Jorge Alexandre Fonseca Almeida

ANÁLISE E VALIDAÇÃO DE INFORMAÇÃO NA WEB PARA MITIGAÇÃO DE EVENTOS DE EMERGÊNCIA

Trabalho de Projeto de Mestrado em Tecnologias de Informação Geográfica, área de especialização em Ciências e Tecnologias de Informação Geográfica, orientada por Professor Doutor José Paulo Elvas Duarte de Almeida, Professor Doutor Alberto Jorge Lebre Cardoso e Mestre Joaquim António Saraiva Patriarca e apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

2019

· U C ·

Jorge Alexandre Fonseca Almeida

ANÁLISE E VALIDAÇÃO DE INFORMAÇÃO NA WEB PARA MITIGAÇÃO DE EVENTOS DE EMERGÊNCIA

Trabalho de Projeto de Mestrado em Tecnologias de Informação Geográfica, área de especialização em Ciências e Tecnologias de Informação Geográfica, orientada por Professor Doutor José Paulo Elvas Duarte de Almeida, Professor Doutor Alberto Jorge Lebre Cardoso e Mestre Joaquim António Saraiva Patriarca e apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

2019

UNIVERSIDADE DE COIMBRA

"Uma longa viagem começa com um único passo"

Lau Zi

Agradecimentos

Agradeço aos meus pais e à minha irmã, por fazerem parte desta jornada, por terem estado sempre presentes durante todos estes anos e por todo o esforço que fizeram por mim. Por todo o apoio, por todos os momentos, e principalmente pela ajuda, pelo significado que deram à palavra família, na sua essência máxima. Sem vocês, todo este percurso não teria tido o mesmo valor, e por isso tomo-vos como um exemplo.

Aos meus amigos de sempre, um agradecimento profundo por terem estado comigo durante todo o meu percurso, pelo apoio e pela confiança, mesmo que por vezes a distância fosse um obstáculo.

Aos amigos que esta cidade me presenteou, agradeço pelo companheirismo, também pelos ótimos anos que passámos juntos, e principalmente pelas alegrias.

Por fim, agradeço aos meus orientadores, pelo incentivo, pela disponibilidade e pelo acompanhamento ao longo deste longo e árduo ano e pela oportunidade de trabalhar neste projeto.

Nº do aluno: 2013168808

Nome: Jorge Alexandre Fonseca Almeida

Título do Trabalho de Projeto:

Análise e validação de informação na Web para mitigação de eventos de emergência

Palavras-Chave:

- Mitigação de eventos de emergência
- Proteção Civil
- Redes Sociais
- Validação de informação
- Web

Resumo

Com o surgimento da WEB 2.0, a quantidade de informação disponível online aumentou, surgindo consequentemente novos serviços/plataformas, como Redes Sociais, blogs, aumentando o seu potencial de utilização da web em vários domínios. Conforme o aumento da quantidade de informação, também a quantidade de dados por validar se expandiu, de vários tipos e formatos, gerando novos problemas associados à utilização destes. Desta maneira, para conseguirmos criar uma aplicação de mitigação de eventos de catástrofe com base em dados disponíveis online (redes sociais, noticias online) que ajude uma entidade como a Proteção Civil na tomada de decisão em casos de eventos de catástrofe, é necessário validar e analisar todas a informações recolhidas para a sua devida utilização e conseguir colocar toda esta informação catalogada numa base de dados.

Para realizar esta operação de validação destes dados recolhidos online, é necessário utilizar uma metodologia que consiga identificar publicações que sejam relacionadas ou não com o evento de catástrofe a ser estudado, existindo várias possibilidades, sendo a metodologia de Laylavi (2016), que desenvolve uma metodologia de recolha, pré-processamento e avaliação de relação ao evento de publicações na rede social Twitter, uma excelente base metodológica para o desenvolvimento deste trabalho.

No que respeita a este trabalho de projeto, a recolha de dados e o pré-processamento são baseados em laylavi (2016), no entanto, a metodologia para a validação da informação, será com base na filtragem com palavras de interesse de acordo com um evento (incêndios) e Furacão Leslie para a macro validação temática e a utilização da junção de palavras de interesse com um topónimo para micro validação temática.

Abstract

With the emergence of WEB 2.0, the amount of information available online has increased, resulting in new services / platforms, such as Social Networks, blogs, increasing their potential for web use in various domains. As the amount of information increases, the amount of data to be validated has expanded, of various types and formats, creating new problems associated with their use. Thus, in order to be able to create a disaster event mitigation application based on data available online (social networks, online news) that helps an entity such as Civil Protection in decision making in case of disaster events, it is necessary to validate and analyze all information collected for its proper use and be able to put all this cataloged information into a database.

To perform this validation operation of this data collected online, it is necessary to use a methodology that can identify publications that are related or not to the disaster event being studied, and there are several possibilities, being the methodology of Laylavi (2016), which develops a methodology for collecting, preprocessing and evaluating the event of publications on the social network Twitter, an excellent methodological basis for the development of this work.

For this project work, data collection and preprocessing are based on laylavi (2016), however, the methodology for information validation will be based on filtering with words of interest according to a event (fires) and Hurricane Leslie for thematic macro validation and the use of joining words of interest with a toponym for thematic micro validation.

Índice

1. Introdução

1.1. Enquadramento geral e motivação

A quantidade de informação disponível na web cresce a uma velocidade enorme a cada dia que passa, o que faz com que estes dados sejam de extrema importância para aplicações que requeiram uma grande quantidade de dados. Dito isto, a quantidade de informação disponível na web tem um grande potencial para a coleção e armazenamento de informação.

Para criar uma aplicação de mitigação de eventos de catástrofe (acedendo a publicações de usuários online, ou de jornais) é necessária muita desta informação para se conseguir mitigar um evento de catástrofe, o que gera a possibilidade de se conseguir tomar uma decisão em tempo real, sendo uma mais valia para uma entidade como a Proteção Civil ou os Bombeiros. No entanto nem toda a informação tem relevância para um evento em questão, o que torna muita desta informação irrelevante, daí a necessidade de haver um processo de validação da informação. Para se conseguir obter este resultado existem algumas iterações a realizar primeiro: a recolha dos dados para uma base de dados, o pré-processamento (limpeza dos dados: eliminar carateres especiais, emojis, espaços, maiúsculas, links) para que se torne ainda mais rápido e eficaz no seu processamento, só depois se poderá realizar a validação destes dados, sendo considerados relevantes ou não em relação a um evento de catástrofe.

Em caso de desastres extremos, como inundações, terremotos incêndios florestais / urbanos, planos rápidos de segurança e ações de mitigação são necessários. Informações relevantes sobre um desastre ou acidente são a localização geoespacial, as condições ambientais envolventes da área e os riscos associados a áreas ou infra-estruturas que podem aglomerar um grande número de pessoas. (Fontes et al., 2017).

Uma das principais fontes de informação na web são as redes sociais, onde são partilhados vários tipos de informação. Apesar do grande volume de dados, estes têm algumas limitações, como por exemplo o seu acesso reduzido por parte das API's das próprias redes sociais, isto se se quiser obter estes dados de forma gratuita, em que basicamente o número de acessos aos dados disponíveis publicamente é limitado por um determinado tempo, sendo necessário ativar planos para ser possível obter um maior acesso aos dados. Apesar destas limitações, um dos principais

problemas é o nível qualitativo da informação, sendo que maior parte destes dados são disponibilizados pelos próprios utilizadores nas redes sociais.

O enquadramento deste trabalho de projeto insere-se num projeto chamado GeoTimeline desenvolvido no Instituto de Engenharia de Sistemas e Computadores de Coimbra (INESCC) que consiste no desenvolvimento de uma aplicação para a mitigação de eventos de emergência (Resposta mais rápida e eficaz, tendo como principal fonte a informação disponível online). Estas fontes de informação consistem fundamentalmente de Redes Sociais/Jornais online, que possuem uma grande potencialidade como fonte de informação de forma gratuita apesar do acesso limitado e em tempo real.

No entanto existe um enorme volume de informação que é necessário tratar e validar a informação recolhida, passando por dois processos importantes: o pré-processamento e o relacionamento ou não com o fenómeno pretendido (Validação e a classificação de texto).

Concluindo, a utilização desta metodologia para uma aplicação de ajuda na tomada de decisão e mitigação de eventos de catástrofe torna-se numa grande motivação para o desenvolvimento da mesma, já que relativamente a eventos de catástrofe de incêndios recentes em Portugal, a necessidade de uma aplicação que consiga ajudar as entidades a mitigar este tipo de eventos é urgente.

1.2. Objetivos do Trabalho

O principal objetivo deste trabalho é a conceção e implementação de uma metodologia que consiga fazer a validação da enorme quantidade de informação disponível na WEB tendo em conta o critério final, que é a mitigação de eventos de emergência.

Este objetivo considera que esta metodologia que possa vir a ser útil a uma entidade como a proteção civil, na deteção de um evento de catástrofe em concreto que esteja a acontecer em tempo real, ou para servir como um processo de análise de eventos passados.

Definir uma metodologia, se possível em tempo real, para a deteção de contributos que tragam valor de informação acrescentada para um evento, que possa ser utilizado por entidades de proteção civil para cada evento em específico (Incêndios, furacões, ou outros tipos de catástrofe).

Desenvolver uma metodologia que consiga realizar a macro validação (relacionado com um evento) e micro validação (relacionado com um evento e um topónimo).

Identificação e extração de informação útil sobre fenómenos extremos, e determinar melhor metodologia para um desempenho eficaz (exatidão na identificação de contribuições úteis) e eficiente (tempo) na classificação automática de texto.

1.2.1. Estrutura do trabalho de projeto

Este trabalho de projeto é composto por cinco capítulos.

O primeiro capítulo refere-se à introdução, enquadramento geral e à motivação, onde são apresentados os principais objetivos previstos com este trabalho de projeto.

No segundo capítulo é apresentada a revisão bibliográfica, onde é apresentada a importância da Web e a sua grande potencialidade como fonte de informação. É também apresentado o fundamento deste trabalho de projeto, a metodologia de Laylavi (2016), assim como as suas características e importância para o desenvolvimento deste trabalho.

O Terceiro capítulo é o capítulo onde se descreve a componente prática deste trabalho, que diz respeito à apresentação da metodologia e todos os processos e métodos usados assim como a sua descrição mais aprofundada para desenvolver os testes realizados no capítulo seguinte.

No capítulo quatro apresenta-se os resultados obtidos com a metodologia apresentada no terceiro capítulo, que vai de encontro a cinco testes realizados.

Finalmente, no capítulo cinco, são expostas as considerações finais deste trabalho de projeto e os possíveis desenvolvimentos futuros, no que diz respeito aos resultados apresentados.

2. Enquadramento do trabalho

Muitas situações de catástrofe requerem a utilização de informação confiável e atualizada sobre os eventos, para mitigação em caso de acidente em grande escala, qualquer situação de emergência ou monitoramento de segurança.

A emergência da Web 2.0 disponibilizou uma enorme quantidade de informações, de uma ampla variedade de fontes, que podem ser potencialmente usadas pelas autoridades, no entanto recolher estes dados diretamente da Web e colocá-los numa base de dados não é o suficiente, se esta informação não for validada respetivamente ao evento.

2.1. Importância da WEB como fonte de dados

Os dados apresentados na tabela 1, indicam o uso da internet a nível mundial no mês de junho de 2019, que separa o evento por categorias, ou seja, Regiões Mundiais, e que se subdivide em População, em Percentagem de População Mundial, em Usuários de Internet, em Percentagem do Índice de Utilização, em Crescimento 2000-2019, e em Percentagem de Internet Mundial.

Tabela 1 - Utilização Mundial de Internet e Estatística Populacional em junho de 2019 (Internet World Stats, 2019)

WORLD INTERNET USAGE AND POPULATION STATISTICS JUNE, 2019 - Updated						
World Regions	Population (2019 Est.)	Population % of World	Internet Users 30 June 2019	Penetration Rate (% Pop.)	Growth 2000-2019	Internet World %
Africa	1,320,038,716	17.1 %	521,614,944	39.5 %	11,454 %	11.6 %
Asia	4,241,972,790	55.0 %	2,275,469,859	53.6 %	1,891 %	50.5 %
Europe	829,173,007	10.7 %	727,559,682	87.7 %	592 %	16.1 %
Latin America / Caribbean	658,345,826	8.5 %	453,702,292	68.9 %	2,411 %	10.1 %
Middle East	258,356,867	3.3 %	175,502,589	67.9 %	5,243 %	3.9 %
North America	366,496,802	4.7 %	327,568,628	89.4 %	203 %	7.3 %
Oceania / Australia	41,839,201	0.5 %	28,636,278	68.4 %	276 %	0.6 %
WORLD TOTAL	7,716,223,209	100.0 %	4,510,054,272	58.4 %	1,149 %	100.0 %

Com uma população mundial estimada, em 2019, de 7,716,223,209 de pessoas, em que o número de usuários da Internet em junho de 2019 foi de 4,510,054,272, tendo um índice de utilização de 58,4%, ou seja cerca de 60% da população mundial acedeu, de alguma maneira, à internet. Em relação à taxa de crescimento entre 2000 e 2019 é de 1,149% a nível global, o que representa a evolução e crescimento da WEB, fundamentando a possibilidade de utilizar

contribuições de publicações para o desenvolvimento de um conceito de mitigação de eventos de catástrofe.

Parafraseando Hakimpour (2009), "a quantidade de dados digitais disponíveis para pesquisadores e profissionais do conhecimento aumentou tremendamente nos últimos anos. Isto é especialmente verdade no domínio geográfico. Como a quantidade de dados cresce, problemas de relevância e sobrecarga de informações tornam-se mais graves, o uso da tecnologia semântica é necessário para resolver esses problemas", revelando a grande expansão que a Web 2.0 teve após o seu surgimento a partir de 2004, em que se passou a utilizar a web como uma plataforma, aparecendo novos serviços, redes sociais, blogs de tecnologia de informação e wikis e a necessidade da validação semântica para classificar a informação.

Todo o conteúdo da Web é, até certo ponto, de natureza social, sendo principalmente criado e compartilhado por outras pessoas que não a pessoa que consome o conteúdo (Morris e Teevan, 2010), o que por si revela a potencialidade que a Web tem para disponibilizar informação "crowdsourced", ou seja, originada pelos utilizadores, revelando a importância que esta informação pode ter para o desenvolvimento de uma aplicação para a mitigação de eventos, ou até mesmo para outro tipo de aplicação que exija uma grande quantidade de dados originados por pessoas.

2.2. Metodologia de Laylavi

Como se irá perceber durante este trabalho de projeto, a tese de dissertação de Laylavi (2016) é a principal base metodológica deste trabalho de projeto, não só pela sua conceptualização, mas também pela semelhança de vários processos, essencialmente a coleção das publicações.

A arquitetura da metodologia de Laylavi (2016), foi projetada para satisfazer os requisitos de resposta de emergência para informações de caráter crítico, atualizadas, relevantes e referenciadas espacialmente. Neste caso esta estrutura computacional executa uma avaliação do grau de relação com eventos e inferência da localização de mensagens a partir da rede social "Twitter".

Contextualizando, uma "framework" pode ser definido como uma conceção abstrata e a sua implementação numa aplicação num determinado problema específico (Mattsson et al., 1997).

