INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Deteção de palavras emergentes em tweets portugueses e análise do seu percurso na redes sociais

Afonso do Carmo Marques Mendes Pinto

Mestrado em Gestão de Sistemas de Informação

Orientador:

Doutor Fernando Manuel Marques Batista, Professor Associado,

Iscte – Instituto Universitário de Lisboa

Outubro, 2020

TECNOLOGIAS
E ARQUITETURA

Deteção de palavras emergentes em tweets portugueses e análise do seu percurso na redes sociais

Afonso do Carmo Marques Mendes Pinto

Mestrado em Gestão de Sistemas de Informação

Orientador:

Doutor Fernando Manuel Marques Batista, Professor Associado,

Iscte – Instituto Universitário de Lisboa

Outubro, 2020

Agradecimentos

Um trabalho de mestrado é uma longa viagem, que inclui um caminho permeado por inúmeros desafios, tristezas, incertezas, alegrias e muitos percalços pelo caminho, mas apesar do processo a que qualquer investigador está destinado, reúne contributos de várias pessoas, que demonstram ser indispensáveis para encontrar o melhor rumo em cada percalço e em cada momento da caminhada.

Em primeiro lugar, não posso deixar de agradecer ao meu orientador, Professor Doutor Fernando Manuel Marques Batista, por toda a paciência, empenho, disponibilidade e apoio com que me orientou neste trabalho. Nesta viagem presenteou-me sempre com um rigoroso nível cientifico, um interesse permanente, uma visão critica e oportuna, os quais contribuíram para enriquecer, com grande dedicação, passo por passo, todas as etapas e desafios subjacentes ao trabalho realizado.

Desejo igualmente agradecer aos meus colegas do Mestrado em Gestão de Sistemas de Informação, especialmente à Beatriz Albuquerque, à Bruna Dantas e ao Simão Zanatti Saraiva, cujo apoio e a motivação incondicional ajudaram a tornar este trabalho e todo o percurso do mestrado uma agradável experiência de aprendizagem.

Ao António Madureira e à Teresa Madureira, amigos de sempre, agradeço o apoio e a confiança que sempre depositaram em mim, sem toda a disponibilidade que me foi oferecida esta viagem teria sido muito mais difícil.

À minha irmã Constança, pelos conselhos preciosos e críticas oportunas, um especial agradecimento pela paciência demonstrada e total disponibilidade.

À minha mãe, pelo apoio incondicional, generosidade e alegria que me brindou constantemente, contribuindo para chegar ao fim deste difícil percurso, bem como pela leitura crítica e atenta das versões preliminares da dissertação, contribuindo para o seu aperfeiçoamento e por ser um modelo de coragem.

Finalmente, o meu profundo e sentido agradecimento a todas as pessoas que contribuíram para a concretização desta dissertação, estimulando-me intelectual e emocionalmente.

Lisboa, 30 de Junho de 2020

Afonso Pinto

Resumo

Este trabalho aborda o problema da deteção de palavras emergentes numa língua, com base em conteúdos de redes sociais. Propõe uma abordagem para a deteção de novas palavras no Twitter, e relata os resultados alcançados para um *dataset* com dados geolocalizados recolhidos entre Janeiro de 2018 e Junho de 2019 e publicados em território português com um total de 8 milhões de tweets. Os primeiros seis meses de dados foram utilizados para definir um vocabulário inicial, a partir do qual foram identificadas novas palavras nos 12 meses seguintes. O conjunto de palavras resultante foi analisado manualmente, revelando uma série de eventos distintos e sugerindo que o Twitter pode ser um recurso valioso para pesquisar a dinâmica do vocabulário de uma língua.

É proposta uma metodologia para o mapeamento da propagação das palavras anteriormente identificadas como emergentes, onde é localizada a origem da emergência e a propagação das mesmas por Portugal através de diferentes meios sociais e geográficos. Foram identificados padrões para a emergência, sejam eles religiosos, musicais, etc. Com base nos resultados, foi identificada a cidade de Lisboa como a principal região para a emergência das palavras seguida a cidade do Porto, onde também está representada a maioria dos utilizadores do Twitter.

Com o objetivo de disponibilizar os resultados alcançados neste trabalho recorreu-se ao desenvolvimento de um *website,* onde é possível de uma forma facilitada visualizar as palavras emergentes e a sua representação geográfica, assim como estatísticas relacionadas com as mesmas.

Palavras-chave: Palavras emergentes, Twitter, vocabulário, redes sociais, propagação, mapeamento da propagação

Abstract

This work tackles the problem of detecting emerging words on a language, based on social networks content. It proposes an approach for detecting new words on Twitter, and reports the achieved results for a collection of 8 million Portuguese tweets. This study uses geolocated tweets, collected between January 2018 and June 2019, and written in the Portuguese territory. The first six months of the data were used to define an initial vocabulary, from which new words were identified on the following 12 months. The set of resulting words were manually analyzed, revealing a number of distinct events, and suggesting that Twitter may be a valuable resource for researching the vocabulary dynamics of a language.

A methodology is proposed for mapping the propagation of the previous words identified as emerging, where the source of the emergency is located and the propagation by Portugal through different social and geographical. Were identified patterns for the emergency, be they religious, musical, etc. Based on the results, the city of Lisbon was identified as the main region for the emergence of words and followed by the city of Porto, where the majority of Twitter users are also represented.

To make the results achieved in this work accessible, a website was developed, where it is possible to visualize in an easy way the emerging words and their geographical representation, as well as statistics related to them.

Keywords: Emerging words, Twitter, vocabulary, social networks, linguistic propagation, vocabulary development

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

API *Application Programming Interface*

BPMN Business Process Model and Notation

CTT Correios e Telecomunicações de Portugal

DB *Database*

DCBD Descoberta de Conhecimento em Base de Dados

DCT Descoberta de Conhecimento em Textos

DM *Data Mining*

IA Inteligência Artificial

LSA *Latent Semantic Analysis*

PLN Processamento de Linguagem Natural

SQLite *Structured Query Language Lite*

NLP *Natural Language Processing*

NLTK *Natural Language Toolkit*

TM *Text Mining*

WIP *World Internet Project*

1 Introdução

Este capítulo descreve as motivações para a realização deste estudo, analisa as questões e objetivos de investigação e descreve a abordagem metodológica adotada.

1.1 Enquadramento do tema

As redes sociais são um facilitador para as pessoas comunicarem e trocarem ideias entre si, por isso é natural que elas desempenhem um papel importante na evolução da escrita, na leitura e é expectável que permitam o aparecimento de novas palavras e expressões.

Nos últimos anos, as redes sociais tornaram-se cada vez mais parte da vida quotidiana da sociedade portuguesa. O Twitter começou como uma sala de chat com um número limitado de pessoas, onde hoje em dia é um lugar «barulhento» [1], onde as pessoas, independentemente da idade, sexo ou classe social, produzem todo o conteúdo possível sobre os aspetos da sua vida quotidiana, sendo assim o «lugar» ideal para o estudo da linguagem.

De acordo com Harmer [2], se a estrutura da linguagem é necessária e é considerada a base da aprendizagem de qualquer língua, então o papel do vocabulário também não pode ser negligenciado, uma vez que fornece os meios necessários para a aprendizagem. Tradicionalmente, dois tipos de variação lexical foram identificadas [3]. A variação semasiológica refere-se à variação no significado das palavras, tal como acontece com a palavra "cedo", uma vez que pode denotar vários conceitos em função do contexto da frase em que se encontra, pode referir-se em termos temporais (por exemplo, "é antes do tempo"), assim como pode ser usado na sua forma verbal, por exemplo "ceder o lugar", significando oferecer o lugar a alguém. A variação onomasiológica refere-se à variação da forma como os conceitos são identificados, como, por exemplo, a variação na palavra para identificar o local de espera pelo autocarro, que em português de Portugal seria uma "paragem de autocarro" e em português do Brasil seria ser um “ponto de ônibus”. De acordo com Grondelaers, Geeraerts e Speelman [4] as variações semasiológicas envolvem mudanças no significado das palavras, enquanto a variação onomasiológica envolve mudanças em como são identificadas as palavras, o que inclui a formação de novas palavras.

Quanto à natureza das palavras e distinções de significado, Shahid [5] apresenta quatro ideias diferentes. A primeira é que, exceto para palavras técnicas, duas palavras diferentes

contêm diferentes significados, querendo com isto dizer que, uma tradução poderá distorcer parcialmente o verdadeiro significado que se pretende transmitir. A segunda ideia, diz que em muitas linguagens o número de palavras com apenas um significado é muito reduzido. A terceira ideia será que uma palavra adquire o seu significado de acordo com o contexto em que se insere. Por fim, a quarta ideia considera que na mesma língua não existe uma palavra que possa substituir outra, como por exemplo oceano e mar, ainda que sejam sinónimos, são utilizadas para diferentes efeitos. Assim, duas questões surgem:

- Qual o papel das redes sociais no desenvolvimento do vocabulário de uma língua?
- Serão as redes sociais vistas como um meio adequado para o desenvolvimento do vocabulário?
- Dado um vocabulário será possível traçar o seu percurso nos *social media*?

De acordo com Blood [6], Dyrud, Worley e Flatley [7] e Kajder, Bull e Van Noy [8], existem blogues onde os utilizadores (estudantes) conseguem obter conhecimento consultando-os. Mutum e Wang [9] destacam a ideia que os blogues e as redes sociais podem ser extremamente didáticos, uma vez que através do *chat* ou pelos comentários, interagindo com utilizadores, poderão estar aperfeiçoar as suas competências nos assuntos discutidos.

1.2 Motivação e relevância do tema

O desenvolvimento dos sistemas de informação têm permitido o armazenamento de grandes quantidades de dados e informação das mais variadíssimas áreas, onde as redes sociais têm um papel importante. Com a facilidade na comunicação que as redes sociais oferecem às pessoas, cada vez mais existe interação por via destes sistemas, com isto, diariamente existe um aumento da disponibilidade dos dados e das possibilidades de estudo dos dados inseridos. Foi entre os anos de 2010 e 2012 que se sentiu o maior crescimento da adesão às redes sociais, que, segundo um estudo recente, 6.2 milhões de portugueses utilizam redes sociais ativamente, em primeiro lugar surge o Facebook, seguido do Instagram, Linkedin e em quarto lugar o Twitter. O Twitter foi eleito como principal rede social para efeitos de análise e de estudo, uma vez que consegue em termos relativos ser a rede mais jovem, com 56% da população com menos de 24 anos. Deste modo, a motivação para a elaboração deste trabalho surgiu da necessidade de compreender de que formas as redes sociais refletem a existência de novas palavras, já numa fase posterior compreender o que motiva a emergência de palavras através de um método que possa ser aplicado em qualquer sistema.

Têm existido ao longo do tempo investigações empíricas sobre a evolução lexical, usando variadíssimas perspetivas, em grande parte as investigações têm sido no âmbito do processo de formação das palavras ou em como o significado e a utilização das palavras tem

vindo a ser modificado ao longo de relativos longos períodos de tempo. Existe pouca informação sobre a propagação das palavras ao longo do tempo desde que surgem pela primeira vez. Este tipo de emergência lexical tem sido difícil de analisar uma vez que há alguns anos atrás os investigadores não possuíam a quantidade suficiente de dados num largo período de tempo de modo a traçar a propagação. Geralmente, a variação lexical é um campo de alta dificuldade para serem efetuados estudos, isto porque a maioria das palavras são de caráter raro. Por exemplo, segundo Grieve [10] fora das 100,000 formações de palavras mais frequentes num *corpus* de 450 milhões de palavras 50 mil formações ocorrem somente 124 vezes, isto significa o surgimento de uma nova formação de palavra a cada 3.6 milhões de palavras referidas. Ainda assim, a análise da emergência léxical, que envolva formações especialmente raras e que contenham um aumento na frequência relativa de forma muito rápida em pequenos períodos de tempo requer acesso a um *corpora* que seja extremamente diversificado ao longo do tempo.

A compilação de um *corpora* que vá de encontro aos requisitos necessários tem vindo a ser possível através da obtenção de *data* da Internet, esta abordagem é habitualmente designada de *corpus*. Mas especificamente, a partir de 2011 os investigadores têm analisado *db/corpora* com dados obtidos através das redes sociais, especialmente do Twitter, uma vez que facilita a obtenção dos dados para efeitos académicos através da sua API interna.

Assim, pretende-se desenvolver uma metodologia que assente nas dificuldades apresentadas e que consiga ajudar organizações ou indivíduos que tenham como objetivo analisar um conjunto vasto de palavras.

1.3 Questões e objetivos de investigação

O objetivo principal desta dissertação é apresentar um estudo no formato de documento que proporcione uma metodologia aos centros de investigação na área de processamento de língua naturais (PLN), facilitando a compreensão das principais características das palavras emergentes. Para a concretização foram definidos objetivos mais concretos, assim os principais objetivos da investigação serão:

- Será possível identificar palavras emergentes através de tweets?
- Quais fenómenos estarão por detrás da emergência das palavras?
- Identificada uma nova palavra, será possível traçar o seu percurso nas redes sociais?

Para isto será proposto uma metodologia para a deteção das palavras emergentes com base na sua frequência ao longo dos 18 meses cobertos pela base de dados de tweets, com o universo de palavras emergentes levantados, estes serão comparados com um dicionário lexical da língua portuguesa com o intuito de identificar novas palavras que tenham surgido nesse período temporal.

Assim, o segundo objetivo passará por compreender temas em comum que essas palavras possam conter, tentando assim identificar padrões propícios para o surgimento de palavras novas e emergentes.

Por fim, o terceiro objetivo passará pelo mapeamento do universo de palavras identificadas em território português, onde para cada palavra será identificado o local do surgimento pela primeira vez, assim como a densidade destas pelos diversos distritos. Espera-se que igualmente seja possível identificar padrões para o surgimento como para a propagação.

1.4 Abordagem metodológica

O desenvolvimento deste trabalho científico passará por dois tipos de abordagem, numa fase inicial seguirá uma abordagem quantitativa, onde serão analisados os dados obtidos através da API interna do Twitter recorrendo a ferramentes como SQLite ou Python para tratar e estudar, será através desta abordagem permitirá possível obter os resultados pretendidos.

1.5 Estrutura e organização da dissertação

No primeiro capítulo foi introduzido o tema para que o leitor perceba a relevância assim como o enquadramento nos dias de hoje. Foram ainda apresentados os objetivos a concretizar com a realização deste trabalho, assim como o que motivou a escolha do tema aqui apresentado.

O capítulo seguinte consiste na discussão dos fundamentos que estão subjacentes à realização deste trabalho, onde serão apresentadas as áreas que estão presentes neste trabalho e ainda as técnicas e métodos que serão utilizados.

No terceiro capítulo são apresentados trabalhos relacionados com o trabalho aqui desenvolvido, onde será possível compreender melhor a importância do estudo textual assim como verificar os resultados obtidos num contexto diferente ao de este trabalho.

No quarto capítulo será apresentada toda a metodologia que virá a ser utilizada descrevendo pormenorizadamente cada fase do trabalho.

O quinto capítulo, pretende-se apresentar e discutir os resultados alcançados. Por fim, no quinto e último capítulo serão apresentadas as principais conclusões, quais os contributos para a comunidade científica, implicações ocorridas e propostas para futuras investigações.

Por fim, no sexto e último capítulo são apresentadas as várias conclusões que foram possíveis obter com base nos resultados obtidos neste trabalho, assim como as etapas seguintes que se pretendem realizar num trabalho futuro.

Enquadramento

2

O principal objetivo deste capítulo consiste na apresentação dos principais temas que serão abordados e metodologias existentes atualmente na literatura e que serão utilizadas ou referenciadas no âmbito desta dissertação.

2.1 Redes sociais

As redes sociais são facilitadores de conexões sociais entre pessoas, grupo ou organizações que compartilham dos mesmo valores ou interesses, interagindo entre si. Nesta secção serão apresentados os principais temas relacionados com as redes sociais, uma vez que o número de utilizadores ativos tende a crescer de dia para dia, com a dinâmica de melhorias constantes dessas plataformas e o surgimento de novas.

2.1.1 Da comunicação à partilha de informação

As primeiras redes sociais surgiram há uns anos atrás, com o aparecimento do *Sixdegrees* em 1997, surgiu a possibilidade de criar um perfil individual virtual, tendo sido substituído por outros, três anos mais tarde, devido à falta de financiamento para a sua continuidade. Mais tarde, a nível internacional surgiram o *Live Journal, Asian Avenue, Fotolog* entre outros. Nos dias de hoje, o número de redes sociais quase não tem fim, onde as que possuem um maior impacto e com o maior número de utilizadores serão o Facebook e o Youtube (Clement [11]).

O verdadeiro desenvolvimento das redes sociais deu-se após a transição da web 1.0 para a web 2.0, especialmente para efeitos de lazer e de partilha de diversos conteúdos como fotos, vídeos, dados pessoas e estados de espírito.

Em termos de *web sites* ou comunidades de relacionamentos, pode referir-se que, no decorrer do tempo, muitos destes foram evoluindo. A rede que se sentiu uma maior evolução foi o *MySpace* especialmente nos estado Estados Unidos da América e na Europa, o *Friendster* teve um marco importante nas ilhas do pacífico, o *Orkut* tornou-se o *web site* mais visitado no Brasil, o Hi5 conseguiu um uniformidade por todo o mundo, assim como o *Facebook*.

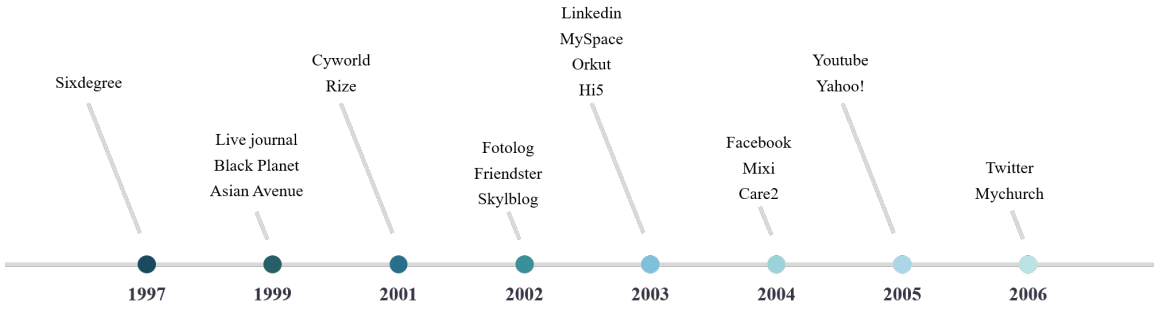

Figura 2.1: *Cronologia da evolução das Redes Sociais em 2014*

Porter [12] assume que as comunidade virtuais e as redes sociais podem ser divididos em dois grupos: as comunidades que são iniciadas por membros individualmente, e as comunidades iniciadas por organizações de ordem comercial ou de outro tipo, com ou sem fim lucrativo.

