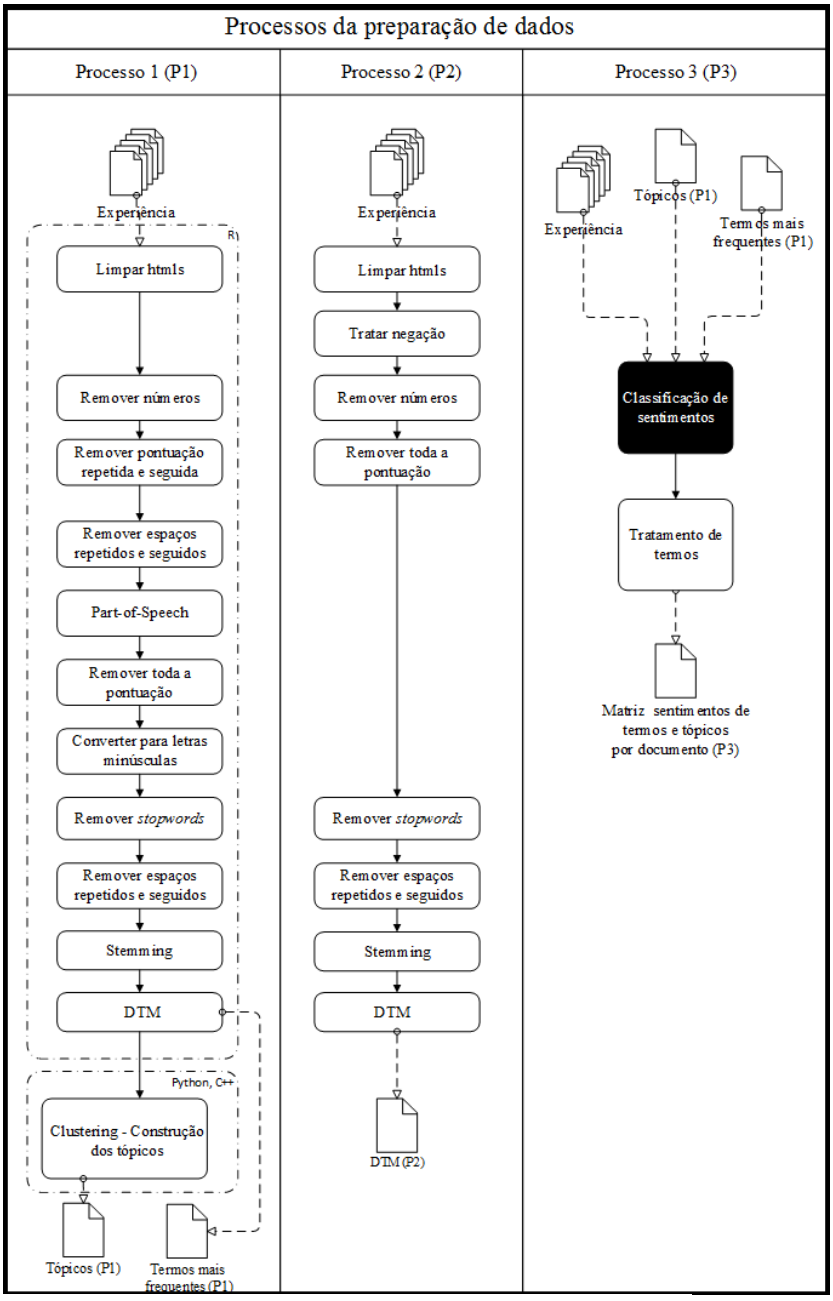


Figura 17 Processos da preparação de dados

4.3.1 Processo 1 (P1) - Construção dos tópicos

Como já explanado, este processo tem como objetivo aumentar o conhecimento sobre a informação presente no conjunto de dados através da exploração dos tópicos mais enfatizados pelas *reviews.*

O P1 tem início com a remoção de htmls recorrendo a expressões regulares e ao pacote gsub do R.

De seguida, foi removida toda a numeração. Este passo foi feito substituindo os números por um espaçamento de maneira a evitar palavras coladas, caso não exista nenhum espaçamento entre os números e as palavras seguintes.

Procedeu-se também à remoção de pontuação repetida (Tabela 12) e espaçamentos multiplos. Na remoção de pontuação, foram preservados os primeiros sinais de pontuação pois, como será aplicada a técnica *Part-of-Speech*, a remoção total da pontuação poderia influenciar a classificação da classe gramatical da palavra.

Tabela 12 Exemplo de tratamento

Comentário original	"it was an amazing experience!!!!I really recommend it"
Comentário tratado	"it was an amazing experience! I really recommend it"

Com o propósito de obter uma contextualização mais objetiva dos aspetos que abordavam as *reviews,* à semelhança de Hu & Liu (2004b), o POS foi utilizado com o propósito de extrair apenas os substantivos presentes e ter uma melhor contextualização. Para a aplicação do POS ao conjunto de dados foi utilizado o *parser* fornecido pelo pacote openNLP e criado um ciclo que percorresse os comentários classificando todos os termos consoante a sua classe gramatical (Tabela 13). Após a classificação de todos comentários foram extraídos todos termos cujas *tags* correspondiam a /NN (nomes no singular), /NNS (nomes no plural).

Tabela 13 Excerto de comentário classificado pelo POS

Review classificada	"the/DT accommodation/NN at/IN this/DT university/NN is/VBZ something/NN outstanding/JJ because/IN firstly/RB it/PRP 's/VBZ cheap/JJ and/CC friendly/JJ environment/NN is/VBZ their/PRP$ among/IN students/NNS with/IN different/JJ cultures/NNS and/CC safetywise/NN"
Review após extração de substantivos	accommodation, university, something, environment, students, cultures, safetywise

Partindo do conjunto de dados obtido até ao momento, foi retirada a restante pontuação ainda existente pois, a partir deste momento, a sua presença iria apenas criar ruído no conjunto de dados.

Tendo em vista o objetivo deste processo, considerou-se que manter termos com letras maiúsculas e minúsculas iria aumentar esparsidade do conjunto de dados desnecessariamente. Optou-se, então, por converter todas as letras do conjunto de dados para minúsculas, reduzindo a dimensionalidade do *corpus* e concentrando assim as frequências dos termos.

No que trata às *stopwords,* através da lista de *stopwords* em inglês, fornecida pelo pacote tm, foram removidos do conjunto de dados alguns termos menos relevantes. Foi necessário retirar os apóstrofos das palavras existentes na lista do pelo pacote, pois como tinha sido já removida toda a pontuação essas palavras não iriam ser detetadas. Por exemplo, a palavra *don't* não seria removida pois, após o tratamento anterior as palavras *don't* do *corpus* ficaram como *dont.* Apesar deste tratamento ter removido bastantes palavras, outras sem relevância para a

análise continuaram a permanecer no conjunto de dados por não se encontrarem na lista pelo que algumas foram removidas recorrendo à função *RemoveWords.*

Face às substituições e remoções efetuadas após o POS, limparam-se novamente eventuais múltiplos espaçamentos que pudessem existir no conjunto de dados.

Após todos os tratamentos indicados, o *corpus* ficou com um total de 1925 documentos e 5818 termos. Para reduzir a dimensionalidade do conjunto de dados assim como a sua esparsidade, foi aplicada a técnica de *stemming*. Uma vez aplicado o *stemming*, o conjunto de dados reduziu significativamente a sua quantidade de termos passando de 5818 para 4629.

Finalmente, foi então construída a DTM. A DTM teve apenas em conta unigramas[15] e a medida *term-frequency* para cálculo do peso de cada termo (Figura 18).

```
<<DocumentTermMatrix (documents: 1925, terms: 4629)>>
Non-/sparse entries: 41499/8869326
Sparsity           : 100%
Maximal term length: 23
Weighting          : term frequency (tf)
```

Figura 18 Primeira DTM do P1

De maneira a tentar reduzir a esparsidade e dimensionalidade, definiu-se que apenas permaneceriam no *corpus* termos com uma frequência mínima de dois e, à semelhança de Grün e Hornik (2011), foi aplicado um tamanho mínimo de caracteres por termo de três e aplicada a medida TF-IDF com o propósito de remover alguns dos documentos e termos mais irrelevantes. Segundo os autores, a TF-IDF deve ser calculada para a DTM e, de seguida, devem ser removidos os termos cujo valor de TF-IDF sejam ligeiramente inferiores à mediana de maneira a garantir que os termos mais frequentes são omitidos[16]. Foram também removidos todos os documentos cuja soma dos valores de TF-IDF dos seus termos fosse inferior ou igual a zero.

Como se observa na Figura 19, os resultados da segunda DTM melhoraram existindo uma redução significativa quer de termos (de 4629 para 1799) quer de documentos (de 1925 para 1858). Embora continue bastante esparsa, a matriz teve também uma redução na esparsidade de 1% assim como no tamanho máximo dos termos (de 23 para 14).

```
<<DocumentTermMatrix (documents: 1858, terms: 1811)>>
Non-/sparse entries: 28980/3335858
Sparsity           : 99%
Maximal term length: 14
Weighting          : term frequency (tf)
```

Figura 19 Segunda DTM do P1

[15] Foi também testada a DTM com unigramas, bigramas e trigramas no entanto, os resultados foram bastante mais confusos. Além de existirem bastantes repetições de termos entre tópicos, existiam também tópicos que continham muitas repetições de termos por tópicos por exemplo termos como “tip”, “travel” e “travel tip”.

[16] O valor mínimo da *tf-idf* foi adaptado várias vezes uma vez que, quanto mais próximo da mediana fosse definido o *threshold,* mais termos repetidos existiam entre tópicos o que os tornava muito difíceis de nomear. Posto isto, foi necessário permitir mais termos de maneira a conseguir tópicos mais distintos.

Finalmente, uma vez construída a DTM, os dados encontravam-se já com uma estrutura consistente, o que permitia que fossem feitas algumas análises descritivas. Assim, a partir da DTM foi gerada uma *wordcloud* e exploradas algumas coocorrências de termos. A *wordcloud* foi gerada tendo em conta os termos com uma frequência mínima de 130 (Figura 20). A informação apresentada sugere que os comentários incidem essencialmente sobre o campus (*campus*), a acomodação (*accommod*) e a escola (*school*). Num olhar mais minucioso observam-se também termos que sugerem alguma importância dada às despesas (por exemplo *cost, price, money*), à cultura (*cultur, travel*), a aspetos académicos (*teacher, class*), lazer (*pub, cafe, parti*) e até mesmo à informação (*inform*).

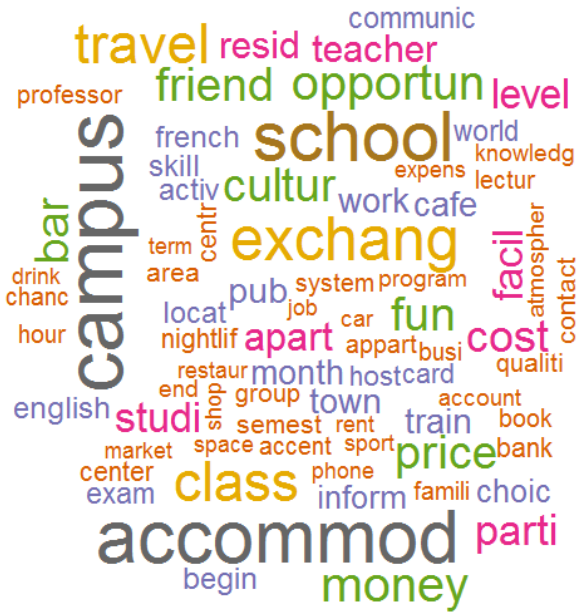


Figura 20 Wordcloud termos mais frequentes do P1

Tabela 14 Correlações dos 6 termos mais frequentes do P1

Termos mais frequentes	Termos mais correlacionados	Correlação
campus	bathroom	0.23
	pocket	0.21
	toilet	0.19
accommod	matricul	0.25
	fax	0.22
	applic	0.2
school	busi	0.39
	stone	0.31
	aspir	0.3
exchang	gesthous	0.26
	semest	0.26
	prepar	0.23
class	art	0.31
	neighbor	0.28
	sceneri	0.28
travel	tip	0.21
	kittel	0.19
	card	0.17

Por sua vez, no que diz respeito às correlações, verificaram-se os três termos mais correlacionados com os seis termos mais frequentes (Tabela 14). Segundo a tabela percebe-se que as correlações existentes entre os termos são bastante baixas sendo que nenhum deles chega a atingir um patamar de correlação de 0.5.

Finalmente, partiu-se para a exploração de possíveis tópicos latentes. Inicialmente, devido à sua popularidade na revisão literária foi aplicada a técnica CTM porém, considerou-se que os resultados obtidos não eram muito satisfatórios, como também já sugeria a literatura. Desse modo, foi aplicado o BTM sendo que os seus resultados pareceram ser mais consistentes. Em todo o caso, serão apresentados os resultados obtidos em ambas as técnicas. Ressalta-se que para ambas as técnicas, os resultados apresentados não foram obtidos na primeira tentativa pelo que foi necessário fazer-se alguns ajustes ao longo do processo decorrido até ao momento com o intuito de otimizar os resultados[17].

[17] As principais alterações foram devidamente elucidadas enquanto cada uma das técnicas utilizadas foi abordada

Para aplicar técnicas de modelação de tópicos é frequentemente necessário definir à priori a quantidade de tópicos que se pretende criar. Desse modo, o número de tópicos é comumente calculado a partir de medidas como a *perplexity* e o *log-likelihood* (Blei & Lafferty, 2007; Guerreiro et al., 2015).

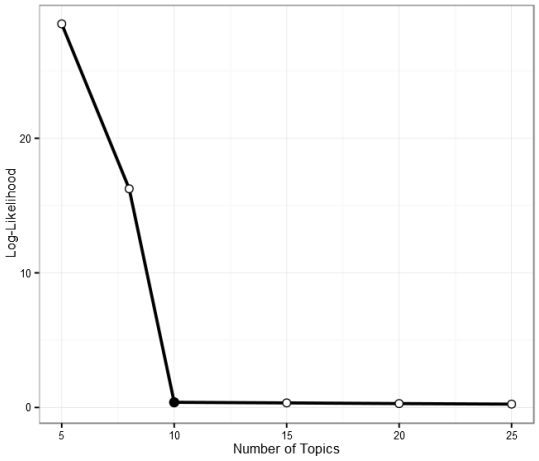


Figura 21 Log-Likelihood

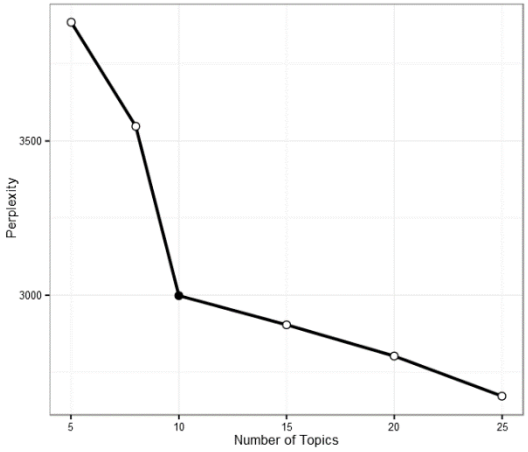


Figura 21 Perplexity

O gráfico apresentado nas Figura 21 e 22 sugere que o número de tópicos a construir são dez. Foi então aplicada a técnica CTM com 10 tópicos. Na Figura 23 é possível observar-se os resultados dos tópicos, produzidos pelo CTM, com as dez palavras mais prováveis de cada um. Os tópicos apresentados pelo CTM demonstraram, à primeira vista, ser um pouco confusos e difíceis de classificar pois alguns termos eram bastante similares entre tópicos e abordavam vários aspetos, tonando difícil distingui-los entre si e dar-lhes um nome mais consistente. Ainda assim, numa tentativa de nomear os tópicos fornecidos pelo CTM, verificou-se qual o tópico mais correlacionado para cada *review* e procedeu-se à leitura de algumas *reviews* por tópico. Após uma análise exaustiva das *reviews* tornou-se bastante árduo classificar alguns dos tópicos fornecidos pelo CTM.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
campus	opportun	campus	school	accommod	campus	school	exchang	bar	travel
accommod	cultur	pub	year	exchang	hour	class	money	year	money
parti	school	center	cost	facil	accommod	bar	travel	friend	campus
begin	level	parti	travel	cultur	level	year	resid	class	school
studi	part	fun	card	money	pub	friend	campus	ticket	class
facil	busi	town	inform	price	cost	famili	class	kitchen	price
cafe	exchang	lectur	parti	french	class	professor	fun	travel	cost
money	teacher	year	choic	school	town	cafe	appart	sport	teacher
locat	price	system	resid	contact	price	fun	friend	accommod	opportun
price	year	appart	friend	work	semest	chanc	teacher	studi	train
appart	friend	train	studi	fun	exchang	busi	cultur	restaur	work

Figura 23 Tópicos detetados através do CTM

Posto isto, foi então aplicado o BTM. De acordo com a literatura, o BTM apontaria resultados melhores para o tipo de conteúdo em questão uma vez que, embora existam comentários relativamente longos, a média de palavras por comentário é relativamente pequena.