2.3. Arquitetura da Metodologia

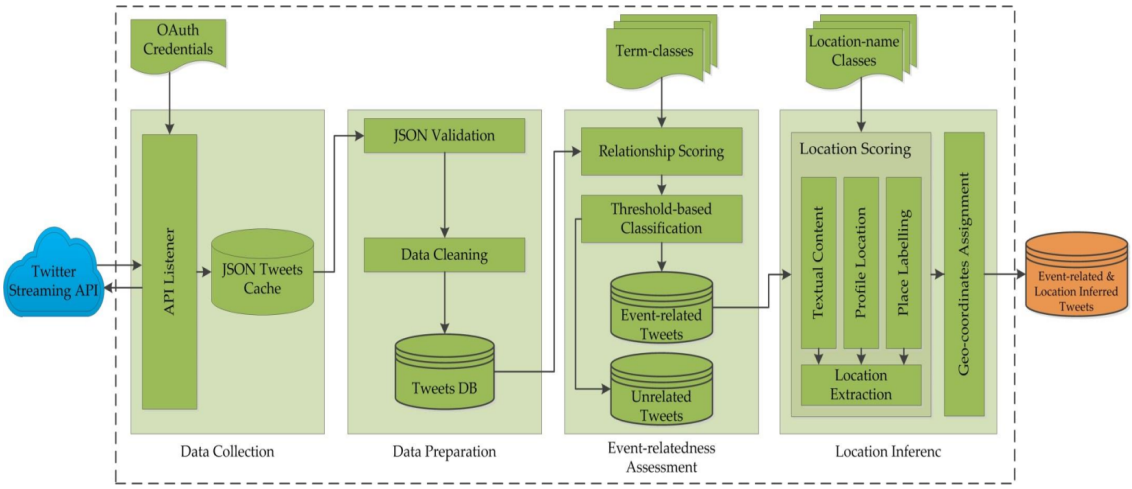

Figura 1: Arquitetura da metodologia de Laylavi (Laylavi, 2016)

Na figura 1, é possível observar a conceção da metodologia todas as componentes necessárias para o seu funcionamento, interagindo primeiramente com a Twitter Streaming API para colecionar *tweets* produzidos num evento de emergência, passando de seguida pela coleção de dados, preparação, avaliação de grau de relação ao evento, inferência de localização, chegando por fim ao objetivo, que é uma base de dados composta por *tweets* relacionados com o evento com a devida localização aproximada, inferida pela última componente. É importante referir que a base de dados criada poderá ser uma mais valia para a análise da situação e para o processo de tomada de decisão na resposta de emergência.

Respetivamente à coleção de dados, este modelo utiliza a Twitter Streaming API, embora existam outras alternativas, esta é a que disponibiliza uma abordagem mais simples e eficiente para a coleção de *tweets* em tempo real.

Outras alternativas exemplificadas, utilizadas por alguns investigadores são nomeadamente ferramentas open source, como por exemplo: o YourTwapperKeeper (Bruns et al., 2012; Larsson et al., 2012), DMI Twitter and Analysis Toolsets (DMI-TCAT) (Bruns et al., 2014; Trice, 2015) e o NodeXL (Hansen et al., 2010; Yep et al., 2014).

2.4. Limpeza de dados

Uma das etapas fundamentais para qualquer processo de análise de dados em qualquer campo computacional é a preparação dos dados a serem utilizados, neste caso a preparação de dados consiste basicamente na preparação da qualidade dos dados para a análise devido ao pré-processamento dos dados originais (S. Zhang et al., 2003) que inclui técnicas relacionadas com a validação, transformação, limpeza e redução dos dados "em bruto", isto é, não pré-processados (J. Han et al., 2011). Com o pré-processamento dos dados é possível melhorar a qualidade dos dados originais e facilitar a análise e processamento com a devida estruturação dos dados e a redução de ruído associado aos dados "em bruto".

2.5. Avaliação do grau de relação com o evento

Relativamente à validação das publicações, Laylavi (2016) utiliza um método de avaliação do grau de relação com o evento, componente que permite determinar a probabilidade de um *tweet* ser relacionado a um tipo de emergência específico (por exemplo: incêndios, inundações, furacões, tempestades). Para determinar se um *tweet* é relacionado ou não, é feita uma comparação do *tweet* colecionado com uma representação predefinida de *tweets* relacionados ao evento. Com base nesta comparação direta, é dada uma pontuação a cada *tweet* baseado no grau de relação com o evento alvo. Feito este processo, a pontuação relacionada ao evento de cada *tweet* colecionado é comparada com um valor de limiar predeterminado (limiar), que permite que a componente decida se o *tweet* é de facto relacionado com o evento ou não, sendo que se o valor da pontuação for superior ao limiar, será obviamente relacionado com o evento em questão.

3. Metodologia utilizada para a conceptualização

A metodologia deste trabalho de projeto baseia-se essencialmente na metodologia de Laylavi (2016) utilizado para extração, preparação e caracterização de dados de texto disponíveis na plataforma Twitter para a mitigação de eventos de catástrofe, como podemos verificar na figura 2.

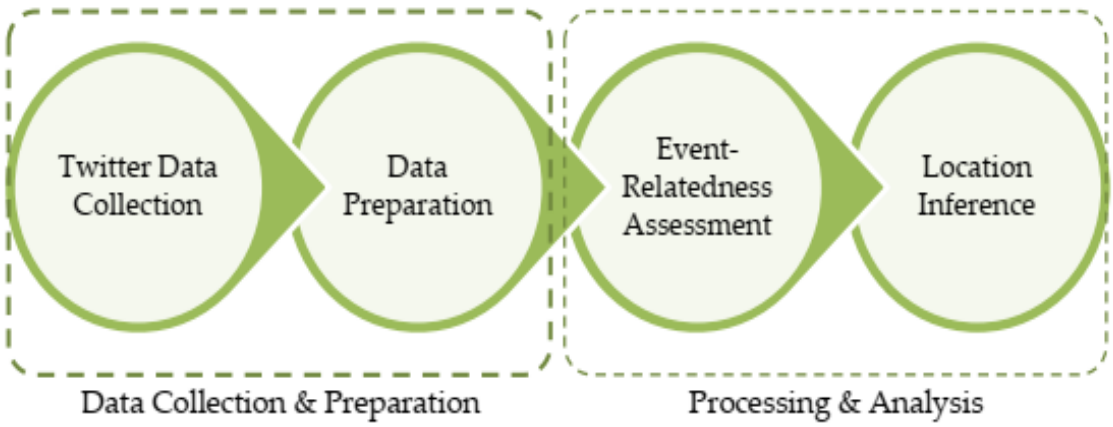

Figura 2: Estrutura conceptual da metodologia (Laylavi, 2016)

Apesar de Laylavi desenvolver um método de classificação para as publicações do Twitter, nesta metodologia utilizou-se a filtragem direta com palavras de interesse relativas a cada evento testado, com o objetivo de criar uma metodologia com resultados de filtragem razoáveis, sem a necessidade de criar um conceito complexo para o efeito, visto que à partida, ao filtrar publicações com palavras de interesse relativas a um evento iriamos ter resultados positivos, pois apenas estamos a filtrar publicações com as palavras de interesse que definirmos.

Seguindo o conceito de Laylavi (2016), o conceito deste trabalho de projeto começa também por colecionar dados de várias fontes na web (redes sociais/ noticias online) armazenando esta informação numa base de dados. No entanto existe a necessidade de pré-processar e validar toda esta informação para a tornar utilizável para análise. Para isso é necessário proceder ao pré-processamento, tendo como base Laylavi (2016), removendo acentos, carateres especiais, letras minúsculas, links, emojis (tudo o que seja desnecessário e que torne o conceito mais rápido e eficaz).

A maior diferença entre estes dois conceitos é que, ao contrário do conceito de Laylavi que utiliza uma referência de classe de palavras para classificar *tweets* sendo relacionados ou não com um evento, este utiliza uma metodologia de filtragem direta de publicações que contenham palavras de interesse relativamente a um evento, escolhidas manualmente pelo operador.

3.1. Arquitetura da Metodologia

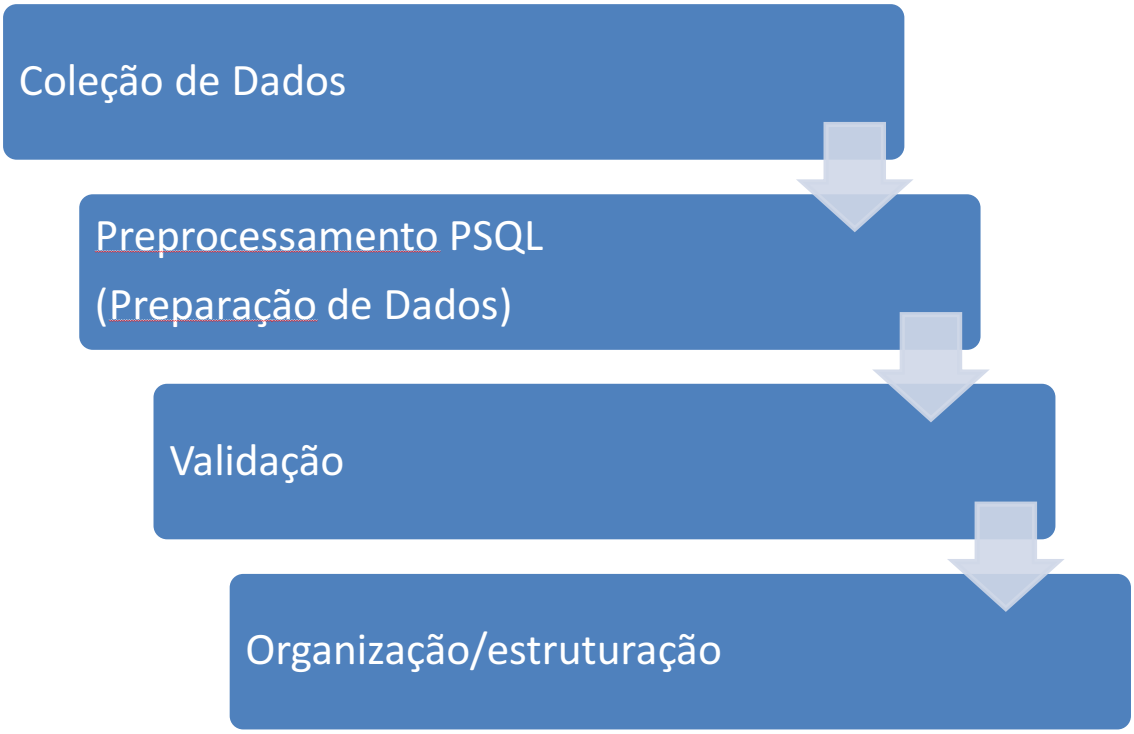

Figura 3 – Arquitetura do conceito

A arquitetura desta metodologia desenvolvida representada na figura 3, assemelha-se à estrutura conceptual da metodologia do modelo desenhado por Laylavi (2016) conforme representado na figura 2. No entanto, o modelo de Laylavi baseia-se apenas em dados disponibilizados na rede social Twitter e apenas na língua Inglesa, já que o caso de estudo se realizou na cidade de Sidney. Ao contrário deste exemplo, este conceito foi testado apenas para dados disponibilizados pelo Facebook, nomeadamente por notícias de jornais online nacionais (24 Sapo, Jornal de notícias, Jornal Expresso, CM Jornal, jornal ionline, diário de notícias, publico, tvi24, diário de Coimbra, RTP noticias, observador, diário as beiras, diário de leiria, SIC noticias, noticias ao minuto, e jornal sol).

Para uma melhor compreensão da estrutura conceptual, está disponível no anexo D (formato físico) o código integral em Python, desenvolvido para esta metodologia, assim como todos os ficheiros Excel e resultados utilizados para os mesmos testes.

Este conceito baseia-se apenas na língua portuguesa, podendo os resultados variar de acordo com a língua utilizada.

A metodologia consiste em quatro etapas distintas, conforme podemos ver na figura 3, na coleção de dados, na preparação dos dados, na validação destes e, por fim, no seu armazenamento.

Em primeiro lugar temos a coleção dos dados, que consiste na utilização da API da rede social Facebook para a coleção e armazenamento de publicações numa base de dados. Esta informação depois de ser armazenada, necessita de ser tratada para uma melhor otimização da sua utilização, realizando-se um pré-processamento das publicações colecionadas, preparando assim os dados para o processo de validação.

Este processo de preparação de dados consiste na utilização de ferramentas do software PostgreSQL para a remoção de qualquer tipo de informação desnecessária que esteja presente nas publicações (emojis, acentuação, links, assim como a remoção de stopwords e a transformação de todas as letras em letras minúsculas, espaços em branco, assim como a redução das palavras para a palavra a sua respetiva palavra de origem (stemming)). O processo de pré-processamento foi inicialmente desenvolvido em código python com a utilização de alguns módulos e scripts para a remoção de stopwords (módulo NLTK), e a para a remoção da restante informação não desejada, no entanto devido à sua baixa performance em relação ao tempo de resposta, chegou-se à conclusão que a utilização das ferramentas embutidas no PostgreSQL seriam a escolha mais adequada, pois conseguem realizar este procedimento com melhor performance.

Após a preparação dos dados, é necessário validar esta informação, utilizando também a filtragem do PostgreSQL. Para a filtragem desta informação foi utilizado a filtragem direta das publicações num determinado espaço de tempo, relativamente a cada evento selecionado. Neste caso foram escolhidos 3 eventos em espaços de tempo distintos. No que diz respeito à filtragem, foram selecionados dois termos diretamente relacionados com cada evento, como será mencionado no capítulo 3.4.3.

No caso dos eventos de incêndios foram utilizadas as palavras "fogo" e "incêndio", que correspondem a "fog" e "incendi" na posterior análise dos dados visto que no pré-processamento dos dados todas as palavras foram reduzidas à sua palavra de origem. No caso do evento do Furacão Leslie, foram utilizadas também duas palavras diretamente relacionadas com este evento: furacão ("furaca") e Leslie ("lesli").

No processo de validação foram criados dos tipos de validação, macro validação (relativa ao evento) para conseguir validar um evento de emergência e a micro validação (relativa ao evento e a um topónimo) para identificar o evento e a sua localização.

3.2. Coleção de Dados

Como a coleção de dados não contempla este trabalho de projeto, não é possível desenvolver sobre a sua construção. No entanto, esta é constituída por dados recolhidos através das redes sociais Facebook, Twitter, Flickr, contendo a mensagem da publicação, a data e hora de publicação e localização do usuário sempre que possível. No que diz respeito à coleção de dados da rede social Facebook, foram recolhidas publicações de jornais online publicados nesta rede social, como foi referido anteriormente.

3.3. Preparação dos Dados (Pré-processamento):

3.3.1. Problemas de desempenho

O desenvolvimento do código deste conceito começou por ser estruturado na linguagem Python, em conjunto com o módulo "pandas" para estruturar o dataframe extraído a partir da base de dados do PostgreSQL, realizando também o processo de limpeza e filtragem. No entanto não demorou muito tempo para perceber que o desempenho do pré-processamento e tratamento dos dados era muito lento, havendo a necessidade de adicionar um indicador de computação em como o código estava em processamento, pois este demorava muito tempo a terminar a tarefa. Após esta análise, fomos obrigados a alterar o código do pré-processamento e filtragem, passando a utilizar as ferramentas embutidas no PostgreSQL para realizar este mesmo processamento, melhorando a performance do conceito, pelo simples facto de que o pré-processamento do Python é baseado na iteração de cada linha de uma coluna, tornando-o muito lento para bases de dados com milhares de linhas, como é este caso.

Tabela 2 – Resultados de experiência entre SQL e Pandas (adaptado de Medium, 2017)

	Mais lento (segundos)		Mais rápido (segundos)		Mediana (segundos)	
	SQL	Pandas	SQL	Pandas	SQL	Pandas
Join	21.2	27.9	19.8	26.4	20.0	27.0
Groupby	8.9	38.6	8.6	35.7	8.6	37.8
Filter	10.2	27.5	9.5	25.0	9.7	25.3
sort	30.9	30.1	28.2	28.0	28.7	28.9

Para fundamentar esta alteração, podemos observar a tabela 2 (Medium, 2017) onde foram realizados 30 testes entre SQL e o módulo Pandas do Python com uma simples comparação de quatro ferramentas: join, groupby, filter e sort. Estes resultados demonstram uma excelente performance do SQL em relação ao Pandas no que diz respeito às ferramentas join, filter e groupby. No entanto a ferramenta sort o SQL foi apenas 0.2 segundos mais rápido.

É possível observarmos o código produzido inicialmente no anexo A, tendo como base um tutorial online adaptando-o a este conceito (machine learning mastery, 2017) relativamente ao pré-processamento, onde era utilizado o módulo NLTK no método "def preprocess" para transformar letras maiúsculas em minúsculas, para transformar as palavras em objetos (tokenizer) e para remover stopwords, ou palavra vazia em Português ("a", "o", "em", "no", etc). Para remover acentuação foi utilizado o método "def remove_accents" para devolver os carateres equivalentes sem a sua acentuação. O método "def clean" apenas aplica os dois métodos acima descritos à dataframe.