Segundo Oliveira [13], as crianças, os jovens e os utilizadores dos espaços são os verdadeiros protagonistas da sua vivência ativa com outras pessoas e objetos, que possibilitam descobertas pessoais num espaço onde será realizado um trabalho individual ou em pequenos grupos.

Uma das questões mais importantes no estudo da dinâmica das redes sociais pretende-se com o grau da centralidade. Isto é, compreender de que forma a informação flui e o seu grau de intercomunicação.

Grande parte das dinâmicas das redes sociais depende diretamente das interações que a própria rede compreende na sua constituição e influencia a sua estrutura [14]. De acordo com o autor, não existem redes sociais paradas mas sim redes dinâmicas que estão constantemente em transformação. Não obstante Nicolis, Prigogine, Freeman et al. [15], os processos dinâmicos das redes sociais resultam das diversas interações dos seus utilizadores.

Indubitavelmente, um dos fatores que transmite dinamismo a uma rede social é a sua emergência, que de acordo com Johnson e Duberley [16], envolve o aparecimento de padrões de comportamentos distintos e muito variados na sua constituição, e que são calculados em larga escala, deste modo, serão consideradas as propriedades emergentes, aquelas que o próprio sistema constitui, e são construídas coletivamente. A emergência surge como representação de comportamentos coletivos e não centralizados.

Os elementos dinâmicos das redes sociais são a competição e a cooperação, como processos sociais. Os autores Ogburn e Nimkoff [17] reconhecem a competição como «a forma fundamental de luta social». A competição é um fenómeno que pode gerar cooperação entre os utilizadores numa mesma rede. Tanto a cooperação como a competição são fenómenos emergentes e naturais nas redes sociais, resultam num impacto diferenciado na estrutura social. A cooperação surge em *weblogs* coletivos, produzidos por diversas pessoas, que

dependem da cooperação entre todos os seus utilizadores para que continuem a existir, através de opiniões e relatos. Desta forma, estes elementos dinâmicos são fundamentais para a perceção das redes sociais no tempo, e a sua compreensão enquanto elementos não estáticos.

2.1.2 Os portugueses nas redes sociais

Atualmente, é possível afirmar que as redes sociais fazem parte do dia a dia dos portugueses, exercendo um papel essencial na sua vida. As ações que as redes sociais possibilitam são infindáveis, sendo que, são fundamentalmente os utilizadores, que fazem das suas páginas de perfil o ponto de partida para navegar *online*. A força dos laços aumenta tanto com comunicação individual (posts, comentários e mensagens), mas também pela absorção do conteúdo exportado pelo amigos (estados, *updates*, fotografias, vídeos, etc.).

De acordo com o inquérito realizado pelo OberCom, associação sem fins lucrativos pioneira na investigação das redes sociais, em 2008, cerca de 53% dos internautas em Portugal utilizavam redes sociais, números que inevitavelmente cresceram desde então. Logo em 2010, de acordo com o estudo desenvolvido no âmbito do *World Internet Project* (WIP), através da aplicação de um questionário a uma amostra populacional de 1255 pessoas a nível nacional, avançam que 56,4% dos internautas utilizam redes sociais. Já em 2013 o número de participantes das redes sociais subiu para 70,6% de acordo com a OberCom.

Segundo a fonte de dados *Statista* e indicadores da *Marktest*, em 2019 a liderança de utilizadores ativos nas redes sociais pertence ao Facebook, ainda que seja o Instagram a rede que mais tem «crescido». Cerca de 90% dos 9 milhões de utilizadores com perfil ativo nas redes sociais utiliza ativamente o Facebook em Portugal traduzindo-se assim em 7,2 milhões de contas ativas. Também o Youtube possuí uma percentagem elevada de utilizadores ativos, perto dos 90%, traduzindo-se assim em 7,2 milhões de contas ativas. O Instagram tem crescido em Portugal, contando com 61% dos utilizadores da Internet em Portugal com conta ativa ativa na rede social, ou seja, 4,6 milhões de utilizadores. O Whatsapp, pertencente ao Facebook, contém 61% dos utilizadores ativos da Internet em Portugal, atingindo assim os 4,9 milhões de utilizadores ativos. Relativamente ao LinkedIn, esta plataforma é utilizada sobretudo para a construção de uma «teia» de relações profissionais, contando com 35% do utilizadores da Internet, tendo cerca de 2,9 milhões de utilizadores em Portugal. O Pinterest, uma rede social para partilha de fotografias conta com cerca de 32% de utilizadores, o que se traduz em 2,7 milhões de contas ativas. O Twitter que é particularmente popular entre o jovens, com menos de 24 anos tem um alcance de 31% entre os utilizadores de Internet em Portugal, o que equivale a 2,5 milhões de contas ativas. Por fim, a plataforma Twitch sendo o local ideal para quem pretenda ver *gameplays* e competições de *e-sports* possui cerca de 1,12 milhões de utilizadores em Portugal o que equivale a 14% dos utilizadores da Internet. Através da Figura 2.2 consegue-se compreender melhor a dimensão de utilizadores por plataforma aqui mencionadas.

Figura 2.2: *Redes sociais utilizadas pelos portugueses em 2019*

2.1.3 A partilha de informação e os estudantes nas redes sociais

Desde o princípio da humanidade que o Homem procura organizar-se em grupos, com os quais estabelece vínculos, comunica, partilha informação e produz conhecimento. Nesse sentido, as possibilidades de interação entre indivíduo foram ampliadas.

A expansão do acesso ao mundo virtual proporcionada pela Internet, possibilitou uma expressiva popularização dos meios de comunicação, permitindo a democratização do acesso à informação e o surgimento de um fenómeno sem precedentes - as Redes Sociais.

Deste modo, as redes sociais vêm aproveitar o novo mundo de oportunidades, permitindo o registo de novos utilizadores, oferencendo-lhes a possibilidade de editar, partilhar conteúdo e informação, bem como convidar amigos para criar uma rede de contactos. Adjacente a estas funcionalidades, diversos estudos e investigações foram proporcionadas tendo por base os dados que as redes sociais armazenam, exemplo disso foram os estudos realizados no âmbito da análise de sentimentos, deteção de *fake news*, deteção de sinais de sofrimento psíquico em utilizadores de redes sociais e como é o caso deste estudo em concreto a deteção de palavras emergentes em redes sociais assim como a sua propagação.

Segundo Ellison, Steinfield e Lampe [18], as redes sociais têm como objetivo, além de permitir a ligação entre as pessoas e facilitar os laços sociais, atuar como como facilitador da construção do capital social, ou seja, os utilizadores passam a dispor de meios e ferramentas inovadoras e interativas para interagir com outras pessoas, para criar relações e comunidades, para escrever, partilhar informação e conteúdos multimédia (de forma

virtual).

As trocas de informação, a partilha de ideia e opiniões, experiências e sentimentos, tornaram-se presença frequente na interação digital, através das inúmeras formas e redes sociais existentes em todo o mundo.

No estudo apresentado por Miranda, Morais, Alves et al. [19], relativo à utilização das redes sociais por estudantes do ensino superior, os motivos que os levam a aceder às redes são maioritariamente a vontade de contactar amigos (98%), para fins de entretenimento (90%), apoio à aprendizagem (62%), discussão de temas de interesse (50%), promoção de eventos (44%) e contactos profissionais (40%). De acordo com as categorias definidas pelos autores, a distribuição em termos percentuais, das unidades de análise pelas categorias referidas é a seguinte: recursos disponíveis (31%), contactos (19%), facilidade de utilização (17%), sem interesse (12%), construção de conhecimento (10%), partilha de conteúdos (9%), outras (2%).

Contudo, ao utilizarem as redes sociais como ferramenta académica, os estudantes universitário tornam-se responsáveis pela sua própria aprendizagem. Macedo [20] refere que ao agir sobre a informação que obtém, modificando-a e anexando-a a seu repertório, cada utilizador pode potencializar o seu uso e retirar lucro.

2.1.4 Breve apresentação do Twitter

A rede social Twitter é considerada um micro-blogue que permite aos utilizadores fazerem publicações ou *tweets* até 280 caracteres, onde podem incluir *links* relevantes. Através da plataforma os utilizadores podem «seguir» outros utilizadores, para que os *tweets* publicados pelas pessoas que se encontra a seguir apareçam na *timeline*. O *Twitter* tornou-se especialmente utilizado entre celebridades, políticos, empresas e até presidentes de diversos países (Simões [21]). É comum que num primeiro período de utilização desta rede social exista algum esforço mas compreender o seu funcionamento e como podem utilizar esta rede social. O foco desta rede pretende que seja especialmente utilizada a partir de *smartphones* para que a leitura das curtas publicações seja de fácil visionamento.

É o local onde as pessoas compartilham as suas opiniões e descobrem o que está acontecer no mundo. Atualmente são publicados mais de 500 milhões de «Tweets» todos os dias na plataforma.

Na Figura 2.3 são levantadas as principais funcionalidades do Twitter para melhor compreensão de como é apresentada uma publicação / *tweet*.

1. Tweet: Um tweet é uma mensagem publicada no Twitter, que pode conter textos, fotos, links e vídeos.

2. Menção: Pretende captar a atenção da pessoa mencionada para o Tweet. É comum

Figura 2.3: *Componentes fundamentais do Twitter*
Fonte: Twitter do ISCTE-IUL *(2019)*

ser utilizado em perguntas, agradecimentos ou simplesmente para dar destaque a um determinado conteúdo.

3. Resposta: Permite responder a um tweet de qualquer pessoa.

4. *Retweet*: Consiste em compartilhar um Tweet de outra pessoa. Pode ser compartilhado AS-IS ou pode ainda ser acrescentado um comentário.

5. *Favorite*: Demonstra ao autor da publicação que gostou do conteúdo.

6. *Hashtag*: Pode ser apenas uma palavra ou frase, sem espaços e precedida pelo símbolo «#». habitualmente é utilizado para organizar conversas e facilitar a localização de todo o conteúdo relacionado com um determinado assunto.

2.2 *Text Mining*

De acordo com Berry [22], a área de TM habitualmente conhecida por *Text Mining* traduz-se na capacidade computacional de descobrir novas informações através da extração de informação textual de diferentes e variadas fontes. A principal dificuldade que surge é na identificação da informação que é útil e relevante para dar resposta às necessidades da investigação. Assim o principal objetivo passará por descobrir a informação desconhecida, algo que ainda não se saiba.

Segundo Berry [22], TM e *data mining* são similares, porém de acordo com Elmasri e Navathe [23] as ferramentas de *data mining* estão preparadas para trabalhar com bases de dados estruturadas, já os mecanismos de *Text Mining* conseguem ser mais dinâmicos e trabalhar dados não estruturados ou semi-estruturados como documentos de texto, emails, etc.

A área de *natural language processing* (NLP) já obteve ferramentas que transmitem às maquinas linguagem natural para que a estas consigam efetuar uma análise, compreensão e geração de textos. Algumas dessas ferramentas que foram desenvolvidas traduzem-se em informação extraída, no rastreio de tópicos, na sumarização, na categorização, no *clustering*, no conceito de ligação, na visualização de informação e ainda na resposta de questões [24].

A Figura 2.4 descreve as fases do *Text Mining*.

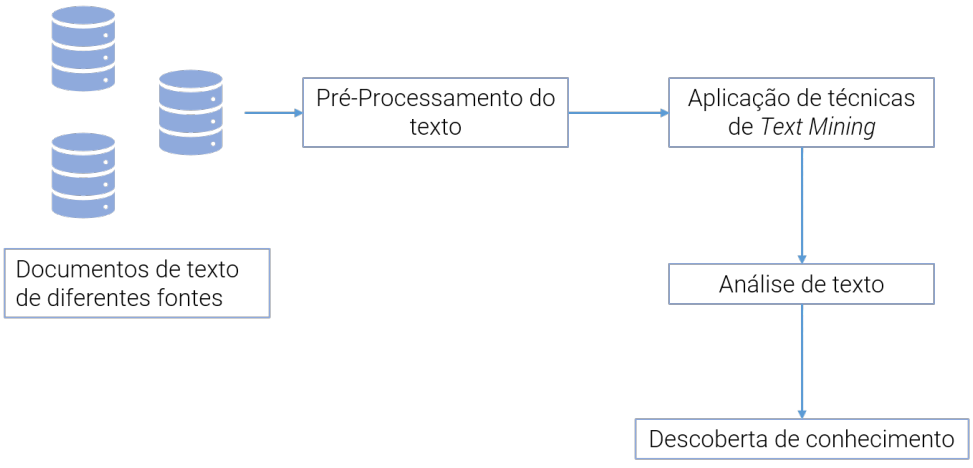

Figura 2.4: *Etapas do Text Mining*

2.2.1 Extração e recolha da informação

Na recolha dos dados várias palavras-chave vão sendo identificadas e com elas são criadas relações com o texto. Esta metodologia é extremamente útil quando é alvo de análise um volume extenso de dados.

De forma a evitar uma sobrecarga de dados desnecessária é possível nesta fase realizar uma recolha com base em palavras-chave que vão de encontro ao objetivo do estudo, podendo estas ser, nomes de pessoas, moradas, emails, etc.

2.2.2 Categorização

A categorização dos dados recolhidos envolve identificar os principais temas que as palavras obtidas transmitem. As relações entre as palavras e os temas (ex. laranja, amarelo, preto - cores) são identificadas através de sinónimos ou termos relacionados [25].

2.2.3 *Natural Language Processing*

A área de investigação de NLP explora a forma como os computadores podem ser utilizados para compreender e manipular/transformar a linguagem textual. Os investigadores nesta área recolhem conhecimento na forma como o ser humano comunica e utiliza a linguagem textual para que as ferramentas e técnicas de tratamento por parte das máquinas fiquem com a *performance* mais adequada e real possível [26].

Assim, procedendo às técnicas de NLP será mais fácil a organização e a estrutura de texto complexo em dados úteis para análise.

2.2.4 Remoção de *Stop Words*

As palavras denominadas como «*Stop words*» são as quais foram identificadas como tal por não transmitirem valor à análise que se pretende efetuar e como consequente deverão ser removidas do *dataset.* As *stop words* devem ser removidas uma vez que resultará numa diminuição da dimensão do *dataset*. Habitualmente, as «*stop words*» mais comuns consistem em artigos, preposições, pronomes, etc. estes não adicionam valor ao seu tratamento no decorrer da análise. Exemplo de «*stop words*»: «minha», «dela», «tive». De acordo com Porter [27] as palavras «*stop words*» são removidas dos documentos por essas não serem identificadas como «palavras-chave» nas aplicações de *Text Mining*.

2.2.5 Clustering

Segundo Liritano e Ruffolo [28], a técnica de *clustering* é especialmente utilizada para agrupar documentos idênticos, difere de categorização uma vez que os documentos são indexados de imediato em vez de serem primeiramente agrupados em tópicos . A utilização desta técnica em documentos permite ao analisado visualizar o documento em vários subtópicos relacionados com o conteúdo do documento, o que é uma mais valia uma vez que o mesmo não deixará de ocorrer em tópicos ao qual está relacionado. Cada documento é alvo de processamento por algoritmos de *clustering* onde estes irão criar um vetor de tópicos e

medir o nível de compatibilidade com cada tópico. Este tipo de tecnologia é especialmente útil em investigações onde o volume de documentos é muito extenso.

2.3 Utilização de *word embeddings*

Word embeddings são um tipo de representação que permitem palavras com significados similares terem uma similar representação.

Segundo Goldberg [29], um dos benefícios na utilização de vetores de baixa dimensão é computacional: a maioria dos conjuntos de ferramentas de redes neuronais não desempenha bem o seu papel com vetores muito densos e de dimensão reduzida. O principal benefício das representações densas é o poder de generalização: se acreditarmos que algumas características podem fornecer indicações semelhantes, vale a pena fornecer uma representação que seja capaz de captar estas semelhanças.

Word embeddings são de facto, um conjunto de técnicas em que palavras individuais são representadas como vetores de valor real num espaço vetorial pré-definido. Cada palavra é mapeada para um vetor e os valores vetoriais são aprendidos de uma forma que se assemelha a uma rede neuronal.

De acordo com Bengio, Ducharme e Vincent [30], associando cada palavra do vocabulário a um vetor de características de palavras distribuído, o vetor irá representar diferentes aspetos de cada palavra, associada a um ponto no espaço vetorial.

2.3.1 *Embedding Layer*

A camada de *embedding*, é uma incorporação de palavras que é aprendida em conjunto com um modelo de rede neuronal numa tarefa específica de processamento de linguagem natural, como a classificação de documentos.

É necessário que o texto do documento esteja «limpo» e preparado de forma a que cada palavra seja codificada. A dimensão do espaço vetorial é especificada como parte do modelo, tais como 50, 100, ou 300 dimensões. Os vetores são inicializados com pequenos número aleatórios. A camada de *embedding* é usada na parte frontal de uma rede neuronal e é encaixada de forma supervisionada através do algoritmo de retro-propagação.

Quando o *input* para uma rede neuronal contem símbolos, é comum associar cada valor de características possível a um vetor d-dimensional para alguns d. Estes vetores são então considerados parâmetros do modelos e são treinados em conjuntos com outros parâmetros Goldberg [29].

Esta abordagem de aprendizagem de uma camada de *embedding* requer muitos dados de formação e pode ser lenta, mas irá aprender uma integração tanto direcionada para os dados de texto específicos como para a tarefa NLP.

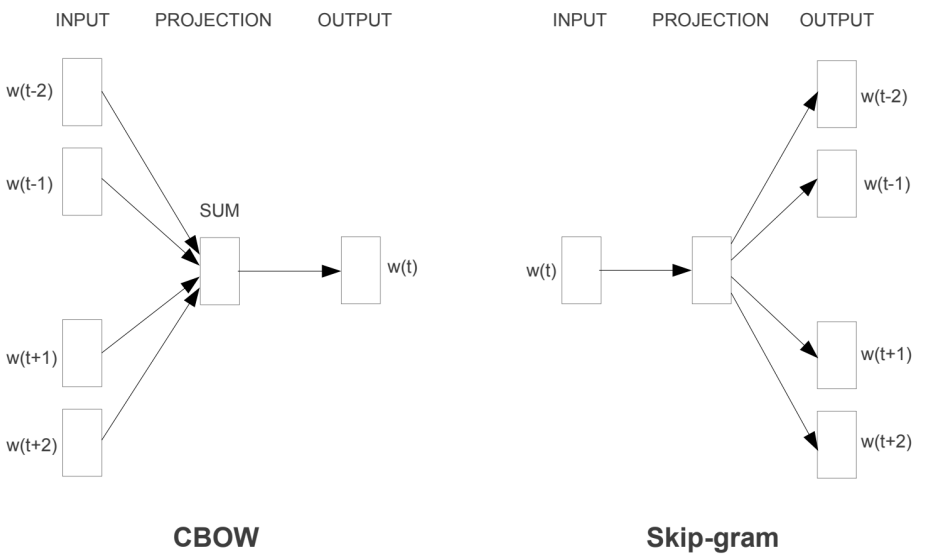

Figura 2.5: *Modelos de treino* .
Fonte: «Efficient Estimation of Word Representations in Vector Space» *(2013)*

2.3.2 Word2Vec

Word2Vec é um método estatístico para a aprendizagem eficiente de uma palavra incorporada num corpus textual.