Para a aplicação do BTM, foi utilizado o código fornecido pelo autor[18] e adaptado de acordo com os parâmetros pretendidos, nomeadamente o número de tópicos e a quantidade de palavras a apresentar. O código foi executado a partir de um *Shell Script* que invoca outros *scripts* desenvolvidos em *Python* e C++.

Os resultados do BTM são apresentados na Figura 24. Os termos apresentados pelo BTM aparentaram, por tópico, ter uma relação com mais sentido entre si e, verificou-se também uma diminuição de palavras menos descritivas nos primeiros cinco termos. Por exemplo o CTM apresenta em cinco tópicos de dez, termos tais como *begin* (tópico 1)*, year* (tópico 4, tópico 7 e tópico 9) e *hour* (tópico 6) ao passo que, no BTM, nos primeiros cinco termos não existe nenhum termo aparentemente menos descritivo pelo que as palavras são mais objetivas.

À semelhança do que foi feito na análise de tópicos do CTM, analisaram-se algumas das *reviews* por tópico com o propósito de atribuir a designação mais adequada a cada um. Os resultados dos tópicos criados pelo BTM são analisados na secção "5 – Implementação Análise dos resultados".

Housing	Campus Facilities	Culture enrichment	Host city life	Academic	Expenses	Turism	Student life	Skills development	International environment
apart	campus	cultur	campus	question	card	travel	school	exchang	class
school	kitchen	school	accommod	exam	money	exchang	accommod	studi	exchang
bar	bathroom	friend	pub	professor	train	parti	campus	degre	exam
town	toilet	famili	bar	class	bank	book	class	program	bar
resid	sport	host	centr	system	travel	sport	exchang	level	drink
centr	shower	opportun	town	note	account	price	travel	facil	parti
campus	opportun	cafe	parti	posit	ticket	ticket	money	year	hour
exchang	point	french	cafe	test	cost	tip	cost	world	teacher
accommod	wash	money	accent	lectur	discount	activ	friend	dutch	semest
cafe	facil	accommod	drink	semest	bike	summer	price	qualiti	beer

Figura 24 Tópicos detetados através do BTM

4.3.2 Processo 2 (P2) - Construção do input com matriz frequências TF-IDF

O P2 tem em vista a construção do *input* 1 que serve como um dos experimentos base para os modelos de classificação na fase de modelação através da construção de uma DTM com frequências TF-IDF. No final, recorrendo à informação obtida em P1, foi adicionada uma coluna com a identificação do tópico mais correlacionado com cada *review.*

À semelhança do processo supracitado, o processo de tratamento para construção de tópicos iniciou-se com a remoção de htmls através de expressões regulares.

Uma vez que neste processo se pretende analisar os sentimentos, implementou-se o tratamento das negações pois, caso não exista, pode perder-se informação substancial que pode fazer a diferença entre uma classificação positiva ou negativa. Por exemplo observando o comentário original da Figura 25, sem o tratamento da negação, a palavra *safe*, entraria para a DTM como uma entidade positiva, no entanto, a verdade é que no comentário está escrito que a localização da acomodação não é segura pelo que a palavra *safe* deve entrar na DTM como uma entidade negativa. Assim, foi construída uma expressão regular que detetasse termos que indicassem uma negação e, aplicasse um prefixo "NOT_" a todos até ao sinal de pontuação seguinte, como apresentado no comentário tratado da Figura 25 (Das & Chen, 2007; Pang et

[18] https://github.com/xiaohuiyan/BTM

al., 2002), ou até uma conjunção adversativa. Com este tratamento, a palavra *safe* entrará na DTM como uma entidade negativa (not_safe).

comentário original	the housing location wasn't safe.
comentário tratado	the housing location wasn't not_safe.

Figura 25 Tratamento da negação

De seguida procedeu-se à remoção da numeração bem como da pontuação. Neste tratamento, tal como no P1, substituíram-se os números e sinais de pontuação por espaços ao invés de simplesmente apagá-los. Esta substituição evita que as palavras se colem caso não exista nenhum espaçamento entre palavras depois de algum número ou sinal de pontuação (Figura 26). Os apóstrofos foram simplesmente retirados para que palavras como *can't* ou *don't* ficassem como *cant* ou *dont* em vez de *can t* ou *don t*.

Comentário original	The university is great.Travel a lot!
Comentário mal tratado	The university is greatTravel a lot
Comentário bem tratado	The university is great Travel a lot

Figura 26 Tratamento de pontuação

No que respeita às *stopwords*, à semelhança de P1, foi criada uma lista personalizada, partindo do pacote tm, onde foram removidos os apóstrofos de todos os termos. No entanto, devido ao tratamento da negação, algumas das *stopwords,* ficaram com um prefixo NOT_ pelo que foi necessário acrescentar-se novas palavras à lista de palavras a remover para lidar com estes casos. Desse modo, a lista de termos a remover continha todas as palavras em inglês sem apóstrofos, fornecidas pelo pacote tm, e ainda as mesmas palavras do pacote mas com o prefixo NOT_ adicionado a cada uma delas.

Seguidamente, quis garantir-se que não existiam espaços acumulados tendo sido criada uma expressão regular que substituísse múltiplos espaços seguidos apenas por um.

Por último, aplicou-se o *stemming* de maneira a reduzir um pouco a esparsidade e dimensão aquando da construção da DTM.

A DTM foi construída com unigramas, bigramas e trigramas, utilizando a *term frequency* para ponderação de peso de cada termo (Figura 27). A DTM produziu um total de 196.716 termos, com 378.352.122 entrada esparsas e com termos cujo tamanho chegava a atingir os 44 caracteres.

```
<<DocumentTermMatrix (documents: 1925, terms: 196716)>>
Non-/sparse entries: 326178/378352122
Sparsity           : 100%
Maximal term length: 44
Weighting          : term frequency (tf)
```

Figura 27 DTM depois stemming do P2

Para reduzir a dimensão e a esparsidade da matriz definiu-se que um termo só seria considerado válido se tivesse um mínimo de três caracteres e se se repetisse pelo menos cinco vezes no *corpus.* Adicionalmente foi também aplicada a *TF-IDF.* Posto isto, a DTM melhorou consideravelmente, existindo uma diminuição substancial de termos, (de 196.716 para 4.804), de entidades esparsas (de 378.352.122 para 9.121.188), o tamanho máximo dos termos diminuiu e 44 para 34 assim como percentagem de esparsidade conseguiu ter uma diminuição de 1% (Figura 28).

```
<<DocumentTermMatrix (documents: 1922, terms: 4804)>>
Non-/sparse entries: 112100/9121188
Sparsity           : 99%
Maximal term length: 34
Weighting          : term frequency - inverse document frequency
```

Figura 28 DTM final segundo processo do P2

Para uma observação amigável e clara do que mais é abordado ao longo das *reviews* foi criada uma *wordcloud* (Figura 29) com os 65 termos, em *stemming,* mais frequentes. Da *wordcloud* gerada saltam à vista essencialmente palavras relacionadas com aspetos académicos (*student, cours, univers)*, com despesas (expens, cost, price), acomodação (accommod, apart), língua (french, english, languag), entre outros. Verifica-se também que surgiram dois casos especiais na *wordcloud* nomeadamente um bigrama – "intern student" – e, um termo que tem o impacto do tratamento de negações – "no_problem". Relativamente a este último termo notam-se dois reparos: o primeiro é que durante a fase de *data understanding*, embora tenham sido apagados vários comentários que apenas diziam "no problem", continuaram a persistir vários com este conteúdo no meio. A outra questão prende-se com o tratamento de negações apresentar já uma ajuda notória, pois caso não tivesse sido aplicado o tratamento apenas apareceria o termo "problem", ou seja, como uma entidade negativa. Assim, com o tratamento de negações aplicado, percebe-se que é uma entidade positiva.

Adicionalmente foram observadas as coocorrências dos três termos mais correlacionados com os seis termos mais frequentes (Tabela 15). As correlações não são muito elevadas existindo apenas duas que atingem um patamar de 0.5 nomeadamente o termo student com intern e, o termo french com franc.

Figura 29 Wordcloud top 65 termos do P2

Tabela 15 Correlações termos mais frequentes do P2

Termos mais frequentes	Termos mais correlacionados	Correlação
student	intern	0.51
	intern student	0.5
	cours	0.42
french	franc	0.5
	french student	0.42
	french speak	0.35
cours	student	0.42
	lot	0.32
	place	0.29
univers	student	0.4
	home univers	0.35
	home	0.34
lot	student	0.33
	cours	0.32
	travel lot	0.31
good	realli good	0.33
	good place	0.31
	good univers	0.26

Após o desenvolvimento da DTM, foi construída a primeira matriz de *input* para a fase de modelação. Para tal, foi colocado todo o conteúdo da DTM e adicionada uma nova coluna com os tópicos mais correlacionados por review (desenvolvidos no P1) bem como a variável

target ov_stars (Tabela 16). Adicionalmente, foram também adicionadas algumas variáveis demográficas.

Tabela 16 Dataset input 1

Nome variável	Tipo de variável	Descrição
Entidades	Independente	4804 termos presentes na DTM com peso TF-IDF
Tópicos	Independente	Identificador do tópico mais correlacionado por *review.*
Pais_Destino	Independente	País de destino
Universidade	Independente	Universidade anfitriã
Pais_Origem	Independente	País de origem
OV_stars	Dependente	Classificação dada ao campo *ov_stars*

4.3.3 Processo 3 - Construção de matriz com valores de sentimento (Semantria)

O P3, visa a construção do *input 2.* Como já mencionado, o *input 2* representa uma alternativa às tradicionais técnicas de *text mining* para análise de sentimentos, recorrendo a um *plug-in* para o Excel – Semantria – que analisa os sentimentos do conjunto de dados automaticamente. À semelhança do *input 1* pretende utilizar-se esta informação para prever a satisfação do aluno face à IES, ou seja, para prever a variável ov_stars.

Antes de submeter o conjunto de dados no Semantria, foram acrescentados os tópicos detetados pelo BTM às categorias do *plug-in*, bem como adicionados os substantivos mais frequentes da DTM do P1 às entidades. O intuito deste passo era tentar impor que o Semantria detetasse e analisasse forçosamente os sentimentos destes. De seguida, foi então construído o *input* 2 para testar na fase de modelação.

A quantidade de variáveis produzida pelo Semantria foi bastante superior comparativamente à quantidade de termos presentes na DTM final do P2. Isto acontece porque o Semantria, além de manter termos muito irrelevantes como datas, faz a deteção de vários tipos de atributos tais como entidades e temas. Além destes fatores, o Semantria não produz *stemming* pelo que existem casos tais como: *account, accounted, accounting.* Estas questões representam um problema a nível de construção do conjunto de dados uma vez que, os dados nestas condições, levam à construção matriz muito esparsa.

De maneira a tentar reduzir a dimensionalidade do *dataset* foram selecionadas as entidades mais frequentes da DTM do P1 e foi criada uma macro no Excel que fizesse uma espécie de *stemming.* No fundo, a macro procurava os termos mais frequentes fornecidos pela DTM e assinalava aqueles que começavam da mesma forma mas que tinham terminações diferentes. Por exemplo, a lista de entidades mais frequentes tinha o termo "*account*" e, a macro, assinalava como sendo o mesmo termo as palavras "*accounted*" e "*accounting*". Após todas as palavras estarem assinaladas, verificou-se manualmente cada uma delas para evitar que termos como *"bar"*, "*bars*" e "Barcelona" fossem considerados o mesmo termo.

Após a verificação das palavras assinaladas, uma nova macro foi criada para combinar as várias palavras consideradas como sendo entidades idênticas na mesma coluna. Por vezes, palavras como *account* e *accounting* ou *bar* e *bars* estavam escritas na mesma *review* associadas a sentimentos distintos pelo que, o sentimento associado a uma não podia substituir

o valor de sentimento da outra. Nesse sentido, na construção das novas colunas foi calculada uma média de maneira a ser possível lidar com estes casos.

Para a construção da matriz do *input 2,* além das entidades preparadas através das macros, foram também colocados os *entity types* e os temas cuja frequência fosse superior ou igual a cinco (Tabela 17). Adicionalmente foi também acrescentada uma coluna com o sentimento geral do documento e o tópico mais associado. À semelhança do *input* 1 criado no P2 foram também acrescentadas variáveis demográficas.

Tabela 17 Dataset input 2

Variável	Tipo	Descrição
Entidades	Independente	98 entidades
Entity types	Independente	20 *entity types* (tipos de entidade)
Temas	Independente	160 temas
Tópico	Independente	Identificador do tópico mais correlacionado
Sentimento do documento	Independente	Sentimento geral da *review*
Pais_Destino	Independente	País de destino
Universidade	Independente	Universidade anfitriã
Pais_Origem	Independente	País de origem
Ov_stars	Dependente	Classificação *ov_stars*

4.4 MODELAÇÃO

Na fase de modelação recorreu-se à ferramenta IBM Modeler 17[19]. Previamente à fase de modelação propriamente dita, é importante notar que ao longo dos processos, alguns documentos do *input* 1 e do *input 2* não foram analisados pelas ferramentas. Como verificado no P2, três documentos foram removidos na construção da DTM aquando da aplicação da TF-IDF. Por sua vez, no Semantria, uma vez que estava a ser testada a versão gratuita, quatro documentos não foram analisados devido ao limite de caracteres. Desse modo, optou-se por retirar estes sete documentos de ambos os conjuntos de dados para que seja possível fazer-se uma comparação mais viável. A Figura 30 tem uma síntese da quantidade de *reviews* por *input.*

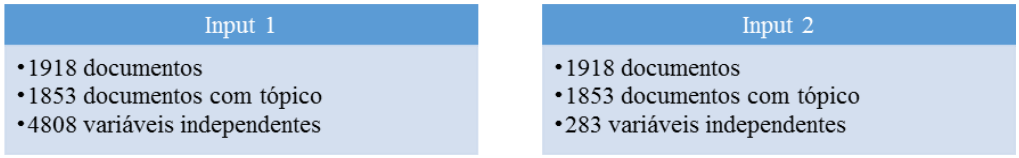
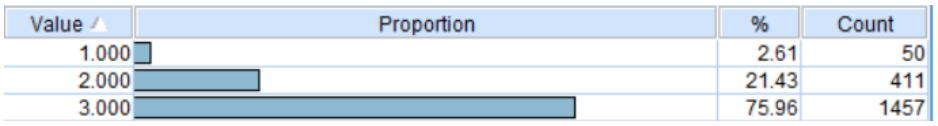


Figura 30 Síntese dos conjuntos de dados

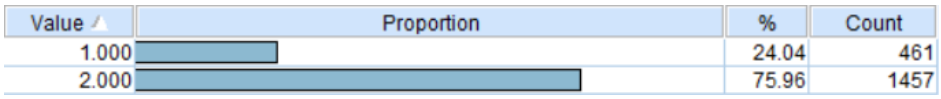
Para a aplicação de algoritmos de classificação com o objetivo de prever a variável *ov_stars* considerou-se que seria melhor agrupar-se esta variável numa escala distinta, uma vez que a escala original variava entre 1 e 5 estrelas e, algumas das estrelas tinham frequências muito baixa (Figura 14). Assim, através do nó *Reclassify* foram criadas duas variáveis *target* a partir da *ov_stars.* A uma delas deu-se o nome de *stars*, sendo a escala desta dividida em três classes: 1 (compreende as estrelas de 1 a 2.5 inclusive), o 2 (estrelas 3 e 3.5) e o 3 (estrelas 4, 4.5 e 5) (Figura 31). A outra variável chama-se *starsbinaria* e, como o próprio nome indica, caracteriza-se por ser binária. Na *starsbinaria,* pontuações até às 3.5 estrelas foram agrupadas na classe 1 e, a partir das 4 estrelas inclusive foram agrupadas na classe 2 (Figura 32).

[19] http://www-01.ibm.com/software/analytics/spss/products/modeler/



Value	Proportion	%	Count
1.000		2.61	50
2.000		21.43	411
3.000		75.96	1457

Figura 31 Distribuição da variável stars



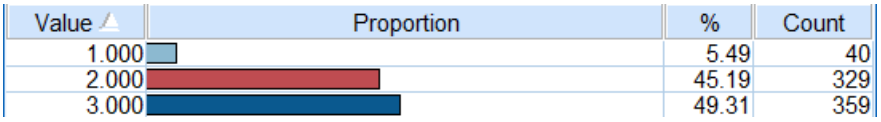
Value	Proportion	%	Count
1.000		24.04	461
2.000		75.96	1457

Figura 32 Distribuição da variável starsbinaria

Embora exista uma redução de estrelas com frequências inferiores a 10 comparativamente à distribuição original, a variável *stars* é bastante pouco balanceada sendo que a classe 1 não chega a 3% do conjunto de dados e a classe 3 ocupa mais de 75% da amostra. Por sua vez, na *starsbinaria,* embora as duas classes tenham quantidades mais significativas do que as classes 1 e 2 da variável *stars*, continua a existir uma grande diferença entre as classes um e dois.