Em relação à filtragem para validação automática, também foi utilizado o Python inicialmente, no entanto tinha a limitação de filtrar apenas uma palavra, como é possível observarmos também no Anexo A, com o método "def Train", onde filtrava cada palavra linha a linha, tendo uma performance consideravelmente baixa. No entanto os três primeiros testes (incêndio de Pedrógão, Incêndios de Outubro de 2017 e furacão Leslie) foram feitos utilizando este método, filtrando o período de tempo de cada evento por cada palavra de interesse e depois fazendo a análise diretamente no ficheiro Excel.

3.3.2. SQL para pré-processamento e filtragem automática

Como já foi referido anteriormente, o processo inicial seria utilizar o Python para pré-processar os dados e para a sua filtragem, mas como foi referido anteriormente, devido a problemas de performance baixa realizou-se a migração para a integração de SQL para realizar este processamento diretamente a partir da base de dados, tornando o processo muito mais rápido. Como é possível observarmos no Anexo B, temos a expressão completa da query, que consegue fazer o pré-processamento, filtragem de palavras e escolha de período de tempo para a filtragem, conseguindo alterar estes procedimentos com a utilização da parametrização disponibilizada. Como é possível verificar foram utilizados comandos embutidos no PostgreSQL para o pré-processamento dos dados, o que melhorou a performance significativamente. Estes comandos, como é possível observarmos no Anexo B, consistem de comandos básicos como o "to_tsvector " que devolve a lista de cada palavra mãe e as suas posições no documento, o comando "Lower"

que devolve todas as palavras em letras minúsculas, o comando "unaccent" que remove a acentuação e o comando "regexp_replace" que remove carateres especiais e espaços. Para a filtragem de palavras foi utilizado o comando básico em que "message LIKE cada palavra", ou seja, cada publicação presente na coluna "message" será filtrada se estiver presente uma das palavras inseridas no parâmetro "Palavras".

Posteriormente, e tendo como vista trabalhos futuros de tratamento destes dados, realizou-se a extração de duas colunas a partir destas publicações, uma coluna chamada "Order Expression" em que a mensagem da publicação está ordenada e com palavras repetidas, e uma outra coluna "No_duplicated" onde não estão presentes palavras repetidas e não está a mensagem ordenada, como é possível observarmos na seguinte tabela 3.

Tabela 3 – Demonstração de output de uma publicação após pré-processamento e validação manual.

	Fid	Datahora	Message	Order_express	No_duplicated	Valid_tema
0	1128940 1891234 3_77941 5218926 883	2017-10-15 17:57:09	Há Taça no Calhabé… ACADÈMICA – PAÇOS DE FERREIRA	ha tac calhab academ pac ferreir	academ calhab ferreir ha pac tac	0
1	1128940 1891234 3_77946 0658922 339	2017-10-15 20:23:48	No escurinho do estádio Cidade de Coimbra. Nova falha na iluminação. É por causa dos incêndios, informa a a direcão da AAC/OAF.Estamos no Prolongamento do ACADÉMICA 1 - Paços de Ferreira 1.	escurinh estadi cidad coimbr nov falh iluminaca caus incendi inform direca aac oaf prolong academ 1 pac ferreir 1	1 aac academ caus cidad coimbr direca escurinh estadi falh ferreir iluminaca incendi inform nov oaf pac prolong	1

A partir destas duas colunas existe a possibilidade de comparação entre as duas, possivelmente para testes de desempenho, já que como estas estão pré-processadas, o processamento é mais rápido do que utilizando a coluna "message" que contém mais informação desnecessária (acentuação, pontuação). Como podemos ver também na tabela 3, o FID da mensagem original é mantido para a possibilidade de restaurar a publicação original se necessário, assim como a data e hora da publicação e a mensagem original.

3.4. Validação Semântica

3.4.1. Macro validação

No que diz respeito à macro validação, esta serve para validar as publicações em relação a um evento, ou seja, utilizando as palavras de interesse designadas para um evento de catástrofe, estas publicações serão validadas em relação ao evento. Laylavi (2016) tem um procedimento diferente, onde utiliza várias palavras de interesse (designadas anteriormente por uma equipa de proteção civil tendo em conta um tipo de evento) utilizando um algoritmo para classificar cada *tweet*, tendo em conta se possuem as palavras de interesse ou não. Após esta classificação, cada *tweet* passa por um limiar onde serão classificados como relacionados com um evento ou não relacionados com um evento, conforme podemos observar na figura 4.

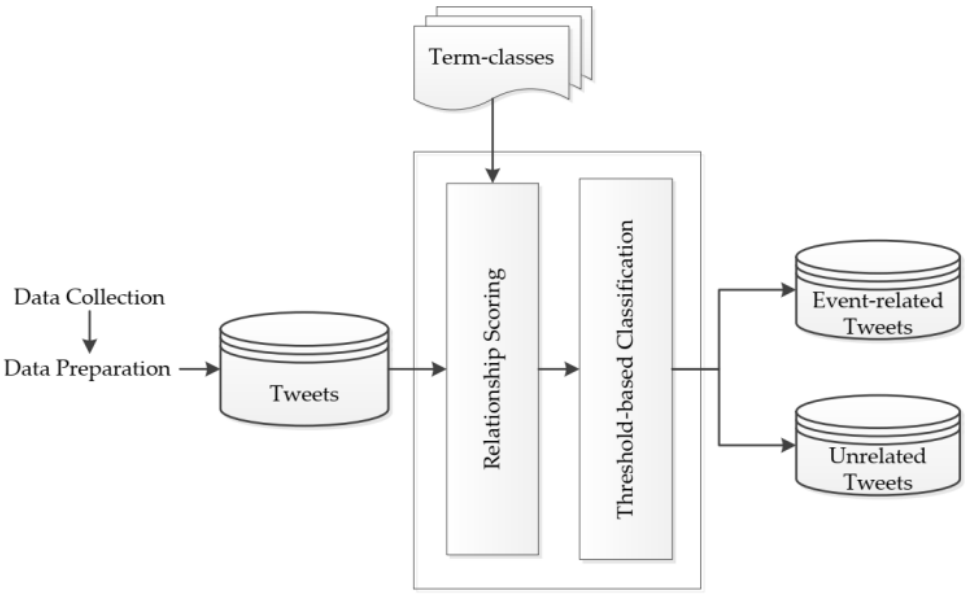

Figura 4: Arquitetura da componente de avaliação do grau de relação ao evento (Laylavi, 2016)

No entanto, devido à complexidade deste processo, tivemos que optar por algo não tão complexo que consiga resultados satisfatórios, tendo como objetivo a "afinação" posterior do conceito de validação.

3.4.2. Micro validação

O objetivo da micro validação é, desde o inicio, localizar o mais aproximadamente a publicação do evento real, ou seja, segundo Laylavi (2016) utilizando técnicas de inferir a localização do utilizador, através de informação disponível através da conta de perfil do usuário, ou se possível utilizar as coordenadas caso o utilizador tenha os serviços de localização ativados no seu dispositivo. Uma outra técnica utilizada por Laylavi no seu conceito é a inferência da localização da publicação em relação ao evento.

Contudo, neste trabalho de projeto apenas foi utilizada a filtragem direta com topónimos visto que os dados disponíveis pelas publicações do Facebook não contêm informação geográfica, ou quando contêm são sobre a conta de Facebook utilizada para a publicação. Como a única fonte de dados utilizada para os testes neste trabalho pertencem a publicações de jornais online, a localização será a localização onde a sede do jornal em destaque se localiza. Para contornar esta situação, utilizou-se um topónimo de modo a conseguir testar o conceito, utilizando um topónimo de interesse relacionado com cada evento, em conjunto com a junção de duas palavras de interesse (exemplo: "fogo" ou "incêndio" e "topónimo").

3.4.3. Palavras de Interesse

Para cada evento foram definidas duas palavras de interesse. O critério de seleção destas palavras de interesse foi a ligação direta com o evento, de forma a testar a metodologia. No entanto, segundo Laylavi (2016), é aconselhável que estas palavras sejam escolhidas por uma entidade como a proteção civil de forma a assegurar uma maior taxa de probabilidade de recolha de publicações com um maior contributo para a mitigação do evento.

Como apenas foram realizados testes em apenas 2 cenários diferentes, eventos de incêndios ou o caso do furacão Leslie (apesar da tentativa de testar outros eventos como cheias em Coimbra e em Lisboa, no entanto o número de publicações era consideravelmente reduzido, acabando por se realizarem apenas os testes respetivos aos eventos descritos posteriormente), a escolha de palavras de interesse foi reduzida, considerando-se apenas dois termos para cada evento. Em Relação aos eventos de Incêndio, foram consideradas as palavras de interesse "fogo" e "incêndio". No que diz respeito ao evento do Furacão Leslie, foram utilizadas outras palavras de interesse, diretamente relacionadas com o evento, havendo maior probabilidade de contribuições relacionadas com o evento: "Furacão" e "Leslie".

Para efeito de testes, como poderemos observar no capítulo seguinte, foram efetuados diferentes testes para verificar qual a melhor metodologia a ser utilizada no que diz respeito a

palavras de interesse. Estes testes incluem cada uma das palavras de interesse analisadas em separado e a junção das duas palavras com "e" (ou seja, cada publicação teria que incluir ambas as palavras) e a junção das duas palavras com "ou" (ou seja, cada publicação teria que incluir uma das duas palavras).

Como no pré-processamento se utilizou o processo de "stemming", ou seja, a remoção do sufixo de todas as palavras, obtendo a palavra de origem, por exemplo, todas as palavras designadas por "foguete", "fogo", "fogueira" foram reduzidas para a palavra "fog" que é a palavra de origem, o mesmo se aplica para a palavra incêndio (que corresponde a "incendi"), Leslie como "lesli", furacão como "furaca"(assim como na utilização de filtragem de topónimos realizada nos testes efetuados), Lousã como "lous", Pedrógão como "pedrog" e Coimbra como "coimbr".

Deste modo, ao filtrar uma publicação com a palavra de interesse "fog", serão filtradas todas as publicações que contenham a mesma palavra pré-processada, conseguindo filtrar publicações com erros de acentuação, que ao ser pré-processada utilizando o método de stemming, será devolvida a palavra de origem.

3.5. Armazenamento dos Dados

No que diz respeito ao processamento e armazenamento de dados filtrados, utilizaram-se 2 "outputs" do pré-processamento (sem ordem e sem repetição / com ordem e com repetição), mantendo-se o ID principal de cada publicação assim como a sua mensagem original e a hora de publicação, conforme é possível ver na tabela 3.

3.6. Critérios de validação manual

- Critério 1:

Relativamente ao critério 1 de validação manual, originalmente utilizado nos testes dos eventos dos incêndios de Pedrógão Grande, incêndios de outubro de 2017 e Furacão Leslie, assim como no teste da amostra orientada final, baseia-se no relacionamento direto ou indireto com o evento. Ou seja, contribuições com informação que indique qualquer tipo de informação relacionada com o evento, não estabelecendo relação temporal direta (podendo ser notícias sobre o evento vários meses depois, ou até relações indiretas como concertos de solidariedade.

Veio-se justificar na amostra de treino orientada que este critério apesar de aumentar os verdadeiros positivos consideravelmente, não é o melhor critério para ser utilizado, visto que

muitas destas contribuições não estão diretamente relacionadas com o evento, apresentando publicações sem interesse para a entidade que a utilizar.

- Critério 2:

No que diz respeito ao critério 2, este corresponde a toda a informação que indique que um incêndio aconteceu ou está a acontecer num determinado lugar e num determinado momento. Assim sendo, na contribuição tem de existir alguma referência a uma localização, ainda que não seja concreta e/ou absoluta.

Para além disso, o momento da contribuição tem de coincidir com o momento da ocorrência (contudo, pode haver uma tolerância de alguns dias).

Consideram-se contribuições relativas a qualquer tipo de incêndio (florestal ou urbano).

3.7. Primeiros testes

Para realizar os primeiros testes foi necessário escolher primeiro alguns eventos que certamente teriam publicações suficientes para realizar estes testes. Devido à limitação temporal da recolha de publicações para a base de dados, apenas tínhamos dados recolhidos a partir da data 13-05-2010 (data da primeira publicação do Facebook recolhida), não poderíamos recuar muito no tempo, daí terem sido testados os eventos de catástrofe mais recentes que aconteceram em Portugal Continental, o Incêndio de Pedrógão de 2017, que provocou muitos danos e mortes, assim como os Incêndios de Outubro de 2017. Para não considerar apenas testes com eventos de catástrofe de incêndios, utilizou-se o evento de catástrofe do Furacão Leslie de 2018.

Dito isto, delimitou-se o período de tempo de cada evento para proceder à filtragem e validação do conceito.

Testes de Filtragem com 3 eventos distintos em períodos de tempos diferentes:

- Incêndios de Pedrógão Grande (17 junho de 2017 – 24 junho de 2017);
- Incêndios de Outubro de 2017 (15 outubro 2017 – 17 outubro 2017);
- Furacão Leslie (13 outubro 2018 – 14 outubro 2018).

Como já tinha sido referido anteriormente, as palavras de interesse utilizadas para cada um dos respetivos eventos foram:

- Incêndios Outubro 2017:
    - Palavras utilizadas: “incêndio”, “fogo”, “Lousã”;

- Incêndios Pedrógão Grande;
    - Palavras utilizadas: “incêndio”, “fogo”, “Pedrógão”;
- Furacão Leslie:
    - Palavras utilizadas: “furacão”, “Leslie”, “Coimbra”.

Para a criação dos resultados destes três testes foram gerados 3 ficheiros Excel para cada evento, visto que neste ponto, o código ainda não tinha sido alterado para SQL e ainda não tinha sido criado o script de comparação automática, estando apenas limitado à filtragem de uma palavra por filtragem. Dito isto, para a apresentação dos resultados foi necessário extrair uma folha Excel com todas as publicações dentro do período de tempo de cada evento, de seguida foram selecionadas aleatoriamente 118 publicações desta tabela (a referência de 118 publicações para todos os testes refere-se à inexistência de mais publicações dentro do período de tempo do evento dos Incêndios de outubro de 2017, estabelecendo a referência de 118 publicações para cada teste de forma a manter o mesmo número de amostra para cada teste).

Sendo selecionadas as 118 publicações de cada período de tempo para cada evento, procede-se então à filtragem de cada palavra de interesse relativamente a cada evento durante o período de tempo do evento, ou seja, em alguns eventos existem mais do que 118 posts, portanto vão existir mais publicações filtradas com uma palavra de interesse do que a amostra total, para isso foi necessário cruzar o FID de cada publicação filtrada pelas publicações validadas manualmente, conforme será explicado posteriormente.

Após obtermos as 118 publicações do período de tempo total de cada evento, procedeu-se então à validação manual das 118 publicações (utilizando o critério 1 de validação manual, visto que o critério 2 surgiu no teste de amostra posterior a estes primeiros testes).

Após a validação manual de cada publicação estar terminada, foi necessário cruzar a informação de cada filtragem de cada uma das palavras de interesse com as 118 publicações aleatórias, de modo a analisar a presença de verdadeiros positivos, verdadeiros negativos, falsos positivos ou falsos negativos. Para isto, foi utilizado o FID original de cada publicação filtrada, cruzando a informação com o FID das 118 publicações aleatórias, se o FID de uma palavra de interesse corresponde a algum FID das 118 publicações, significa que este é positivo, ou seja, a publicação contém a palavra de interesse em questão.

Por fim, após estarem ambas as palavras de interesse de cada evento identificadas, realizou-se a filtragem para obter a junção de publicações que contenham “fogo” e “incêndio” (publicações que contenham fogo e incêndio) e “fogo” ou “incêndio” (publicações que contenham a palavra

fogo ou a palavra incêndio), assim como a filtragem da combinação de "fogo" ou "incêndio" com um topónimo.