Foi desenvolvido por Tomas Mikolov, et al. no Google em 2013 como resposta para tornar mais eficiente a formação baseada na rede neural da incorporação e, desde então, tornou-se o padrão de facto para o desenvolvimento da incorporação de palavras préformadas.

Foram introduzidos dois modelos de aprendizagem diferentes que podem ser utilizados como parte da abordagem do *Word2Vec* para aprender a palavra, são eles:

- «Bag-of-Words» ou modelo CBOW.
- Modelo «Skip-Gram».

Como se pode verificar na Figura 2.5, o modelo CBOW aprende através da previsão da palavra atual com base no seu contexto. O modelo «Skip-Gram» aprende através da previsão das palavras mais próximas no espaço vetorial de uma dada palavra.

Ambos os modelos focam-se na aprendizagem das palavras de acordo com o seu contexto, onde este é definido através das palavras que se encontram na proximidade. A proximidade é um parâmetro configurável nos modelos.

2.3.3 GloVe

Global Vectors ou GloVe é uma extensão do método word2vec para a aprendizagem eficiente de vetores de palavras.

As representações clássicas do modelo de espaço vetorial de palavras foram desenvolvidas utilizando técnicas de fatorização matricial como a análise semântica latente (LSA), que utilizam estatísticas globais de texto, ainda que não obtenham resultados tão relevantes quanto os obtidos com o métodos aprendidos como o word2vec, na captura do significado e demonstração em tarefas como o cálculo de analogias.

GloVe é uma abordagem que combina tanto as estatísticas globais de fatorização matricial como LSA, com a aprendizagem baseada em word2vec (Cothenet [31]).

Traduz-se num modelo de regressão log-bilinear global para aprendizagem não supervisionada de representações de palavras que super outros modelos de analogia de palavras, similaridade de palavras e tarefas Pennington, Socher e Manning [32].

2.3.4 fastText

FastText é uma biblioteca criada por uma equipa de investigadores do Facebook com o intuito de obterem uma aprendizagem eficiente de representações de palavras e classificações de texto (Figura 2.6).

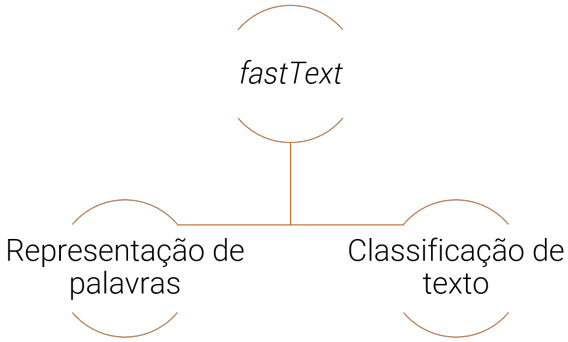

Figura 2.6: *Funcionalidades do fastText* .

Esta biblioteca diferencia-se no sentido que os vetores de palavras (word2vec) trata cada palavra como a mais pequena unidade cuja representação vetorial pode ser visualizada, *fastText* é capaz de atingir níveis de performance muito elevados na presentação de palavras e na classificação de frases. Cada palavra é representada por *n* conjuntos de caracteres, por exemplo a palavra «*matter*», o *n* é igual a 3, a representação seria <*ma, mat, att, tte, ter, er*>, então se a palavra <*mat*> faz parte da representação ajuda assim a preservar o significado e permite adicionar prefixos e sufixos.

Tabela 2.1: *Precisão / velocidade do fastText* .
Fonte: Publicação realizada pela equipa de investigação do Facebook

	Yahoo		Amazon full		Amazon polarity	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
char-CNN	71.2	1 day	59.5	5 days	94.5	5 days
VDCNN	73.4	2h	63	7h	95.7	7h
fastText	72.3	5s	60.2	9s	94.6	10s

Esta representação de palavras proporciona as seguintes vantagens sobre word2vec ou GloVe.

1. É útil para encontrar a representação vetorial de palavras raras. Uma vez que as palavras raras contém n-gramas de caracteres e partilhar esses n-gramas com as palavras comuns. Por exemplo, para um modelo treinado num conjunto de dados de notícias, os termos médicos podem ser considerados raros.

2. Apresenta representações vetoriais para as palavras que não constam no dicionário, uma vez que estas também podem ser decompostas em caracteres n-grams. Tanto o word2vec como o GloVe não fornecem quaisquer representações vetoriais para as palavras que não constem no dicionário.

3. Possui um *desempenho* superior ao word2vec ou ao GloVe [31].

Vantagens

1. A biblioteca é extremamente rápida em comparação com outros métodos para alcançar a mesma precisão (Figura 2.1).

2. Vetores de frases (supervisionados) podem ser facilmente processados.

3. fastText funciona bem em pequenos *datasets.*

Desvantagens

1. Esta não é uma biblioteca autónoma de NLP, uma vez que depende de outra biblioteca para as etapas de pré-processamento.

2. Esta biblioteca tem uma implementação em *python* que não é oficialmente apoiada.

3 Trabalho Relacionado

Este capítulo apresenta os trabalhos e estudos já realizados que estão de alguma forma relacionados com o trabalho aqui desenvolvido.

3.1 Emergência léxical

Os autores Grieve, Nini e Guo [33] introduzem um método quantitativo para a identificação de palavras emergentes num longo período de tempo e descrevem a análise léxical aplicada em território americano. Este estudo tem por base um *corpus* com mais de mil milhões de palavras obtidas através de Tweets entre outubro de 2013 e novembro de 2014. No total foram detetadas 29 palavras emergentes, que representam várias classes semânticas, partes gramaticais da fala e processos de formação de palavras, identificados no decorrer da análise. As 29 formações identificadas foram examinadas em diversas perspetivas com o intuito de melhor compreender o processo da emergência léxical.

Neste estudo o Twitter foi selecionado com fonte de dados por fornecer uma quantidade elevada de dados num curto espaço de tempo e por ser uma rede social informal de linguagem natural onde milhões de pessoas por toda a América participam, incluindo pessoas com uma idade reduzida e pessoas de uma baixa classe social, que presumivelmente serão responsáveis por introduzirem novas palavras, especialmente palavras consideradas «calão». Os autores preveem que utilizando os dados do Twitter para análise que seja possível identificar palavras numa fase inicial do seu desenvolvimento e que seja possível traçar a sua utilização ao longo do tempo.

O *corpus* analisado no estudo contem 8,9 mil milhões de palavras geo-localizadas e com data e hora da publicação, totalizando, assim, 980 milhões de tweets escritos/publicados por 7 milhões de utilizados únicos. Todos os dados existentes no *corpus* foram obtidos através da API interna do Twitter.

Afim de analisar os padrões temporais da emergência léxical, o *corpus* foi dividido em 397 *subcorpora* diários. Obtendo, assim, uma média de 22 milhões de *tokens* por dia, ainda que a amplitude de *subcorpora* diário varie entre 10 e 29 milhões de *tokens.*

Para identificar palavras emergentes no *corpus*, todas as formações de palavras que ocorreram pelo menos mil vezes foram extraídas para análise. De forma a medir quais for-

mações de palavras contêm uma frequência que tenha crescido exponencialmente ao longo do tempo, foi aplicada a correlação de *Spearman* para cada formação, através da correlação da sua frequência por dia e o seu *rank*. Uma correlação positiva entre a frequência diária e o dia do período (*rank*) indica que o uso dessa formação léxical específica aumentou ao longo do tempo.

Após a obtenção de 131 formações de possíveis palavras emergentes, estas foram analisadas à «mão», onde vários problemas foram detetados, como o surgimento de nomes de pessoas, produtos ou nomes de empresas e assim, estas foram removidas do universo de palavras emergentes, finalizando assim com 29 palavras emergentes.

A um nível mais geral, o estudo demonstrou que é possível identificar padrões na emergência léxical no Inglês Americano Moderno através de uma análise quantitativa e manual. Ainda que não tenham sido identificados um número extenso de palavras emergentes e que não seja um *corpus* completo com todas as palavras inseridas no Twitter no espaço de tempo descrito, muito menos no Inglês Americano Moderno. O principal motivo por não ter sido identificado um número extenso de palavras emergentes deve-se ao facto de que, ainda que tenha sido utilizado um *corpus* de grandes dimensões não foi grande o suficiente para compreender a variação léxical.

As mudanças lexicais foram também examinadas a partir das perspetivas complementares de lexicalização, Brinton e Traugott [34] e gramaticalização Hopper e Traugott [35]. Por exemplo, Naya [36], analisou a lexicalização com base nos registos do *Oxford English Dictionary*. A investigação sobre a lexicalização sobrepõe-se à investigação sobre a mudança onomasiológica, especificamente no que diz respeito à formação de palavras, embora a lexicalização se concentre geralmente em certos tipos de processos de formação de palavras e em padrões de mudança lexical a longo prazo. A lexicalização também é frequentemente contrastada com institucionalização. Por exemplo, Bauer, Bauer, Laurie et al. [37] viu a lexicalização como um processo que se seguiu à institucionalização. Alternativamente, a gramaticalização é o processo através do qual os itens lexicais perdem significado à medida que desenvolvem gradualmente em palavras que expressam informação gramatical. Por exemplo, Krug [38] acompanhou a transformação de vários verbos principais (e.g. «want»), em semi-modais (e.g. «wanna») numa variedade de *corpora*. A investigação sobre a gramaticalização sobrepõe-se, portanto, à investigação sobre a mudança semasiológica, embora se concentre especificamente no desenvolvimento do significado gramatical.

A investigação quantitativa sobre a mudança léxical também analisou como as frequências relativas das palavras e unidades de palavras múltiplas subiram e desceram ao longo do tempo (e.g. Krug [38], Nevalainen e Raumolin-Brunberg [39], Gries e Hilpert [40], Geeraerts e Cuyckens [41], Siemund [42]). Na investigação sobre a gramaticalização e a mudança semasiológica, a análise das frequências relativas das palavras é principalmente interessante por poder ajudar na identificação, descrição e explicação das alterações no significado das palavras. Por exemplo, ao traçar a frequência de vários semi-modais desde

o século XVII, Krug [38] mostrou que houve um aumento exponencial na sua utilização à medida que se tornaram gramaticalizados. Na investigação sobre institucionalização, lexicalização e mudança onomasiológica, a análise das frequências relativas das palavras é principalmente interessante porque permite a ascensão de novos itens lexicais e a competição entre sinónimos a ser rastreada ao longo do tempo. Por exemplo, Nevalainen e Raumolin-Brunberg [43] traçaram a frequência do subjetivo «You» relativamente a «Ye» entre o século XV e o século XVII, onde descobriram que a ascensão seguiu uma clara curva em forma de «S», com a sua frequência a aumentar gradualmente no início, depois rapidamente e depois novamente gradualmente, à medida que a mudança se aproximava da conclusão.

3.2 Mapeamento léxical

No artigo realizado pelos autores Grieve, Nini e Guo [44], estes introduzem um método para o mapeamento de inovações lexicais, onde posteriormente a sua origem e propagação será identificada. Este estudo pretende complementar e dar continuidade ao anterior referido, assim, terá igualmente como *corpus* para o estudo em questão um universo de milhões de palavras armazenadas entre 2013 e 2014. As palavras classificadas foram analisadas, o que permitiu detetar cinco principais padrões regionais de inovação léxical no Tweets, as regiões detetadas foram então, Costa Oeste, o Nordeste, o Médio-Atlântico, o Sul profundo, e a Costa do Golfo.

O mapeamento procedeu-se através da frequência relativa acumulada de cada palavra, desde o início do surgimento da palavra até oito pontos temporais. Este mapeamento para uma maior facilidade de compreensão foi descrito em mapas de calor, onde locais do mapa mais escuros indicam uma maior frequência da palavra e locais mais claros, uma menor frequência da palavra. Assim, foi possível verificar que, por exemplo, a palavra «*Baeless*» surgiu numa fase inicial no sul dos Estados Unidos da América especialmente na Georgia, com o decorrer do tempo foi-se propagando pelo centro-oeste até chegar ao nordeste no fim de 2014. Verificou-se então que terá alcançado praticamente o país na sua totalidade mas que manteve uma maior concentração a sul.

O procedimento repetiu-se igualmente para as restantes palavras emergentes, o que permitiu identificar quatro padrões em comum nas inovações lexicais, sendo eles:

1. Os padrões regionais de inovação lexical podem ser observados na comunicação escrita *online*, apesar de a maioria destas palavras possivelmente não terem ocorrido pela primeira vez *online*.

2. As palavras emergentes no Twitter tendem a ter origem em áreas urbanas, mas a influência cultural de uma área urbana parece ser mais importante do que o seu tamanho.

3. A difusão das palavras emergentes é afetada pela geografia e densidade populacional porém, o principal fator são as regiões culturais.

4. O inglês afro-americano é a principal fonte de inovação lexical no Twitter americano.

Deste modo, o estudo forneceu um ferramenta metodológica para futuras pesquisas sobre a análise da inovação linguística, demonstrando como a origem e a propagação das palavras emergentes podem ser medidas e mapeadas.

Os autores Grondelaers, Speelman e Geeraerts [45], referem que contribuição da linguística cognitiva para a lexicologia diacrónica e descreve como os estudos lexicais no âmbito da linguística cognitiva estão a evoluir gradual e naturalmente para uma abordagem sócio-lexicológica que se relaciona com a sociolinguística. Assume a distinção entre a semasiologia e a onomasiologia como o seu princípio organizador básico. Com base nesta distinção, o artigo traça o campo da onomasiologia (provavelmente o menos conhecido dos dois subcampos da lexicologia) e examina a contribuição da linguística cognitiva para esse campo. Também ilustra a importância da sócio-lexicologia para o estudo da variação e mudança onomasiológica, com referências a estudos sociolinguísticos em geral no âmbito da linguística cognitiva. Há duas formas de a linguística cognitiva contribuir para a semasiologia diacrónica: empregando tais mecanismos de mudança semântica como metáfora e metonímia, que a linguística cognitiva tem vindo a lançar nova luz, e explorando a estrutura baseada em protótipos da polissemia. Este artigo considera ainda a teoria dos protótipos, modulações sobre os casos centrais, o desenvolvimento de conjuntos radiais, poligénese semântica, mudança semântica a partir de subconjuntos, e tipos referenciais e não referenciais de significado.

No artigo escrito por Radford, Atkinson, Britain et al. [46] referem que a variação na língua é multidimensional. Analisaram como a variação na estrutura social se reflecte nos padrões de som da língua e como esta variação é muitas vezes indicativa da mudança linguística em curso. Verificou-se também como a variação geográfica da língua é causada por diferentes níveis de contacto entre diferentes povos em diferentes momentos. Nesta secção, focam-se especialmente na variação das palavras e nas suas origens, significados e contextos de utilização. Por fim, é ainda examinado a mudança tanto na escolha das palavras como nos significados dessas palavras.

3.3 Tendências emergentes na linguagem das redes sociais

O estudo realizado por HenryGrieve [47] examina a linguagem nas redes sociais e a emergência de tipos específicos de palavras que permitem uma comunicação corrente. A evolução na comunicação tem de acompanhar a evolução que se verifica no mundo, por isso certamente a evolução irá afetar os temas abordados e a forma de comunicar através da «linguagem virtual».

O uso da social media no mundo está a crescer de uma forma «drástica» como é verificado por investigações realizadas em África, através do uso de redes sociais como *Facebook*, *Twitter* e *WhatsApp*. O uso das redes sociais estará a afetar o vocabulário, neste estudo são analisados os principais motivos que causam a utilização desta recente variedade de termos na língua e a forma como são utilizadas na comunicação nas redes sociais.

Primeiramente, os estratos de comunicação estão a aumentar, uma vez que um nigeriano em média tem mais de 500 amigos no *Facebook*, ainda que possivelmente desses 500 amigos não sejam mais do que 50 amigos ativos. Considerando a velocidade que é necessária para interagir com mais de 20 desses amigos com diferentes tópicos de conversação e a necessidade para uma resposta num prazo limitado, esta situação de rapidez na conversação deu origem a uma nova classe de pessoas chamas «*speed freaks*», a utilização de mensagens nas redes sociais não é, portanto, possível a menos que o codificador utilize formas abreviadas das palavras comuns, a fim de recuperar o atraso no ritmo da conversa, isto deu ao mundo uma nova necessidade de desenvolver uma forma mais rápida de enviar mensagens enquanto conversamos, habitualmente na Nigéria no momento de dar inicio a uma conversa, a maioria dos jovens começa com «*hwfr*», que é a forma abreviada de «*How far*» ou «*How is it going*», há ainda outro exemplo, «I» o que significa «*Hi*», todas estas variações utilizadas na Nigéria para darem inicio a uma conversa devem-se ao fato de a interação ter de ser iniciada a grande velocidade, uma vez que há mais alguém do outro lado à espera de uma resposta. As conversas nas redes sociais é uma forma de discurso multimodal, que é um tipo de comunicação que não é feita cara a cara, por esse motivo, existem muitas pessoas para se falar ao mesmo tempo, esta é a razão para o desenvolvimento desta nova forma de comunicação.

Em segundo lugar, a razão para a utilização desta nova forma de linguagem na comunicação é a necessidade de mostrar emoções por isso o mais habitual é recorrer-se a emoticons ou *emojis*, isto porque verificou-se que transmitir emoções em forma de texto pode ser difícil para o utilizador, pelo que a utilização dos emoticons se tornou extremamente popular entre os utilizadores de redes sociais na Nigéria.

O uso de estas novas formas de comunicação virtual tem recebido várias críticas, ainda assim, certamente existem vantagens na sua utilização, tais como o fato de tornar a comunicação mais rápida e fluída não só na Nigéria, mas num mundo em que a rapidez é importante para manter a «sobrevivência» de uma conversa.

Numa investigação realizada por Zhang, Zhao e Xu [48], dedicada à compreensão dos mecanismos subjacentes às tendências de popularidade. Nas últimas décadas, o foco principal tem sido em particular as propriedades da atenção coletiva e os princípios subjacentes à difusão de artigos novos. Lehmann, Lalmas, Yom-Tov et al. [49] concentram-se na localização de *hashtags* no Twitter e identificam classes discretas de *hashtags* de acordo com a sua evolução popular ao longo do tempo. Os autores também descobrem que os fatores exógenos são mais importantes do que a propagação epidémica no estabelecimento da

popularidade do *hashtag*. JafariAsbagh, Ferrara, Varol et al. [50] propõem um *streaming* quadro de deteção e agrupamento de memes (expressão que descreve um conceito relacionados com humor) em redes sociais em linha. Romero, Galuba, Asur et al. [51] estudam a mecânica de difusão da informação comparando o processo de difusão através de diferentes tópicos no Twitter. Por fim, Bauckhage, Kersting, Hoppe et al. [52] investigam os padrões de adoção de 175 serviços de comunicação social e empresas Web utilizando dados do *Google Trends* e argumentam que a atenção coletiva em quase todos os serviços experimenta uma fase de crescimento acelerada seguida de saturação e declínio prolongado.