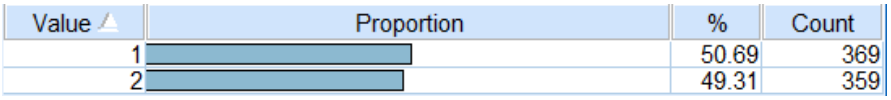
Após terem sido criadas as variáveis, foi criada uma partição de treino e de teste. A partição foi gerada aleatoriamente tendo em conta 80% de registos para treino e 20% para teste para o *input 1*. Para uma melhor comparação, a partição utilizada no *input 1* foi sempre a mesma em todas as experiências e *inputs* de maneira a garantir que os algoritmos eram treinados e testados sempre sobre o mesmo conjunto de documentos respetivamente.

De seguida, devido ao desequilíbrio existente entre as frequências das variáveis *stars* e *starsbinaria*, foi aplicada uma estratégia de balanceamento dos dados através do nó *balance* à partição de treino de cada uma delas. Segundo, Chawla (2005) quando as classes dos dados a prever são pouco balanceadas entre si não deve proceder-se à aplicação de modelos de classificação sem que antes se proceda a técnicas de balanceamento pois estas são benéficas para obtenção de melhores resultados. Desse modo, na variável *stars,* uma vez que a classe dominante é a 3 optou-se por retirar aleatoriamente registos desta variável para que esta ficasse com uma frequência mais semelhante à classe 2 (Figura 33). Por sua vez, na variável *starsbinaria,* foram retirados aleatoriamente registos à classe 2 (Figura 34).



Value	Proportion	%	Count
1.000		5.49	40
2.000		45.19	329
3.000		49.31	359

Figura 33 Distribuição variável stars - partição de treino balanceada



Value	Proportion	%	Count
1		50.69	369
2		49.31	359

Figura 34 Distribuição variável starsbinaria - partição de treino balanceada

Finalmente, procedeu-se à aplicação de algoritmos sobre os dois *inputs.* Inicialmente o *input 1* tinha um total de 4808 variáveis independentes, no entanto, devido à elevada quantidade de variáveis assim como à sua esparsidade, os algoritmos demoravam bastante tempo a ser executados. Desse modo, foi necessário aplicar-se o nó *feature selection* à partição de treino. Este nó tem como intuito selecionar as melhores *features* (variáveis) para a classificação de uma variável *target* descartando as restantes variáveis. Neste caso, o nó *feature selection* tem em

conta os valores de significância (*p-value*), de acordo com o coeficiente de *pearson.* Por seu turno, o *input 2* tinha apenas 283 variáveis independentes e, como tal, o tempo de processamento era curto pelo que não existia a necessidade de ser aplicada a seleção de *features.* Ainda assim, por uma questão de comparação optou-se também por analisar os resultados com e sem *feature selection.*

A Tabela 18 apresenta a quantidade de variáveis independentes que foram utilizadas para as diferentes experiências apresentadas.

Tabela 18 Quantidade de variáveis por experiência

1918 documentos	***input 1*** (4808 variaveis)		***input 2*** (283 variaveis)	
	stars	***starsbinaria***	***stars***	***starsbinaria***
feature selection	305	222	22	19
Sem feature selection	-	-	283	283

Na fase de modelação foram feitas várias análises a ambos os *inputs* no entanto, apenas serão apresentados os resultados dos algoritmos de alguns algoritmos mais utilizados na literatura (Rushdi Saleh et al., 2011) – *support vector machines* e *naive bayes* – bem como a regressão logística, árvores de decisão, redes neuronais e o algoritmo KNN. A Tabela 19 apresenta as parametrizações utilizadas nos algoritmos para ambos os *inputs* e variáveis target.

Tabela 19 Parametrização dos algoritmos

Modelo	Parâmetros
Naive bayes	Struture Type: TAN
	Paramther learning method: Maximum Likehood
SVM	Mode: Simple
Regressão logística	Method: Stepwise
	Procedure **stars**: Multinomial; Procedure **starsbinaria**: binomial
KNN	Objective: Speed and accuracy
	Normalize Range inputs
	Neighbors: Automatically set K (min: 3, max:5)
	Distance computation: Euclidian metric
	Predictions: Mean of nearest neighbor values
	Cross validation: Randomly assign folds (10 folds)
Redes Neuronais	Model: Multilayer perceptron (MLP)
	Hidden Layers: Automatically compute numers of units
	Stopping rule: Use maximum training time (15 min)
	Boosting: Favor model accuracy (10 Trials)
C5.0	Output Type: Decision tree
	Mode: Simple
	Boosting: Favor model accuracy (10 Trials)

4.5 AVALIAÇÃO

De modo a analisar a qualidade dos resultados obtidos pelos modelos de classificação são consideradas algumas medidas de avaliação nomeadamente: *accuracy* (exatidão), *precision* (precisão), *recall ou sensitivity* (cobertura), F-*measure* (medida-F) e *specificity* (especificidade). De maneira a compreender de forma mais clara, os conceitos serão explanados com o auxílio de uma matriz de confusão de duas classes (Tabela 20).

Tabela 20 Matriz de confusão

		Valor Real	
Valor previsto		Positivo	Negativo
	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Quando o modelo prevê acertadamente que uma determinada palavra tem um sentimento positivo, estamos perante uma situação de verdadeiro positivo (VP). Por outro lado, se o modelo previr erradamente que o sentimento corresponde a um sentimento negativo, está-se presente de uma situação de falso negativo (FN). A situação de falso positivo (FP) surge quando um sentimento é classificado pelo modelo como positivo mas é na realidade negativo. Por sua vez, quando um sentimento negativo é classificado corretamente como negativo está-se perante um verdadeiro negativo (VN).

A ***accuracy*** (Goadrich, Oliphant, & Shavlik, 2006) indica a percentagem de acertos face à totalidade. Desse modo, quanto maior for a quantidade de acerto maior será a exatidão do modelo. É importante usar esta medida com precaução uma vez que caso uma das classes tenha uma quantidade de dados bastante superior, o modelo provavelmente terá a tendência para classificar todos os registos como pertencentes a essa classe. Caso esta situação ocorra a *accuracy* pode ser até bastante alta porém, não será considerado um bom modelo pois apenas está a ser tendencioso. A fórmula correspondente ao cálculo da *accuracy* é a seguinte:

$$Accuracy = (VP + VN) / (VP + VN + FP + FN). \quad (4)$$

Para colmatar o problema da *accuracy* devem ter-se em conta outras medidas tais como a ***precision*** e o ***recall*** (Goadrich et al., 2006). A *precision* permite identificar, tendo em conta todas as previsões positivas efetuadas, qual foi a percentagem de acerto da classe verdadeira positiva. Ou seja, uma alta *precision* indica que a maior parte dos documentos classificados como positivos, foram corretamente classificados. Assim, a fórmula para o cálculo da *precision* é a seguinte:

$$Precision = VP / (VP + FP). \quad (5)$$

Por contraste à *precision*, o ***recall*** permite perceber, tendo em conta a quantidade real de verdadeiros positivos, qual a percentagem de documentos que foi corretamente prevista como positiva. Esta medida indica a capacidade que o modelo tem de conseguir detetar eficazmente documentos pertencentes à classe positiva. Assim, a fórmula de cálculo do *recall* é a seguinte:

$$Recall = VP / (VP + FN). \quad (6)$$

A *precision* e o *recall* devem ser analisadas cuidadosamente uma vez à medida que uma aumenta a outra tem tendência a diminuir pelo que dependendo da análise em causa, poderá não ter problema sacrificar uma pela outra. Tipicamente, pretende-se que exista um ponto de equilíbrio entre as duas pelo que a medida **F-*measure*** fornece informação sobre tal. A F-*measure* combina então a *precision* e o *recall* numa única só medida através de uma média harmónica tipicamente definida pela seguinte fórmula:

$$F\text{-}measure = (2 \times precision \times recall) / (precision + recall). \quad (7)$$

Por último, a **specificity** tem em conta a quantidade de acertos de verdadeiros negativos sendo a sua fórmula a seguinte:

$$Specificity = VN / (VN + FP) \qquad (8)$$

4.5.1 Avaliação do *input 1* – **Matriz TF-IDF**

Tabela 21 Resultados input 1

		stars					starsbinaria				
		Accuracy	Precision	Sensivity (Recall)	Specificity	F-measure	Accuracy	Precision	Sensivity (Recall)	Specificity	F-measure
Naive bayes	treino	61.21%	62.96%	47.45%	74.17%	54.11%	67.03%	67.32%	66.93%	66.93%	67.13%
	teste	58.44%	36.44%	36.65%	70.67%	36.54%	49.37%	53.33%	51.87%	54.63%	52.59%
SVM	treino	94.37%	96.03%	93.81%	96.43%	94.91%	96.15%	96.24%	96.13%	96.13%	96.19%
	teste	57.93%	43.15%	46.70%	69.76%	44.85%	61.46%	54.30%	55.56%	55.56%	54.92%
Regressão logistica	treino	50.96%	70.95%	39.67%	67.49%	50.89%	78.85%	79.46%	78.74%	78.74%	79.10%
	teste	75.82%	35.34%	33.49%	66.43%	34.39%	53.40%	58.31%	61.32%	61.32%	59.78%
KNN	treino	67.45%	46.39%	47.91%	79.16%	47.14%	79.40%	46.06%	79.20%	80.21%	58.25%
	teste	46.35%	34.95%	36.46%	68.60%	35.69%	41.31%	40.24%	51.56%	52.80%	45.20%
Redes neuronais	treino	71.94%	48.42%	50.79%	81.76%	49.58%	94.64%	94.66%	94.66%	94.66%	94.66%
	teste	55.14%	36.68%	37.29%	71.56%	36.99%	58.19%	56.60%	56.47%	58.91%	56.53%
C5.0	treino	99.04%	99.31%	97.86%	99.41%	98.58%	99.86%	99.86%	99.86%	99.86%	99.86%
	teste	58.19%	36.51%	36.24%	70.65%	36.38%	56.42%	56.03%	54.94%	58.15%	55.48%

- Na variável *stars,* à exceção do KNN, todos os modelos obtiveram *accuracies* superiores a 50% quer em treino quer em teste. No entanto, à exceção da *specificity,* nas restantes medidas de avaliação, aquando da fase de teste, não foi atingida sequer uma percentagem de 50% em nenhum modelo. Desse modo, considera-se que nenhum dos modelos é bom para a previsão da variável *stars.*
- Para ambas as variáveis, as SVM e as árvores de decisão (C5.0), apresentaram uma queda abrupta na percentagem de acerto da fase de treino para a fase de teste, denotando ter ficado demasiadamente agarrados à amostra de treino criando *overfitting* no modelo. Além destes dois, na classificação da variável *starsbinaria* as redes neuronais apresentam também claramente a criação de um modelo com *overfitting.*
- Apesar dos baixos resultados, a regressão logística para classificação da variável *starsbinaria* foi o modelo que sugeriu obter resultados. Este modelo obteve todas as medidas de avaliação superiores a 50% e com uma diferença entre valores de treino e teste mais próximos entre si. No entanto, apesar de algumas medidas chegarem a atingir os 60%, considera-se que a *accuracy* em teste deste modelo ainda é bastante próxima do aleatório para ser considerada um modelo minimamente válido.
- Ainda assim, percebe-se que a variável *starsbinaria* atingiu resultados ligeiramente superiores aos obtidos com a variável *stars* continuando, contudo, a ser bastante baixos.

4.5.2 Avaliação do input 2 – Matriz Sentimentos

Tabela 22 Resultados input 2 sem feature selection

		stars					starsbinaria				
		Accuracy	Precision	Sensivity (Recall)	Specificity	F-measure	Accuracy	Precision	Sensivity (Recall)	Specificity	F-measure
Naive bayes	treino	64.97%	50.63%	50.85%	76.73%	50.74%	70.98%	70.07%	66.67%	66.67%	68.33%
	teste	47.36%	8.33%	32.98%	67.43%	13.31%	66.75%	58.04%	53.69%	57.10%	55.78%
SVM	treino	85.30%	59.97%	83.02%	90.64%	69.64%	80.76%	80.34%	78.53%	78.53%	79.42%
	teste	51.64%	8.81%	37.77%	62.88%	14.29%	65.49%	54.83%	55.53%	55.53%	55.18%
Regressão logistica	treino	61.95%	57.01%	52.32%	74.60%	54.56%	71.92%	70.91%	68.14%	68.14%	69.49%
	teste	58.19%	35.67%	35.65%	68.96%	35.66%	65.74%	56.14%	55.69%	56.85%	55.92%
KNN	treino	-	-	-	-	-	-	-	-	-	-
	teste	-	-	-	-	-	-	-	-	-	-
Redes neuronais	treino	89.56%	90.40%	90.43%	93.47%	90.42%	94.14%	93.88%	93.83%	93.83%	93.85%
	teste	44.08%	30.58%	27.74%	60.01%	29.10%	52.39%	48.85%	47.38%	48.45%	48.10%
C5.0	treino	96.70%	97.73%	94.69%	97.88%	96.19%	99.26%	99.15%	99.30%	99.30%	99.22%
	teste	53.15%	47.27%	37.58%	70.47%	41.87%	66.25%	58.95%	56.78%	59.41%	57.84%

- O modelo KNN não foi capaz de produzir um modelo preditivo através das variáveis independentes nem para a variável *stars* nem para a *starsbinaria*.
- Na variável *stars,* os resultados apresentados pelos modelos foram francamente baixos. As *accuracies* quer de treino e teste foram bastante baixas sendo que a maioria delas está muito próxima dos 50%, em teste.
- Na variável *stars,* à exceção da *specificity*, não houve nenhum modelo capaz de atingir em teste uma percentagem de 50% nas medidas de avaliação, sendo que as naive bayes e as SVM nem chegaram a 15% de F-*measure*. Assim, à semelhança do *input 1* considera-se que não existe nenhum modelo capaz de prever a variável *stars.*
- Para ambas as variáveis, verifica-se novamente a presença de um claro *overfitting* nas redes neuronais assim como no modelo C5.0.
- Na variável binária, as naive bayes e a regressão logística apresentam os melhores resultados, atingindo *accuracies* com percentagens a partir de 65%. Adicionalmente, a *specificity* assim como a F-*measure* destes modelos obtiveram também melhores resultados com percentagens treino e teste mais próximas entre si. No entanto, embora os resultados obtidos não sofram uma diferença demasiadamente grande entre treino e teste, considera-se que os resultados das medidas de avaliação obtidos na fase de teste são ainda baixos sendo maioritariamente inferiores a 60%.
- As SVM da *starsbinaria* embora tenham, na fase de treino, percentagens mais elevadas em todas as medidas de avaliação, comparativamente às naive bayes e a regressão logística, verifica-se uma queda significativa para os resultados de teste.
- No geral, os modelos obtiveram melhores resultados de previsão para a variável *starsbinaria* comparativamente aos da variável *stars*, tal como nos resultados do *input 1.* No entanto, considera-se que todos os resultados foram baixos.