3.8. Teste de amostra orientada

Para testar esta metodologia foi também gerada uma amostra de referência, tendo como base quatro critérios distintos de forma a assegurar a eficácia da metodologia. É importante referir que este teste foi direcionado para evento de incêndio, utilizando as duas palavras acima referidas para a sua validação automática ("fogo" e "incêndio"). No que diz respeito à criação orientada da amostra de referência, o objetivo é garantir que na amostra de referência existam:

- Linhas com palavras de interesse e topónimos;
- Linhas com palavras de interesse;
- Linhas com topónimos e sem palavras de interesse;
- Linhas sem palavras de interesse e sem topónimos.

Como o principal objetivo da criação orientada desta amostra de treino era verificar se, de facto, esta metodologia consegue filtrar publicações de relevância a um evento em questão, tendo em atenção a utilização de topónimos para possibilitar a localização do evento, utilizando várias publicações que contenham palavras de interesse ("fogo" ou "incêndio"), publicações que contenham palavra de interesse e também topónimos, publicações com topónimos mas sem palavras de interesse e por fim publicações sem topónimos ou palavras de interesse.

O período de tempo desta amostra é mais extenso do que os períodos de tempo utilizados para os testes anteriores (Incêndio de Pedrógão, Incêndios de Outubro de 2017 e Furacão Leslie), utilizando um período de tempo consideravelmente maior (2017-06-01 a 2018-10-31) de forma a obter uma amostra mais alargada.

Após a criação orientada da amostra de teste, foi necessário validar manualmente cada publicação, utilizando o primeiro critério, utilizado nos testes anteriores (Incêndio de Pedrógão, Incêndios de Outubro de 2017 e Furacão Leslie).

No entanto, foi possível perceber que ao se utilizar o critério 1 para a validação manual, incluindo todas as publicações relacionadas com o evento, mesmo após este ter acontecido, ou qualquer tipo de publicação que fosse indiretamente relacionada com o evento, iria gerar erro na filtragem de verdadeiros positivos (aumentando este valor), pois apesar de muitas publicações relacionadas indiretamente com os incêndios conterem uma das palavras de interesse, estas não iriam ajudar uma entidade como a proteção civil, havendo a necessidade de criar um novo critério

com a inclusão e validação de apenas publicações relacionadas diretamente com o evento e que estejam a acontecer no momento (utilizando o critério 1 de validação, muitas publicações referiam-se a eventos terceiros, como a realização de concertos solidários, bombeiros homenageados, entre outras situações irrelevantes para o evento em questão, que por si não iria contribuir para a mitigação de um evento em tempo real).

Contudo, foram realizados dois testes, um para cada critério, conseguindo comparar diretamente os dois critérios entre si. No que diz respeito às palavras de interesse utilizadas, e visto que este processo foi realizado após os testes dos 3 eventos anteriores, apenas foi utilizado a junção de "fogo" ou "incêndio", já que esta foi a que teve melhores resultados anteriormente, não utilizando cada palavra individualmente, assim como a junção de "fogo" e "incêndio".

Por fim, foi também realizado um teste com diferentes modelos de machine learning (Linear Support Vector Machine, Logistic Regression, Random Forest e Naive Bayes) para efeitos de comparação com os testes realizados anteriormente, o que neste caso se revelou inadequado devido à inexistência de uma base de treino suficiente extensa para conseguir treinar os modelos acima referidos e obter resultados no mínimo satisfatórios.

3.8.1. Automatização de validação em Excel

Os primeiros testes realizados (Incêndios de Pedrógão, Incêndios de outubro de 2017 e furacão Leslie) foram analisados manualmente diretamente no ficheiro do Excel utilizando a ferramenta de filtragem de maneira a analisar os seus resultados, disponibilizados no próximo capítulo. No entanto, dado o trabalho demorado deste processo, desenvolveu-se um script para realizar este processo automaticamente, tendo como base a validação manual realizada anteriormente (referência), comparando diretamente com os resultados da filtragem automática.

Podemos observar este script no anexo C, onde o ficheiro de referência onde foi realizada a validação manual é comparado com o ficheiro de resultados filtrados pelo processo automático, criando uma coluna com a respetiva classificação (verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos).

4. Verificação dos resultados da metodologia implementada

Como já foi referido anteriormente, a fonte de informação recolhida para a obtenção destes resultados foi a rede social "Facebook", nomeadamente publicações de jornais de notícias, que apresentam uma maior taxa de credibilidade, evitando contribuições falsas ou não fidedignas.

Apesar deste conceito ter sido testado apenas na rede social Facebook, é um conceito com flexibilidade de plataformas (Twitter, noticias online, Facebook), neste caso para efeitos de testes apenas foi considerado o uso da plataforma Facebook.

Como já foi referido no capítulo anterior, os testes realizados foram feitos em períodos de tempo distintos, sendo dois eventos relacionados com incêndios (incêndio de Pedrógão e incêndios de outubro de 2017) e um evento relacionado com o furacão Leslie para existir um termo de comparação e de validação em como a metodologia é capaz de ser utilizada noutros eventos de catástrofe.

Em primeiro lugar, podemos analisar estes três primeiros testes, referidos no capítulo anterior, conseguindo comparar os resultados de cada um deles. É de notar que estes três primeiros testes utilizaram o critério 1 de validação manual, onde qualquer contribuição relacionada com o evento é considerada como relacionada ao evento.

No teste seguinte, podemos analisar uma amostra de criação orientada com o objetivo de garantir que na amostra de referência existam: linhas com palavras de interesse e topónimos, linhas com palavras de interesse, linhas com topónimos e sem palavras de interesse e linhas sem palavras de interesse e sem topónimos. Como já foi referido no capítulo anterior, esta amostra tem o intuito de testar a metodologia num período de tempo de amostra mais alargado, de forma a validar o conceito. Neste teste chegamos à conclusão que o critério 1 não seria o mais indicado para este conceito, criando um novo critério de validação, no entanto foram feitos os testes com a utilização de ambos os critérios para ser possível efetuar uma comparação direta entre estes dois critérios de validação.

4.1. Caso de estudo: Incêndios de Outubro de 2017

Respetivamente a este caso de estudo, como já foi referido no capítulo 3.7, no período de tempo deste evento apenas foram recolhidas 118 publicações, contudo, para realizar os testes com um número de amostra de publicações de referência igual entre os três testes foi definida uma amostra de 118 publicações para os três eventos, selecionando aleatoriamente as 118 publicações de cada período de tempo no caso do incêndio de Pedrógão e Furacão Leslie, pois estes eventos possuíam mais do que 118 publicações recolhidas.

Evento: Incêndios de Outubro de 2017
Período de Tempo: 15 outubro 2017 – 17 outubro 2017
Número total de publicações que ditou o tamanho de amostra: 118

Tabela 4 – Resultados Incêndios de Outubro de 2017

Método	Processo automático				Referência (Classificação Manual)	
	Filtrados (+)		Não Filtrados (-)			
	Devidamente (base 77)	Indevidamente (base 41)	Devidamente (base 41)	Indevidamente (base 77)	Positivos	Negativos
'fog' (26)	26	0	41	51	77	41
'incendi' (42)	42	0	41	35		
'fog' e 'incendi' (9)	9	0	41	68		
'fog' ou 'incendi' (59)	59	0	41	18		
"incendi" ou "fog" e "lous" (3)	3	0	41	74		

Iniciando a análise com o primeiro teste, os Incêndios de outubro de 2017, podemos observar os resultados na tabela 4. Sendo que o "método" se refere à palavra ou palavras de interesse utilizadas para a filtragem automática, podemos observar que no que diz respeito à primeira palavra de interesse "fog" em que no período de tempo do evento foram recolhidas 26 publicações que continham a palavra de interesse "fog". Neste caso todas as publicações com a palavra de interesse "fog" foram filtradas corretamente, sendo contabilizadas como verdadeiros positivos coincidindo, portanto, com a validação manual realizada anteriormente. No entanto as restantes 51 publicações classificadas como falsos negativos (que correspondem às restantes publicações validadas manualmente que não foram filtradas) pois estas publicações não contêm a palavra "fog". Relativamente às publicações classificadas como falsos positivos, coincidem com a validação manual realizada, ou seja, todas as publicações não relacionadas com o evento foram filtradas.

Relativamente à palavra de interesse "incendi", analisa-se a mesma situação que a palavra "fog", foram filtradas 42 publicações com a palavra "incendi" no período de tempo do evento, sendo que estas 42 foram filtradas, pois continham a palavra de interesse, contudo, 35 publicações que correspondem às restantes publicações validadas manualmente não foram filtradas pela simples razão de não conterem a palavra "incendi" na publicação. As restantes publicações consideradas irrelevantes ao evento validadas manualmente foram corretamente não filtradas como, que acaba por acontecer para todas as palavras de interesse testadas, como se pode observar na tabela 4.

Utilizando a junção de "fog" e "incendi" começamos a obter resultados mais interessantes, com 9 publicações filtradas no período de tempo do evento, foram filtradas 9 publicações que continham ambas as palavras, dentro das 77 publicações validadas manualmente, um valor muito baixo, no entanto é de esperar visto que nem todas as publicações contêm ambas as palavras na mesma publicação. O que diz respeito à junção de "fog" ou "incêndio", o valor de verdadeiros positivos é de 59 publicações filtradas corretamente em 77 publicações validadas manualmente, ou seja, 59 destas publicações continham a palavra de interesse "fog" ou a palavra de interesse "incendi", que corresponde a 77% de publicações filtradas corretamente (tendo como base a classificação manual).

Em relação ao teste com um topónimo, neste caso com a palavra de interesse "Lousã", obtivemos apenas 3 verdadeiros positivos, um número incrivelmente baixo, no entanto é de esperar, já que as publicações recolhidas com a presença de topónimo são muito reduzidas.

Para este caso concreto, percebe-se que a junção das palavras 'fog' ou 'incendi' são suficientes para ter um bom desempenho em termos de verdadeiros positivos, em relação à utilização de apenas uma palavra como modo de filtragem.

Não foram realizados testes com mais palavras (por exemplo: "bombeiro" ou "queimados") porque teoricamente poderia aumentar o número de verdadeiros positivos, mas também não garante que não aumentasse os falsos positivos (por exemplo: "bombeiros fazem juramento de bandeira", entre outros casos semelhantes).

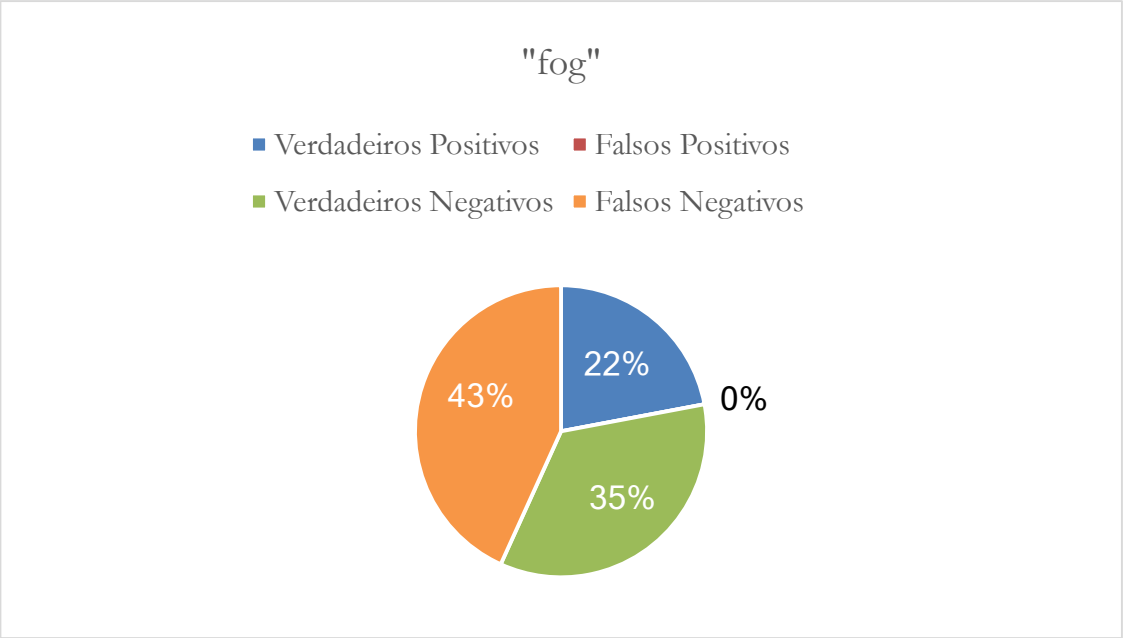

Figura 5: Resultados da palavra filtrada "fog" do evento Incêndios de outubro de 2017

Na figura 5, podemos observar o valor percentual relativamente à palavra de interesse "fog" para o evento dos incêndios de outubro de 2017, sendo este gráfico repartido por 4 secções, uma vermelha referente aos falsos positivos, uma azul remetente aos verdadeiros positivos, uma verde remetente aos verdadeiros negativos e por fim a laranja que corresponde aos falsos negativos.

Analisámos que a maior percentagem pertence à secção laranja, referente aos falsos negativos que o processo automático filtrou, sendo esta percentagem de 43%, correspondente às publicações que não foram filtradas corretamente como sendo verdadeiros positivos pois não continham a palavra de interesse "fog" nas publicações.

Analisámos também que a segunda maior percentagem pertence à secção verde, referente aos verdadeiros negativos, apresentando uma percentagem de 35%, ou seja, foram filtrados corretamente como não sendo relacionados com o evento.

Por fim a última percentagem remete à secção azul, a qual nos apresenta os verdadeiros positivos, com uma percentagem de 22%, que correspondem às publicações que contêm a palavra de interesse "fog".

Neste gráfico não analisámos uma secção que remeta aos falsos positivos, uma vez que, todas as publicações não relacionadas com o evento foram corretamente filtrados como verdadeiros negativos.

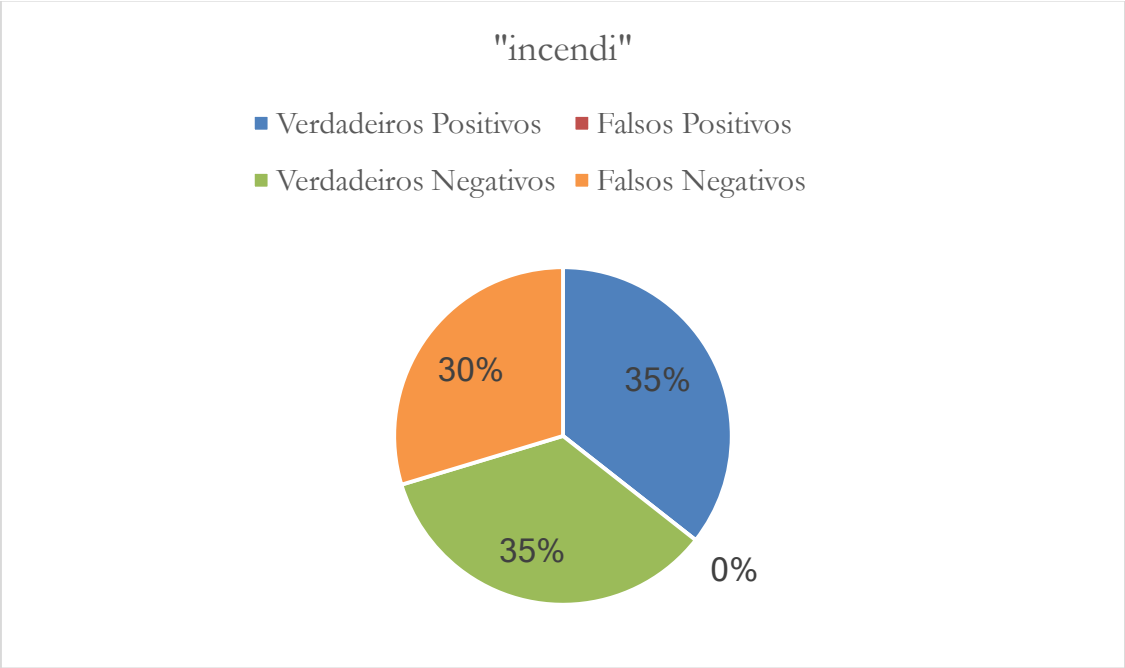

Figura 6: Resultados da palavra filtrada "incendi" do evento Incêndios de outubro de 2017

Na figura 6, observamos o mesmo procedimento que a figura 5, no entanto neste caso a palavra de interesse é "incendi", onde o valor percentual de falsos positivos e verdadeiros positivos é de 35%

Por fim, em relação aos falsos positivos, não existe valor pois nenhuma publicação foi incorretamente classificada como relacionada com o evento.