4 Metodologia

A elaboração deste trabalho, propõe, como referido anteriormente, desenvolver um modelo de deteção de palavras emergentes dentro de um período de tempo, através de um modelo auto proposto. O planeamento de todo o processo revela-se de uma grande importância, uma vez que permite definir com rigor todas as etapas do trabalho, bem como o seu cumprimento no tempo disponível. Deste modo, este capítulo apresenta as etapas do trabalho, assim como toda a metodologia adjacente ao desenvolvimento do mesmo.

4.1 Visão geral da metodologia

As fases do trabalho seguem o modelo de CRISP-DM, composto por 6 fases, sendo elas, *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation* e por fim *Deployment*. Esta metodologia sendo ela um processo flexível foi adaptada para que se «encaixasse» nas necessidades do trabalho, como ilustra a Figura 4.1.

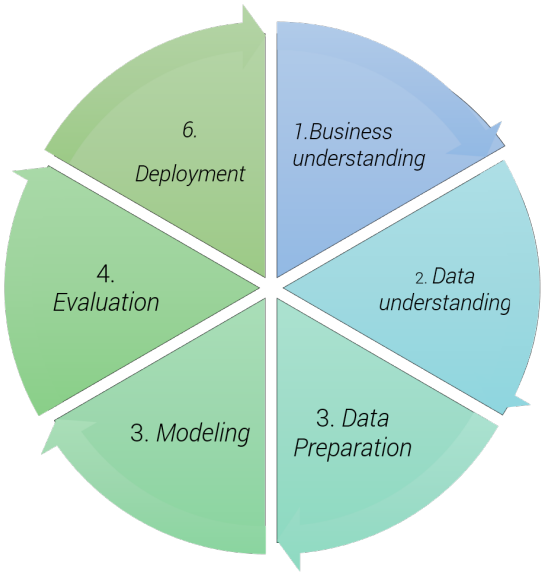

Figura 4.1: *Etapas do trabalho segundo a metodologia CRISP-DM*

1. Na fase do *Business Understanding* foi analisado o mercado em que se insere o estudo e com que tipo de dados poder-se-á resolver o problema, para isso verificou-se que

uma análise a uma base de dados de palavras com um espaço temporal de pelo menos um ano seria o ideal, dado que se verificou que o volume de dados já seria extenso.

2. Na fase de *Data Understanding* os dados contidos na base de dados foram interpretados, onde se verificou que a geolocalização se encontrava em formato *JSON,* e por isso optou-se pela criação de colunas extras de forma a segmentar cada par de nome/valor.

3. Na fase seguinte, *Data Preparation,* sendo ela uma das fases mais essenciais de forma a garantir que os dados se apresentem na qualidade adequada a serem tratados foi realizada a filtragem dos token mencionados mais à frente na secção 4.3.

4. No *Modeling,* ainda que possa ser necessário efetuar alguma atividade de pré-processamento, foi modelada a solução para o problema (secção 4.4), afim de estruturar o mecanismo e os requisitos para identificação das palavras emergentes.

5. Na fase de *Evaluation,* os resultados obtidos foram analisados para garantir que o modelo atinge adequadamente os objetivos do negócio, assim como será mais à frente aprofundado as palavras obtidas foram analisadas com recurso a *word embeddings* e a DELAF, sendo este um formato de dicionários computacionais onde descrevem as palavras simples e também as palavras compostas (Finatto, VALE e Laporte [53]).

6. Por fim, a última fase, o *Deployment,* os resultados obtidos serão disponibilizados ao «*end user*» através de um *website*.

Como já anteriormente foi referido, este trabalho pretende construir e propor um modelo que permita detetar palavras emergentes dentro de um espaço temporal, traçar o mapeamento dessas palavras com a localização geográfica onde as mesmas foram mencionadas e por fim, disponibilizar a visualização dos resultados obtidos através de uma página *web*. Como tal, nesta secção são apresentadas todas as tarefas a elaborar no decorrer do desenvolvimento do modelo, bem como a sua validação.

A Figura 4.2, mostra a metodologia referida na revisão da literatura mas adaptada a este trabalho.

4.2 Extração de dados

A correta seleção da fonte de extração dos dados é de grande importância para a qualidade dos mesmo e para o sucesso do trabalho, uma vez que será através desta que serão compiladas as palavras para o *training* e para o *testing corpus,* deste modo a rede social *Twitter* foi escolhida como meio de *input* para a obtenção dos dados. Assim, uma vez que o focus do trabalho pretende ser aplicado ao contexto português, os critérios de recolha adotados serão os seguintes:

Figura 4.2: *Processo de Desenvolvimento do trabalho*

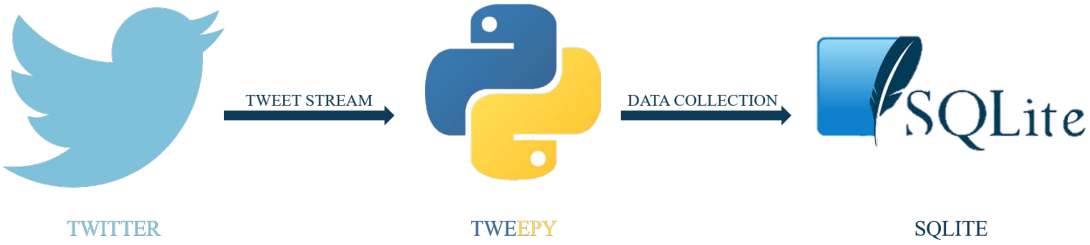

Figura 4.3: *Procedimento de recolha de dados*

- **Geolocalizados**: Serão somente armazenados os *tweets* que contenham informação sobre a localização do local onde foi feita a publicação. Este critério é fundamental para que seja possível analisar a propagação das palavras.
- **Território português**: Serão escolhidos para armazenamento os *tweets* cujo o local de partilha tenha sido em Portugal, uma vez que se pretende apenas efetuar o estudo em território português.

É mais habitual o *tweet* conter dados sobre a localização quando o mesmo foi publicado através de um telemóvel.

Quando os requisitos de recolha a cima mencionados estiverem cumpridos os dados passarão pelo procedimento representado na Figura 4.3.

Tanto para o *training corpus* como para o *testing corpus* o procedimento de extração de dados será o mesmo. Para o *training corpus* incluirá os dados relativos aos primeiros seis meses entre 1 de janeiro de 2018 e 31 de junho do mesmo ano, o período de temporal relativo ao *testing corpus* será entre 1 de julho de 2018 e 31 de junho de 2019.

4.3 Armazenamento e Filtragem

Após a extração dos *tweets* estar concluída para uma base de dados em em SQLite (Figura 4.3) por facilitar a etapa seguinte, ou seja, uma normalização da palavras obtidas. O ar-

mazenamento numa base de dados torna-se realmente importante uma vez que caso não o fosse feito, seria impossível proceder a analises a palavras «em massa». Para além disso, o armazenamento possibilitará a exploração dos dados, de forma a obter estatísticas de palavras (p.e. a frequência ao longo dos meses).

Relativamente à filtragem, esta é uma importante etapa uma vez que é fulcral para a qualidade dos dados que serão posteriormente analisados, neste sentido, decidiu-se remover os seguintes *tokens*:

- Palavras começadas por «http»: Foram removidas as palavras começadas por «http» uma vez que estas representam *websites* e não são serão analisados.
- Palavras começadas por «@»: Tendo sido escolhido o *Twitter* como fonte de dados, nesta plataforma é utilizado o *token* «@» para mencionar outro utilizador, o que não é pretendido obter utilizadores, por este motivo foram removidas.
- Número: Uma vez que o estudo foca-se em palavras foram removidos todos os conjuntos de caracteres que contenham somente números.
- Palavras que contenham @: Igualmente não se pretende analisar endereços de email, por este motivo foram também estes removidos.
- Palavras com uma frequência menor que 15 utilizadores distintos: Para o estudo é importante que a palavra tenha sido referida por diferentes utilizadores, por esse motivo todas as palavras que contivessem a baixo de 15 utilizadores foram removidas. Esta condição pretende «descartar» todas as palavras que tenham sido referidas/publicadas por *bots* (*robots*).
- Número de ocorrências superior a 50: Como mais à frente será explicado, a deteção das palavras emergentes basear-se-á na frequência das palavras e visto que uma palavra pode conter uma frequência elevada e ainda assim ter ocorrido poucas vezes foi proposto o valor 50 como número mínimo de ocorrências que a palavra deverá ter, deste modo, todas as palavras que não cumprirem este critério serão igualmente removidas. Adotou-se assim que ocorrências a cima 50 seriam consideradas como «elevadas».

4.4 Desenvolvimento do Modelo de Deteção de Palavras Emergentes

Uma vez construídos os *training* e *testing corpus*, será o momento de propor um modelo de deteção das palavras emergentes, para isso o conjunto de palavras dos primeiros seis meses (*training corpus*) servirão como base de comparação para detetar no *testing corpus*

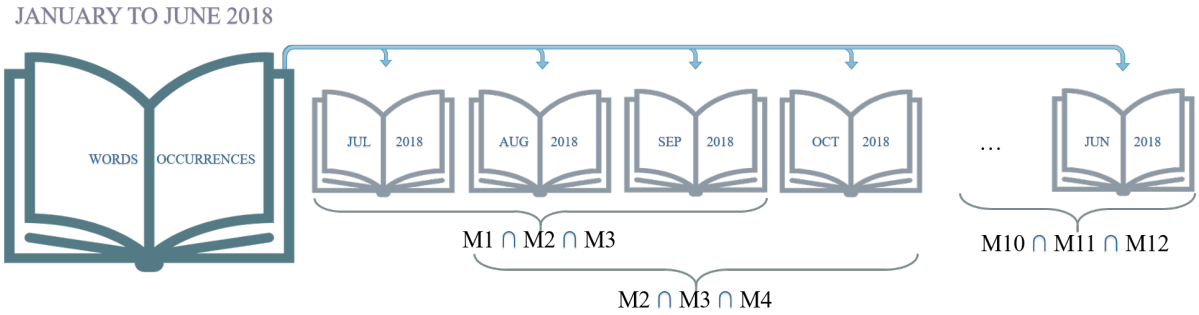

Figura 4.4: *Modelo para identificar palavras emergentes*

palavras que não tenham surgido no *training corpus*. Realizado este procedimento a qualidade dos dados não será a melhor, uma vez que grande parte dos resultados obtidos serão erros ortográficos, por esse motivo e por se pretender obter palavras emergentes será feita uma última filtragem analisando a frequência mensal de cada palavra, ficando apenas com as palavras que ao longo de pelo menos três meses possuam uma frequência elevada comparativamente aos restantes meses. A Figura 4.4 representa o modelo proposto.

4.5 Surgimento das palavras

As palavras emergentes não tiveram necessariamente de ser utilizadas pela primeira vez em 2018 ou 2019. Embora a utilização das palavras tenha aumentado muito ao longo do período analisado, elas podem ter surgido por um período de tempo consideravelmente mais longo. O objetivo da análise, no entanto, não foi para identificar palavras que foram formadas pela primeira vez durante este período de tempo (ou seja, neologismo deteção), mas para identificar formas raras não listadas em dicionários que se estavam a espalhar rapidamente no Twitter entre 2018 e 2019. Presumivelmente, a maioria das palavras não são utilizadas pela primeira vez *online* e portanto tentando datar a formação das palavras através da análise dos dados do Twitter ou qualquer outra variedade de comunicação mediada por computador não é geralmente fiável, à parte, talvez, os termos relacionados com o Twitter e alguns acrónimos. No entanto, continua a ser informativo para considerar o quão novas são estas formas emergentes.

Dada a falta de uma *corpora* suficientemente grande e densa de linguagem falada informal, é provavelmente impossível identificar a hora ou o local exato onde estas formas foram introduzidas. Para testar esta hipótese, cada uma das palavras emergentes será pesquisado no *Google Trends* (`www.google.com/trends`), o que permite que a frequência dos termos de pesquisa no Google seja analisada ao longo do tempo, e no dicionário informal (`www.dicionarioinformal.com.br`), que é um dicionário popular, onde as palavras são definidas pelos utilizadores.

Figura 4.5: *Modelo para identificar novas palavras na língua*

4.6 Deteção de novas palavras

Uma vez detetadas as palavras consideradas como emergentes, será o momento de analisar quais destas serão identificadas como novas, isto é, que não pertençam ao dicionário da língua portuguesa. Para isso isso recorreu-se a recursos de PLN, como o uso de léxicos para verificar a existência das palavras emergentes na língua portuguesa. Foi selecionado um dicionário de palavras simples de português brasileiro com 75 mil palavras e flexões das mesmas.

Assim, com base no dicionário selecionado cada palavra emergente foi analisada para averiguar se a mesma existia no dicionário, caso não se verifique a sua existência no dicionário será então classificada como uma nova palavra (Figura 4.5).

4.7 Mapeamento geográfico

Existem várias maneiras de mapear a origem das palavras assim como a sua propagação. A abordagem adotada é através do número de ocorrências de cada palavra pelos distritos ao longo do tempo de análise.

Ainda que, como já foi referido, para a extração do *tweet* fosse obrigatório que contivesse a localização onde o mesmo foi publicado é necessário que exista um tratamento dos erros.

Tratamentos dos erros

1. Associação dos concelhos ao respetivo distrito: Ainda que todos os dados obtidos contenham localização geográfica, esta encontra-se representada pelo concelho, sendo que o estudo pretende ser efetuado a nível do distrito, esta associação precisa de ser realizada. Para isto recorreu-se um ficheiro disponibilizado pelo *website* da Central de Dados dos CTT (Correios e Telecomunicações de Portugal) que contem quais os concelhos pertencentes a cada distrito (p.e. «Lagos» pertencer a «Faro»).

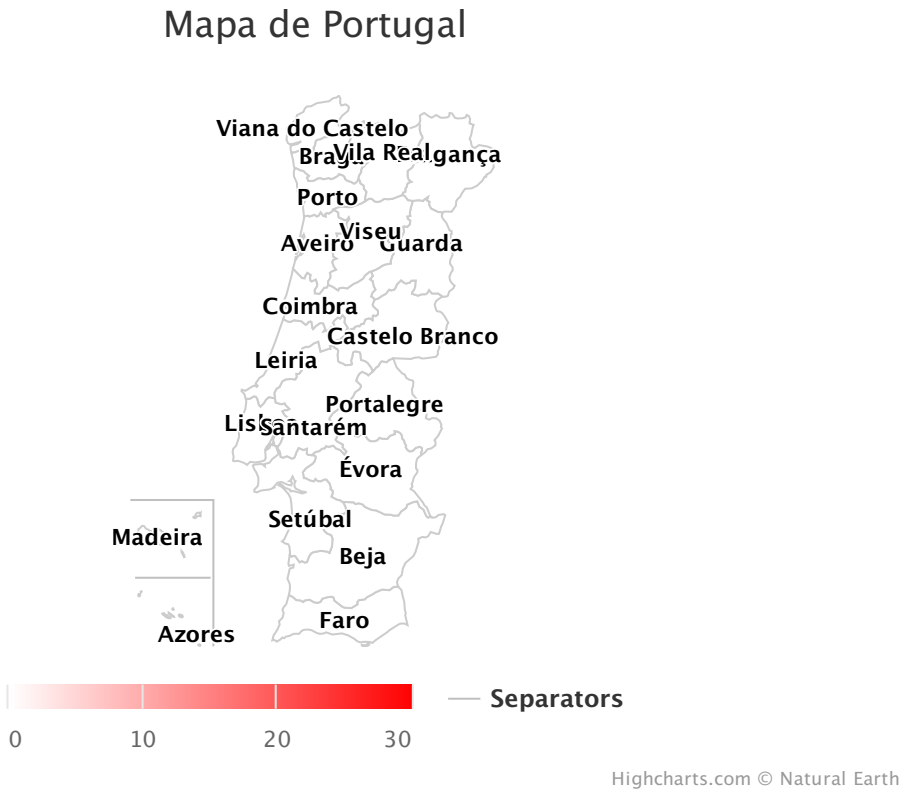

Figura 4.6: *Mapa de Portugal*

2. Erros ortográficos ou estrangeirismos: Foram detetados alguns casos onde o concelho relativo ao *tweet* se encontrava com erros ortográficos (p.e. «Vila Nova de Famalicáo»), nestes casos foi preciso de uma forma manual proceder à correção dos mesmo. Foi ainda detetado alguns casos onde a localização não se encontrava em português como por exemplo, «Oporto» ou «Lissabon», nestes casos foi também necessário prepará-los para poderem ser analisados e assim foram traduzidos para «Porto» e «Lisboa».

Após ter sido tratada a qualidade dos dados cada palavra estas estarão preparadas as condições para estas serem trabalhadas, deste modo pretende-se que cada palavra emergente seja representada num mapa como o da Figura 4.6 onde será possível visualizar o número de ocorrências da palavras por distrito e consequentemente conseguir compreender os locais onde as palavras foram mais mencionadas.

4.8 *Word embeddings - fastText*

Ainda que o treino de fastText seja feito em *multi-thread* a leitura dos dados é feita via *single-thread*.

O uso desta biblioteca pretende ser um mecanismo de explicação do motivo de ori-

gem de uma determinada palavra, para isto serão utilizados dois modelos de palavras pré-treinadas, a primeira com base no *Common Crawl* e um segundo modelo com base no *Wikipédia,* ambos os modelos com o idioma português.

Na tabela 4.1 encontram-se as palavras vizinhas mais próximas da palavra «Informática».

Tabela 4.1: *Palavras vizinhas do exemplo de fastText*

Common Crawl	***Wikipédia***
Infomática	informática
Informáti	geoinformática
informática	microinformática
Informática-	informáticas
Cavuca	teleinformática
Tecnologia	informático
Icompy	neuroinformática
Informatica	quimioinformática
Informátic	performática

Com a utilização desta biblioteca espera-se que o conjunto de palavras obtidas para cada palavra emergente ajude a compreender os principais motivos que levam a que as palavras emerjam em Portugal.

4.9 Desenvolvimento do *website*

O desenvolvimento do *website* pretende disponibilizar livremente o estudo aqui realizado, onde possibilite compreender tanto a área de estudo como o objetivo concreto do trabalho realizado. No *website* o utilizador poderá visualizar as palavras emergentes assim como o mapeamento da frequência de cada palavra pelos distritos de Portugal, poderá ainda ter acesso a mais dados como:

- Número de palavras do *training* e do *testing corpus.*
- Total de palavras analisadas.
- Número de diferentes utilizadores.
- Número total de palavras emergentes.
- Número de palavras novas / já existentes.