Tabela 23 Resultados input 2 com feature selection

		stars					starsbinaria				
		Accuracy	Precision	Sensivity (Recall)	Specificity	F-measure	Accuracy	Precision	Sensivity (Recall)	Specificity	F-measure
Naive bayes	treino	62.64%	66.27%	49.12%	76.22%	56.42%	70.98%	70.07%	66.67%	66.67%	68.33%
	teste	61.96%	37.40%	35.50%	70.57%	36.43%	66.75%	58.04%	53.69%	57.10%	55.78%
SVM	treino	67.57%	66.60%	59.50%	78.83%	62.85%	80.76%	80.34%	78.53%	78.53%	79.42%
	teste	55.92%	34.67%	34.96%	67.22%	34.81%	65.49%	54.83%	55.53%	55.53%	55.18%
Regressão logistica	treino	59.15%	67.67%	45.55%	73.49%	54.45%	71.92%	70.91%	68.14%	68.14%	69.49%
	teste	56.93%	36.43%	37.18%	71.06%	36.80%	65.74%	56.14%	55.69%	56.85%	55.92%
KNN	treino	62.18%	41.45%	45.40%	74.84%	43.33%	71.14%	37.28%	68.28%	69.16%	48.23%
	teste	59.45%	35.30%	37.38%	69.36%	36.31%	74.45%	36.41%	69.11%	69.11%	47.69%
Redes neuronais	treino	83.87%	88.51%	87.44%	90.16%	87.97%	99.26%	99.16%	99.30%	0.00%	99.23%
	teste	52.39%	44.68%	41.66%	69.06%	43.12%	61.71%	0.00%	55.73%	0.00%	0.00%
C5.0	treino	82.44%	83.07%	78.23%	89.04%	80.58%	74.45%	76.55%	69.11%	69.11%	72.64%
	teste	58.69%	42.23%	44.10%	69.93%	43.15%	69.02%	51.32%	50.99%	50.99%	51.16%

- Na variável *stars* verifica-se, à semelhança dos resultados anteriores, que embora existam *accuracies* e *specificities* superiores a 50%, as restantes medidas de avaliação são bastante baixas na fase de teste sugerindo não passar de uma previsões aleatórias.
- Para ambas as variáveis, as redes neuronais e o algoritmo C5.0 são novamente aqueles que sugerem criar um modelo com *overfitting* justificando os bons resultados na fase de treino e a queda abrupta aquando da fase de teste. Na variável *starsbinaria,* o C5.0 apresenta resultados ligeiramente melhores com uma diferença entre treino e teste menor, no entanto, considera-se que é ainda uma diferença bastante significativa e com resultados bastante baixos em teste.
- O algoritmo KNN foi capaz de criar um modelo com uma *accuracy* e *specificity* satisfatórias principalmente na previsão da variável binária. No entanto ao observar a F-*measure* os resultados obtidos não chegam sequer a atingir os 50% em nem nenhuma das variáveis.
- Na variável binária, à semelhança do experimento anterior, os melhores resultados foram obtidos através das naive bayes e da regressão logística. As naive bayes bem como a regressão logística apresentaram *accuracies* superiores a 60% e que não diferem demasiadamente da fase de treino para a fase de teste. Adicionalmente, as restantes medidas de avaliação são ligeiramente superiores aos restantes modelos e também com percentagens relativamente próximas entre treino e teste, o que contribui também para uma F-*measure* mais elevada e estável.
- As SVM, apesar de não apresentar resultados muito maus tem ainda uma diferença significativa entre os resultados de treino e teste.
- As medidas de avaliação dos três primeiros algoritmos obtiveram percentagens iguais às obtidas sem *feature selection* do *input 2.*
- De um modo geral, resultados com *feature selection* na variável *stars,* apesar de fracos, são melhores quando comparados aos resultados da variável *stars* sem *feature selection.* No entanto, verifica-se que, tal como nos restantes *inputs*, a variável *starsbinaria* obteve melhores resultados do que a variável stars.

5 RESULTADOS (IMPLEMENTAÇÃO)

A fase de implementação passa normalmente pela aplicação do conhecimento extraído e modelos criados na fase de modelação no quotidiano das organizações. Uma vez que neste caso não existe uma implementação propriamente dita, nesta fase pretende responder-se às hipóteses desenvolvidas ao longo da revisão de literatura e explorar os resultados obtidos.

A fase de preparação de dados representou um papel muito importante nesta investigação. A aplicação do BTM sobre o conjunto de dados permitiu extrair conhecimento bastante útil sobre o conteúdo que é mais falado pelos alunos, o que levou a uma compreensão dos dados bastante mais clara. Na fase de modelação, tentou ainda construir-se um modelo que conseguisse prever satisfatoriamente uma variável de três classes (*stars*) ou uma variável binária (*starsbinaria*) através de dois *inputs.* No entanto, os resultados obtidos pelos algoritmos de previsão foram bastante baixos, refletindo que a amostra de dados dados não era suficiente para criar um modelo que preveja eficazmente a satisfação dos alunos.

Posto isto, inicialmente, serão explanadas as razões que levaram à denominação de cada tópico e, de seguida serão validadas as hipóteses de acordo com os resultados do BTM (Figura 24). Antes de mais, é importante salientar-se que todas as *reviews* estão associadas a todos os tópicos, apenas difere a distribuição da proporção dos tópicos por *review.* Ou seja, cada *review* aborda vários aspetos da experiência internacional. Deste modo, é natural que existam alguns termos que pareçam nada ter a ver com a designação dada ao tópico simplesmente porque aparecem frequentemente associados a outras palavras que também são mencionadas na *review.* Por outro lado, alguns termos repetem-se nos diferentes tópicos porque, dependendo do contexto, o termo pode ter uma interpretação diferente.

- ***Housing*** Nas cinco primeiras palavras deste tópico encontram-se dois termos relacionados com acomodação nomeadamente: apart derivado da palavra *apartment* e, resid derivado da palavra *residence* – palavra frequentemente associada a residências de estudantes. Nas cinco palavras seguintes aparece o stem accommod derivado de *accomodation.* Ao escreverem sobre a acomodação, os alunos escrevem sobre aspetos associados à mesma, como por exemplo a proximidade da acomodação quer à escola (stem school) quer ao centro da cidade (stems town e centr). São também repetidamente encontrados em *reviews* alunos que preferem uma acomodação no *campus* para poderem estar em maior contacto com alunos internacionais (stem campus e exchang).
- ***Campus facilities*** Este tópico sugere a importância dada às instalações sendo que os primeiros seis termos são todos eles referentes a uma instalação seja ela a cozinha (*kitchen*), casas de banho (*bathroom* e *toilet*) ou até mesmo instalações desportivas (*sport*). Na décima palavra surge ainda o stem facil derivado da palavra *facility/facilities*. Ao ler algumas *reviews* verificou-se que as instalações abordadas são maioritariamente instalações fornecidas pelo *campus*, sendo muitas vezes referentes a instalações da acomodação.
- ***Culture enrichment*** O terceiro tópico tem o stem cultur no topo das dez palavras o que denota a importância dada à oportunidade (stem *opportun)* de conhecimento de uma nova

cultura. Além este termo surgem os stems famili e host sendo que, estes últimos aparecem frequentemente relacionados como "*host family".* Aliás se analisarmos as correlações do termo "*host"* a sua maior correlação, embora baixa (0.23), é exatamente com o termo "famili" sendo que, por vezes, é mencionado nas *reviews* o conhecimento da cultura através da própria família anfitriã. O stem host aparece ainda muitas vezes assoiado a comentários sobre características dos pais anfitrião.

- ***Host city life*** Os termos deste tópico revelam uma relação existente com a vida na cidade anfitriã nomeadamente os vários, cafés (stem café), festas (stem parti) existentes no centro da cidade (stems *town centr)*. As *reviews* mais correlacionadas com este tópico sugerem essencialmente estar relacionadas com a vida noturna no centro da cidade, embora existam também algumas que falam de cafés num âmbito mais diurno. Nas *reviews,* o stem *accent* ocorre muitas vezes associado à comunicação e dificuldades sentidas (ou não) com a população local.
- ***Academic*** Este tópico está claramente relacionado com aspetos académicos percebendo-se facilmente através do stem exam, professor, class, test, lectur, entre outros. Através das *reviews,* verificou-se que o stem system está também relacionado a aspetos académicos frequentemente pelos termos "*education system".*
- ***Expenses*** Nas *reviews* mais associadas a este tópico fala-se essencialmente do custo de vida. Da melhor maneira de gerir o dinheiro, por exemplo através de uma conta bancária (stem *bank* e *account*); de poupar dinheiro, por exemplo através de descontos (stem discount) ou da compra de cartões de comboio para poder viajar (stems card, train, travel).
- ***Tourism*** Muitas das *reviews* incentivavam os alunos aproveitarem a localização do país anfitrião para viajar e conhecer países ao redor. Desse modo considerou-se que este tópico estaria relacionado com os alunos aproveitarem por estar num processo de mobilidade internacional viajarem para países fronteiriços.
- ***Student life*** O tópico vida de estudante é um pouco mais geral, com uma distribuição de stems muito idêntica por vários aspetos. Por exemplo aspetos académicos (school, campus e class), a acomodação (campus e accommod), aspetos financeiros (money, cost, price) e aspetos mais de lazer (travel, friend). Nesse sentido, a designação deste tópico surge por ser um aspeto que tem em conta os vários aspetos principais que um aluno em mobilidade internacional passa ao longo da sua experiência.
- ***Skills development*** Outra questão bastante notada na partilha de experiências internacionais relaciona-se com o desenvolvimento de capacidades. A língua é uma das principais capacidades desenvolvidas sendo que o stem level aparece frequentemente associado a *reviews* cujos autores afirmam ter escolhido determinado país para melhorar o nível (level) de linguagem. Este *stem* aparece também em *reviews* que dizem que ter melhorado fortemente o domínio da língua do país anfitrião ou a inglesa. O stem studi encontra-se nestas *reviews*, por vezes, associado a estudar primeiro a língua antes de viajar para o país por exemplo uma das *reviews* afirma "*before coming to france study french".* O stem program surge também por vezes associado a aspetos de idioma como por exemplo

surge na seguinte *review: "The program is in English, but it is better to learn some French".* Além da língua, várias *reviews* mencionam também a qualidade dos programas e de ensino das instituições nas áreas de estudo.

- ***International Environment*** As *reviews* mais associadas a este tópico têm tendência a abordar o ambiente internacional quer a nível de aulas em que por vezes foram tomados cuidados especiais por existir alunos internacionais na turma, como por exemplo aulas dadas em inglês. Por outro lado, é também abordado o ambiente internacional em atividades sociais por exemplo festas para alunos internacionais.

Tabela 24 Quantidade de reviews por tópico

Tópico	Qtd de Reviews
Housing	8
Campus Facilities	9
Culture enrichment	40
Host city life	22
Academic	3
Expenses	6
Turism	7
Student Life	1747
Skills Development	2
International environment	9
Sem tópico associado	**65**
Total de *reviews*	**1918**

A Tabela 24 representa a distribuição do primeiro tópico mais correlacionado com cada *review.* Nota-se claramente o destaque da quantidade de *reviews* associadas ao tópico *Student Life* comparativamente aos restantes tópicos.

Uma vez exploradas as denominações dos tópicos apresentar-se-ão agora os testes e as analises dos resultados das hipóteses formuladas ao longo do capítulo 2. Começar-se-á pelas hipóteses **H3** e **H2** relacionadas com os tópicos e fatores. Por último, serão testadas as hipóteses **H1a**, **H1b**, **H1c** e **H1d** relacionadas com os termos específicos.

5.1 TESTE DA HIPÓTESE 3

H3	Os critérios de escolha de uma IES apurados na revisão literária são consistentes com o que os alunos mais realçam nos comentários online após a experiência internacional.

Para testar a hipótese **H3** fez-se uma comparação subjetiva a partir do conteúdo dos tópicos detetados pelo BTM (Figura 24) com os critérios e fatores apurados ao longo da revisão de literatura (sintetizados na Tabela 1).

No que diz respeito ao tópico sobre habitação verificou-se, na revisão de literatura, que um dos critérios de escolha de uma IES passava pela importância da disponibilização de alojamento por parte da IES, ou pelo menos, da informação sobre tal (Maringe & Carter, 2007). Nos comentários, embora os alunos falem da forma como lhes foi disponibilizado alojamento (se foi fornecido pelo *campus* ou independente), falam essencialmente das características – sobre as instalações, qualidade da localização, a distância face ao *campus* ou ao centro da cidade – e do ambiente (principalmente sobre o ambiente internacional).

No que diz respeito ao tópico *Campus facilities* verificou-se que a literatura reforça a importância de instalações como bibliotecas e áreas de estudo (Bourke, 2000; Maringe & Carter, 2007; Price et al., 2003) e com menor significância instalações desportivas (Price et al., 2003). Segundo os resultados obtidos, as instalações mais frequentemente observadas nas *reviews* dizem respeito essencialmente a instalações do alojamento fornecido pelo *campus.* No entanto, verificaram-se também bastantes comentários referentes a instalações e atividades desportivas.

Estes resultados sugerem que, além da importância dada às instalações do alojamento, os alunos podem dar importância a uma vida ativa e, consequentemente valorizar a disponibilização de instalações desportivas.

O tópico *Culture enrichment* apresenta uma tendência de comentários que falam sobre a experiência a nível multicultural e fazer-se novas amizades. Assim, este tópico vai de encontro ao que se esperava segundo a revisão de literatura uma vez que o desejo de conhecer novas culturas é apontado como uma das principais motivações para uma experiência internacional (Bourke, 2000; Counsell, 2011; Maringe & Carter, 2007). E, o conhecimento de pessoas novas vai de encontro com o desejo de aumentar a rede de contactos internacional (Cubillo et al., 2006; Maringe & Carter, 2007).

Quanto ao tópio *Host city life*, a literatura apontava já a importância da cidade para a escolha da IES mencionando que as características pelas quais a cidade era reconhecida internacionalmente poderiam atrair estudantes (Cubillo et al., 2006). Contudo, de acordo com os resultados, os comentários parecem falar da cidade numa vertente mais festiva e boémia do que propriamente numa vertente cultural. Adicionalmente, verifica-se uma importância dada a distâncias entre o centro da cidade ao *campus* bem como à acomodação, o que sugere o interesse dado à localização destes face ao centro da cidade.

Ainda em relação à cidade anfitriã, segundo a literatura, um dos critérios de escolha tinha em conta a proximidade de idioma por ser mais fácil comunicar ou uma distância de idioma com o objetivo de desafiar capacidades linguísticas (Beine et al., 2014; Cubillo et al., 2006; Maringe & Carter, 2007). E, efetivamente verificou-se a existência de vários comentários que falavam sobre as dificuldades sentidas em relação a comunicação e sotaques da comunidade local.

Através dos resultados obtidos, o tópico *Academic* denota que existem comentários referentes ao corpo docente bem como relacionados com outros aspetos académicos como as aulas e os exames. Este tópico, vai de encontro ao que se esperava de acordo com a literatura uma vez que esta defendia que uma boa reputação de ensino assim como a qualidade e experiência do *staff,* principalmente o corpo docente, eram critérios decisivos para a escolha da IES (Maringe & Carter, 2007; Mazzarol & Soutar, 2002).

O tópico *Expenses* fez-se "corresponder" completamente a um dos fatores apurados no estado da arte – o fator "F3. Fatores Financeiros". No tópico de despesas apurado pelo BTM notam-se referências essencialmente a custos de deslocação dentro e fora do país, dicas de poupar de como poupar dinheiro e de como o gerir. Contudo, não se verificaram comentários referentes aos custos académicos. De facto, já na revisão de literatura, Beine et al. (2014) apontava como tendo maior peso o custo de vida comparativamente aos custos académicos.

Ao longo da revisão do estado da arte, a referência ao interesse em conhecer novas culturas e à troca de vivências com pessoas (Maringe & Carter, 2007) traduz a coerência do tópico *Tourism*. A análise dos dados refletiu que frequentemente, viagens para países fronteiriços fazem parte da experiência internacional destes alunos. Deste modo, acredita-se que a localização geográfica do país talvez também possa influenciar a escolha do destino de estudo.

Considera-se o *Student life* é bastante transversal acabando por abranger os restantes tópicos equitativamente e representando, de modo geral, a experiência internacional como um todo. Na verdade, um dos comentários mais frequentes ao longo das *reviews* era simplesmente um apelo para que os alunos aproveitassem a experiência independentemente de qualquer outro aspeto mais positivo ou negativo.

No tópico *Skills development* os resultados obtidos referem-se essencialmente a melhorias de *hard skills*, especialmente capacidades linguísticas. Já no estado da arte, uma das principais motivações pessoais consistia em querer melhorar as capacidades linguísticas (Bourke, 2000; Counsell, 2011). Embora não surja como um termo nos tópicos apresentados, a língua inglesa sugere ter bastante relevância, no sentido em que vários comentários falam sobre o idioma dos programa académicos serem ou não inglês.

Na revisão literária vários autores referiram a importância da promoção de um ambiente internacional pois tal funcionava como atração para potenciais alunos internacionais (Beine et al., 2014; Cubillo et al., 2006; Maringe & Carter, 2007). Através dos resultados obtidos no tópico *International environment*, verificou-se que os alunos efetivamente enfatizavam bastante o ambiente internacional não só em atividades sociais mas também nas próprias aulas e no próprio alojamento.

A tabela 25 resume a relação explorada anteriormente entre os tópicos encontrados e os fatores evidenciados na literatura.