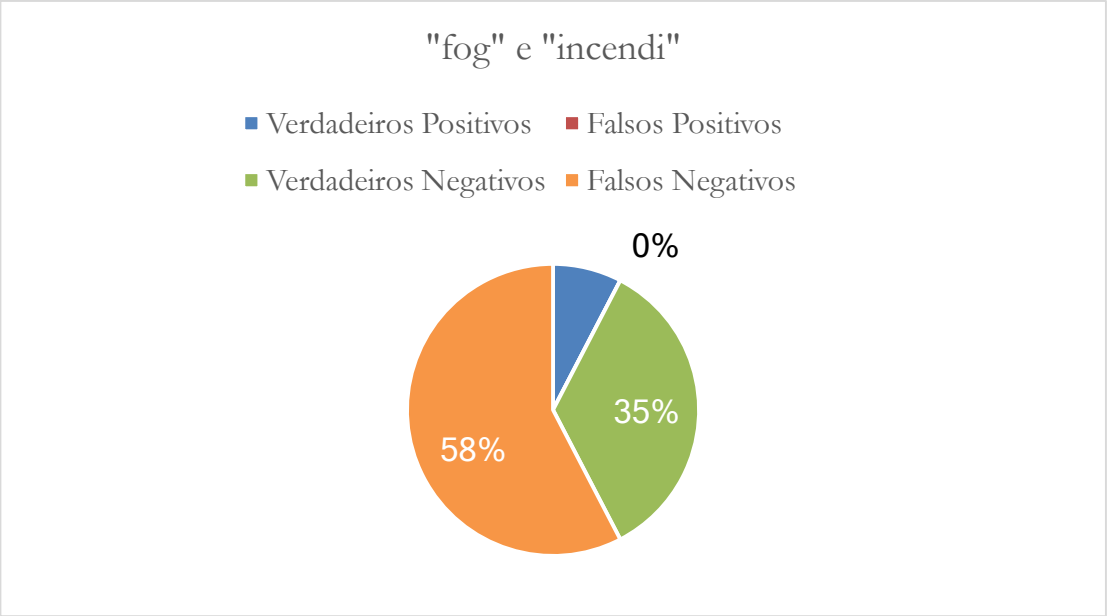

Figura 7: Resultados da palavra filtrada "fog" e "incendi" do evento Incêndios de outubro de 2017

Na figura 7, podemos observar a junção das palavras de interesse "fog" e "incendi", em que o número de falsos negativos é de 58%, que como foi acima referido, significa que estas publicações não contêm as duas palavras na mesma publicação. Em relação aos verdadeiros negativos, têm o valor de 35%.

No que diz respeito aos verdadeiros positivos, apenas 7% das publicações foram corretamente filtradas, que como tinha sido observado na tabela 4, representa um valor percentual muito baixo.

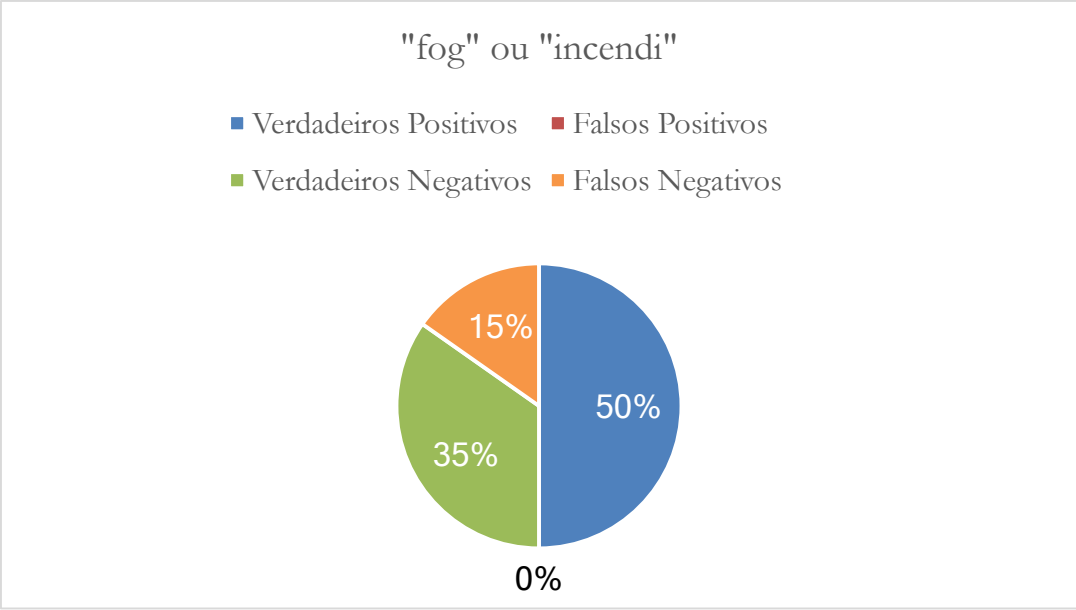

Figura 8: Resultados da palavra filtrada "fog" ou "incendi" do evento Incêndios de outubro de 2017

Na figura 8, podemos observar as percentagens relativamente à junção das palavras de interesse "fog" ou "incendi". Olhando para a figura, conseguimos observar a maior secção de verdadeiros positivos com o valor de 50% de publicações filtradas corretamente, no entanto no que diz respeito aos falsos negativos, 15% das publicações deviam ter sido filtradas como verdadeiros positivos.

Por fim, temos os verdadeiros negativos, corretamente filtrados como não relacionados com o evento, representando 35%.

Neste exemplo podemos observar que os resultados obtidos foram melhores relativamente aos anteriores, uma vez que os resultados não foram tão limitados, pois a utilização da conjunção "ou" permite ao processo automático abranger um leque maior de publicações ao conseguir filtrar publicações que tenham uma das duas palavras de interesse.

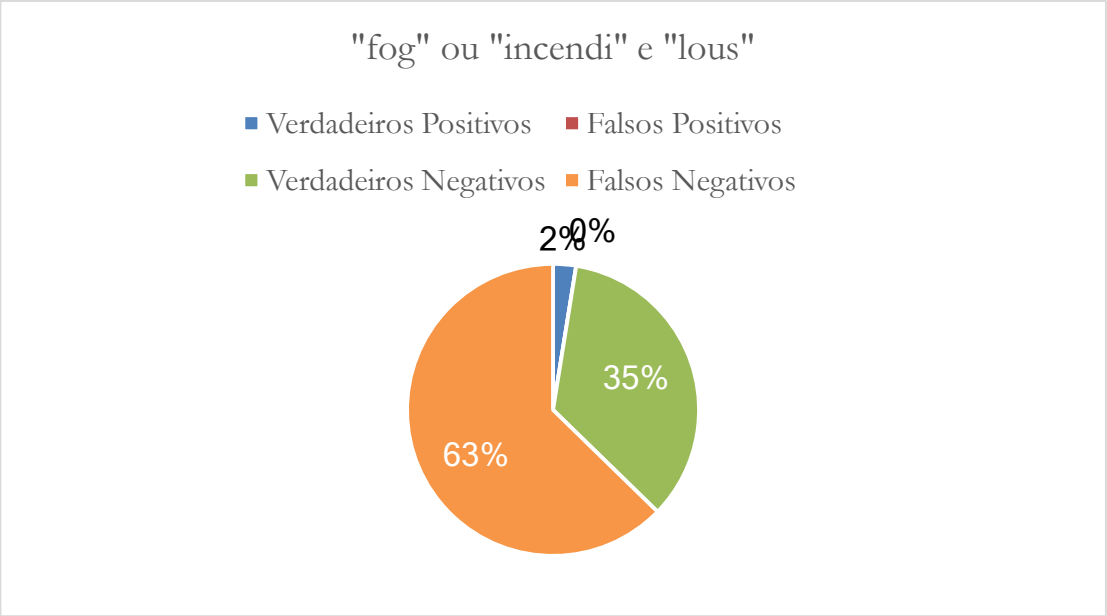

Figura 9: Resultados da palavra filtrada "fog" ou "incendi" e "lous" do evento Incêndios de outubro de 2017

Por fim, observando a figura 9, obtemos apenas 2% de verdadeiros positivos que corresponde a um valor muito baixo em relação ao teste anterior. Isto significa que a micro validação apenas deve ser efetuada em casos específicos de localização já que este vai limitar muito as publicações filtradas.

Sendo assim, obtivemos 63% de falsos negativos e 35% de verdadeiros negativos.

4.1.1. Caso de estudo: Incêndios de Pedrógão de 2017

Evento: Incêndio de Pedrógão

Período de Tempo: 17 de junho de 2017 – 24 de junho de 2017

Total de publicações efetivo no período de tempo: 568

Número total de publicações escolhidas aleatoriamente inseridos no número total de publicações: 118

Tabela 5 - Resultados Incêndios de Pedrógão de 2017

Método	Processo automático				Referência (Classificação Manual)	
	Filtrados (+)		Não Filtrados (-)			
	Devidamente (base 47)	Indevidamente (base 71)	Devidamente (base 71)	Indevidamente (base 47)	Positivos	Negativos
'incendi' (173)	28	0	71	19	47	71
'fog' (103)	8	0	71	39		
'incendi' e 'fog' (2)	2	0	71	45		
'incendi' ou 'fog' (34)	34	0	71	13		
"incendi" ou "fog" e "Pedrog" (18)	18	0	71	29		

Na tabela 5, referente ao evento "Incêndio de Pedrógão", podemos observar que filtrando a palavra "incendi", obtemos 173 resultados de um total de 568 no período de tempo do evento, desses quais o processo automático filtrou como verdadeiros positivos 28 publicações de uma base de 47 validados manualmente.

Como verdadeiros negativos, o processo automático não filtrou nenhuma publicação que fosse relacionada com o evento sem o ser, correspondendo ao valor 0 em todos os métodos.

Porém, como falsos positivos, todas as publicações foram filtradas corretamente como não sendo relacionadas com o evento, dentro das 71 validadas manualmente, apresentando neste caso em particular uma taxa de sucesso de 100%, no entanto, como falsos negativos, foram filtrados 19 numa base de 47, o que indica alguma taxa de erro na que deveriam ter sido filtrados, no entanto, não foram filtrados pois não contêm a palavra incendi.

Filtrando a palavra "fog" obtemos um resultado de 103 palavras dentro dos 568, no entanto apenas foram utilizadas 118 publicações escolhidos aleatoriamente, nas quais o processo automático destaca como verdadeiros positivos 8 publicações, que corresponde a um valor muito baixo.

Por outro lado, o processo automático destaca como falsos positivos 71 publicações de uma base de 71, que é o esperado, significando que todas as publicações não relacionados com o evento foram filtrados corretamente, e como negativos falsos foram filtradas 39 publicações numa base de 47, que deveriam ter sido filtrados como verdadeiros positivos.

Ao filtrarmos mutuamente "incendi" e "fog", o processo automático filtra como verdadeiros positivos somente 2 publicações de uma base de 47, um valor muito baixo. Por outro lado, como falsos positivos, o processo automático filtra 71 posts de uma base de 71, enquanto que nos falsos negativos o processo automático filtra 45 de uma base de referência de 47, representando uma margem de erro muito grande, visto que num universo de 47 publicações relativos ao evento, apenas 2 publicações possuem a combinação de "incendi" e "fog", tornando este método pouco recomendável devido ao seu resultado muito baixo, aconselhando o uso da junção de "palavra1" ou "palavra2".

No entanto, ao filtrarmos a combinação de "incendi" ou "fog", o desempenho dos resultados visualizados é superior relativamente a "incendi" e "fog", uma vez que não limitamos a filtração de resultados por parte do processo automático, sendo possível então observar como filtrados positivos 34 publicações da base validada manualmente de 47, e como verdadeiros negativos 0 publicações de uma base de 71.

No que diz respeito aos falsos positivos, 71 publicações de uma base de 71 foram corretamente filtradas como sendo não relacionadas com o evento, por outro lado, sendo que a filtração dos resultados não é tão limitada, observa-se um número menor falsos negativos, de 13 publicações de uma base de 47, um número muito menor comparados aos resultados anteriores.

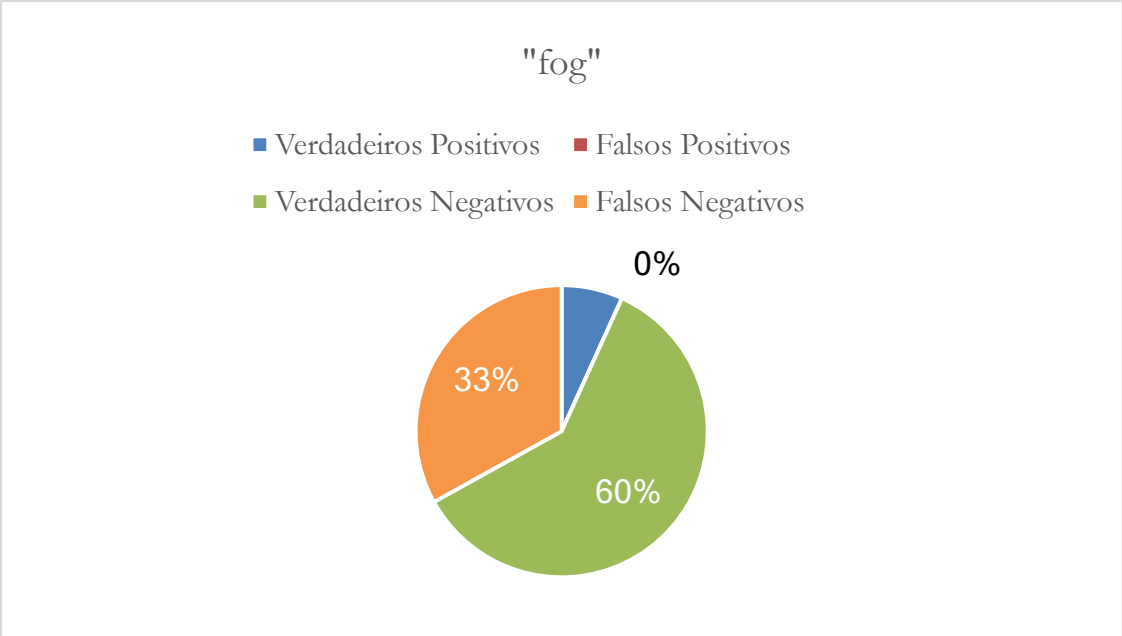

Figura 10: Resultados da palavra filtrada "fog" do evento Incêndio de Pedrógão de 2017

Na figura 10, conseguimos observar o valor percentual relativamente à palavra de interesse "fog" para o evento do incêndio de Pedrógão de 2017.

Analisamos que a maior percentagem pertence é referente aos verdadeiros negativos que o processo automático filtrou, sendo esta percentagem de 60%, correspondente às publicações que foram filtradas corretamente como não sendo relacionadas com o evento.

A percentagem pertence à secção azul, referente aos verdadeiros positivos, apresentando uma percentagem de 7%, ou seja, foram filtradas corretamente como sendo relacionados com o evento tendo como base a validação manual.