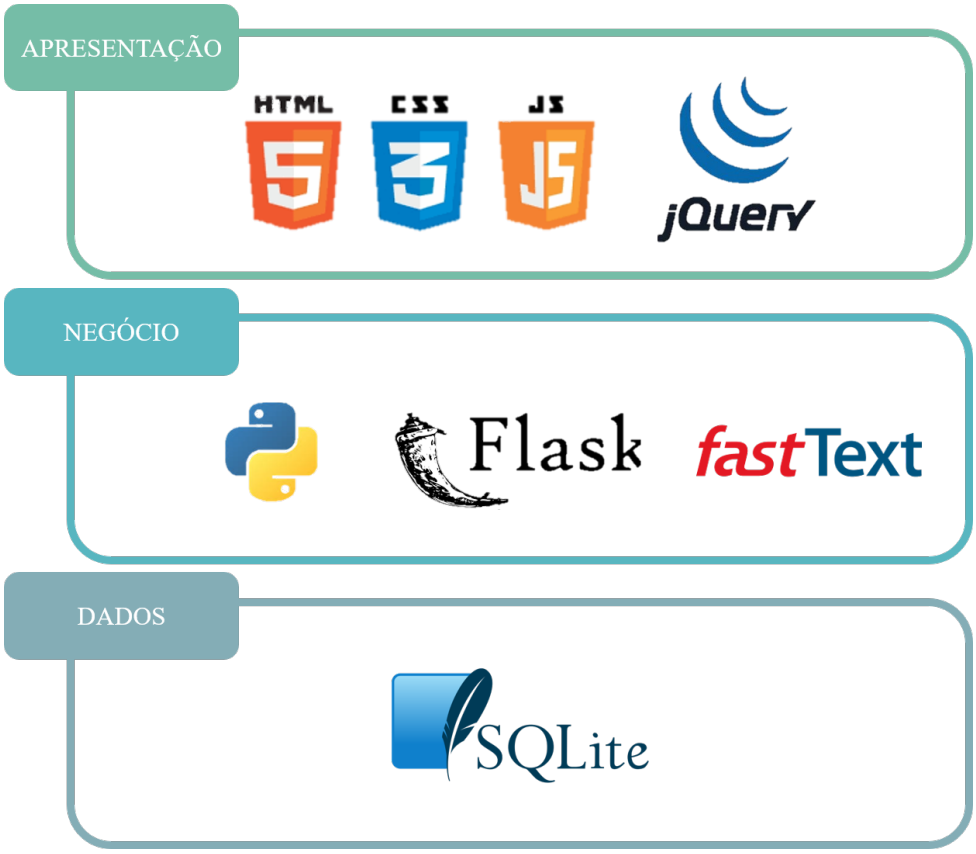

Figura 4.7: *Modelo 3 camadas* (*website*).

- Número de dias analisados.

Para o desenvolvimento do *website* foram seguidas as boas práticas no que diz respeito à organização por *layers* ou camadas, assim como demonstra a Figura 4.7 o *website* encontra-se organizado por 3 camadas. A primeira camada ou camada de apresentação será a que irá interagir diretamente com o utilizador na representação dos dados e esta assenta em ferramentas como HTML5, CSS3, JavaScript e JQuery. A segunda camada ou camada de negócio será a que ficará encarregue de fazer o processamento lógico e tratar os dados que deverão ser apresentados, para isso, e como representado na Figura 4.7 esta foi desenvolvida em Python com recurso à *framework Flask* e ainda com recurso ao fastText. Por fim, a terceira camada, a camada dos dados e esta será o repositório dos dados a serem disponibilizados no *website* que serão então o *tweets* todos os dados agregados cada *tweet* (ex: data, localização, utilizador, etc...).

4.10 Modelação do sistema em BPMN

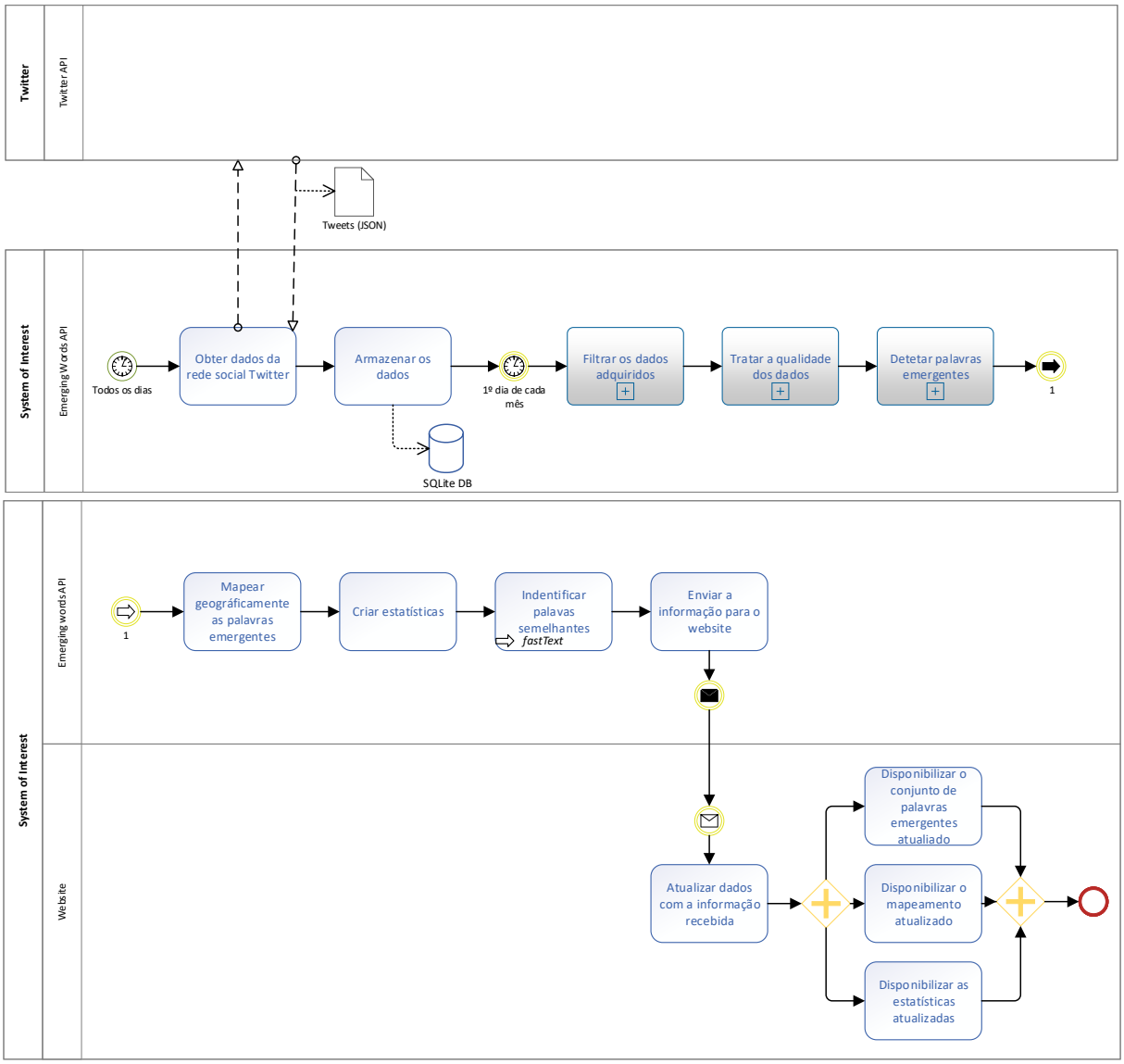

Figura 4.8: *Modelação do sistema (BPMN)*

A Figura 4.8 representa a modelação do sistema com recurso à notação *Business Process Model and Notation* (BPMN), uma vez que esta é uma metodologia de gestão de processos e que através do seu grafismo facilita o entendimento.

Todos os dias o sistema de deteção de palavras emergentes recebe proveniente da API interna do Twitter dados sobre as publicações realizadas pelos utilizadores na rede social onde estes dados serão então armazenados. Aquando o início de cada mês existirá uma atualização na informação apresentada ao utilizador final, para isso os dados armazenados passarão por uma filtragem onde serão descartados consoante as regras descritas na secção 4.3. Terminada a filtragem os dados serão tratados afim de melhoramento da qua-

lidade dos mesmos onde será feita uma uniformização, como por exemplo as palavras que contenham letras maiúsculas passarão apenas a constar em minúsculas. Com a filtragem terminada e uma melhor qualidade dos dados serão então detetadas as palavras emergentes de acordo com as características mencionadas na secção 4.4. Assim que se tenha obtido o conjunto de palavras emergentes será calculada a frequência destas por cada distrito de Portugal e calculadas algumas estatísticas que serão apresentadas no *website* (Figura 5.3) e para cada palavra emergente serão identificadas as palavras vizinhas, descrito na secção 4.8. Assim, após o tratamento lógico da solução os resultados serão transmitidos ao *website* que ficará encarregue de demonstrar todos os resultados obtidos e permitir ao utilizador navegar e realizar a análise que pretender efetuar de acordo com as palavras emergentes disponíveis (descrito na secção 5.9).

5 Análise e discussão dos resultados

Este capítulo visa apresentar os resultados obtidos com o desenvolvimento deste trabalho e respetiva discussão, de acordo com a metodologia que foi proposta no capítulo anterior.

5.1 Extração dos *tweets*

Após a extração ocorrida entre Janeiro de 2018 e Junho de 2019 obteve-se um total de 8860871 *tweets* os quais serão alvo de análise, como anteriormente já foi referido todos contêm a geolocalização do local da sua publicação. Ao longo dos 632 dias de recolha de dados obteve-se informação sobre 96538 diferentes utilizadores. Estima-se que a API interna do Twitter apenas permita a extração de 1% de todos os *tweets* realizados, o que ainda assim se resuma num valor bastante elevado.

De salientar que na coluna da tabela da base de dados que recebeu os *tweets*, encontra-se com dados no formato *json*, o que como mais à frente será explicado mais pormenorizado levará ao tratamento dos dados e à criação de novas colunas.

5.2 Segmentação dos *tweets*

Relativamente à segmentação dos *tweets* para o *training* e *testing corpus* a Tabela 5.1 ilustra os períodos de extração correspondentes, tal como o número de *tweets* por cada uma das amostras.

É importante salientar que a obtenção das datas em cada *tweet* foi fulcral para a realização deste trabalho. A segmentação só foi possível porque era sabido que a segmentação seria feita de acordo com a data de extração.

Tabela 5.1: *Informação das amostras selecionadas*

Corpus	Período de extração	Amostra
Training	01-Jan-2018 > 31-Jun-2018	2.934.963
Testing	01-Jul-2018>31-Jun-2019	5.925.908

Tabela 5.2: *Palavras comuns ao longo de 3 meses*

***Sets* de meses**	**Palavras comuns**
2018-07, 2018-08, 2018-09	43
2018-08, 2018-09, 2018-10	70
2018-09, 2018-10, 2018-11	80
2018-10, 2018-11, 2018-12	83
2018-11, 2018-12, 2019-01	94
2018-12, 2019-01, 2019-02	104
2019-01, 2019-02, 2019-03	115
2019-02, 2019-03, 2019-04	129
2019-03, 2019-04, 2019-05	147
2019-04, 2019-05, 2019-06	146

5.3 Desenvolvimento do modelo

Durante uma fase de pré-processamento, foram removidos uma série de *tokens* irrelevantes para o estudo. Especialmente, foram removidos todos os *tokens* começado por «*http*» ou «@», por corresponderem a *websites* ou menções a utilizadores da plataforma, foram ainda removidos os conjuntos de caracteres que fossem somente números. A *tokenization* utilizada a partir da função «*TweetTokenizer*» da biblioteca *Natural Language Toolkit* (NLTK) foi utilizada com o objetivo de normalizar as palavras com mais de três letras iguais a ocorrerem de forma simultânea, como por exemplo a palavra «looooooove» foi convertida para «looove».

Como anteriormente já foi referido e é possível verificar na Figura 4.4, numa primeira fase começou-se por organizar as palavras dos *tweets* em dicionários dos primeiros seis meses, (entre janeiro e junho de 2018), correspondendo a 2.9 milhões de *tweets*. Numa segunda fase foram identificadas todas as possíveis novas palavra nos 12 meses seguintes. Foi considerada uma frequência mínima de 50 e por pelo menos 15 diferentes utilizadores da plataforma, com o objetivo de minimizar a presença de erros ortográficos. Por fim, assumimos que uma palavra emergente teria que conter uma frequência elevada de ocorrências ao longo de três meses consecutivos, assim foram calculadas as palavras que ocorrem em pelo menos três meses consecutivos. Na tabela 5.2 é possível verificar o volume de palavras candidatas a emergentes e que ocorrem ao longo de 3 meses.

É interessante de verificar que o número de palavras candidatas aumenta consoante o tempo se vai «afastando» do período considerado para o *training corpus*, o que sugere que a utilização das palavras irá se modificando ao longo do tempo.

5.4 Palavras emergentes

Após proceder à análise das palavras e às filtragens obtiveram-se 26 palavras emergentes ao longo de 1 ano de *time-span*. Como é possível verificar na tabela 5.3 destacam-se cinco *tokens* classificados como emojis / emoticons, a presença destes como emergente possivelmente terá sido motivada pelo aparecimento de novas aplicações ou por *updates* a aplicações já existentes. Verificou-se também que alguns *tokens* resultantes são, de facto, palavras já bem conhecidas, que não estavam presentes no *corpus* inicial. Isto acontecerá devido a vários «*major*» fatores:

1. O *training corpus*, foi criado com um limite de dados (imposto pela API do *Twitter*).

2. O vocabulário usado no *tweets* é particularmente distinto de outras fontes de dados, como livros ou jornais, uma vez que esta rede social é de carácter informal.

3. O conteúdo produzido nesta rede social é altamente influenciado por acontecimentos exteriores.

A tabela 5.3 apresenta assim as principais palavras candidatas a emergentes. A maioria das palavras identificadas correspondem a nomes de pessoas, especialmente relacionadas com futebol (e.g. Keizer, Gudelj, Militao, Corchia, Phellype, Castaignos e Manafá). Desta forma as palavras foram agrupadas da seguinte forma:

Emojis / Emoticons: emoticons, retratam o humor de um escritor ou expressões faciais na forma de ícones, geralmente usados em conjunto com uma frase para expressar emoções.

Benfiquistão, minguem, Vagandas: estas palavras emergentes correspondem a derivações de palavras já existentes. As últimas são utilizadas num sentido irónico, ironizando o nome "Varandas", Presidente do Sporting.

Bozo: apesar de corresponder a uma palavra existente, o seu verdadeiro significado atual foi alterado durante o período de tempo em estudo. É frequentemente usado para se referir a Bolsonaro, o atual Presidente do Brasil, de uma forma muito depreciativa e fazendo uso da realização fonética do seu nome como nas suas origens italianas Bol[z]onaro.

Lomotif: corresponde ao nome de uma aplicação *mobile* que viralizou durante um longo período de tempo.

T*aki, sicko, shallow, legaci*es: o aparecimento destas palavras está relacionado com novas músicas ou séries de televisão. Taki, sicko e shallow referem-se ambos ao nome de uma música (Taki Taki, Sicko Mode e Shallow), enquanto *Legacies* é o nome de uma série de televisão.

Tabela 5.3: *Palavras emergentes e sua frequência correspondente ao longo dos 12 meses*

Token	**Freq.**	**Observação**
phellype	125723	Jogador de futebol
corchia	59514	Jogador de futebol
castaignos	52735	Jogador de futebol
bozo	41613	Palavra inglesa para «homem estúpido»
120M	11786	Preço pago por um jogador de futebol
benfiquistão	10381	Analogia ao benfica
trotinetes	9836	Palavra existe em português
trotinetas	9785	O mesmo que «trotinetes»
taki	9710	Nome de uma música (Taki Taki)
militao	9665	Nome de um jogador de futebol
minguem	9635	Minguém o mesmo que «ninguém»
manafá	8594	Jogador de futebol
vagandas	7146	Ironia para Varandas, Presidente do Sporting
shallow	5654	Nome de uma música
sicko	5237	Nome de uma música (sicko mode)
keizer	4147	Treinador de futebol
kbk	4147	Calão para «kill or be killed»
lomotif	3135	Nome de uma app
legacies	3082	Nome de uma série
gudelj	2824	Jogador de futebol
guaidó	2689	Político venezuelano
🥺	525	Emoji de rosto implorando
🥵	271	Emoji de rosto fervendo
🥳	163	Emoji com rosto de festa e chapéu de festa
🥴	125	Emoji com rosto embriagado
🥶	66	Emoji com rosto gelado

Tabela 5.4: Surgimento das palavras emergentes no *Google Trends* e no Dicionário Informal

Token	***Google Trends***	**Dicionário informal**	**Delta**
phellype	2018	-	-
corchia	2018	-	-
castaignos	2012	-	-
bozo	2004	2007	3
120M	2008	-	-
benfiquistão	-	-	-
trotinetes	2004	-	-
trotinetas	2004	-	-
taki	2004	-	-
militao	2004	-	-
minguem	-	-	-
manafá	2013	-	-
vagandas	-	-	-
shallow	2004	2019	15
sicko	2004	-	-
keizer	2005	-	-
kbk	2006	-	-
lomotif	2017	-	-
legacies	2017	-	-
gudelj	2009	-	-
guaidó	2018	-	-

5.5 Surgimento das palavras

Como é possível verificar na Tabela 5.4, a ocorrência mais antiga registada para algumas das palavras analisadas foi no ano 2004, uma vez que o *website Google Trends* só começou a fazer o registo a partir desse mesmo ano. Esta abordagem de datar a introdução de novas palavras ainda que não seja 100% real, uma vez que estas palavras poderão não ter surgido pela primeira vez *online*, permite estabelecer uma previsão aproximada da data real de introdução. De facto, nos casos em que as primeiras palavras identificadas pelo *Google Trends* são anteriores à primeira entrada no Dicionário Informal (p.e. «bozo», «shallow». Parece portanto que a maioria das palavras emergentes na lista foram introduzidas após ou por volta de 2004, como «sicko», «shallow», «taki», «trotinetas», «militao», «bozo» e «trotinetes». É evidente que várias destas palavras existem há muitos anos (p.e. «trotinetes»). Um importante resultado descritivo deste estudo é, portanto, que as novas formas são frequentemente caracterizadas por uma utilização muito pouco frequente durante anos, até acabarem por surgir e obterem uma frequência na sua utilização muito elevada durante um período de tempo.