Tabela 25 Síntese comparativa da revisão literária com os tópicos do BTM

Tópico BTM	Factor	Motivos de escolha da IE (segundo a revisão literária)	Sobre o que mais falam após experiência (recorrendo ao BTM)
Housing	**F2. Instituição**	2.8 Disponibilização de alojamento;	Características das instalações, localização e ambiente do alojamento;
Campus Facilities	**F2. Instituição**	2.6 Disponibilização de bibliotecas e áreas de estudo;	Instalações de alojamento; Instalações desportivas.
Culture Enrichment	**F4. Motivações Pessoais**	4.1 Experiência internacional; 4.3 Conhecer nova cultura; 4.6 Aumentar redes de contactos;	Oportunidade de conhecer novas culturas; Fazer novas amizades;
Host city life	**F1. País** **F4. Motivações Pessoais**	1.2 Proximidade/distância de idioma; 1.5 Características pelas quais a cidade é reconhecida; 4.1 Experiência internacional;	Vida boémia na cidade; Distância entre o centro da cidade ao *campus* ou do alojamento; Dificuldades de comunicação com os locais;
Academic	**F2. Instituição**	2.1 Reputação de ensino; 2.2 Reputação e experiência do *staff*;	Corpo docente; Características das aulas e dos métodos de avaliação;
Expenses	**F3. Fatores Financeiros**	F3. Fatores financeiros;	Custo de vida; Custos de viagens; Dicas de como poupar e gerir dinheiro;
Tourism	**F4. Motivações Pessoais**	4.1 Experiência internacional; 4.3 Conhecer nova cultura; 4.6 Aumentar redes de contactos;	Aproveitar para viajar para países fronteiriços;
Student life	**F1. País** **F2. Instituição** **F3. Fatores Financeiros** **F4. Motivações Pessoais**	2.5 Vida social dentro e fora da instituição; 4.1 Experiência internacional;	Vida social; Despesas; Experiência internacional;
Skills development	**F4. Motivações Pessoais**	2.3 Disponibilização e qualidade do curso; 4.5 Melhorar capacidades linguísticas;	*Hard skills*, principalmente capacidades linguísticas;
International Environment	**F1. País** **F2. Instituição**	1.7 Ambiente internacional;	Ambiente internacional dentro e fora da instituição;

Embora nem todos os critérios apurados no estado da arte se tenham destacado nas *reviews* escritas pelos alunos após a sua experiência, nota-se uma grande coerência com os resultados obtidos pelo BTM, o que confirma a **H3**. Verifica-se que as motivações pessoais são

bastante enfatizadas nas *reviews,* principalmente o conhecimento de novas culturas, o desenvolvimento de capacidades linguísticas e a vontade de querer participar numa experiência internacional. Embora as *reviews* abordem aspetos académicos, observou-se um maior destaque dado ao lazer e à vida noturna. As *reviews* sugerem também que os alunos valorizam a disponibilização de alojamento bem como um bom ambiente internacional no alojamento, na instituição e na própria cidade anfitriã. No que diz respeito aos fatores financeiros, observaram-se bastantes referências às despesas mais num contexto de custo de vida do que num contexto académico. Adicionalmente, verificaram-se várias referências a viagens quer dentro da própria cidade quer para países fronteiriços como era também já expectável a partir da revisão literária. Por último, embora com menos ênfase na literatura observou-se que era dada alguma importância à prática de atividades desportivas.

Verificou-se ainda que os fatores, apesar de não se encontrarem delineados exatamente como no estado da arte, enquadravam-se nos vários tópicos apurados pelo BTM. Observando a Tabela 25 percebe-se que a maioria dos fatores decompõe-se em mais do que um tópico. De acordo com a revisão literária, o fator "F1. País" abrange aspetos relacionados com a cidade e com o ambiente internacional e, como tal, abrange essencialmente o tópico *Host city life* e em parte o *International environment.* Por sua vez o fator "F2. Instituição" diz respeito tanto a fatores académicos como a instalações do *campus* desse modo, abrange fortemente os tópicos *Housing* (uma vez que este é essencialmente referente à acomodação fornecida pelo *campus*)*, Campus facilities, Academic* e, o *International environment* numa vertente respeitante ao ambiente internacional promovido pela instituição. O fator "F3. Fatores financeiros" teve uma correspondência direta com um tópico, o tópico *Expenses*. O surgimento deste tópico denota a importância que tem o custo de vida e outras despesas associadas à vida do estudante ao longo da experiência. Por último, os tópicos *Culture enrichment, Tourism, Student life* e *Skills development* encontram-se bastante relacionados com o tópico "F4. Motivações Pessoais".

Uma vez feita uma análise descritiva e comparativa entre os tópicos e a informação extraída ao longo da revisão literária, pretende analisar-se a relação destes com a satisfação geral dos alunos internacionais em relação a uma IES, por forma a testar a hipótese **H2**.

No entanto, antes de mais, é importante referir que, devido à abrangência de aspetos que englobava, o tópico *Student life,* estava correlacionado em primeiro lugar com a maioria das *reviews* pelo que foi suprimido das restantes análises. Tomou-se esta decisão para ser possível distinguir melhor as *reviews* entre si através de uma distribuição de tópicos mais equitativa. Como já explanado, todos os tópicos estão correlacionados com as todas *reviews,* simplesmente em diferentes proporções. Assim, as *reviews* que estavam correlacionadas com o tópico *Student life* em primeiro lugar foram associadas ao segundo tópico mais correlacionado. E, por sua vez, as

restantes *reviews* que não tinham o tópico *Student life* associado em primeiro lugar, permaneceram com tópico mais correlacionado segundo o BTM.

Posto isto, a distribuição de tópicos passou a ser a apresentada na Tabela 26.

Tabela 26 Quantidade de reviews por tópico sem o Student life

Tópico	Qtd de Reviews
Housing	243
Campus Facilities	27
Culture enrichment	608
Host city life	386
Academic	29
Expenses	216
Turism	114
Skills development	54
International environment	176
Sem tópico associado	**65**
Total	**1918**

5.2 TESTE DA HIPÓTESE 2

H2	A satisfação dos alunos face a uma IES difere de forma significativa consoante o fator.

Para testar a hipótese **H2** teve-se em conta a variável dependente ov_stars, pois como já explicado, considera-se que esta representa a satisfação do aluno de um modo geral. Os fatores referidos na literatura são apenas teóricos pelo que não é possível testá-los estatisticamente sem que se faça uma recolha prévia de dados que meçam cada fator da literatura. No entanto, como já explorado, os fatores vão bastante de encontro aos tópicos apurados pelo BTM e, desse modo, a variável independente será, não o fator da literatura, mas sim o tópico mais correlacionado com cada *review* consoante o BTM.

Recorreu-se então ao SPSS Statistics[20] com o propósito de fazer um teste de hipóteses que analisasse as variações na amostra, sendo primeiramente essencial tomar a decisão se este seria paramétrico ou não-paramétrico. Os testes paramétricos caracterizam-se por ser mais robustos no entanto, necessitam de respeitar três pressupostos base para que possam ser aplicados: (1) a amostra de dados deve seguir uma distribuição aproximadamente normal, (2) deve haver homogeneidade de variância e, (3) devem ser aplicados em variáveis contínuas e com intervalos bem delineados. Por sua vez, os testes não paramétricos caracterizam-se por ser menos rígidos não implicando que os dados sigam uma distribuição normal e permitindo que as variáveis a analisar possam ser ordinais ou categóricas.

Além da variável dependente não ser contínua, através dos testes à normalidade, Kolmogorov-Smirnoff e do Shapiro-Wilk, verificou-se que também não segue uma distribuição normal em nenhum dos tópicos, uma vez que a significâncias dos testes têm todas $p<0.05$ (Figura 35). Posto isto, conclui-se que a amostra de dados a analisar não pode ser testada através de um teste paramétrico.

[20] http://www-01.ibm.com/software/analytics/spss/products/statistics/

		Kolmogorov-Smirnov			Shapiro-Wilk		
	Tópico	Statistic	df	Sig.	Statistic	df	Sig.
OV_Stars	Housing	0.234	0.000	0.000	0.872	243	0.000
	Campus Facilities	0.312	0.000	0.000	0.796	27	0.000
	Culture Enrichment	0.265	0.000	0.000	0.843	608	0.000
	Host City Life	0.257	0.000	0.000	0.846	386	0.015
	Academic	0.215	0.001	0.001	0.907	29	0.000
	Expenses	0.237	0.000	0.000	0.849	216	0.000
	Tourism	0.223	0.000	0.000	0.881	114	0.000
	Skills development	0.223	0.000	0.000	0.832	54	0.000
	International Environment	0.249	0.000	0.000	0.849	176	0.000

Figura 35 Testes de normalidade OV_Stars

Assim, face à questão a analisar, o teste não-paramétrico Kruskal-Wallis é o mais indicado uma vez que este visa analisar variâncias nas distribuições tendo em conta mais do que dois grupos e variáveis ordinais.

Embora o Kruskal-Wallis permita que a distribuição não seja normal, continua a existir como requisito que exista homogeneidade de variâncias nos vários grupos, isto é, a variabilidade dos tópicos deve ser aproximadamente a mesma. O teste de Levene é frequentemente utilizado para analisar a homoscedasticidade no entanto, como o pressuposto da normalidade não se verificou, foi necessário fazer-se um teste de Levene não-paramétrico (Figura 36) (Nordstokke & Zumbo, 2010; Nordstokke, Zumbo, Cairns, & Saklofske, 2011). Desse modo, consideram-se as seguintes hipóteses de teste:

H0: A variância é igual entre os diferentes tópicos.

H1: Existe pelo menos um tópico cuja variância difere dos restantes.

Uma vez que p>0.05, a hipótese nula foi aceite assumindo-se que a variância é igual entre os diferentes tópicos.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	85157,942	8	10644,743	0.235	0.984
Within Groups	83510511,77	1844	45287,696		
Total	83595669,71	1852			

Figura 36 One-way Anova - Teste Levene não paramétrico

Uma vez verificados os pressupostos do Kruskal-Wallis, procedeu-se ao teste propriamente dito, considerando as seguintes hipóteses de teste:

H0: A variância da satisfação geral é igual em todos os tópicos.

H1: Existe pelo menos um tópico cuja variância da satisfação geral difere dos restantes.

Ranks

	Tópico	N	Mean Rank
OV_Stars	Housing	243	843,92
	Campus Facilities	27	847,74
	Culture Enrichment	608	973,53
	Host City Life	386	949,57
	Academic	29	746,21
	Expenses	216	917,36
	Tourism	114	843,56
	Skills development	54	885,35
	International Environment	176	952,08
	total	1853	

Test Statistics

	OV_Stars
Chi-Square	20,681
df	8
Asymp. Sig.	0.008

Figura 37 Resultados do teste Kruskal Wallis

Os resultados do Kruskal-Wallis (Figura 37) apresentaram *mean ranks* distintos sendo o *Culture Enrichment* e o *Academic* aqueles que apresentam maior e menor *mean rank* respetivamente. Adicionalmente, com p=0.008, a hipótese nula foi rejeitada indicando que a satisfação da experiência dos alunos difere significativamente entre pelo menos dois tópicos, confirmando a **H2**. A rejeição da hipótese nula, embora permita perceber que existem diferenças entre os tópicos, não indica quais são os tópicos que efetivamente diferem entre si. Desse modo, é necessário recorrer-se a um teste *post-hoc* que faça uma comparação entre pares.

O SPSS, ao detetar um teste de Kruskal-Wallis significativo, realiza automaticamente o teste *post-hoc* Dunn-Bonferroni que compara os diferentes grupos dois a dois.

Ao analisar os resultados (Tabela 27[21]), verificou-se que existia uma diferença significativa entre o tópico *Housing* e o tópico *Culture Enrichment*. Uma vez confirmada a H2, os resultados demonstram que os alunos que abordam essencialmente aspetos relacionados com o enriquecimento cultural sentem-se tendencialmente mais satisfeitos com a sua experiência internacional quando comparados aos alunos que abordam aspetos essencialmente relacionados com a habitação, uma vez que o *mean rank* do tópico *Culture Enrichment* é superior ao do tópico *Housing*. Por sua vez, no que diz respeito aos restantes tópicos, as diferenças existentes nos *mean ranks* não foram consideradas significativas, assumindo-se desta forma que a satisfação geral é semelhante entre os alunos que abordam fundamentalmente qualquer um dos restantes tópicos.

Tabela 27 Significâncias ajustadas do post hoc teste Dunn Bonferroni

	Academic	Tourism	Housing	Campus Facilities	Skills Development	Expenses	Host city life	International environment	Culture Enrichment
Academic		1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.664
Tourism			1.000	1.000	1.000	1.000	1.000	1.000	0.435
Housing				1.000	1.000	1.000	0.396	1.000	0.028
Campus Facilities					1.000	1.000	1.000	1.000	1.000
Skills Development						1.000	1.000	1.000	1.000
Expenses							1.000	1.000	1.000
Host city life								1.000	1.000
International environment									1.000

[21] Vide anexo B

Considerou-se ainda interessante observar a mediana dos vários tópicos. Segundo a Figura 38 verifica-se que todos eles têm uma mediana positiva rondando as quatro estrelas. Contudo, três dos tópicos apresentam ter uma mediana ligeiramente superior nomeadamente o *Culture enrichment,* o *Host city life* e o *international environment.* Além da mediana ser superior, verifica-se que a dispersão destes tópicos à sua volta, embora exista, é menor comparativamente à maioria dos restantes. Estes tópicos sugerem que satisfação dos alunos deve-se muito a aspetos relacionados com lazer e multiculturais.

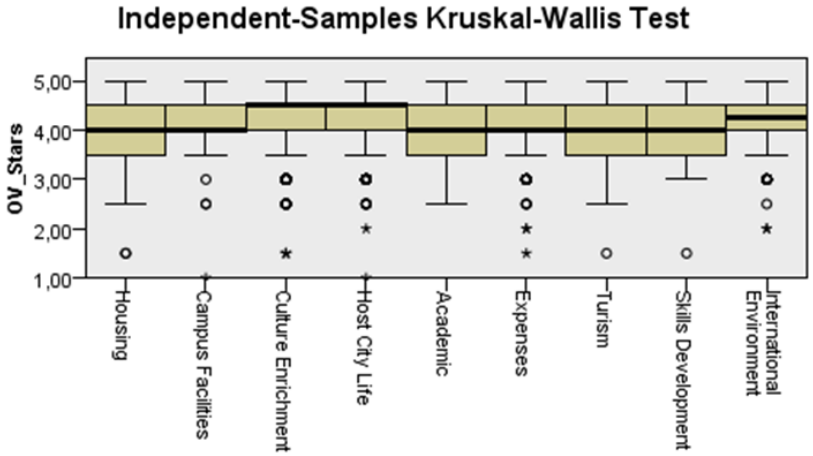


Figura 38 Mediana dos tópicos

Por fim, serão agora analisadas as hipóteses associadas ao sentimento dos termos com a satisfação dos alunos face às IES. Para tal, as próximas análises serão feitas tendo em conta o valor de sentimento dos termos ponderado pelo Semantria.

Inicialmente, criou-se uma *wordcloud* de temas para ter uma ideia geral dos sentimentos do conjunto de dados de acordo com o Semantria. Através da *wordcloud* (Figura 39) surge como tema principal os alunos internacionais (*international students)*. E, analisando os vários temas de

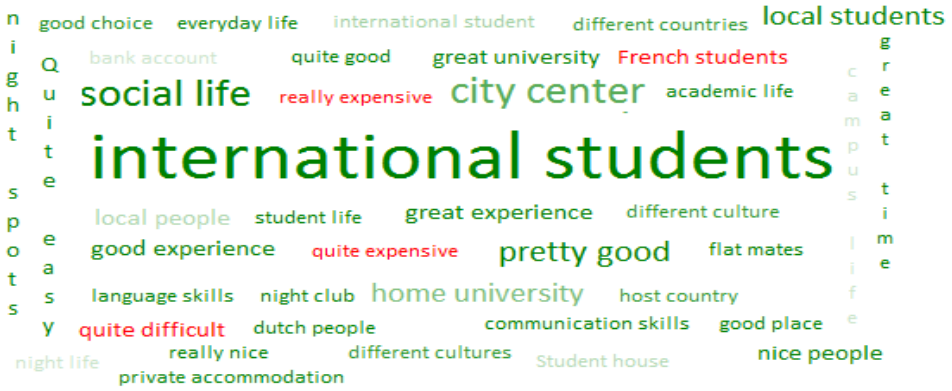


Figura 39 Wordcloud de temas detetados pelo Semantria

um modo geral percebe-se que estes se enquadram com a maioria dos tópicos detetados pelo BTM sendo que existem várias referências a questões relacionadas com despesas (por exemplo *really expensive, quite expensive, bank account*), enriquecimento cultural (*different culture*, *local people*), acomodação (*student house, private accomodation),* desenvolvimento de capacidades (*language skills, communication skills*), aspetos académicos (*academic life, great university*), a vida na cidade (*city center*, *night spots, night life*), a vida do estudante (*social life, student life*), turismo (*different countries*) e ambiente internacional (*international students*). Os sentimentos de

cada tema e a sua força são apresentados através de um espectro de cores desde o vermelho (sentimento negativo), ao verde (sentimento positivo) passando pelo cinza (sentimento neutro) no meio. Desse modo, percebe-se que essencialmente temas relacionados com despesas e dificuldades estão mais associados a sentimentos negativos; E, temas mais relacionados com um ambiente internacional e vida social estão mais associados a sentimentos positivos.