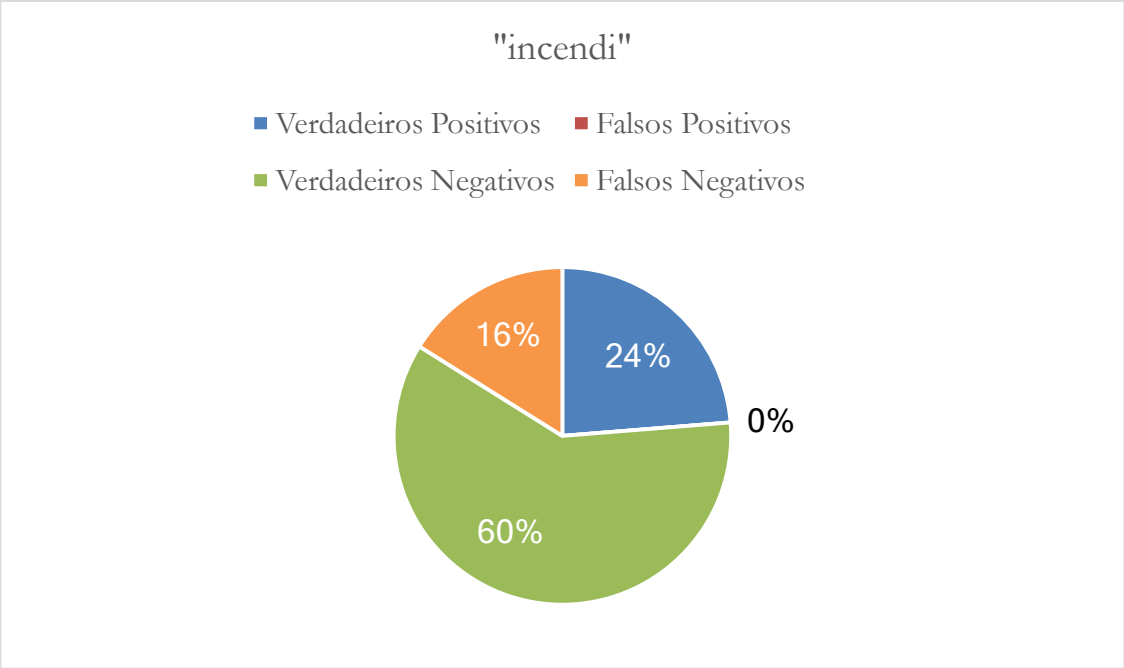

Figura 11: Resultados da palavra filtrada "incendi" do evento Incêndio de Pedrógão de 2017

À semelhança da figura 10, a figura 11 representa a palavra de interesse "incendi" para o evento do incêndio de Pedrógão de 2017.

Em relação aos verdadeiros positivos, obtivemos 24% de filtragens corretas e 60% de verdadeiros negativos.

No entanto, 16% corresponde a falsos negativos, que deveriam ter sido relacionados com o evento, mas como não possuem a palavra de interesse não foram filtrados.

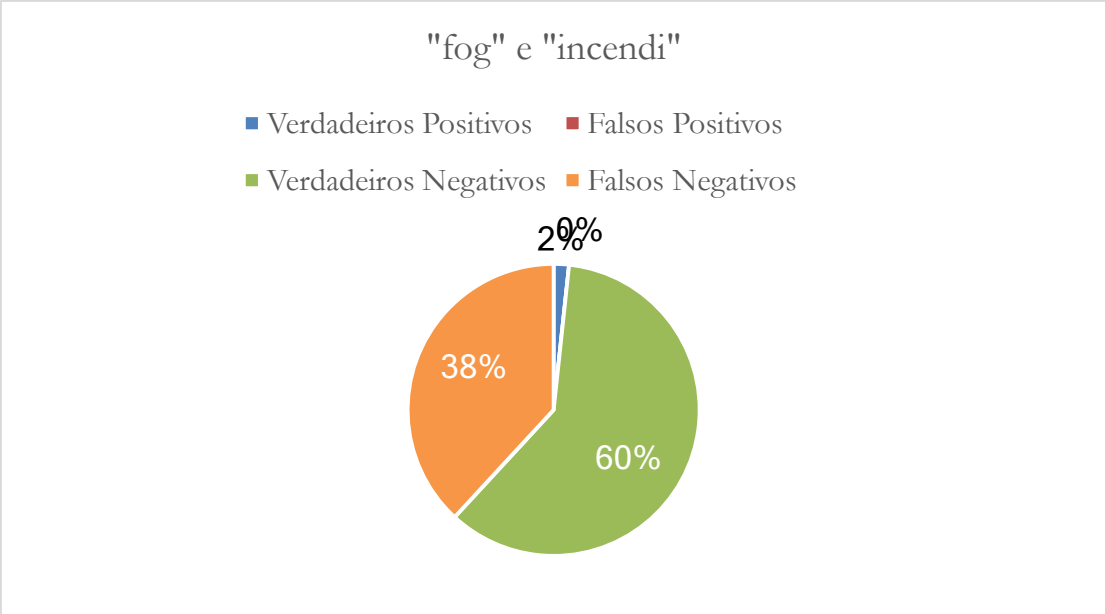

Figura 12: Resultados da palavra filtrada "fog" e "incendi" do evento Incêndio de Pedrógão de 2017

Na figura 12, podemos observar a junção das palavras de interesse "fog" e "incendi", em que o número de falsos negativos é de 38%, que significa que estas publicações não contêm as duas palavras na mesma publicação.

Em relação aos verdadeiros negativos, estes têm o valor de 60%, sendo estes corretamente filtrados como não sendo relacionados com o evento.

No que diz respeito aos verdadeiros positivos, apenas 2% das publicações foram corretamente filtradas.

Como já foi referido anteriormente, este método não é o melhor a ser utilizado neste conceito devido às suas características limitativas.

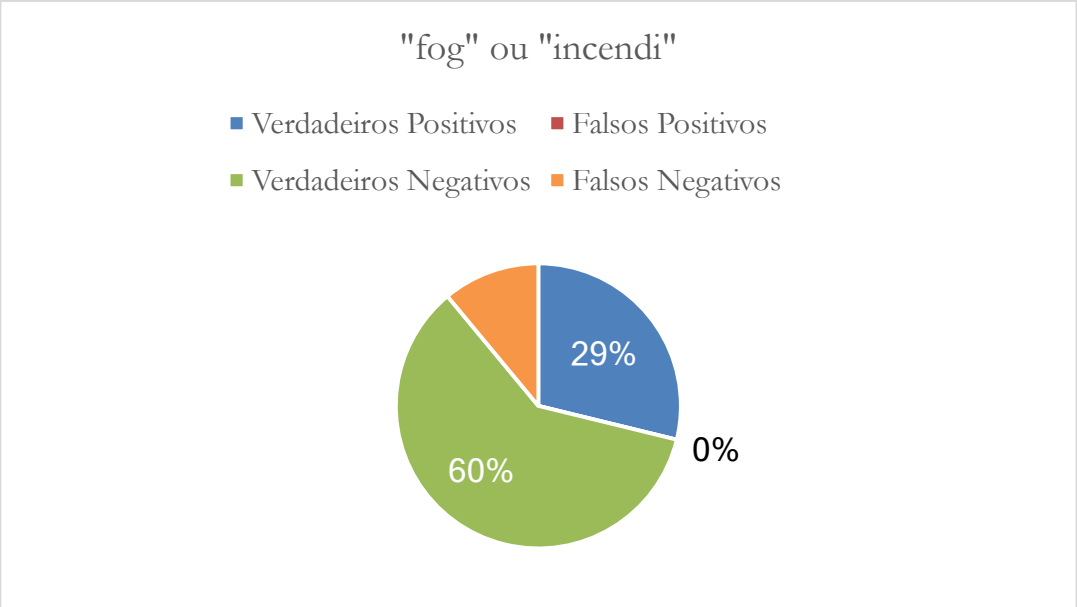

Figura 13: Resultados da palavra filtrada "fog" ou "incendi" do evento Incêndio de Pedrógão de 2017

Na figura 13, podemos observar as percentagens relativamente à junção das palavras de interesse "fog" ou "incendi".

Olhando para a figura, conseguimos observar a maior secção de verdadeiros positivos com o valor de 60% de publicações filtradas corretamente.

No que diz respeito aos falsos negativos, 11% das publicações deviam ter sido filtradas como verdadeiros positivos.

Por fim, temos os verdadeiros positivos, que representam 29% de publicações corretamente filtradas como relacionados com o evento.

Neste exemplo podemos observar que os resultados obtidos foram melhores relativamente aos anteriores, confirmando que a junção de "palavra1" ou "palavra2" consegue ter melhores resultados em termos de verdadeiros positivos do que a utilização de apenas uma palavra e do que com a junção "e".

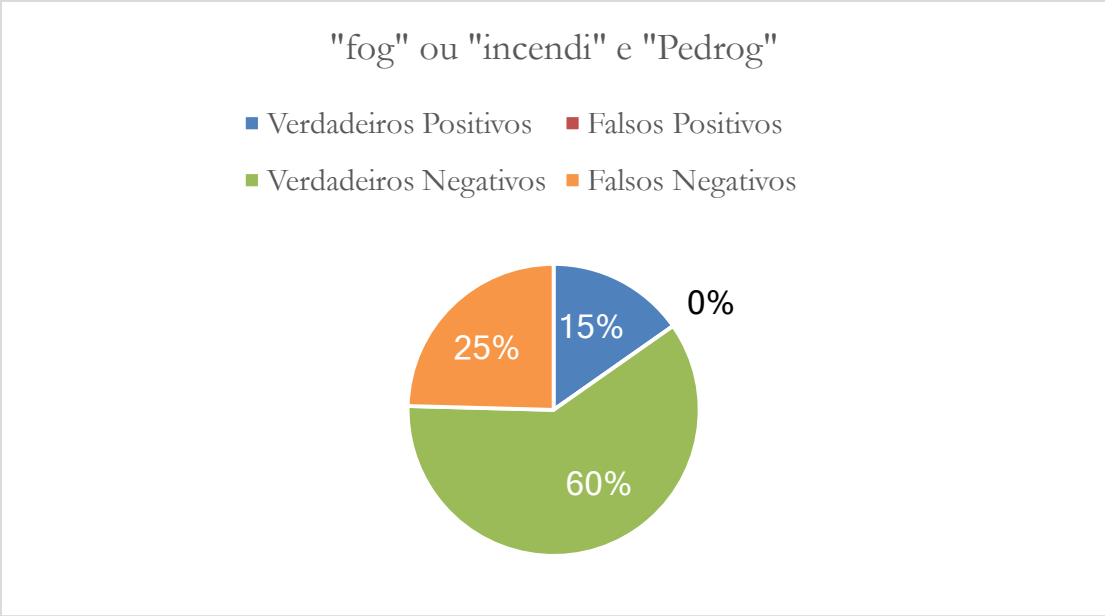

Figura 14: Resultados da palavra filtrada "fog" ou "incendi" e "pedrog" do evento Incêndio de Pedrógão de 2017

Por fim, no que diz respeito à utilização de topónimo para se conseguir localizar o evento, obtivemos 15% de verdadeiros positivos, um número maior do que o teste anterior dos incêndios de outubro de 2017 (figura 9), em que apenas se obteve 2% de verdadeiros positivos, demonstrando que dependendo do evento e do tipo de publicações, nem sempre os topónimos conseguem ser bons métodos de filtragem.

Relativamente aos verdadeiros negativos, 60% das publicações foram filtradas corretamente, enquanto que 25% representa falsos negativos, que deveriam ter sido filtrados como verdadeiros positivos.

4.2. Caso de estudo: Furacão Leslie em 2018

O próximo evento serve como caso de estudo para demonstrar a possibilidade de utilização desta metodologia para outro tipo de eventos, para além dos incêndios, como é o caso do Furacão Leslie, tendo acontecido em outubro de 2018, que afetou Portugal no período de tempo de 13 de outubro a 14 de outubro.

As duas palavras termo escolhidas para designar a filtragem foram "leslie" e "furacão", sendo ambas diretamente relacionadas com o evento e facilmente filtradas pela metodologia, identificando facilmente as publicações relacionadas com o evento.

Evento: Furacão Leslie em 2018
Período de Tempo: 13 outubro 2018 – 14 outubro 2018
Total de publicações efetivo no período de tempo: 1898
Número total de publicações escolhidas aleatoriamente inseridos no número total de publicações: 118

Tabela 6 - Resultados Furacão Leslie 2018

Método	Processo automático				Referência (Classificação Manual)	
	Filtrados (+)		Não Filtrados (-)			
	Devidamente (base 40)	Indevidamente (base 78)	Devidamente (base 78)	Indevidamente (base 40)	Positivos	Negativos
'lesli' (288)	37	0	78	3	40	78
'furaca' (151)	18	1	77	21		
'lesli' e 'furaca' (18)	18	0	78	22		
'lesli' ou 'furaca' (38)	37	1	77	3		
"lesli" ou "furaca" e "Coimbr" (9)	9	0	78	31		

Na tabela 6 acima representada, podemos observar que no evento "Furacão Leslie", filtrando a palavra "lesli", obtemos 288 resultados de um total de 1898 no período de tempo do evento, desses quais o processo automático filtrou como verdadeiros positivos 37 publicações das 40 de referência (classificados manualmente), um resultado extremamente positivo. Relativamente aos verdadeiros negativos, nenhuma publicação foi filtrada incorretamente como sendo relacionada ao evento, apresentando uma taxa de sucesso de 100%.

Por outro lado, quando filtrada, essa mesma palavra apresentou como falsos positivos 78 publicações das 79 de base de referência manual, e apresentou como falsos negativos 3 publicações das 40 de base, ou seja, não continham a palavra "lesli".

Filtrando uma outra palavra, também relacionada com o evento, "furaca", obtemos 151 resultados dos 1898 totais no período de tempo do evento, sendo que desses resultados o processo automático filtrou como sendo verdadeiros positivos 18 publicações das 40 de base, um número bastante inferior que a palavra "lesli".

Contudo, o processo automático filtrou uma publicação como sendo verdadeiro negativo, que corresponde a um caso extraordinário de uma publicação que se refere a um furacão que no mesmo período de tempo aconteceu nos Estados Unidos da América, sendo que um jornal online referenciou esta notícia. Ao descobrir este caso é possível afinar a metodologia, filtrando apenas noticias sobre eventos em Portugal.

No entanto, o processo automático classificou como falsos positivos 77 publicações dos 78 de base, sendo que a publicação em falta foi filtrada como sendo relacionada com o evento incorretamente. A respeito dos falsos negativos, 21 das publicações das 40 de base foram filtradas pois não possuem a palavra "furaca" na publicação, o que relativamente à observação anterior, apresenta uma taxa de insucesso mais elevada.

Ao filtrarmos mutuamente as palavras "lesli" e "furaca", o processo automático apresenta-nos como verdadeiros positivos 18 publicações das 40 de base, e como verdadeiros falsos 0 publicações das 78 de base. Em termos de falsos positivos 78 publicações de 78 foram filtradas corretamente como não sendo relacionadas com o evento. Em relação aos falsos negativos 21 de 40 de base foram filtradas.

Ao filtrarmos as palavras "lesli" ou "furaca", o processo automático apresenta-nos como verdadeiros positivos 37 publicações de 40, (sendo assim a taxa de sucesso mais elevada de entre todos os casos observados nesta tabela), e como verdadeiros negativos 1 de 79 de base, que se refere ao furacão nos Estados Unidos da América. Consequentemente o processo automático

filtrou como falsos positivos 77 publicações das 78 de base, e filtrou como falsos negativos 3 publicações das 40 de base, igual à filtragem com apenas a palavra "lesli", o que demonstra que dependendo do evento, é possível ter resultados melhores tendo em conta a palavra de interesse em uso. Visto que este evento é um evento muito específico com um nome concreto, facilmente se identificam publicações relacionadas com o evento utilizando o próprio nome do Furacão para filtragem automática.