Tabela 5.5: *Palavras mais próximas tendo em conta o modelo fastText treinado com Common Crawl*

trotinetes	**minguem**	**militao**	**bozo**
trotinetas	minguém	Militao	fiuk
trotinete	ninguem	militará	buneco
trotineta	mingue	militara	yudi
Trotinetes	ninquem	parlamentario	raxo
patinetes	niguem	milite	vsf
brinquedosTrotinetes	preucupam	parlamenta	veei
trotinette	ningem	ministrativo	adogo
triciclos	nimguem	parlamentaria	veii
Trotinete	nunguem	estudanti	affz
velocípedes	preucupa	habitao	zuei

5.6 Novas palavras

Foi utilizado outro léxico, nomeadamente o léxico DELAF com o objetivo de identificar quais as palavras identificadas como emergentes. Para isso foi escolhido um dicionário DELAF que contêm as palavras da língua portuguesa bem como as suas flexões. Desta forma, com o recurso a este léxico identificou-se que apenas a palavra «trotinetes» já constava, assim todas as outras 25 serão novas ou derivações de palavras já existentes. O que poderá induzir a que rede social *Twitter* demonstre ser uma fonte de deteção de novas palavras.

5.7 *Embedding - fastText*

De forma a compreender o motivo pelo qual as palavras identificadas na tabela 5.3 emergiram estas foram submetidas uma análise com recurso à biblioteca de fastText, para tal foram selecionados 2 modelos de aprendizagem com o idioma em Português. Como anteriormente já foi mencionado na secção 4.8, o primeiro modelo será o *Common Crawl* e o segundo modelo baseado no *Wikipédia.*

Como é possível verificar nas tabelas 5.5 e 5.6 foram selecionadas 4 palavras para serem analisadas.

A palavra «trotinetes» como já referido terá sido a única palavra a surgir que já consta no vocabulário português e aplicada análise do *fastText* vem confirmar que esta refere-se ao velocípede.

Os resultados obtidos através da análise à palavra «minguem» demonstra que consistem da derivação da palavra já existente «ninguém». Este poderá revelar-se um caso de evolução da língua portuguesa.

Também a palavra «militao» sugere que seja o resultado da derivação da palavra já existente «militar» uma vez que se obteve como resultado «militará», «militancia», «militava»

Tabela 5.6: *Palavras mais próximas tendo em conta o modelo fastText treinado com Wikipédia*

trotinetes	**minguem**	**militao**	**bozo**
jinetes	xinguem	militancia	zombozo
quitinetes	minguet	militava	esbozo
pedetes	minguez	milita	dozo
namnetes	lêmingues	taitao	calabozo
garçonetes	lemingues	otao	jozo
ginetes	ninguem	militam	kozo
gabinetes	aterrorizam	militar	bozomal
menetes	minguados	vitao	yozo
patinetes	minguou	militara	bozon
cinetes	percebiam	kitao	logozo

e «militar».

Por fim, a quarta palavra, «bozo», sugere que a mesma terá derivado da palavra «zombozo», este será o nome dado a um desenho animado retratado por um palhaço sombrio e cruel com um sentido de humor negro do mundo de *Ben 10*, o que sugere que no contexto onde a palavra «bozo» se encontrar terá uma conotação de palhaço, no entanto, segundo Soares [54] afirmou no jornal «*Correio Braziliense*» que «Bolsonaro deve-se «orgulhar» da comparação feita entre eles e que a esquerda elogia o presidente quando assim o chama.», afirmou ainda que «O Bozo foi bom para as famílias. O Bozo era a continuação do lar das pessoas, o Bozo não fazia nada de maldades. Portanto, Bolsonaro, tenha orgulho quando te chamam de Bozo, porque estão te chamando de uma pessoa boa...».

5.8 Propagação geográfica

Uma vez realizada a análise do *fastText* será importante compreender quais os locais em território português onde as palavras mais surgiram, serão assim alvo de análise o mesmo conjunto de palavras que terão sido submetidas à análise na secção anterior.

Após o tratamento dos dados, estes foram assim analisados ao longo de Portugal Continental e e ilhas com base no número de ocorrências por distrito.

Na imagem da Figura 5.1 estão representadas as ocorrências das 4 palavras, no apêndice encontram-se os mapas das restantes palavras emergentes.

No mapa relativo à palavra «trotinetes» destaca-se claramente a cidade de Lisboa com o maior número de ocorrências, com o domínio praticamente absoluto das ocorrências, isto terá sido resultado do fenómeno que foi na região das trotinetes elétricas. Já no mapa da palavra «minguem» é possível verificar que esta ocorreu num maior número de distritos, a capital manteve-se como a cidade com o maior número de ocorrências mas também se verificou uma quantidade elevada na cidade do Porto, tendo como foco ambas as cidades é

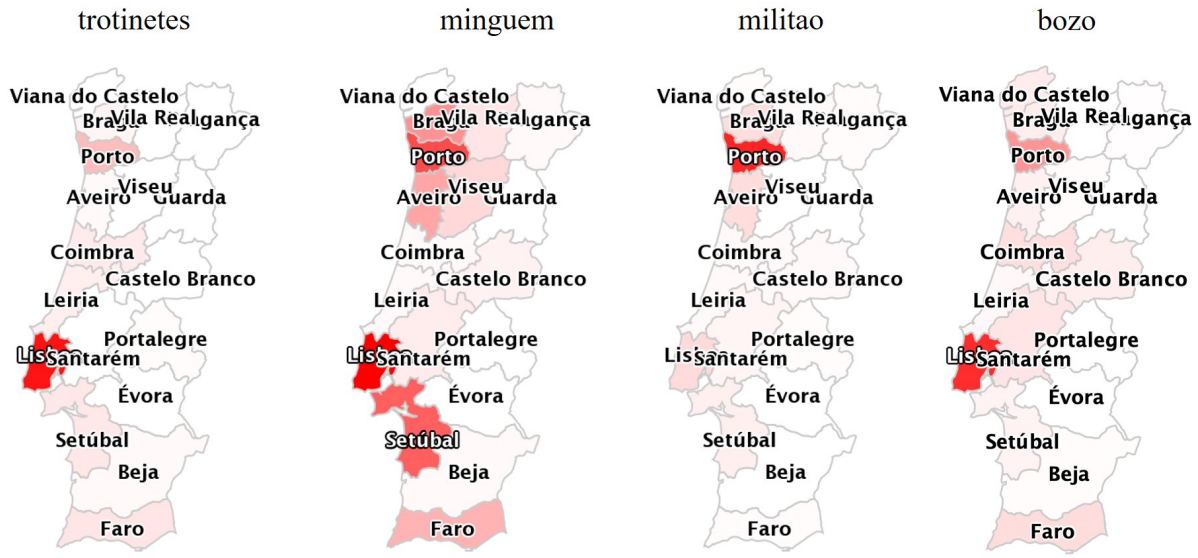

Figura 5.1: *Frequência por distrito das palavras: trotinetes, minguem, militao, bozo*

visível que os distritos em redor obtiveram também ocorrências das palavras, o que sugere que a palavra se tenha propagado diretamente para as cidades mais próximas. Relativamente à palavra «militao», esta claramente obteve o principal foco no distrito do Porto com uma muito reduzida propagação para os restantes distritos, existindo apenas uma breve propagação para duas cidades ao redor do Porto. Por fim, a palavra «bozo», obteve um comportamento idêntico à palavra «minguem», onde é possível verificar que os epicentros terão sido os distritos de Lisboa e do Porto e que existiu uma propagação para as cidades mais próximas das mesmas.

Talvez o padrão geral mais claro seja a distinção das palavras que estão associadas ao centro do país (e.g. keizer, trotinetas ou castaignos) e as que estão associadas com o norte (e.g. lomotif, militao ou manafá). Grande parte dos mapas, tanto os representados na Figura 5.1 como os restantes nos anexos demonstram padrões regionais na propagação, com as palavras alcançarem a grande maioria dos distritos até ao final do período de análise.

Ainda que a distância física afete a emergência de novas palavras no Twitter, existem muitas variáveis que é preciso ter em conta, como por exemplo o surgimento pela primeira vez de uma palavra num determinado distrito no Twitter poderá não corresponder à realidade e essa palavra ter sido originada pela primeira vez noutro local. Claramente outros fatores afetam a propagação e o surgimento das palavras. Estes fatores aparentam incluir a densidade populacional, visto que na grande maioria das palavras a sua dominância persiste nas cidades com maior população de Portugal. Contudo, identificar padrões regionais de inovação léxical no Twitter é muito complexo, revela-se difícil de generalizar padrões comuns de propagação e de identificar o que motiva a propagação.

O principal padrão regional de inovação léxical identificado neste estudo está primeira relacionado com a capital, uma vez que demonstrou ser o local mais comum para ser o

centro das ocorrências das palavras, o que devido a esse fator, de forma natural as cidades diretamente adjacentes com a cidade de Lisboa encontram-se com valores elevados de utilização da palavra. Também o norte demonstrou ter um comportamento semelhante com a cidade do Porto como principal foco. A exceção deste padrão identificado será no sul uma vez que não demonstrou registos do mesmo acontecer.

5.9 *Website*

Como fase final e integrante do modelo de CRISP-DM foi desenvolvido um *website* com o objetivo de partilhar com investigadores dentro da área de AI e NLP a metodologia adotada e os resultados obtidos. No *website* será possível compreender o que motivou o estudo, estatísticas resultantes da análise e uma análise completa a cada palavra emergente com a sua frequência por distrito, exemplos de publicações no Twitter e os resultados obtidos com a aplicação de *embeddings* (*fastText*).

No momento em que os utilizadores acederem ao *website* irão visualizar a página inicial («*Home*»), Figura 5.2, onde, em primeira instância terão uma breve explicação do tema e do estudo aplicado, com imagens ilustrativas que vão alternando entre si, identificas por A), B) , C) e D).

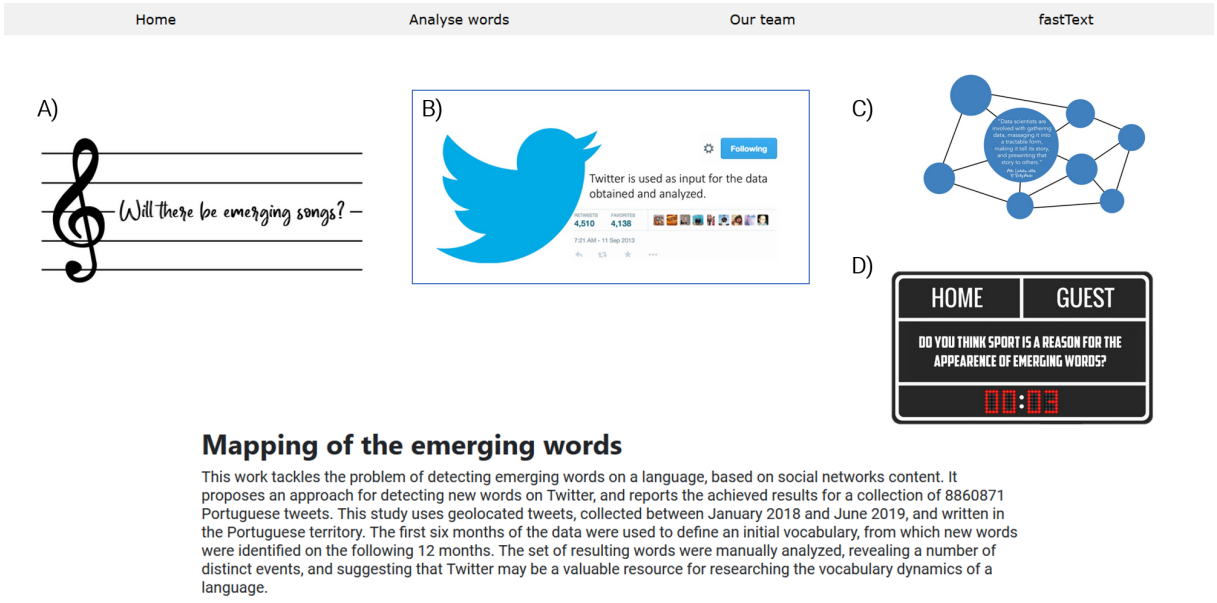

Figura 5.2: *Website - Página inicial («Home»)*

Ainda na página inicial são apresentados um conjunto de dados para compreender mais facilmente a dimensão de dados analisados no estudo, assim como se pode verificar na Figura 5.3, através de um gráfico circular é possível visualizar o top 10 das palavras emergentes que mais ocorreram e ainda informação relativa a:

- Número de tweets que constam no *training corpus* (*Initial vocabulary* (tweets)).

- Número de tweets que constam no *testing corpus* (*One year tweets*).
- Número de palavras analisadas (*Words*).
- Número utilizadores únicos (*Users*).
- Número de palavras emergentes (*Emerging words*).
- Número de palavras emergentes que não constam no dicionário (*New words*).
- Número de palavras emergentes que constam no dicionário (*Existing words*).
- Número de dias do período de dados analisado (*Days analysed*).

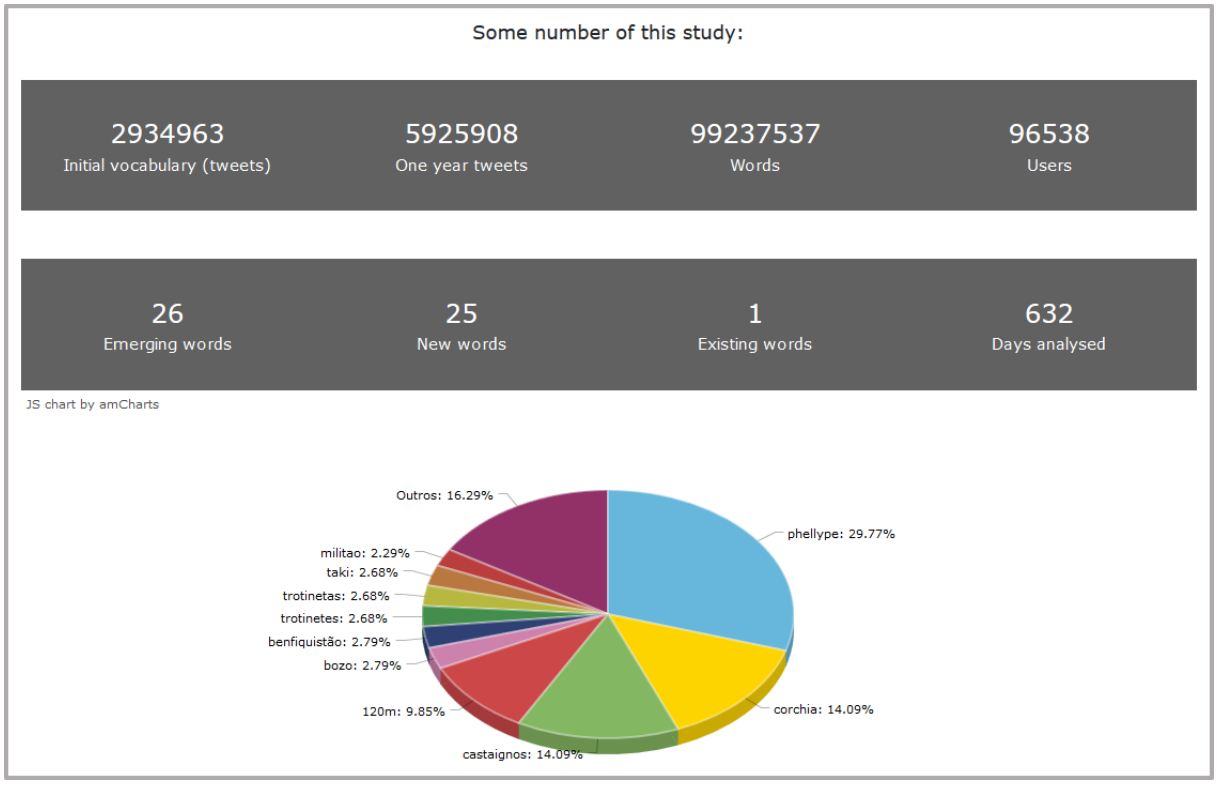

Figura 5.3: *Website - Página inicial («Some number of this study»)*

Por fim, para completar a página principal existe uma secção onde é apresentada a equipa que desenvolveu o trabalho assim como as principais ferramentas utilizadas para a criação do *website.*

Acedendo através do cabeçalho à secção «*Analyse words*», lá será possível proceder à análise das palavras emergentes num mapa do território como o representado na, Figura 5.1, onde em primeira instância o *end user* poderá preencher o campo «*Word to analyse*», e caso a palavra introduzida conste na lista de palavras identificadas como emergentes o mapa será preenchido conforme os dados relativos à palavra. O mapa será somente relativo a Portugal, incluindo os arquipélagos. De forma a representar a intensidade de ocorrências

da palavra selecionada em cada distrito o mapa utilizará a gradação da cor vermelha para representar as ocorrências da palavra.

Assim fará todo o sentido que o mapa seja complementado imediatamente abaixo com um gráfico temporal com o número de ocorrências ao longo do tempo (Figura 5.4), identificando o valor médio de ocorrências com base no período temporal. O gráfico permite ainda compreender qual o mês onde a palavra começou a surgir assim como os meses que obteve um número de ocorrências elevada (pelo menos três meses consecutivos) que permitisse assim a palavra ser emergentes segundo a metodologia apresentada.

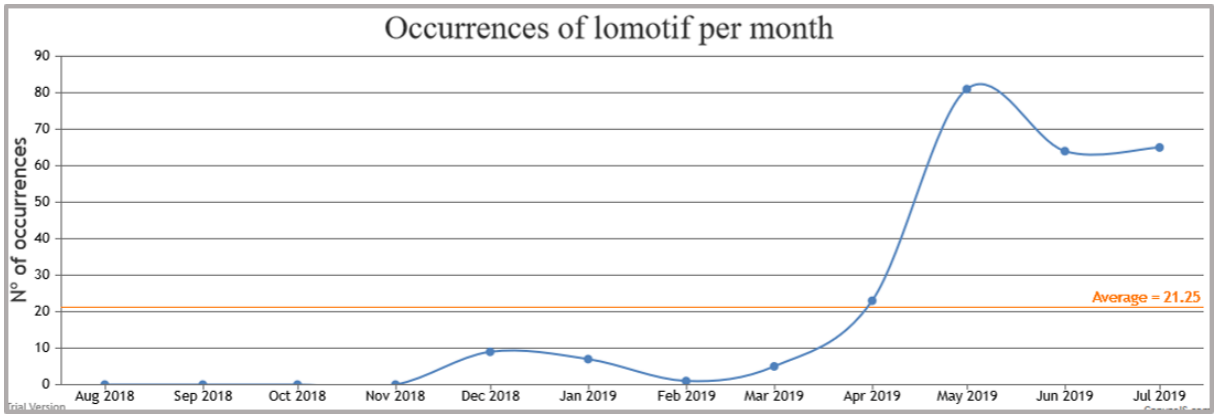

Figura 5.4: *Website - Analyse words («Occurrences of the word per month»)*

Porém será expectável que na maioria dos casos o utilizador ainda não tenha conhecimento das palavras consideradas emergentes, por esse motivo imediatamente a baixo encontra-se uma tabela (Figura 5.5) onde todas as palavras emergentes constarão e o utilizador poderá proceder à analise das palavras diretamente clicando nela. As palavras contidas na tabela encontram-se organizadas de forma ordenada, da palavra com maior número de ocorrências para a com menor número de ocorrências.