Além dos temas, exploraram-se ainda as entidades através de uma *wordcloud* (Figura 40). Na *wordcloud* de entidades observa-se que a entidade que mais se destaca, e pela positiva, é *English.* Adicionalmente, verifica-se também a existência de um termo ligeiramente negativo relacionado com despesas – *costs.*

Após a análise exploratória dos dados classificados com o Semantria, procedeu-se aos testes das hipóteses **H1a, H1b, H1c** e **H1d**.

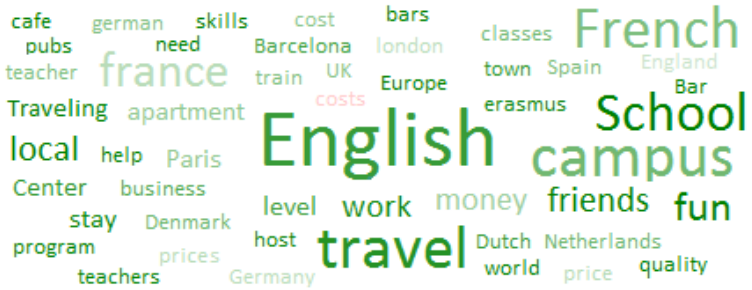


Figura 40 Wordcloud de entidades detetadas pelo Semantria

5.3 TESTE DA HIPÓTESE 1

H1	**H1a.** Com base na revisão literária, o sentimento dos termos relacionados com o país está positivamente correlacionado a satisfação dos alunos face a uma IES.
	H1b. Com base na revisão literária, o sentimento dos termos relacionados com a instituição está positivamente correlacionado com a satisfação dos alunos face a uma IES.
	H1c. Com base na revisão literária, o sentimento dos termos relacionados com questões financeiras está positivamente correlacionado com a satisfação dos alunos face a uma IES.
	H1d. Com base na revisão literária, o sentimento dos termos relacionados com motivações pessoais está positivamente correlacionado com a satisfação dos alunos face a uma IES.

A fim de analisar a correlação entre o sentimento dos termos com a satisfação dos alunos (ov_stars), recorreu-se ao teste não-paramétrico Kendall-tau. Este teste sugere ser bastante indicado para a análise em causa pois é não-paramétrico, não tendo como pressuposto a existência de uma distribuição aproximadamente normal das variáveis nem tão pouco implica uma relação de linearidade entre as variáveis, como acontece com o teste paramétrico pearson. Por outro lado, o teste de Kendall-tau é também mais indicado existem amostras de tamanho inferior a 20 (Pal, 1998, p. 433), o que devido aos dados desta amostra serem muito esparsos acontece em alguns casos. De facto, o trabalho de Long & Cliff (1997) mostra que o Kendall-tau

tem um bom desempenho quando o N é pelo menos 10 e, Pal (1998, p.429) sugere mesmo a eficácia deste teste para amostras a partir de seis elementos.

Os resultados do teste Kendall (Tabela 28) apresentaram algumas correlações

Tabela 28 Resultados do teste Kendall-tau

Termos	Correlation Coefficient	Sig.	N
campus_accomodate	0.791*	0.028	7
good_english	0.708*	0.024	8
different_country	0.429*	0.013	24
job	0.315**	0.008	44
german	0.208*	0.026	65
world	0.206*	0.020	74
money	0.147*	0.029	131
School	0.104*	0.041	218
ET_Place	0.091**	0.001	808
ET_Experience	0.067*	0.028	606
ET_Communication	0.054*	0.029	942
ET_Academic	0.051*	0.037	936

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

significativas entre os sentimentos de alguns termos, temas e *entity types* com a ov_stars. As correlações caracterizam-se por ser todas positivas mas, embora significativas, não são muito fortes, à exceção do tema *different_country* que têm uma correlação moderada (r_T=0.429, p=0.013) e, dos temas *good_english* e *campus_accomodate* com uma correlação relativamente forte (r_T=0.708, p=0.024; r_T=0.791, p=0.028), principalmente este último. Por sua vez, em relação aos *entity types* verifica-se que as correlações são francamente baixas (não atingindo sequer um coeficiente de correlação de 0.1). No que diz respeito ao sinal das correlações, é natural que todas elas sejam positivas uma vez que está a analisar-se um valor de sentimento, ou seja, é normal que à medida que o sentimento sobre algo aumente a sua satisfação aumente consequentemente, embora a correlação possa não ser significativa. Observando, desde já, estas correlações, julga-se que as fracas correlações existentes foram um dos motivos para não ter sido possível construir um modelo preditivo eficaz.

Após verificar-se quais as palavras com correlação significativa, verificou-se se estas tinham uma relação de monotonicidade com a variável ov_stars, uma vez que este é um dos pressupostos do teste de Kendall para que os resultados possam ser corretamente interpretados[22]. Ao analisar a relação monotónica entre as variáveis, percebe-se que nem todas têm uma relação monotónica com a ov_stars. Na verdade, apenas o termo *job* e o tema *different_country,* embora com poucos casos, têm uma relação monotónica significativa com a satisfação geral. Os temas *campus_accomodate* e até mesmo o *good_english* poderiam eventualmente ter uma correlação monotónica com a ov_stars mas, considera-se que a quantidade de casos é muito baixa pelo que não é possível ter-se maior firmeza acerca dessa relação. Relativamente às restantes variáveis não se considera que estas tenham uma relação monotónica significativa com a ov_stars.

[22] Vide anexo C

No que diz respeito às hipóteses a analisar, quer os termos, quer os temas, quer as *entity* confirmam vários aspetos salientados pela literatura existente. No entanto, os *entity types* não serão muito explorados uma vez que têm um coeficiente de correlação muito baixo e não se considera que exista uma relação monotónica com a ov_stars significativa.

Em relação à hipótese **H1a** considerou-se *entity type Place* está essencialmente associado ao país. Alguns dos termos mais frequentes que dizem respeito ao *entity type Place* são "*France*", "Paris", "*Spain*" e "UK". Existe ainda uma expressão que dado à sua frequência foi, toda ela, considerada uma *entity type* nomeadamente "*Scotland where the rain never stops*". Nas *reviews*, os termos que têm o *Place* como *entity type* descrevem, normalmente, aspetos relacionados com as características e a cultura o país onde estão a viver. Estes aspetos vão de encontro àquilo que a literatura já sugeria ao realçar o interesse dos estudantes em querer ir para determinado país ou cidade devido às suas características e/ou devido à distancia/proximidade cultural desse mesmo país com o seu país de origem.

O termo *School*, o tema *campus_accomodate* e o *entity type Academic* permitem avaliar a validade da **H1b**. No que diz respeito ao *entity type*, alguns dos termos mais frequentes são *"class"*, "*teacher"* e "*exam*". Quanto aos termos, o termo *School* tem uma correlação com o aumento da satisfação em relação à IES, que embora esta relação esteja de acordo com o espectável, a monotonicidade entre esta variável e a satisfação é baixa. O tema *campus_accomodate* é aquele com maior coeficiente de correlação (r_T=0.791, p=0.028) referindo-se à acomodação do *campus.* Efetivamente, a literatura salientou a importância da disponibilização de acomodação por parte das IES (Maringe & Carter, 2007) e desse modo, faz bastante sentido que exista uma correlação significativa entre este termo e a satisfação face à IES. Considera-se que a quantidade de *reviews* associada a este termo é bastante baixa. No entanto, Pal (1998, p.429) sugere que o Kendall-tau tem um desempenho aceitável quando N>5.

A hipótese **H1c**, referente às motivações pessoais, foi aquela que teve mais correlações associadas. Alguns dos termos mais frequentes do *entity type Experience* são: "*travel"*, "Erasmus", "*world", "trips".* Alguns dos termos mais frequentes do *entity type Communication* são: "*English*", "*german*" e "*accent"*. Quanto aos termos estatisticamente correlacionados com as motivações pessoais, o termo *world* aparece nas *reviews* essencialmente relacionado o conhecimento de pessoas de todo o mundo. E, efetivamente, na revisão literária foi muito enfatizada a vontade de aumentar a rede de contactos (Cubillo et al., 2006). Por sua vez, o termo *german* surge mais ligado às motivações pessoais num sentido de desenvolvimento de capacidades, neste caso, linguísticas ou dificuldades que sentiram por ser uma língua complexa. Novamente destaca-se que estes termos, embora já com um coeficiente de correlação superior ao dos *entity types,* continuam a ter coeficientes ainda baixos.

Com correlações mais fortes e ainda associado com a hipótese **H1c** surge tema *good_english* e o *different_country*. Em relação ao *good_english* foram vários os autores a referir a importância das línguas para os estudantes, especialmente o Inglês (Bourke, 2000; Counsell, 2011). O interesse em encontrar uma instituição que lecionasse cursos em Inglês, ou apenas o facto de existir uma diferença cultural que permitisse o desenvolvimento do Inglês durante a

experiência foram critérios que surgiram na literatura com bastante relevância. Considera-se que a quantidade de *reviews* (N=8) é francamente baixa. No entanto, esta variável encontra-se bastante alinhada com a literatura e vai bastante de encontro com o que se esperava obter. O tema *different_country* é outro cuja correlação faz todo o sentido de acordo com a literatura. Embora não esteja diretamente relacionado com características das IES, está bastante relacionado com interesse em aumentar a rede de contactos aquando de uma experiência internacional. Nas *reviews* este tema, à semelhança do termo *world,* aparece essencialmente associado ao conhecimento de novas pessoas de países diferentes como, por exemplo, ocorre no excerto da seguinte *review:* "*you have the opportunity to meet a lot new people from many different countries*". O tema *different_country* tem um coeficiente de correlação moderado (r_T=0.429, p=0.013) assim como um N e uma relação de monotonicidade aceitáveis, permitindo extrair uma relação de que quanto maior for o sentimento em relação ao termo *different_country* mais a satisfação geral aumenta, suportando a **H1c**.

Por último, o termo *job* (r_T=0.315, p=0.008) encontra-se associado a fatores financeiros e às motivações pessoais. O valor do coeficiente indica a existência de uma correlação moderada entre o aumento do sentimento em relação ao termo job e a satisfação dos alunos face a uma IES. Com um N>30, considera-se razoável a quantidade de reviews que refere o termo job. Através da leitura de reviews, verificou-se que o termo job reflete consistentemente a importância de dois aspetos mencionados por autores ao longo da revisão de literatura. Um deles, mais correlacionado com motivações pessoais, suportando a **H1d**, centra a importância das oportunidades de trabalho no futuro, ou seja, numa vertente pós-estudos como por exemplo exprime o excerto da seguinte *review*: "*They should try and promote more to international students and should provide internships and full time jobs after completion of studies*". O outro aspeto mais correlacionado com fatores financeiros, suportando a **H1c,** no sentido de os alunos quererem arranjar um part-time durante a experiência, observe-se por exemplo o excerto de uma review com o termo *job*: "*It's a good university, but it must start some Placement Committee for the international students to work, for part time and for internships. I enjoyed a lot this university, but the only thing I didn't like was that I was not able to get a part time job for 3 months*"[24]. Acompanhando o *job*, o termo *money* (r_T=0.147, p=0.029) relacionado com custos e como gerir o dinheiro confirma a hipótese **H1d**.

Verificou-se que para todas as hipóteses em teste se confirma que existem variáveis cujo sentimento influencia positivamente a satisfação dos alunos face a uma IES. A Figura 41 apresenta uma síntese das correlações mais associadas por cada hipótese. As hipóteses que sugerem ter uma relação mais forte com a satisfação dos alunos face a uma IES são a **H1c** (motivações pessoais) e a **H1d** (fatores financeiros). Considera-se que as hipóteses **H1a** (país) e **H1b** (instituição) apesar de suportadas são menos fortes.

[24] Alguns erros ortográficos foram corrigidos para uma leitura mais clara

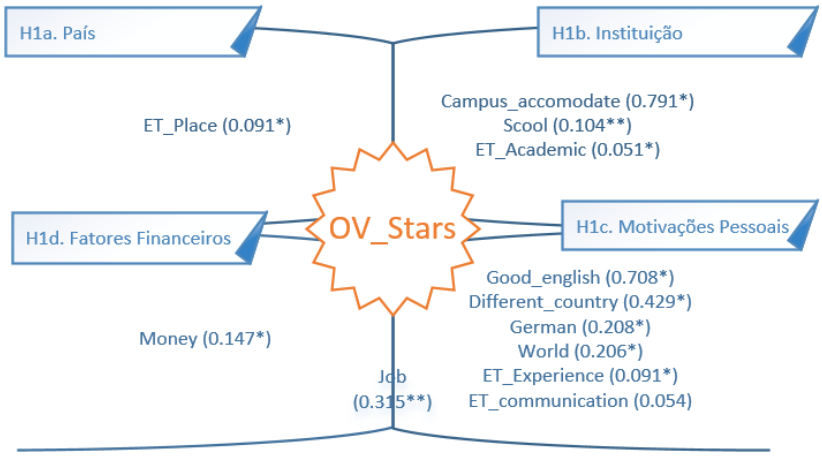


Figura 41 Síntese de correlações com os fatores

Os resultados demonstraram que as motivações pessoais que têm uma relação mais forte com a satisfação dizem respeito essencialmente ao aumento da rede de contactos com pessoas de culturas diferentes bem como ao aumento das oportunidades de trabalho pós-estudos. Ainda relacionado com as motivações pessoais mas menos expressivos, os resultados sugerem que existe uma relação significativa entre a satisfação dos alunos com a língua, especialmente o Inglês. Os fatores financeiros revelaram ter um papel de destaque na satisfação dos alunos principalmente no que diz respeito ao custo de vida do país de destino e na necessidade de conseguir arranjar um *part-time* durante a experiência.

Estes resultados sugerem que a satisfação dos alunos face a uma IES recai essencialmente sobre duas vertentes. Por um lado, uma vertente mais financeira associada aos custos de vida e precisar de trabalhar para suportar a experiencia. Por outro lado, uma vertente associada a motivações pessoais em que o aluno espera enriquecer-se culturalmente e desenvolver capacidades aspirando um bom futuro profissional.

6 CONCLUSÕES, LIMITAÇÕES E TRABALHO FUTURO

Ao escolher uma IES num país estrangeiro, os estudantes recorrem frequentemente a fontes *online*, uma vez que não podem deslocar-se fisicamente à escola e porque consideram as opiniões de ex-alunos mais verdadeiras (Gomes & Murphy, 2003). Estas fontes são capazes de influenciar fortemente a reputação de uma IES e, consequentemente, a tomada de decisão dos alunos. Assim, é essencial que as IES estejam atentas às opiniões dos alunos nos *social media* para que consigam eficazmente gerir sua reputação e ir de encontro às necessidades dos estudantes internacionais.

No entanto, uma vez que comunicação na Internet é expressa numa quantidade enormíssima e tem tantas características próprias, nos dias de hoje, as tradicionais técnicas de *marketing* dificilmente têm capacidade de lidar com toda a informação existente de forma eficiente e objetiva. Torna-se então fundamental recorrer a técnicas capazes de analisar automática e objetivamente este tipo de dados, mais precisamente, técnicas de *text mining*.

Até à data presente, não se encontrou nenhuma investigação que aplicasse técnicas automáticas de análise textual em *reviews online* com o propósito de potenciar a atratividade das IES. Nesse sentido, utilizando técnicas de *text mining* e análise de sentimentos, a presente dissertação debruçou-se em analisar os aspetos que mais se destacaram em *reviews online* que alunos internacionais escreveram sobre a sua experiência, visando ajudar as IES a centrar os seus esforços em aspetos que são realmente importantes para este público-alvo.

Como contributos académicos, os resultados obtidos refletiram que são principalmente fatores financeiros e algumas motivações pessoais influenciam a satisfação dos alunos em relação a uma IES. Em relação aos fatores financeiros, os custos que mais se salientaram dizem respeito essencialmente ao custo de vida do país anfitrião e a deslocações. As *reviews* não revelaram referências significativas relativas ao custo das propinas das IES. Considera-se que esta situação pode dever-se às ajudas financeiras das IES cobrirem o valor das propinas ou, no caso de não cobrirem, o valor alto das propinas poder ser percebido como um sinal de qualidade da IES, como sugere Beine et al. (2014). No entanto, as *reviews* revelaram a necessidade dos alunos quererem trabalhar em *part-time* ao longo da experiência como já evidenciado por (Maringe & Carter, 2007).

Por outro lado, a nível de motivações pessoais, os resultados demonstraram que existe uma preocupação profissional por parte dos alunos a nível de carreira e prospeções futuras tendo-se verificado que era dada bastante importância às oportunidades de trabalho no país anfitrião numa fase pós-estudos. Foram ainda encontradas evidências da importância dada ao desenvolvimento de capacidades principalmente linguísticas, como evidenciado por Bourke (2000) e Counsell (2011). Neste caso, a língua inglesa sugere ter um destaque primordial. Passar por uma experiência multicultural quer a nível de turismo quer a nível de contacto com pessoas de culturas diferentes são também critérios que sugerem influenciar bastante a satisfação. Estar inserido num ambiente internacional é também um dos aspetos com maior destaque nas *reviews* dos alunos como já sugerido por Cubillo et al. (2006).