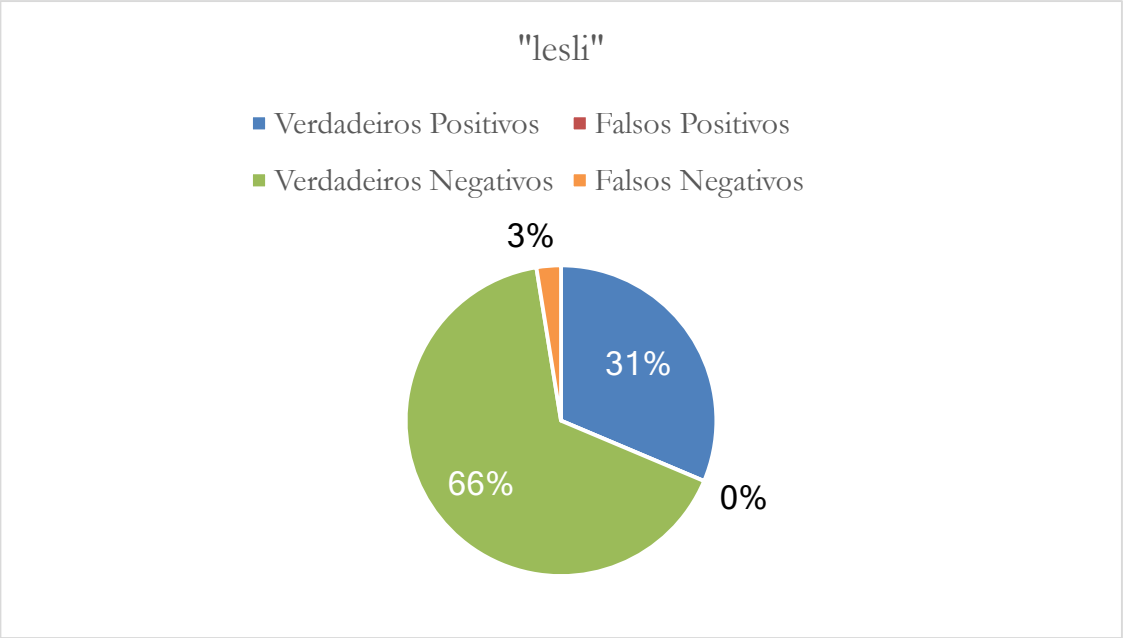

Figura 15: Resultados da palavra filtrada "lesli" do evento Furacão Leslie em 2018

Na figura 15, podemos observar as percentagens relativamente à palavra de interesse "lesli" relativamente ao evento Furacão Leslie em 2018.

A respeito de verdadeiros positivos, temos 31% de publicações filtradas corretamente, e apenas 3% de falsos negativos que não foram filtradas corretamente.

No entanto, o número de verdadeiros negativos é de 66%, correspondendo a 100% de publicações filtradas corretamente não sendo relacionadas com o evento.

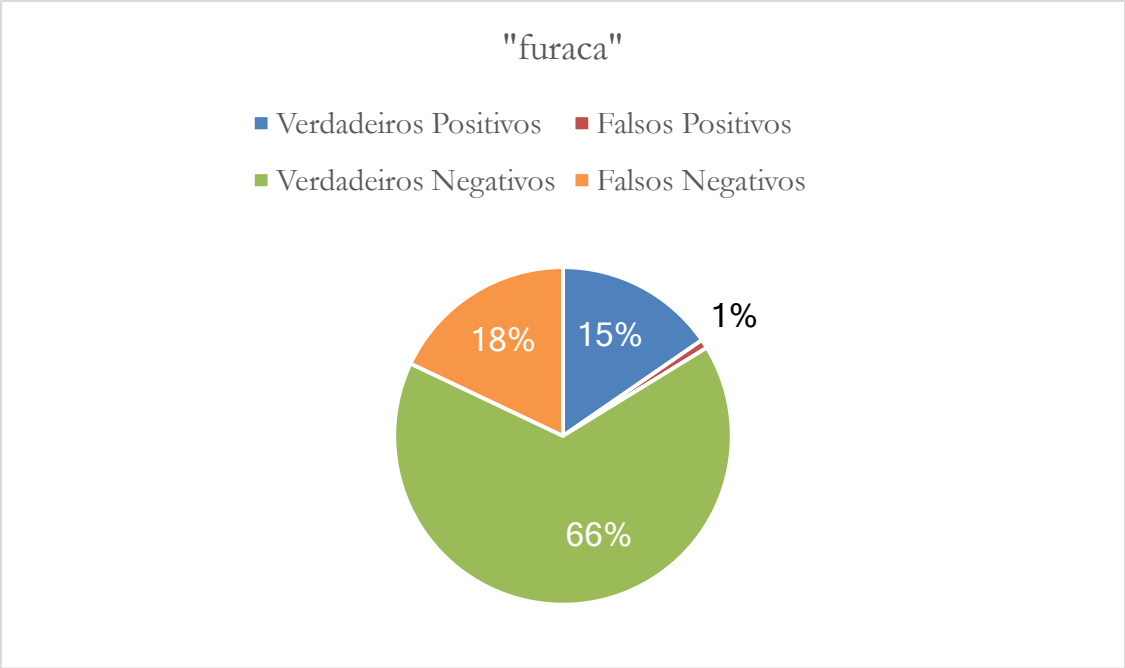

Figura 16: Resultados da palavra filtrada "furaca" do evento Furacão Leslie em 2018

À semelhança da figura 15, a figura 16 representa a percentagem de filtragens da palavra de interesse "furaca".

Em relação aos verdadeiros positivos, obtivemos 15% de filtragens corretas e 66% de verdadeiros negativos.

No entanto, 18% corresponde a falsos negativos, que deveriam ter sido relacionados com o evento, mas como não possuem a palavra de interesse não foram filtrados, enquanto que 1% corresponde aos verdadeiros negativos.

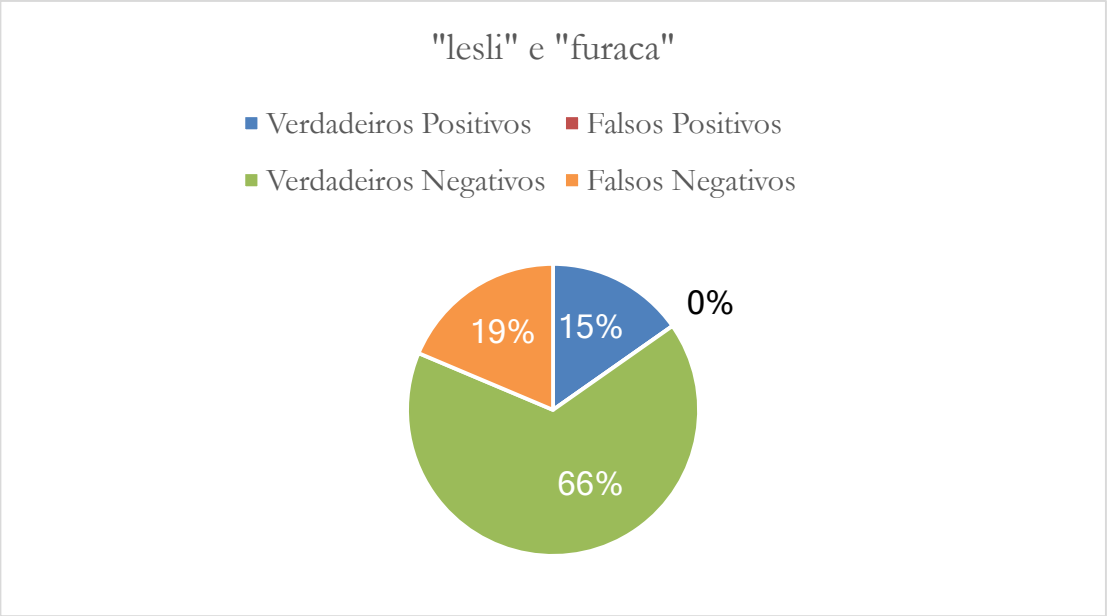

Figura 17: Resultados da palavra filtrada "lesli" e "furaca" do evento Furacão Leslie em 2018

Na figura 17, podemos observar a junção das duas palavras de interesse "lesli" e "furaca", obtendo 15% de publicações verdadeiras positivas filtradas corretamente. Em relação aos falsos negativos, 19% das publicações foram filtradas incorretamente como não sendo relacionadas com o evento, sendo que estas publicações não possuem a junção "lesli" e "furaca" na publicação.

No que diz respeito aos verdadeiros negativos, 66% publicações foram filtradas corretamente como não sendo relacionadas com o evento.

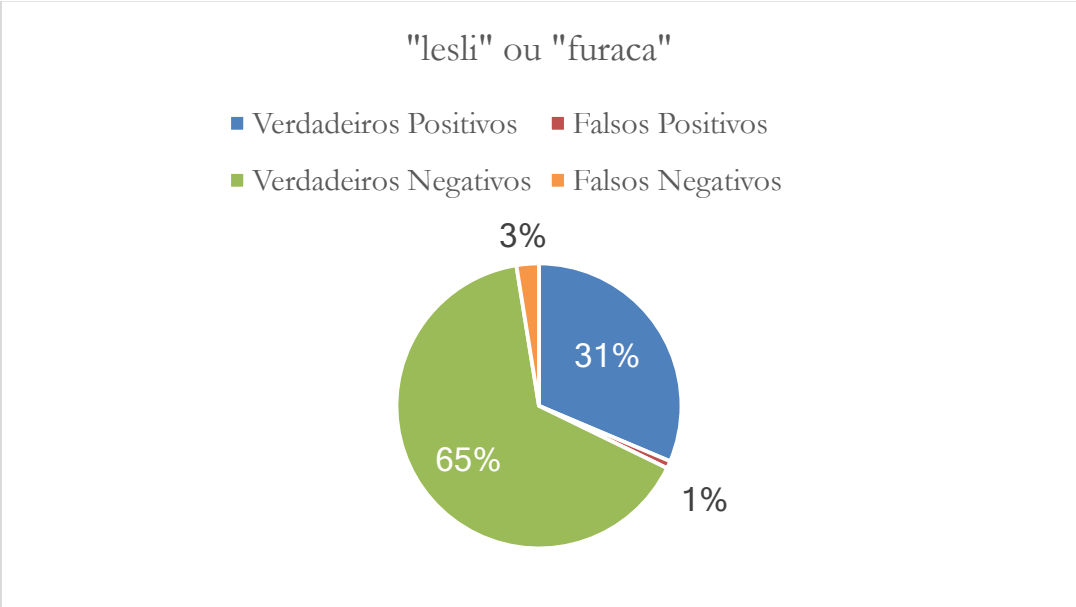

Figura 18: Resultados da palavra filtrada "lesli" ou "furaca" do evento Furacão Leslie em 2018

Na figura 18, podemos observar as percentagens relativamente à junção das palavras de interesse "lesli" ou "furaca".

Observando a figura, temos 31% de verdadeiros positivos, no entanto no que diz respeito aos falsos negativos, apenas 3% das publicações deviam ter sido filtradas como verdadeiros positivos.

Por fim, temos os verdadeiros negativos, corretamente filtrados como não relacionados com o evento, representando exatamente 66%.

Este caso é igual ao da figura 15, só que neste caso existe 1% de verdadeiros negativos, referente à publicação do furacão nos Estados Unidos da América.

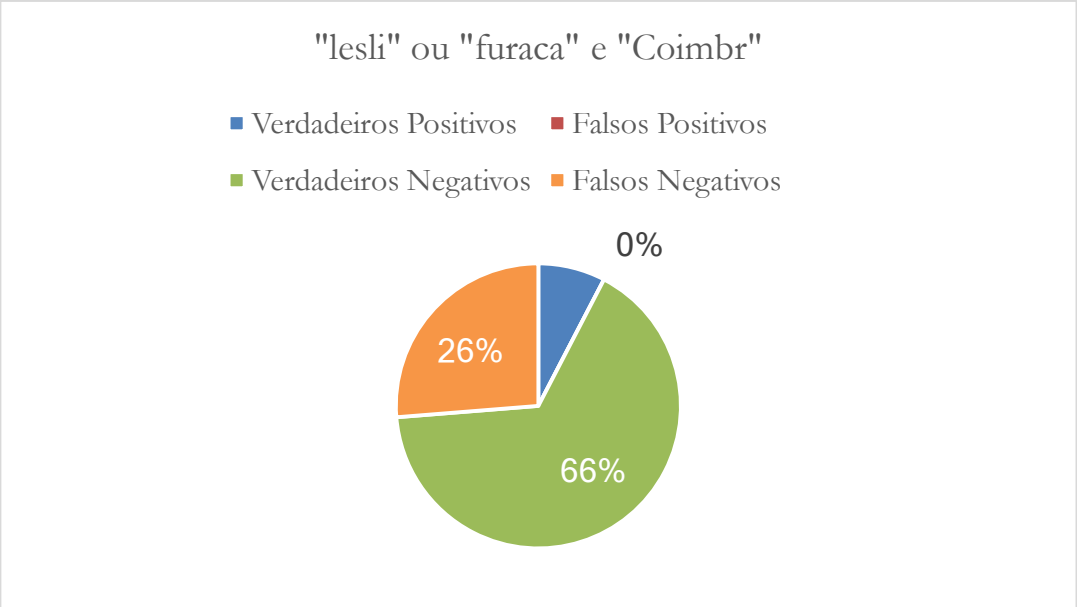

Figura 19: Resultados da palavra filtrada "lesli" ou "furaca" e "coimbr" do evento Furacão Leslie em 2018

Relativamente à figura 19, onde evidenciamos a utilização das duas palavras de interesse "fog" ou "furaca" e o topónimo "coimbr", relativamente ao evento Furacão Leslie em 2018, observamos apenas 8% de verdadeiros positivos filtrados como sendo relacionados com o evento em relação à base de referência manual. Contudo, no que diz respeito aos falsos negativos, 26% das publicações foram filtradas como não relacionadas, devido ao facto de que não possuem o topónimo da publicação.

Por fim, podemos observar 66% das publicações filtradas corretamente como verdadeiros negativos, ou seja, não relacionadas com o evento.

Após estes testes, é evidente que, dependendo do evento em destaque, utilizar apenas uma palavra de interesse ou a junção de "palavra1" e "palavra2" ou "palavra1" ou "palavra2" e "topónimo" são as opções que obtém menores resultados positivos, daí os próximos testes apenas utilizarem o método "palavra1" ou "palavra2" como macro validação do evento e como teste adicional, utilizar também o método de adição do topónimo para micro validação para uma localização mais específica.

4.3. Caso de estudo: Amostra orientada de referência

Visto que os resultados dos casos de estudo anteriores não correspondiam a eventos em períodos de tempo reduzidos, gerou-se uma amostra de treino, conforme acima referido no capítulo 3.7, sendo esta construída com o objetivo de testar a metodologia com uma amostra de treino maior de modo a obter um termo comparativo com os casos de estudo anteriores, focando os testes em eventos de incêndios, como nos testes anteriores (à exceção do furacão Leslie).

Neste teste final, a junção 'fog' e 'incendi' não é utilizada porque a quantidade filtrada não é relevante, confirmando-se pelos testes anteriores, utilizando-se apenas a junção 'fogo' ou 'incêndio', que revelou ter resultados mais positivos em relação a palavras de interesse isoladas ou com a junção "e", assim como esta junção de ambas as palavras, com a adição de um topónimo, como foi referido anteriormente.

É importante relembrar que ao criar esta amostra, gerou-se também um novo critério (critério 2) como referido no capítulo 3.6, com o objetivo de restringir a validação para publicações diretamente relacionadas com o evento, assim como num curto prazo de tempo após o evento acontecer.

Evento: Amostra orientada (critério 1)

Período de Tempo: 1 de junho de 2017 – 31 de outubro de 2018

Total de publicações efetivo no período de tempo: 844

Tabela 7 – Amostra orientada (Critério 1)

Método	Processo automático				Referência (Classificação Manual)	
	Filtrados (+)		Não Filtrados (-)			
	Devidamente (base 233)	Indevidamente (base 611)	Devidamente (base 611)	Indevidamente (base 233)	Positivos	Negativos
'incendi" ou 'fog" (233)	194	39	572	39	233	611
'incendi" ou 'fog" e "Coimbr" (19)	18	1	571	215		

Em primeiro lugar, podemos observar o primeiro teste com a amostra orientada utilizando o critério 1 de classificação, classificando manualmente todas as publicações que sejam diretamente ou indiretamente relacionadas com o evento.

Ao filtrarmos a combinação de "incendi" ou "fog", podemos observar que existem 233 publicações com uma das duas palavras de interesse inseridas nas 844 publicações da amostra total, sendo que 233 publicações também foram classificadas manualmente como relacionadas com o evento de incêndio utilizando o critério 1.

O processo automático filtrou como verdadeiros positivos 194 publicações dos 233 de base de referência (classificados manualmente), e filtrou como verdadeiros negativos 39, ou seja, 39 publicações contêm a palavra, mas não são relacionadas com o evento (por exemplo: publicações que contenham palavras como "fogo-de-artificio" são filtradas, havendo possibilidade de afinar a metodologia para não considerar certos conjuntos de palavras).