O significado de algumas palavras poderá não ser intuitivo, por esse motivo proceder-se-á a dois tipos de análise, na Figura 5.6, verifica-se que serão apresentados dez exemplos de publicações feitas por utilizadores que contenham a palavra emergente escolhida para análise. Por fim, como se pode verificar na Figura 5.7, será ainda aplicada uma análise com a utilização de *embeddings*, como foi aprofundado em 2.3, nesta análise os resultados obtidos serão tanto relativos ao modelo do *Wikipédia* como do *Common Crawl.*

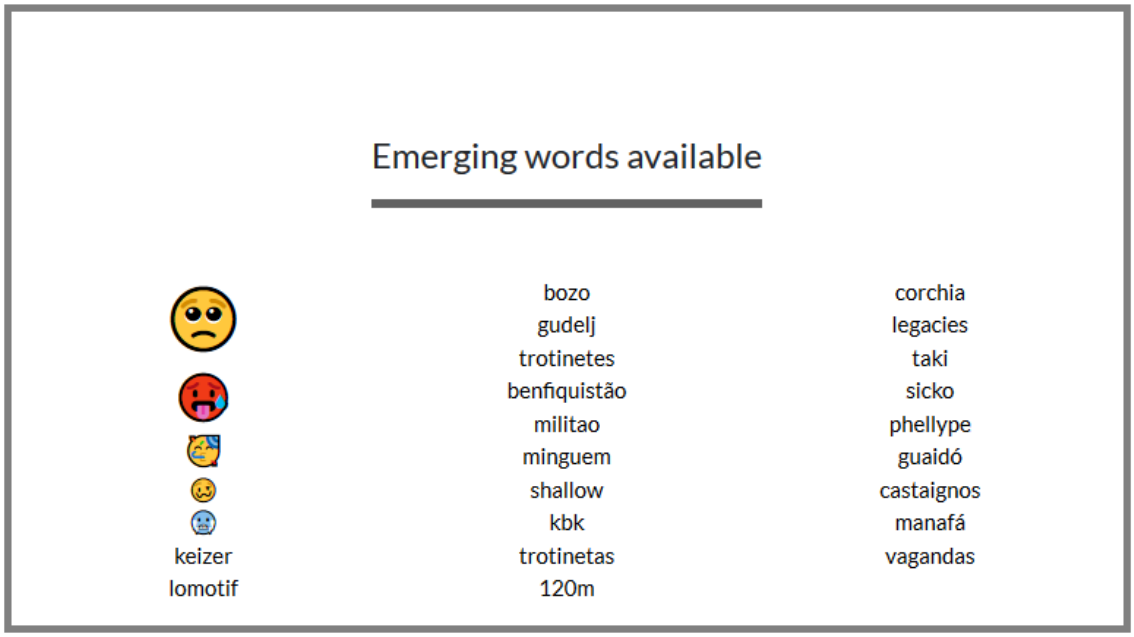

Figura 5.5: *Website - Analyse words («Emerging words available»)*

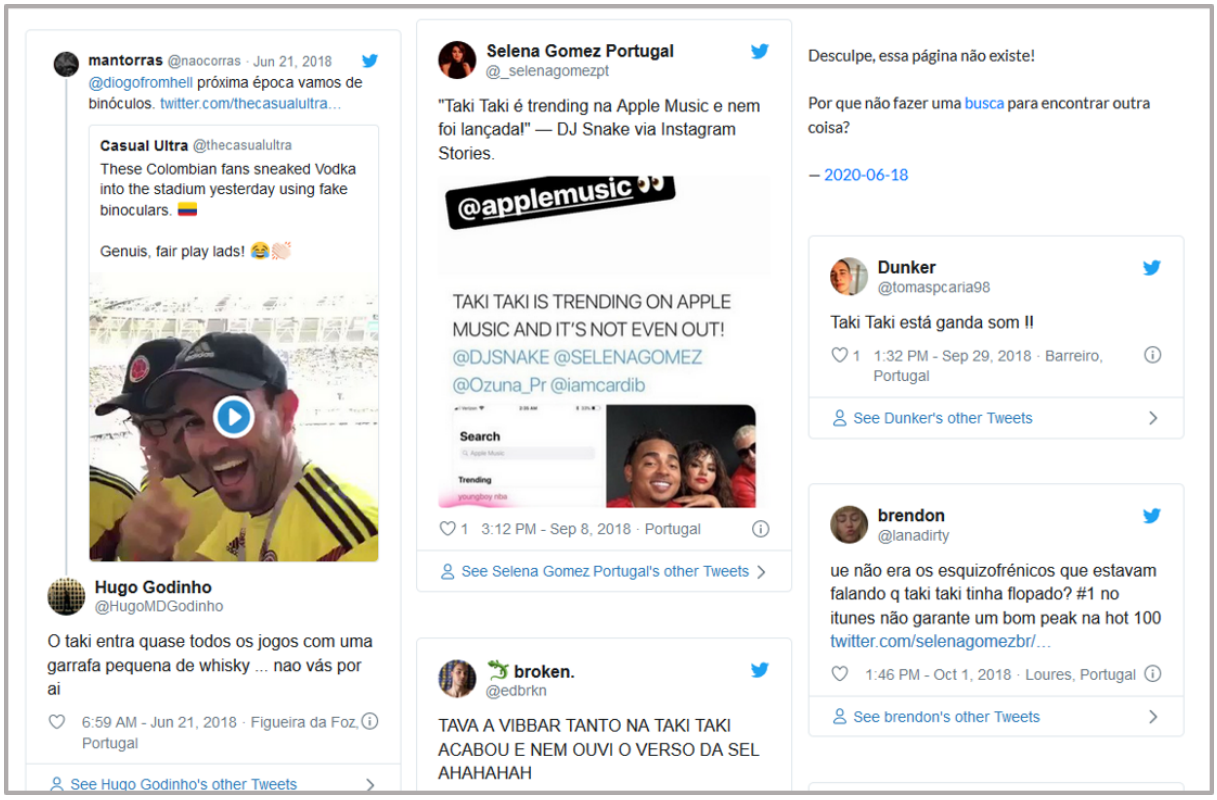

Figura 5.6: *Website - Analyse words («Examples of Tweets»)*

Figura 5.7: *Website - Analyse words («fastText analysis»)*

6 Conclusões e trabalho futuro

Este estudo aborda o conceito de emergência léxical e descreve métodos estatísticos tendo em conta os dados disponíveis para a identificação de palavras emergentes durante um período de tempo. Este método permitiu obter 26 formações léxicas emergentes na língua portuguesa baseado num corpus de 99 mil milhões de palavras obtidas através de publicações no Twitter, armazenados entre 2018 e julho de 2019. As formações identificadas foram inspecionadas de diferentes perspetivas de forma a compreender o que motiva a emergência léxical.

O trabalho demonstrou que é possível identificar padrões na emergência das palavras da língua portuguesa através da metodologia aplicada. É preciso ter em consideração que a análise apresentada não identificou um número muito vasto de palavras emergentes, isto porque a API do Twitter limita na obtenção dos dados, impedido que sejam armazenadas todas as publicações realizadas. Ainda assim, a principal razão para não se ter obtido um número maior de palavras emergentes está relacionado com o volume do *corpus*, ainda que possua uma dimensão elevada, ainda assim não é elevada o suficiente para compreender todos os casos de emergência léxical no espaço de tempo analisado. À medida que mais dados vão ser disponibilizados, será possível conduzir estudos mais completos e mais detalhados no campo da emergência léxical. Ainda, um parâmetro que poderá possibilitar a deteção de um maior número de palavras é a frequência mínima exigida, baixando o valor necessário da frequência mínima das palavras será possível obter palavras mais diversificadas.

Apesar destas limitações, a análise permitiu identificar um número de palavras emergentes com uma variedade de diferentes tópicos, o que permitiu inúmeras observações interessantes sobre natureza da emergência destas palavras. Algo de grande interesse foi a identificação de palavras que já tinham sido introduzidas há vários anos atrás e que ainda assim emergiram. Outro caso identificado foi a derivação de palavras já existentes, estas demonstraram ser interessantes por em vários casos trazer uma conotação positiva ou negativa à qual está a derivar, ainda que a palavra identificada possivelmente venha a ficar esquecida. Porém, este trabalho apresenta evidências de que a maioria das palavras já terá surgido anteriormente, por esse motivo a identificação dos fatores que terão servido com «*trigger*» para e emergência das palavras será importante numa investigação futura.

É igualmente importante ter em conta que o método introduzido não identifica necessariamente as formações lexicais que darão origem a palavras correntes do dia a dia. Será

interessante dar continuidade ao estudo, de modo a que o espaço temporal de análise e o volume de dados seja muito superior para ser possível verificar as palavras que se tornaram habituais no quotidiano os portugueses.

Para além destes resultados descritivos e teóricos, este estudo mostrou também como a adoção de uma abordagem baseada num *corpus* com um grande conjunto de dados pode permitir investigações em novas áreas, especialmente relativamente à variação das palavras, que requer grandes quantidades de dados linguísticos. Existem certamente dificuldades inerentes ao tratamento de dados que foram recolhidos de forma automática, mas como foi demonstrado estes permitem sempre serem melhorados e trabalhados conseguindo trabalhar a qualidade dos dados, proporcionando assim uma visão mais realista do dinamismo da língua portuguesa.

Foi também introduzido um método para mapear padrões comuns na inovação léxical ao longo do período de tempo e com um *corpora* de dados geolocalizados. Com base nos mapas das 26 palavras emergentes, foram identificados duas principais regiões para a inovação léxical, a primeira e principal, na cidade de Lisboa e igualmente com grande foco, a cidade do Porto.

A análise realizada ao mapeamento das palavras permitiu identificar quatro contributos principais para a compreensão da mudança linguística:

- É possível observar padrões regionais de propagação de palavras, mesmo que não seja possível afirmar que as palavras ocorreram pela primeira vez nas redes sociais.
- As palavras tendem a seguir um caminho consistente, como é possível ver na Figura 5.1 que Lisboa e o Porto serão as cidades com o habitual centro de propagação, registando uma propagação para as cidades circundantes, com uma diminuição natural do número de ocorrências das palavras.
- A densidade populacional tem um papel importante na disseminação das palavras e parece ser mais fundamental do que as questões culturais ou religiosas.
- O português brasileiro será uma das principais fontes de inovação lexical.

O grau em que estes resultados podem ser generalizados em diferentes registos, dialetos, eras e línguas, bem como em diferentes níveis de análise linguística, é uma questão em aberto. O Twitter é apenas um canal de expressão dos vários e diferentes que existem, o que não representa uma maioria na percentagem do vocabulário produzido no dia a dia pelas pessoas e presumivelmente não terão ocorridas as palavras pela primeira vez nesta rede social. Por consequentemente, o *corpus* do Twitter só pode refletir parcialmente os padrões de inovação lexical na língua, especialmente atualmente onde existem variadíssimos canais de comunicação. No entanto, dado que estas palavras são utilizadas no discurso quotidianos, e dado que o Twitter é disponibiliza variedade enorme de vocabulário que não

é geograficamente limitada, é possível afirmar que os resultados alcançados podem, de facto, refletir a propagação geral das palavras identificadas.

Finalmente, o trabalho proporcionou uma metodologia para futuras investigações sobre a análise da emergência de palavras e o mapeamento da sua propagação, mostrando como a difusão das palavras emergentes podem ser medidas e cartografadas. Embora este método tenha sido utilizado para o estudo da palavras emergentes, o mesmo pode ser aplicado para analisar o mapeamento da utilização de qualquer forma linguística ao longo do tempo. Em termos mais gerais, este trabalho ilustrou como a análise quantitativa de uma *corpora* de grandes dimensões de comunicação natural permite novas questões de investigação. Não há duvida de que, à medida que mais dados se encontrem disponíveis para serem analisados a nossa compreensão da variação e das mudanças linguísticas continuará a ser enriquecida.

Como seguimento do presente trabalho, pretende-se utilizar outras fontes de informação, tais como jornais, blogues e outras redes sociais como meios para traçar um caminho completo das palavras emergentes, contribuindo assim para uma melhor compreensão da evolução do vocabulário numa determinada língua. Pretende-se caracterizar melhor cada palavra, fornecendo uma análise alargada ao longo do tempo, mapeando a sua utilização, adoção pelas comunidades, e traçando o seu caminho de propagação. Ainda que não existam passos estruturados para abordar esta questão, a maioria das fontes de dados fornecem geolocalização e informação temporal, o que constitui uma vantagem relevante para traçar percursos fiáveis de propagação geo-temporal de uma palavra num futuro próximo.

Bibliografia

[1] J. P. Pereira, “Era uma vez o twitter em portugal”, *Público*, C, vol. 77, nº 3, pp. 95–106, 2016.

[2] J. Harmer, “The practice of english language teaching”, *SERBIULA (sistema Librum 2.0)*, jan. de 2001.

[3] M. L. Murphy, “Theories of lexical semantics by dirk geeraerts”, *Journal of Linguistics*, vol. 47, pp. 231–236, jan. de 2011. DOI: 10.2307/41261748.

[4] S. Grondelaers, D. Geeraerts e D. Speelman, “Lexical variation and change”, em, sér. Lecture Notes in Computer Science, 2007, pp. 988–1011.

[5] S. Shahid, “Teaching of english an introduction”, *Majeed Book Depot Urdu Bazar Lahore*, 2002.

[6] R. Blood. (jan. de 2000). Weblogs: A history and perspective, endereço: http://www.rebeccablood.net/essays/weblog%5C_history.html.

[7] M. Dyrud, R. Worley e M. Flatley, “Blogging for enhanced teaching and learning”, *Business Communication Quarterly*, vol. 68, mar. de 2005. DOI: 10.1177/108056990506800111.

[8] S. Kajder, G. Bull e E. Van Noy, “A space for "writing without writing."”, em, vol. 31, 2004, pp. 32–35.

[9] D. Mutum e Q. Wang, “Consumer generated advertising in blogs.”, em. jan. de 2010, pp. 248–261. DOI: 10.4018/978-1-60566-792-8.ch013.

[10] J. Grieve, “Dialect variation”, em *The Cambridge Handbook of English Corpus Linguistics*, D. Biber e R. Reppen, eds., sér. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2015, pp. 362–380. DOI: 10.1017/CBO9781139764377.021.

[11] J. Clement, *Most used social media 2020*, nov. de 2020. endereço: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[12] M. Porter, *Estrategia Competitiva*. ELSEVIER EDITORA, 2004, ISBN: 9788535215267. endereço: https://books.google.pt/books?id=SxvCKIh706gC.

[13] S. de Oliveira, *Tratado de metodologia científicafica: Projetos de pesquisas, TGI, TCC, monografias, dissertações e teses*. Pioneira Thomson Learning, 2001. endereço: https://books.google.pt/books?id=7RFuRAAACAAJ.

[14] D. J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton, NJ, USA: Princeton University Press, 2003, ISBN: 0691117047.

[15] G. Nicolis, I. Prigogine, W. H. Freeman e Company, *Exploring Complexity: An Introduction*. W.H. Freeman, 1989, ISBN: 9780716718598. endereço: https://books.google.pt/books?id=blt5QgAACAAJ.

[16] P. Johnson e J. Duberley, "Reflexivity in management research*", *Journal of Management Studies*, vol. 40, nº 5, pp. 1279–1303, 2003. DOI: 10.1111/1467-6486.00380. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-6486.00380. endereço: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-6486.00380.

[17] W. Ogburn e M. Nimkoff, *Technology and the changing family*. Westport, CT: Greenwood Press, 1955.

[18] N. B. Ellison, C. Steinfield e C. Lampe, "The benefits of facebook "friends:" social capital and college students' use of online social network sites", *Journal of Computer-Mediated Communication*, vol. 12, nº 4, pp. 1143–1168, jul. de 2007, ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2007.00367.x. eprint: http://oup.prod.sis.lan/jcmc/article-pdf/12/4/1143/22316419/jjcmcom1143.pdf. endereço: https://doi.org/10.1111/j.1083-6101.2007.00367.x.

[19] L. Miranda, C. Morais, P. Alves e P. Dias, *Redes sociais na aprendizagem : Motivação e utilização dos estudantes do ensino superior*. Westport, CT: Whitebooks, 2014, ISBN: 978-989-8765-01-7. endereço: http://hdl.handle.net/1822/33629.

[20] M. Macedo, "A teoria dos usos e gratificações nas entidades do terceiro setor no brasil.", *Razón y palabra, ISSN 1605-4806, Nº. 70, 2009*, jan. de 2009.

[21] P. Simões, "O twitter em contexto académico/profissional: Estudo de caso", 2013.

[22] M. Berry, *Survey of Text Mining*. Springer-Verlag New York, 2004, ISBN: 978-1-4757-4305-0. DOI: 10.1007/978-1-4757-4305-0.

[23] R. Elmasri e S. Navathe, *Fundamentals of Database Systems, 3rd Edition*. jan. de 2000, ISBN: 978-0-8053-1755-8.

[24] S. Bolasco, A. Canzonetti, F. M. Capo, F. della Ratta-Rinaldi e B. K. Singh, "Understanding text mining: A pragmatic approach", em *Knowledge Mining*, S. Sirmakessis, ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 31–50, ISBN: 978-3-540-32394-5.

[25] A. Mohammad, T. Alwadan e O. Almomani, "Arabic text categorization using support vector machine, naïve bayes and neural network", *GSTF Journal on Computing (JoC)*, vol. 5, set. de 2016. DOI: 10.7603/s40601-016-0016-9.

[26] S. Jusoh e H. Alfawareh, "Techniq techntechn techniques, applications and challenging issue in text mining", *IJCSI International Journal of Computer Science Issues*, vol. 9, pp. 431–436, nov. de 2012.

[27] M. Porter, “An algorithm for suffix stripping”, *Program*, vol. 14, nº 3, pp. 130–137, 1980. DOI: 10.1108/eb046814. eprint: http://www.emeraldinsight.com/doi/pdf/10.1108/eb046814. endereço: http://www.emeraldinsight.com/doi/abs/10.1108/eb046814.

[28] S. Liritano e M. Ruffolo, “Managing the knowledge contained in electronic documents: A clustering method for text mining”, *IEEE*, pp. 454–458, 2001.

[29] Y. Goldberg, *Neural Network Methods for Natural Language Processing*, sér. Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool, 2017, vol. 37, ISBN: 978-1-62705-298-6. DOI: 10.2200/S00762ED1V01Y201703HLT037.

[30] Y. Bengio, R. Ducharme e P. Vincent, “A neural probabilistic language model”, em *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich e V. Tresp, eds., MIT Press, 2001, pp. 932–938. endereço: http://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf.