Outro aspeto que foi também significativamente salientado nos resultados diz respeito à disponibilização de acomodação por parte das IES, tal como indicava Maringe & Carter (2007). As *reviews* sugerem que os alunos têm especial interesse em ficar em acomodações fornecidas pelo *campus* pois querem viver a experiência convivendo com uma maior quantidade de alunos de diferentes países e culturas.

Tendo em conta todo o conhecimento adquirido ao longo dos resultados bem como as análises estatísticas iniciais com os dados estruturados, acha-se que os alunos embora possam passar por alguns aspetos mais negativos acabam por dar uma pontuação tendencialmente mais elevada às IES sugerindo que a experiência de mobilidade internacional vale por si mesma.

No que se refere a contribuições práticas, considera-se que as IES podem melhorar a sua atratividade internacional caso ajam em conformidade com os resultados encontrados. Assim, é importante que as IES forneçam um maior suporte financeiro a nível de custos de vida. No caso de não ser possível fornecer maiores ajudas financeiras, é crítico que suportem as colocações dos alunos internacionais em trabalhos *part-time* ao longo das suas experiência. Além de

trabalhos em *part-time,* as IES devem proporcionar aos alunos internacionais mais ajuda na entrada no mercado de trabalho do país anfitrião após os estudos estarem concluídos.

No que diz respeito a questões académicas, julga-se que o destaque da língua inglesa existe devido ao facto de ser a língua oficial, tornando crucial a disponibilização de cursos em Inglês. É também vantajoso que as IES disponibilizem acomodação aos alunos internacionais. Caso não seja possível, as IES têm o dever de disponibilizar informação e ajuda para encontrar acomodação.

Além das IES darem valor a estes aspetos e efetivamente gastarem esforços em proporcionar aos alunos internacionais aquilo que aqui foi sugerido, é crítico que estas informações sejam facilmente percetíveis pelos alunos numa fase em que estes estão a escolher a IES onde vão estudar. É fundamental que as IES disponham estas informações facilmente e invistam numa boa comunicação e marketing para as evidenciar, como já sugerido por Bourke (2000). Adicionalmente, é importante que o marketing e comunicação tenha em conta não só a vertente académica mas, que promovam também fortemente a cultura do país e o ambiente internacional quer dentro, quer fora da IES como sugere Cubillo et al. (2006) e os resultados da presente investigação.

Apesar dos resultados transmitirem conhecimento interessante sobre as experiências dos alunos internacionais, julga-se que este estudo tem algumas limitações. Considera-se que a fonte de dados de onde foram extraídos os dados não é o tipo de fonte ideal para esta análise. Como o *iagora* funciona como um questionário, os alunos são “obrigados” a responder a perguntas concretas e, como tal, “obrigados” a escrever sobre aspetos que possivelmente não teriam escrito caso fosse um campo de texto aberto livre sem perguntas associadas. Além deste problema, como alguns dos campos eram obrigatórios e os alunos não queriam escrever nada sobre eles, muitas vezes escreviam apenas palavras aleatórias para poder passar para o campo seguinte. Isto trouxe bastante ruido ao conjunto de dados.

Outra limitação refere-se à variável dependente, ov_stars. Embora esta variável seja representativa de uma classificação dada à experiência geral (*overall*), o facto de ser uma ponderação poderá também ter influenciado os resultados.

Devido à limitação do âmbito bem como algumas remoções de *reviews* que foram necessárias fazer-se*,* apenas foi possível analisar 1918 *reviews.* Embora 1918 *reviews* seja uma quantidade de *reviews* significativa, muitas vezes o conteúdo textual era muito pequeno e, a quantidade de erros ortográficos bem como a diversidade de palavras fizeram com que o conjunto de dados fosse mesmo muito esparso. Este problema levou a bastantes dificuldades na deteção de tópicos e com certeza, foi um dos motivos por não se ter conseguido obter um modelo preditivo com melhores resultados. Seria importante no futuro extrair novamente *reviews* com vista a aumentar o conjunto de dados a analisar. No entanto, considera-se também pertinente que se procurem outras fontes *online* com o mesmo conteúdo de informação aqui analisado mas que não seja uma fonte tipo questionário, proporcionando assim uma análise de conteúdo mais livre.

A nível de técnicas utilizadas, julga-se que deveriam ser aprimorados os parâmetros do BTM com o objetivo de aperfeiçoar os tópicos.

A nível de resultados, devido à quantidade de referências quer a viagens para países vizinhos, quer às distâncias entre o campus ou a acomodação ao centro da cidade, acredita-se que a localização da acomodação e do *campus* em relação às cidades assim como a localização geográfica dos países possam ser potenciais critérios relevantes para a escolha de uma IES. Além destes, alguns resultados obtidos sugeriram também que a disponibilização de áreas desportivas poderia ser atrativo para estudantes internacionais. Neste sentido, considera-se que, como trabalho futuro, seria interessante explorar melhor se estas variáveis podem efetivamente ser consideradas critérios de escolha.

BIBLIOGRAFIA

Altbach, P. G., & Knight, J. (2007). The Internationalization of Higher Education: Motivations and Realities. *Journal of Studies in International Education*, *11*(3–4), 290–305.

Aranha, C., & Passos, E. (2008). Automatic NLP for Competitive Intelligence. *Emerging Technologies of Text Mining : Techniques and Applications* (pp. 54–76).

Barnett, M. L., Jermier, J. M., & Lafferty, B. a. (2006). Corporate Reputation: The Definitional Landscape. *Corporate Reputation Review*, *9*(1), 26–38.

Beine, M., Noël, R., & Ragot, L. (2014). Determinants of the international mobility of students. *Economics of Education Review*, *41*(2014), 40–54.

Bennett, R., & Kottassz, R. (2000). Practitioner perceptions of corporate reputation: an empirical investigation. *Corporate Communications: An International Journal*, *5*(4), 224–234.

Binsardi, A., & Ekwulugo, F. (2003). International marketing of British education: research on the students' perception and the UK market penetration. *Marketing Intelligence & Planning*, *21*(5), 318–327.

Blake, C. (2011). Text Mining. *Annual review of information science and technology*, *45*(1), 121–155.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*(2), 1–30.

Blei, D. M., & Lafferty, J. D. (2006). Correlated Topic Models. *Advances in Neural Information Processing Systems*, *18*, 147–154.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Bourke, A. (2000). A Model of the Determinants of International Trade in Higher Education. *The Service Industries Journal*, *20*(1), 110–138.

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the workshop on Speech and Natural Language - HLT '91* (p. 112).

Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. *International Conference RANLP*.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, *28*(2), 15–21.

Caro, L. Di, & Grella, M. (2013). Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, *35*(5), 442–453.

Chaney, A., & Blei, D. (2012). Visualizing Topic Models. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 419–422).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rüdi. (2000). *Crisp-Dm 1.0 Step-by-step data mining guide*. *The CRISP-DM Consortium*. SPSS.

Chapman, R. G. (1986). Toward a theory of college sellection: A model of college search and choice behavior. *Advances in Consumer Research*, *13*(1), 246–250.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*, 853–867.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941.

Choudaha, R., & Chang, L. (2012). *Trends in international student mobility*. *World Education News & Reviews*.

Chowdhury, G. G. (2003). Natural Language processing. *Annual review of information science and technology*, *37*(1), 51–89.

Chun, R. (2005). Corporate reputation: Meaning and measurement. *International Journal of Management Reviews*, *7*(2), 91–109.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Counsell, D. (2011). Chinese Students Abroad: Why They Choose the UK and How They See Their Future. *China: An International Journal*, *9*, 48–71.

Cruz, F. L., Troyano, J. A., Enríquez, F. J. O., & Vallejo, C. G. (2013). "Long autonomy or long delay?"The importance of domain in opinion mining. *Expert Systems with Applications*, *40*(8), 3174–3184.

Cubillo, J. M., Sánchez, J., & Cerviño, J. (2006). International students' decision-making process. *International Journal of Educational Management*, *20*(2), 101–115.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, *53*(9), 1375–1388.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*(3), 1707–1720.

Edgett, S., & Parkinson, S. (1993). Marketing for Service Industries - A Review. *The Service Industries Journal*, *13*(3), 19–39.

Europeia, C. (1999). Declaração de Bolonha.

Europeia, C. (2001). A caminho da área europeia de ensino superior.

Europeia, C. (2013). *O Ensino Superior Europeu no Mundo*.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 77–82.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82.

Gao, S., Hao, J., & Fu, Y. (2015). The Application and Comparison of Web Services for Sentiment Analysis in Tourism. *12th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1–6). IEEE.

Garvin, D. A. (1980). *The Economics of University Behavior*. Academic Press.

Goadrich, M., Oliphant, L., & Shavlik, J. (2006). Gleaner: Creating Ensembles of First-Order Clauses to Improve Recall-Precision Crurves. *Machine Learning*, *64*(1–3), 174–212.

Goda, K., Hirokawa, S., & Mine, T. (2013). Correlation of Grade Prediction Performance and

Validity of Self-Evaluation Comments. *Proceedings of the 14th annual ACM SIGITE conference on Information technology education* (pp. 35–42). ACM.

Gomes, L., & Murphy, J. (2003). An exploratory study of marketing international education online. *International Journal of Educational Management and Marketing*, *17*(3), 116–125.

González, C. R., Mesanza, R. B., & Mariel, P. (2011). The determinants of international student mobility flows: An empirical study on the Erasmus programme. *Higher Education*, *62*, 413–430.

Gray, E. R., & Balmer, J. M. T. (1998). Managing Corporate Image and Corporate Reputation. *Long Range Planning*, *31*(5), 695–702.

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30.

Guerreiro, J., Rita, P., & Trigueiros, D. (2015). A Text Mining-Based Review of Cause-Related Marketing Literature. *Journal of Business Ethics*, 1–18.

Hassler, M., & Fliedl, G. (2006). Text preparation through extended tokenization. *Data Mining VII: Data, Text and Web Mining and their Business Applications*, *37*, 13–21.

Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference on Computational linguistics-Volume 1* (pp. 299–305). Association for Computational Linguistics.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, *29*(1), 90–102.

Hearst, M. (1999). Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3–10).

Hemsley-Brown, J. (2012). The Best Education in the World: reality, repetition or cliché? International students' reasons for choosing an English university. *Studies in Higher Education*, *37*(8), 37–41

Herbig, P., & Milewicz, J. (1993). The relationship of reputation and credibility to brand success. *Journal of Consumer Marketing*, *10*(3), 18–24.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*, 19–62.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International journal of medical informatics*, *74*(2), 289–298.

Hu, M., & Liu, B. (2004a). Mining and summarizing customer reviews. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177).

Hu, M., & Liu, B. (2004b). Mining opinion features in customer reviews. *Proceedings of the National Conference on Artificial Intelligence* (pp. 755–760).

Inniss, T., Lee, J., & Light, M. (2006). Towards applying text mining and natural language processing for biomedical ontology acquisition. *Proceedings of the 1st international workshop on Text mining in bioinformatics* (pp. 7–14). ACM.

Isik, M., Öztaysi, B., & Fenerci, K. H. (2012). A Sentiment Analysis as a Tool to Identify The Status

Of Universities: The Case of ITU. *Procedings of the 2012 International Conference on Industrial Engineering and Operations Management* (pp. 1118–1126).

Ivy, J. (2001). Higher education institution image: acorrespondence analysis approach. *International Journal of Educational Management*, *15*(6), 276–282.

Jackson, P., & Moulinier, I. (2007). *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization*. (J. B. Publishing, Ed.) (2nd ed.).

Jones, B., Temperley, J., & Lima, A. (2009). Corporate reputation in the era of Web 2.0: the case of Primark. *Journal of Marketing Management*, *25*(9–10), 927–939.

Kaplan, R. M. (2005). A Method for Tokenizing Text. *Complexity and Education: Inquiries into Words, Constraints and Contexts* (pp. 55–64).

Kotler, P., & Fox, K. (1995). *Strategic Marketing for Educational Institutions* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lawrence, P. (2014). *Reliability of Sentiment Mining Tools: A comparison of Semantria and Social Mention*.

Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An Empirical Comparison of Four Text Mining Methods. *Proceedings of the 43rd Hawaii International Conference on System Sciences* (pp. 1–10). IEEE.

Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications*, *39*(3), 2584–2589. Elsevier Ltd.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *Machine learning: ECML-98* (pp. 4–15). Springer Berlin Heidelberg.

Lexalytics. (2013). Why Lexalytics Is The Best For Social Media Sentiment Analysis Lexalytics.

Lexalytics. (2015a). *Entity Extraction*.

Lexalytics. (2015b). *Categorization of Text*.

Lexalytics. (2015c). *What is Sentiment Scoring?* Retrieved from https://semantria.com/sentiment-analysis

Lexalytics. (2015d). *Sentiment Extraction*.

Lexalytics. (2015e). *Lexalytics Loves Machine Learning*.

Liaw, S.-S., & Huang, H.-M. (2006). Information retrieval from the World Wide Web: a user-focused approach based on individual experience with search engines. *Computers in human behavior*, *22*(3), 501–517.

Lin, C., He, Y., Everson, R., & Rüger, S. (2012). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 1134–1145.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *Mining Text Data* (pp. 415–460). Springer US.

Long, J. D., & Cliff, N. (1997). Confidence intervals for Kendall's tau. *British Journal of Mathematical and Statistical Psychology*, *50*, 31–41.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *An Introduction to Information Retrieval*. Cambridge: Cambridge university press.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A Data Mining & Knowledge Discovery Process

Model. *A data mining & knowledge discovery process model* (pp. 1–17).

Maringe, F. (2006). University and course choice: implications for positioning, recruitment and marketing. *International Journal of Educational Management*, *20*(6), 466–479.

Maringe, F., & Carter, S. (2007). International students' motivations for studying in UK HE: insights into the choice and decision making of African students. *International Journal of Educational Management*, *21*(6), 459–475.

Mazzarol, T. (1998). Critical success for factors for international education marketing. *International Journal of Educational Management*, *12*(4), 163–175.

Mazzarol, T., & Soutar, G. N. (2002). "Push-pull" factors influencing international student destination choice. *International Journal of Educational Management*.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Minami, T., & Ohura, Y. (2013). Lecture Data Analysis towards to Know How the Students ' Attitudes Affect to their Evaluations. *The 8th International Conference on Information Technology and Application* (pp. 164–169).

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (1st ed.). Academic Press.

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, *18*(3), 11–20.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated Data Collection with R.A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley.

Nicholls, J., Harris, J., Morgan, E., Clarke, K., & Sims, D. (1995). Marketing higher education: the MBA experience. *International Journal of Educational Management*, *9*(2), 31–38.

Nordstokke, D. W., & Zumbo, B. D. (2010). A new nonparametric levene test for equal variances. *Psicologica*, *31*(2), 401–403.

Nordstokke, D. W., Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research & Evaluation*, *16*(5), 1–8.

OECD. (2014). *Education at a Glance 2014: OECD Indicators*. Education at a Glance. OECD Publishing.

Pal, S. K. (1998). *Statistics for geoscientists: techniques and applications*.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (pp. 1–8). Computation and Language.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *1*(2), 91–231.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79–86). Association for Computational Linguistics.

Peffers, K., Tuunanen, T., Rothenberger, M. a., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Peisenieks, J., & Skadiņš, R. (2014). Uses of Machine Translation in the Sentiment Analysis of Tweets. *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT* (Vol. 268, pp. 126–131).

Pontil, M., & Verri, A. (1998). Properties of support vector machines. *Neural Computation*, *10*(4), 955–974.

Porter, M. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3), 130–137.

Price, I., Matzdorf, F., Smith, L., & Agahi, H. (2003). The impact of facilities on student choice of university, 212–222.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, *37*(1), 9–27.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., et al. (2006). Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, *76*(2–3), 195–200.

Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, a., & Ureña-López, L. a. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, *38*(12), 14799–14804. Elsevier Ltd.

Sánchez, D., Martín-Bautista, M. J., Blanco, I., & Torre, C. J. D. La. (2008). Text Knowledge Mining: An Alternative to Text Data Mining. *2008 IEEE International Conference on Data Mining Workshops* (pp. 664–672). Ieee.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38.

Šontaitė-Petkevičienė, M. (2013). The view of students towards corporate reputation of Lithuanian universities. *Management of Organizations: Systematic Research*, *66*(66), 115–127.

Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2015). A Predictive Model to Evaluate Student Performance. *Journal of Information Processing*, *23*(2), 192–201.

Srikatanyoo, N., & Gnoth, J. (2002). Country image and international tertiary education. *Journal of Brand Management*, *10*, 139–146.

Tan, A. (1999). Text Mining : The state of the art and the challenges Concept-based. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (pp. 65–70).

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised

Classification of Reviews. *Proceedings of the 40th annual meeting on association for Computational Linguistics* (pp. 417–424).