[31] C. Cothenet, *The general ideas of word embeddings*, mai. de 2020. endereço: https://towardsdatascience.com/short-technical-information-about-word2vec-glove-and-fasttext-d38e4f529ca8.

[32] J. Pennington, R. Socher e C. D. Manning, “Glove: Global vectors for word representation.”, em *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[33] J. Grieve, A. Nini e D. Guo, “Analyzing lexical emergence in modern american english online”, English, *English Language and Linguistics*, vol. 21, nº 1, pp. 99–127, 2017. DOI: 10.1017/S1360674316000113.

[34] L. Brinton e E. Traugott, *Lexicalization and Language Change*, sér. Research Surveys in Linguistics. Cambridge University Press, 2005, ISBN: 9780521540636. endereço: https://books.google.pt/books?id=uRHvnQEACAAJ.

[35] P. Hopper e E. Traugott, *Grammaticalization*, sér. Cambridge Textbooks in Linguistics. Cambridge University Press, 2003, ISBN: 9781139935463. endereço: https://books.google.pt/books?id=EWZuBAAAQBAJ.

[36] M. Naya, *On the history of downright. English Language and Linguistics*. Cambridge University Press, 2008, pp. 267–87.

[37] L. Bauer, P. Bauer, B. Laurie, S. Anderson, J. Bresnan, B. Comrie, W. Dressler e C. Ewen, *English Word-Formation*, sér. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983, ISBN: 9780521284929. endereço: https://books.google.pt/books?id=yGfUHs6FCvIC.

[38] M. Krug, *Emerging English Modals: A Corpus-based Study of Grammaticalization*, sér. Topics in English linguistics. Mouton de Gruyter, 2000, ISBN: 9783110166545. endereço: https://books.google.pt/books?id=TavYdjPiSDkC.

[39] T. Nevalainen e H. Raumolin-Brunberg, *Historical Sociolinguistics*, sér. Longman Linguistics Library. Taylor & Francis, 2014, ISBN: 9781317882176. endereço: https://books.google.pt/books?id=BWmuBAAAQBAJ.

[40] S. Gries e M. Hilpert, "The identification of stages in diachronic data: Variability-based neighbour clustering", *Corpora*, vol. 3, pp. 59–81, mai. de 2008. DOI: 10.3366/E1749503208000075.

[41] D. Geeraerts e H. Cuyckens, "Introducing cognitive linguistics", *The Oxford Handbook of Cognitive Linguistics*, jan. de 2012. DOI: 10.1093/oxfordhb/9780199738632.013.0001.

[42] P. Siemund, *Varieties of English: A Typological Approach*. Cambridge University Press, 2013, ISBN: 9780521764964. endereço: https://books.google.pt/books?id=ZuBpRxizN3QC.

[43] T. Nevalainen e H. Raumolin-Brunberg, *Historical Sociolinguistics: Language Change in Tudor and Stuart England*, sér. Longman linguistics library. Longman, 2003, ISBN: 9780582319943. endereço: https://books.google.pt/books?id=PLpZAAAAMAAJ.

[44] J. Grieve, A. Nini e D. Guo, "Mapping lexical innovation on american social media", English, *SAGE*, vol. 46, nº 4, pp. 293–319, 2018. DOI: 10.1177/0075424218793191.

[45] S. Grondelaers, D. Speelman e D. Geeraerts, "Lexical variation and change", English, 2012. DOI: 10.1093/oxfordhb/9780199738632.013.0037.

[46] A. Radford, M. Atkinson, D. Britain, H. Clahsen e A. Spencer, "Lexical variation and change", em *Linguistics: An Introduction*, 2ª ed. Cambridge University Press, 2009, pp. 224–241. DOI: 10.1017/CBO9780511841613.021.

[47] O. A. HenryGrieve, "Emerging trends in the language of social media in nigeria", English, 2018.

[48] L. Zhang, J. Zhao e K. Xu, "Who creates trends in online social media: the crowd or opinion leaders?", *Journal of Computer-Mediated Communication*, vol. 21, nº 1, pp. 1–16, dez. de 2015, ISSN: 1083-6101. DOI: 10.1111/jcc4.12145. eprint: https://academic.oup.com/jcmc/article-pdf/21/1/1/22316317/jjcmcom0001.pdf. endereço: https://doi.org/10.1111/jcc4.12145.

[49] J. Lehmann, M. Lalmas, E. Yom-Tov e G. Dupret, "Model of user engagement", vol. 7379, jul. de 2012. DOI: 10.1007/978-3-642-31454-4_14.

[50] M. JafariAsbagh, E. Ferrara, O. Varol, F. Menczer e A. Flammini, "Clustering memes in social media streams", *Social Network Analysis and Mining*, vol. 4, pp. 1–13, 2014.

[51] D. Romero, W. Galuba, S. Asur e B. Huberman, "Influence and passivity in social media", *Inf. Syst. J.*, pp. 1–9, jan. de 2011. DOI: 10.1145/1963192.1963250.

[52] C. Bauckhage, K. Kersting, F. Hoppe e C. Thurau, "Archetypal analysis as an autoencoder", em *Workshop New Challenges in Neural Computation*, out. de 2015, p. 8.

[53] M. Finatto, O. VALE e É. Laporte, “Reconhecimento do vocabulário de jornais populares brasileiros por um dicionário computacional de acesso livre”, pt, *Alfa: Revista de Linguística (São José do Rio Preto))*, vol. 63, pp. 63–80, mar. de 2019, ISSN: 1981-5794. endereço: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1981-57942019000100063&nrm=iso.

[54] I. Soares, “Palhaço bozo diz a bolsonaro que esquerda o elogia quando compara os dois”, *Correio Braziliense*, fev. de 2020.

Apêndice

A

Nesta secção encontram-se os mapeamentos realizados para as restantes palavras emergentes que não se encontram analisadas ao pormenor no capítulo 5.

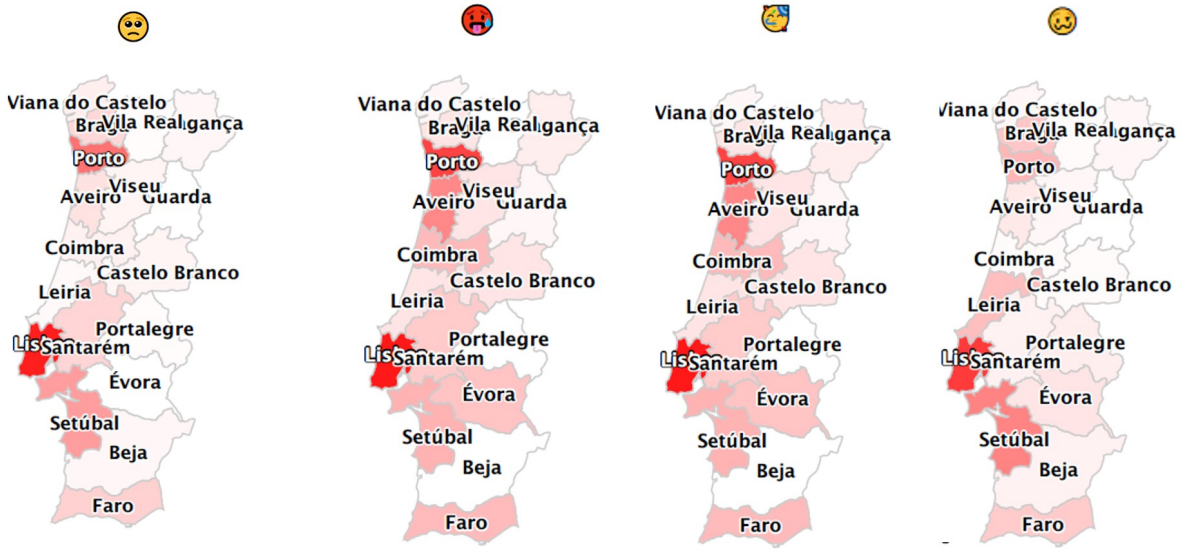

Figura A.1: *Frequência por distrito de emoticons: «Emoji de rosto implorando», «Emoji de rosto fervendo», «Emoji com rosto de festa e chapéu de festa» e «Emoji com rosto embriagado»*

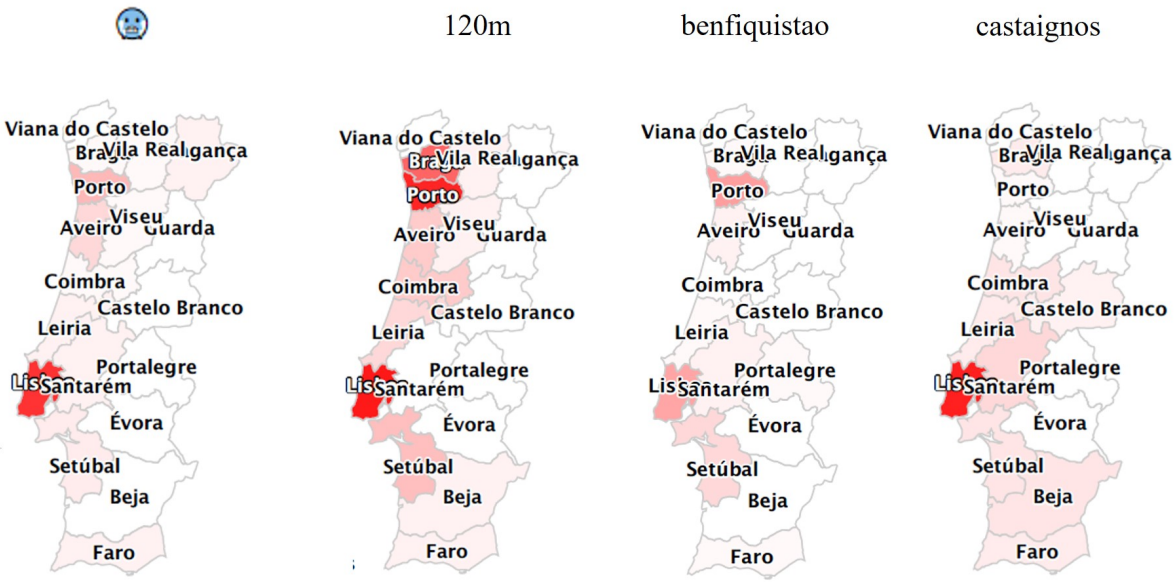

Figura A.2: *Frequência por distrito das palavras: «Emoji com rosto gelado»,* «120m», «benfiquistao» e «*castaignos*»

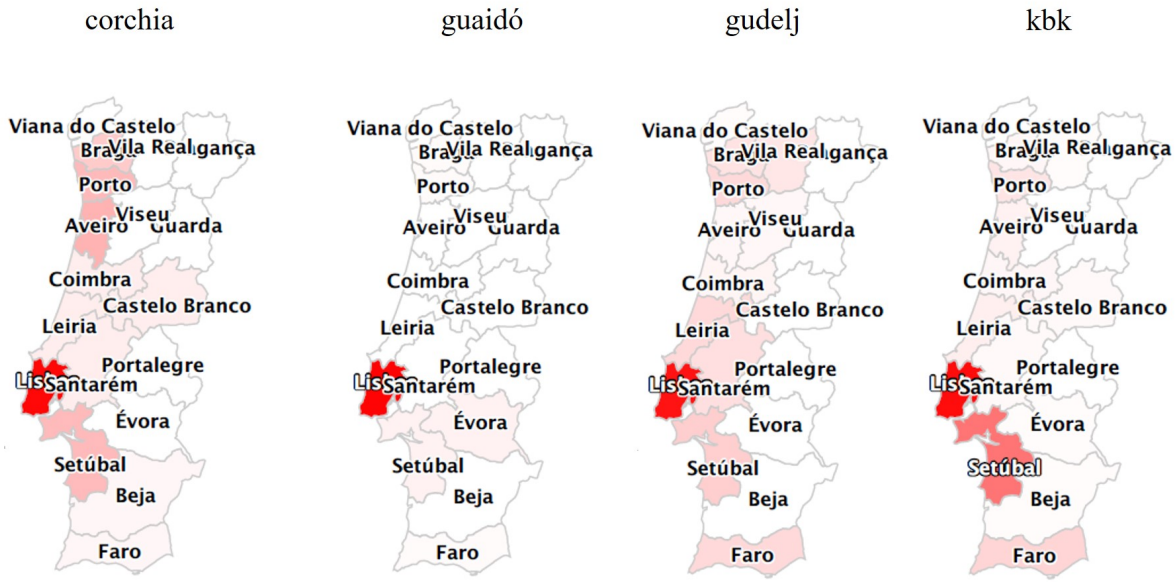

Figura A.3: *Frequência por distrito das palavras: «corchia»,* «guaidó», «gudelj» e «*kbk*»

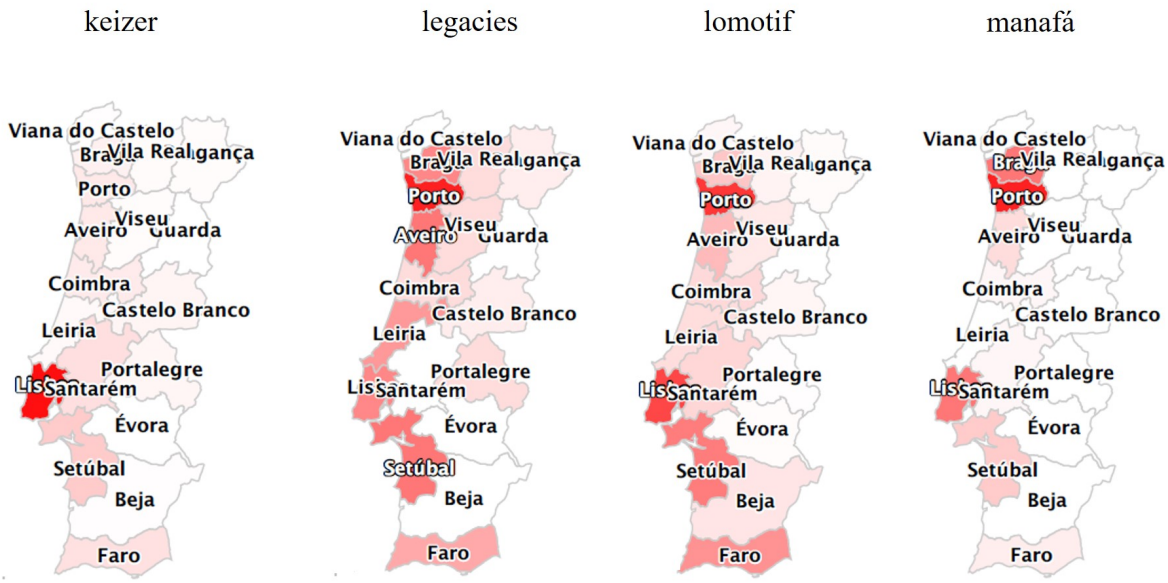

Figura A.4: *Frequência por distrito das palavras: «keizer»,* «legacies», «lomotif» e «*manafá*»

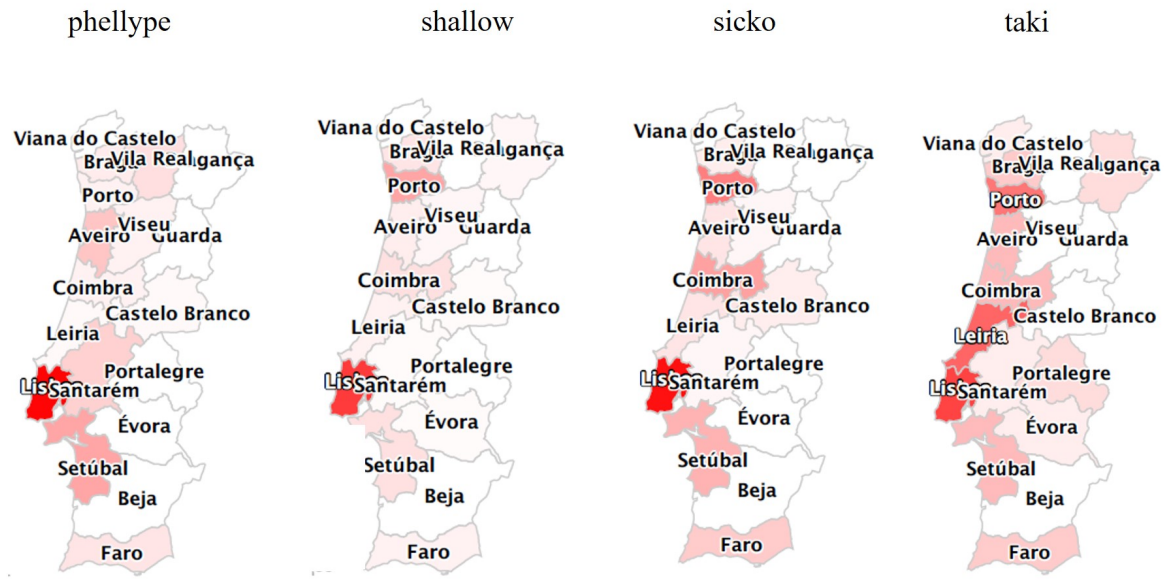

Figura A.5: *Frequência por distrito das palavras: «phellype»,* «shallow», «sicko» e «*taki*»

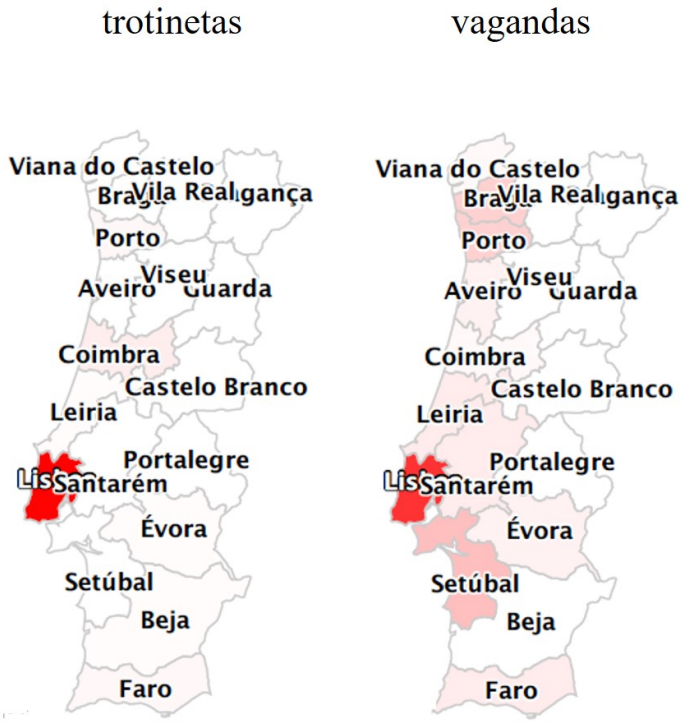

Figura A.6: *Frequência por distrito das palavras: «trotinetas» e* «vagandas»