Walsh, G., Mitchell, V.-W., Jackson, P. R., & Beatty, S. E. (2009). Examining the Antecedents and Consequences of Corporate Reputation: A Costumer Perspective. *British Journal of Management*, *20*(2), 187–203.

Webster, J., & Kit, C. (1992). Tokenization as the initial phase in NLP. *Proceedings of the 14th conference on Computational linguistics-Volume 4* (pp. 6–10).

Wen, M., Yang, D., & Rosé, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 130–137).

Wiebe, J., Bruce, R., Bell, M., Martin, M., & Wilson, T. (2001). A corpus study of evaluative and speculative language. *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16* (Vol. Proceeding, pp. 1–10). Association for Computational Linguistics.

Wilkins, S., & Huisman, J. (2014). Factors affecting university image formation among prospective higher education students: the case of international branch campuses. *Studies in Higher Education*, (December), 1–17.

Williamson, W., & Ruming, K. (2015). Assessing the effectiveness of online community opposition to precinct planning. *Australian Planner*, *52*(1), 51–59.

Yamanishi, K., & Li, H. (2002). Mining open answers in questionnaire data. *Intelligent Systems, IEEE*, *17*(5), 58–63.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *WWW '13 Proceedings of the 22nd international conference on World Wide Web* (pp. 1445–1456).

Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, *38*(6), 7674–7682. Elsevier Ltd.

ANEXOS

ANEXO A - CONJUNTO DE DADOS EXTRAÍDO

Atributo	Tipo	Desambiguação	Descrição
ID	Numérico	ID	Identificador de review
Pais_Destino	Texto fechado	País destino	Nome do país anfitrião
Universidade	Texto fechado	Universidade	Nome da universidade anfitriã
Cidade_Origem	Texto fechado	Cidade origem	Nome da cidade de origem do processo de mobilidade
Estado_Origem	Texto fechado	Estado Origem	Nome do estado origem do processo de mobilidade
Pais_Origem	Texto fechado	País origem	Nome do país origem do processo de mobilidade
HO_Star	Likert	HO (*Housing*)	Pontuação em estrelas dada à acomodação
HO_Type	Texto aberto e fechado	HO (*Housing*)	Tipo de acomodação onde o aluno ficou instalado
HO_Arrangedby	Texto fechado	HO (*Housing*)	Como conseguiu arranjar acomodação
HO_IfreturncHOose	Texto aberto e fechado	HO (*Housing*)	Se voltasse onde ficaria alojado
HO_Why	Texto aberto	HO (*Housing*)	Justificação do atributo HO_Arrangedby
HO_Pcomments	Texto aberto	HO (*Housing*)	Comentários sobre a acomodação
HO_Pa_Cost	Likert	HO_Pa (*Personal assessement*)	Avaliação dos custos de acomodação
HO_Pa_Facilities	Likert	HO_Pa (*Personal assessement*)	Avaliação de instalações de acomodação
HO_Pa_Location	Likert	HO_Pa (*Personal assessement*)	Avaliação da localização da acomodação
HO_Pa_Cleanliness	Likert	HO_Pa (*Personal assessement*)	Avaliação da higiene da acomodação
HO_Pa_Space	Likert	HO_Pa (*Personal assessement*)	Avaliação do espaço da acomodação

SL_Star	Likert	SL (*Student life*)	Pontuação em estrelas do aspeto vida estudantil (*Student Life*)
SL_Hostcitydescr	Texto fechado	SL (*Student life*)	Tipo de vida da cidade. Por exemplo se é dominada por estudantes ou pela comunidade local
SL_Activities	Texto fechado	SL (*Student life*)	Em que âmbito decorrem as principais atividades. Por exemplo se é maioritariamente num âmbito estudantil ou não
SL_Nightlife	Texto fechado	SL (*Student life*)	Em que âmbito decore a vida noturna. Por exemplo se é maioritariamente num âmbito estudantil ou não
SL_Travel	Texto fechado	SL (*Student life*)	Em que âmbito decorrem as viagens. Por exemplo se é maioritariamente num âmbito estudantil ou não
SL_Pcomments	Texto aberto	SL (*Student life*)	Comentários sobre a vida estudantil
SL_Pse_Activities	Likert	SL_Pse (*Personal social experience*)	Classificação sobre as atividades ao longo da experiência
SL_Pse_Nightlife	Likert	SL_Pse (*Personal social experience*)	Classificação sobre a vida noturna ao longo da experiência
SL_Pse_Travel	Likert	SL_Pse (*Personal social experience*)	Classificação sobre as viagens ao longo da experiência
SL_Pse_Overall	Likert	SL_Pse (*Personal social experience*)	Clasificação sobre a opinião geral sobre a vida estudantil

AC_Star	Likert	AC (*Academic*)	Classificação em estrelas dada aos aspetos académicos
AC_CourseRecommend	Texto aberto	AC (*Academic*)	Recomendações sobre os cursos
AC_Pcomments	Texto aberto	AC (*Academic*)	Comentários sobre os aspetos académicos
AC_Mae_Qualityofcourses	Likert	AC_MAE (*Academic My Academic Experience*)	Classificação sobre a qualidade dos cursos
AC_Mae_Varietyofcourses	Likert	AC_MAE (*Academic My Academic Experience*)	Classificação sobre a variedade de cursos
AC_Mae_AvailabAccessRecources	Likert	AC_MAE (*Academic My Academic Experience*)	Classicação sobre os recursos disponíveis

AC_Mae_Interactionwithteachers	Likert	AC_MAE (*Academic My Academic Experience*)	Classificação sobre a interação com professores
AC_Mae_Interactionwithinterstudents	Likert	AC_MAE (*Academic My Academic Experience*)	Classificação sobre a interação entre estudantes internacionais
AC_Mae_Interactionwithlocalstudents	Likert	AC_MAE (*Academic My Academic Experience*)	Classificação sobre a interação com estudantes locais
AC_Moua_ExamsEndOfCourse	Likert	AC_Moua (*My opinion of the university assessement*)	Classificação sobre os exames de final de curso
AC_Moua_ExamsThroughOutTheCourse	Likert	AC_Moua (*My opinion of the university assessement*)	Classificação sobre os exames ao longo do curso
AC_Moua_EssaysAndProjectsEndCourse	Likert	AC_Moua (*My opinion of the university assessement*)	Classificação de trabalhos e projetos no final do curso
AC_Moua_EssaysAndProjectsthroughOutCourse	Likert	AC_Moua (*My opinion of the university assessement*)	Classificação de trabalhos e projetos durante o curso
AC_Moua_OVerall	Likert	AC_Moua (*My opinion of the university assessement*)	Classificação sobre gerais sobre as avaliações

LL_Star	Likert	LL (*Language Learning*)	Classificação em estrelas dada à aprendizagem da língua
LL_Languageofinstruction	Texto fechado	LL (*Language Learning*)	Idioma de instrução
LL_Locallanguagewas	Texto fechado	LL (*Language Learning*)	Idioma do país anfitrião
LL_Instructionlanguagekeydecisionfactor	Texto fechado	LL (*Language Learning*)	Informação sobre se a língua de instrução terá sido um fator de decisão
LL_Locallanguagekeydecisionfactor	Texto fechado	LL (*Language Learning*)	Informação sobre se a língua do país anfitrião terá sido um fator de decisão
LL_Levelbeforeinstructionlanguage	Texto fechado	LL (*Language Learning*)	Nível de desenvolvimento da língua de instrução antes da experiência
LL_Levelafterinstructionlanguage	Texto fechado	LL (*Language Learning*)	Nível de desenvolvimento da língua de instrução depois da experiência
LL_Levelbeforelocallanguage	Texto fechado	LL (*Language Learning*)	Nível de desenvolvimento da língua local antes da experiência
LL_Levelafterlocallanguage	Texto fechado	LL (*Language Learning*)	Nível de desenvolvimento da língua local após experiência
LL_Pcomments	Texto aberto	LL (*Language Learning*)	Comentários sobre o desenvolvimento da língua
LL_Ld_Social	Likert	LL_ld (*Language difficulties*)	Nível de dificuldades na língua a nível social

LL_Ld_Educational	Likert	LL_ld (*Language difficulties*)	Nível de dificuldades na língua a nível educacional
LL_Ld_Admnistrativeinstitutional	Likert	LL_ld (*Language difficulties*)	Nível de dificuldades na língua a nível admnistrativo/institucional
LL_Ld_Overall	Likert	LL_ld (*Language difficulties*)	Nível de dificuldades na língua a nível geral

EX_Stars	Likert	EX (*Expenses*)	Classificação em estrelas dadas às despesas
EX_Mainsourcefunding	Texto fechado	EX (*Expenses*)	Principal fonte de sustento
EX_Othersourcesfunding	Texto aberto e fechado	EX (*Expenses*)	Outras fontes de sustento
EX_Workopportunities	Texto fechado	EX (*Expenses*)	Oportunidades de trabalho. Por exemplo se trabalhou durante a experiência, ou se não era legal trabalhar
EX_PersonalSpendingHabits	Texto aberto	EX (*Expenses*)	Principais habitos de despesas ao longo da experiência
EX_Food	Texto fechado	EX (*Expenses*)	Custos de alimentação comparativamente ao país de origem
EX_Telephone	Texto fechado	EX (*Expenses*)	Custos de telefone comparativamente ao país de origem
EX_HOusing	Texto fechado	EX (*Expenses*)	Custos de habitação comparativamente ao país de origem
EX_Nightlife	Texto fechado	EX (*Expenses*)	Custos de vida noturna comparativamente ao país de origem
EX_Travel	Texto fechado	EX (*Expenses*)	Custos de viagens comparativamente aos paises de origem
EX_Overall	Texto fechado	EX (*Expenses*)	Custos de um modo geral comparativamente ao país de origem
EX_Pcomments	Texto aberto	EX (*Expenses*)	Comentários sobre as despesas
EX_Asn_Secondhandtextbooks	Likert	EX_Asn (*Accessibility of student needs*)	Facilidade em encontrar livros em segunda mão
EX_Asn_SecondhandHousehold items	Likert	EX_Asn (*Accessibility of student needs*)	Facilidade em encontrar itens de casa em segunda mão
EX_Asn_ComputersInternet	Likert	EX_Asn (*Accessibility of student needs*)	Disponibildade de computadores e internet
EX_Asn_Admnistrative	Likert	EX_Asn (*Accessibility of student needs*)	Disponibilidade por parte dos sistemas admnistrativos das universidades
EX_Asn_MoneyfromHome	Likert	EX_Asn (*Accessibility of student needs*)	Facilidade de acesso a dinheiro do país de origem

OV_Stars	Likert	OV (*Overall*)	Classificação em estrelas dada à experiência no geral
OV_WishHadKnown	Texto aberto	OV (*Overall*)	O que o aluno gostaria de ter sabido antes da experiência
OV_Myopinion	Texto fechado	OV (*Overall*)	Opinião sobre a experiência
OV_Precomendation	Texto aberto	OV (*Overall*)	Recomendações sobre a experiência
OV_Icu_Academicreasons	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância das razões académicas para a escolha da instituição
OV_Icu_Culture	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância da cultura para a escolha da instituição
OV_Icu_Costs	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância das despesas para a escolha da instituição
OV_Icu_Activities	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância das atividades para a escolha da instituição
OV_Icu_Campuslife	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância das vida no *campus* para a escolha da instituição
OV_Icu_Partypeople	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância das pessoas e festas para a escolha da instituição
OV_Icu_Weatherlocation	Likert	OV_Icu (*Important choosing this university*)	Classificação da importância do clima e localização para a escolha da instituição
OV_Dmea_Becamefamiliarwithanotherculture	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto se tornou familiar com uma nova cultura durante a experiência
OV_Dmea_Traveled	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto viajou durante a experiência
OV_Dmea_Improvedlanguageskills	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto melhorou as capacidades linguísticas durante a experiência
OV_Dmea_Metpeoplefromothercountries	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto conheceu pessoas de países diferentes
OV_Dmea_Becamemoreindependent	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto se tornou mais independente durante a experiência
OV_Dmea_Partiedalot	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto festejou durante a experiência
OV_Dmea_ExperiencedChangeLife	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto a experiência alterou a sua vida
OV_Dmea_Advancedstudiescareer	Likert	OV_Dmea (*During my experience abroad)*	Informação sobre o quanto a experiência contribuiu para o desenvolvimento dos estudos e carreira

FinalComments	Texto aberto	*Final Comments*	Considerações finais sobre a experiência

Experiencia	Texto aberto	Experiência	Conteúdo dos seguintes atributos: HO_why, HO_Pcomments, SL_Pcoments, Ac_CourseRecommend, AC_Pcomments, LL_Pcomments, EX_PersonalSpendingHabits, EX_Pcomments, OV_WishHadKnown, OV_Precomendation e FinalComments

ANEXO B - RESULTADOS TESTE DUNN BONFERRONI

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Academic-Turism	-97,355	105,540	-,922	,356	1,000
Academic-Housing	97,713	99,697	,980	,327	1,000
Academic-Campus Facilities	101,534	135,710	,748	,454	1,000
Academic-Skills development	-139,145	116,827	-1,191	,234	1,000
Academic-Expenses	-171,154	100,359	-1,705	,088	1,000
Academic-Host City Life	203,363	97,708	2,081	,037	1,000
Academic-International Environment	-205,873	101,700	-2,024	,043	1,000
Academic-Culture Enrichment	227,320	96,454	2,357	,018	,664
Turism-Housing	,358	57,607	,006	,995	1,000
Turism-Campus Facilities	4,179	108,611	,038	,969	1,000
Turism-Skills development	-41,790	83,831	-,499	,618	1,000
Turism-Expenses	73,800	58,746	1,256	,209	1,000
Turism-Host City Life	106,009	54,093	1,960	,050	1,000
Turism-International Environment	-108,518	61,008	-1,779	,075	1,000
Turism-Culture Enrichment	129,966	51,792	2,509	,012	,435
Housing-Campus Facilities	-3,821	102,943	-,037	,970	1,000
Housing-Skills development	-41,432	76,344	-,543	,587	1,000
Housing-Expenses	-73,441	47,454	-1,548	,122	1,000
Housing-Host City Life	-105,650	41,555	-2,542	,011	,396
Housing-International Environment	-108,160	50,228	-2,153	,031	1,000
Housing-Culture Enrichment	-129,607	38,513	-3,365	,001	,028
Campus Facilities-Skills development	-37,611	119,609	-,314	,753	1,000
Campus Facilities-Expenses	-69,620	103,584	-,672	,502	1,000
Campus Facilities-Host City Life	-101,829	101,018	-1,008	,313	1,000
Campus Facilities-International Environment	-104,339	104,884	-,995	,320	1,000
Campus Facilities-Culture Enrichment	-125,786	99,805	-1,260	,208	1,000
Skills development-Expenses	32,009	77,207	,415	,678	1,000
Skills development-Host City Life	64,218	73,728	,871	,384	1,000
Skills development-International Environment	-66,728	78,942	-,845	,398	1,000
Skills development-Culture Enrichment	88,175	72,058	1,224	,221	1,000
Expenses-Host City Life	32,209	43,120	,747	,455	1,000
Expenses-International Environment	-34,718	51,530	-,674	,500	1,000
Expenses-Culture Enrichment	56,166	40,196	1,397	,162	1,000
Host City Life-International Environment	-2,510	46,155	-,054	,957	1,000
Host City Life-Culture Enrichment	23,957	33,025	,725	,468	1,000
International Environment-Culture Enrichment	21,448	43,436	,494	,621	1,000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is ,05.

ANEXO C - RELAÇÃO DE MONOTONICIDADE

			ov_stars
Kendall's tau_b	ov_stars	Correlation Coefficient	1,000
		Sig. (2-tailed)	.
		N	1918
	campus_accommodate	Correlation Coefficient	,791*
		Sig. (2-tailed)	,028
		N	7
	good_english	Correlation Coefficient	,708*
		Sig. (2-tailed)	,024
		N	8
	different_country	Correlation Coefficient	,429*
		Sig. (2-tailed)	,013
		N	24
	job	Correlation Coefficient	,315**
		Sig. (2-tailed)	,008
		N	44
	german	Correlation Coefficient	,208*
		Sig. (2-tailed)	,026
		N	65
	world	Correlation Coefficient	,206*
		Sig. (2-tailed)	,020
		N	74
	money	Correlation Coefficient	,147*
		Sig. (2-tailed)	,029
		N	131
	School	Correlation Coefficient	,104*
		Sig. (2-tailed)	,041
		N	218
	ET_Place	Correlation Coefficient	,091**
		Sig. (2-tailed)	,001
		N	808
	ET_Experience	Correlation Coefficient	,067*
		Sig. (2-tailed)	,028
		N	606
	ET_Communication	Correlation Coefficient	,054*
		Sig. (2-tailed)	,029
		N	942
	ET_Academic	Correlation Coefficient	,051*
		Sig. (2-tailed)	,037
		N	936

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).