Relativamente aos falsos positivos, foram filtradas pelo processo automático 572 publicações das 611 de base de referência manual.

Concluindo, com o processo automático, os falsos negativos contabilizados foram de 39 em 233 da base de referência manual, ou seja, 39 publicações relacionadas com o evento deveriam ter sido filtradas como verdadeiros positivos.

No que diz respeito ao uso da junção das palavras "fog" ou "incendi" e "coimbr" temos um número de resultados menor relativamente ao anterior, sendo que apenas 18 publicações foram filtradas como sendo relacionadas (verdadeiros positivos) com o evento e com o topónimo em 233. No que diz respeito aos verdadeiros negativos, uma publicação foi filtrada, possuindo o topónimo da publicação, no entanto sem qualquer relevância para o evento.

Relativamente aos falsos positivos, foram filtradas 571 publicações corretamente como não sendo relacionadas com o evento, e por fim, nos falsos negativos foram filtradas 215 publicações como não sendo relacionadas com o evento e com o topónimo.

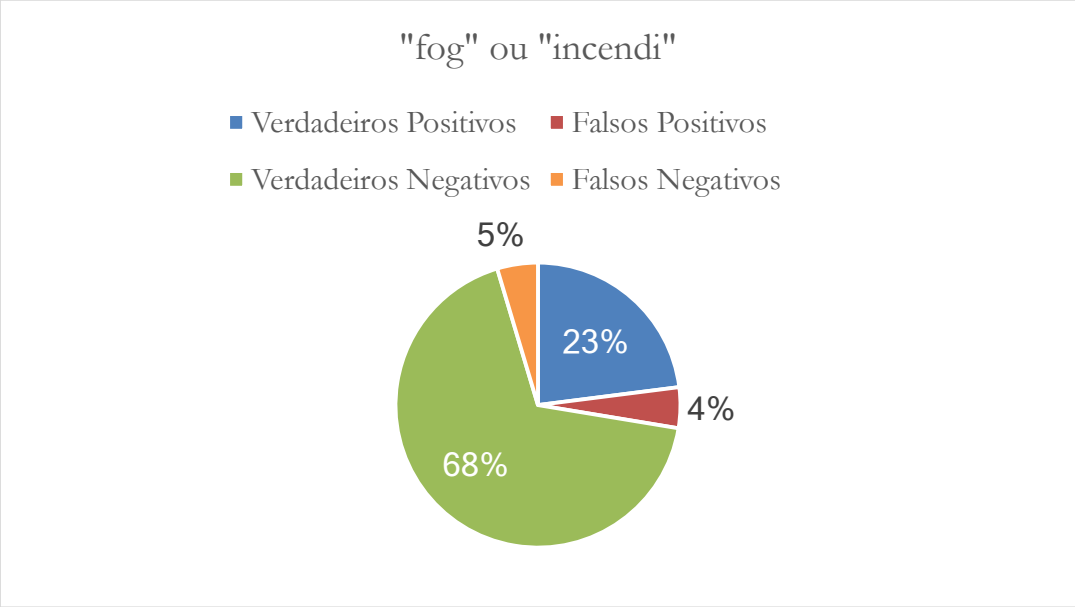

Figura 20: Resultados da palavra filtrada "fog" ou "incendi" da amostra orientada (critério 1)

Na figura 20, podemos observar as percentagens relativamente à junção das palavras de interesse "fog" ou "incendi".

Observando a figura, verificamos que o número de verdadeiros positivos é reduzido, de apenas 23%, enquanto que os falsos negativos são de 5%.

Em relação aos verdadeiros negativos, temo um valor de 68%, que significa que praticamente todas as publicações que deveriam ter sido filtradas como não relacionadas com o evento, foram corretamente filtradas.

Por fim, temos 4% de falsos positivos, que corresponde a publicações filtradas incorretamente como sendo relacionadas com o evento.

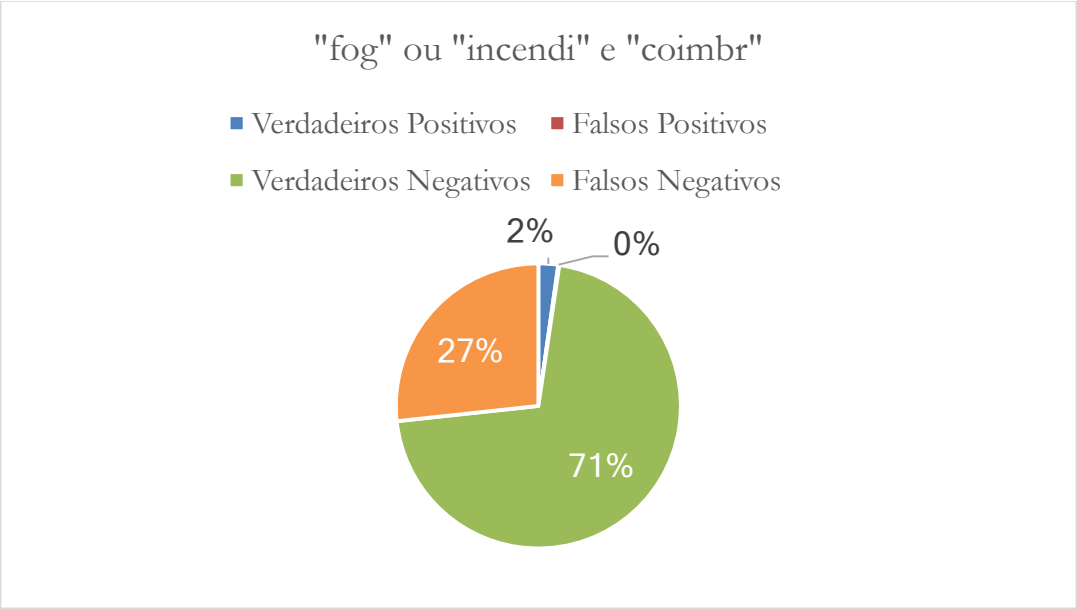

Figura 21: Resultados da palavra filtrada "fog" ou "incendi" e "coimbr" da amostra orientada (critério 1)

Relativamente à utilização das palavras de interesse em conjunto com um topónimo, podemos observar a figura 21, que representa os resultados percentuais deste método utilizado.

Como é de esperar, a percentagem de verdadeiros positivos filtrados é residual, apenas 2% de publicações filtradas como relacionadas com o evento. No que diz respeito aos falsos negativos (incorretamente filtrados como não sendo relacionados com o evento), temos 27%.

Em relação aos verdadeiros positivos, a percentagem é de 71%, o que significa que a metodologia consegue filtrar publicações não relacionadas com o evento com uma boa eficácia. Concluindo, apesar de a tabela 7 revelar 1 publicação como verdadeiro negativo, a percentagem é residual e apresenta-se como 0%.

Evento: Amostra orientada (critério 2)

Período de Tempo: 1 de junho de 2017 – 31 de outubro de 2018

Total de publicações efetivo no período de tempo: 844

Tabela 8 – Amostra orientada (Critério 2)

Método	Processo automático				Referência (Classificação Manual)	
	Filtrados (Verdadeiros)		Não Filtrados (Falsos)			
	Positivos (base 81)	Negativos (base 763)	Positivos (base 763)	Negativos (base 81)	Positivos	Negativos
'incendi" ou 'fog" (233)	70	163	600	11	81	763
"incendi" ou "fog" e "Coimbr" (19)	13	6	757	68		

Procedendo ao segundo teste com a utilização do segundo critério em que cada publicação é classificada manualmente, sendo diretamente relacionada com o evento ou não, podemos observar a partir da tabela 8, a filtragem da combinação das palavras "incendi" ou "fog".

Este processo automático de classificação filtrou como verdadeiros positivos 70 publicações das 81 de base de referência (classificados manualmente), ou seja, em 81 publicações classificadas manualmente como relevantes ao evento, 70 foram filtradas corretamente, enquanto que 11 não foram filtradas (falsos negativos). Uma das razões para existirem falsos negativos é pela simples razão de estas publicações não conterem nenhuma das palavras de interesse utilizadas para a filtragem.

Seguidamente, o processo automático filtrou como falsos positivos 600 publicações das 763 de base de referência, um resultados que poderia ser melhor, no entanto a amostra contém muitas publicações com palavras como "fogo-de-artificio" ou "fogueira" o que vai implicar um desajuste nos resultados, existindo a necessidade de adicionar futuramente um parâmetro que não filtre publicações com certas palavras, e obviamente como o critério 2 a ser utilizado apenas valida publicações de relevância direta com o evento que esteja a acontecer, todas as publicações que

contenham uma das palavras de interesse e que não estejam classificadas como relacionadas com o evento vão ser filtradas como verdadeiros negativos. O mesmo acontece com publicações irrelevantes, que apesar de conterem uma das palavras de interesse, são filtradas erradamente podendo ser, por exemplo, publicações sobre homenagem a bombeiros, ou concertos de solidariedade às vitimas dos incêndios, ou até noticias realizadas meses após o evento ter acontecido, como estas publicações contêm uma das palavras de interesse, são filtradas incorretamente, originando um grande número de verdadeiros negativos, como é possível observar na tabela 8.

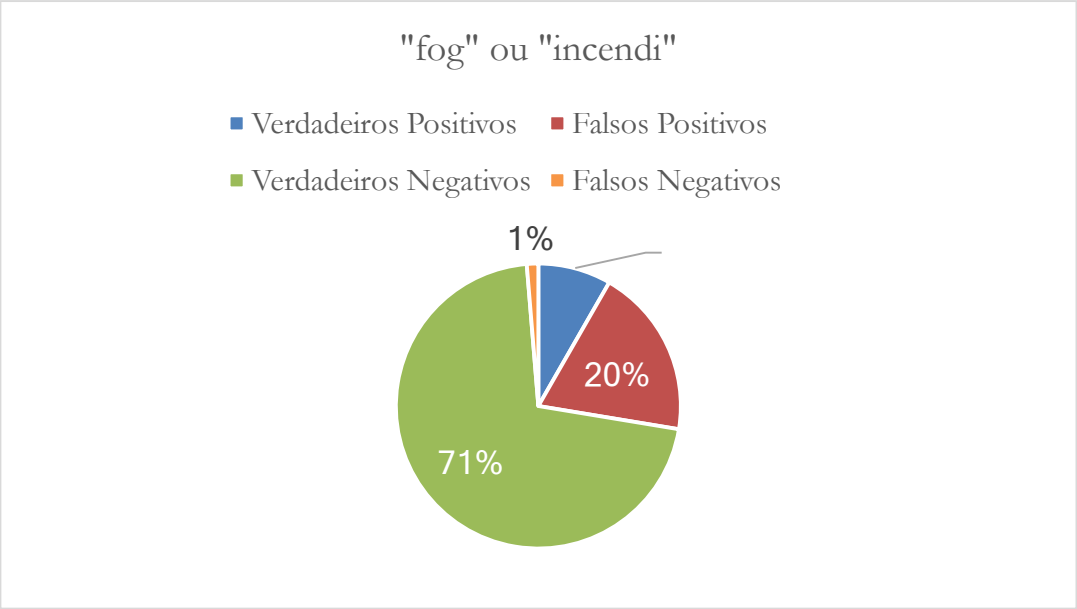

Figura 22: Resultados da palavra filtrada "fog" ou "incendi" da amostra orientada (critério 2)

Relativamente à figura 22, obtemos os resultados relativos ao uso das palavras de interesse "fog" ou "incendi", obtendo 71% de verdadeiros negativos, enquanto que nos falsos positivos obtemos uma percentagem relativamente (20%) grande tendo em conta a base de referência de 81 validados manualmente como sendo relevantes ao evento.

Em relação aos verdadeiros positivos, temos 8% no geral, no entanto como foi possível observar pela tabela 8, 70 publicações em 81 foram filtradas como sendo relacionadas com o evento, o que revela ser um bom método, no entanto é necessário ajustar para não existirem tantos verdadeiros negativos e para a redução de falsos negativos, que passará pela afinação da metodologia em trabalhos futuros, como foi referido anteriormente.

Contudo, nos falsos positivos, temos a percentagem de 20%, que poderia ser melhor pois foram filtrados 20% de posts indevidamente, e 1% de falsos negativos.

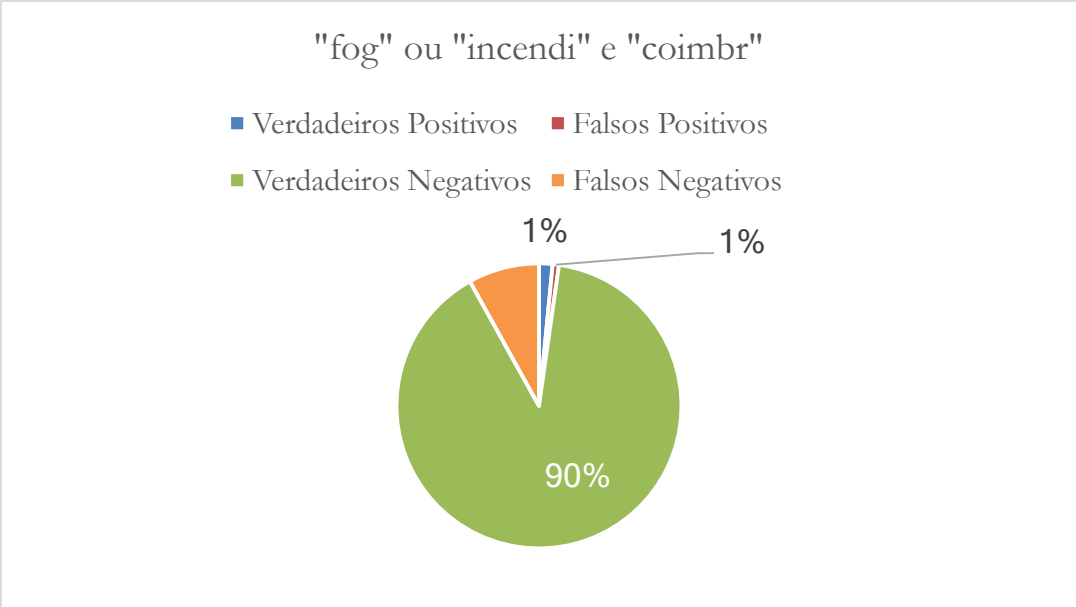

Figura 23: Resultados da palavra filtrada "fog" ou "incendi" da amostra orientada (critério 2)

Finalizando com o último teste, com a utilização das palavras de interesse em conjunto com um topónimo, podemos observar a figura 23, que representa os resultados percentuais deste método utilizado utilizando o critério 2.

A percentagem de verdadeiros positivos filtrados é muito baixa, visto que apenas foram filtradas 13 publicações de acordo com a tabela 8, obtendo apenas o valor de 1% de publicações filtradas como relacionadas com o evento.

No que diz respeito aos falsos negativos (incorretamente filtrados como não sendo relacionados com o evento), obtemos uma percentagem maior, de 8%.

Em relação aos verdadeiros negativos, a percentagem é de 90%, o que significa que a metodologia consegue filtrar publicações não relacionadas com o evento com uma boa eficácia utilizando um topónimo, como já foi confirmado pelo teste anterior utilizando o critério 1, contudo, neste caso o valor subiu de 71% para 90%.

Concluindo com os falsos positivos, apenas 1% (6) das publicações foram filtradas incorretamente como sendo relacionadas com o evento.

5. Conclusões

Apesar dos resultados apurados não conseguirem ter uma capacidade de validação muito boa, este é um conceito que apesar de estar num estado embrionário, pode vir a ser melhorado para conseguir ter melhores resultados, no entanto, existem metodologias atualmente que obtêm melhores resultados, como é o caso do "machine learning", utilizando uma grande amostra de treino para a classificação de texto.

Não foram feitos testes com mais palavras (por exemplo: bombeiro ou queimados) porque teoricamente poderia aumentar o número de verdadeiros positivos, no entanto a sua adição também não garantiria um aumento dos verdadeiros negativos (como um caso de exemplo, a adição da palavra "bombeiro" poderia ser o caso de existir uma publicação: "bombeiros fazem juramento de bandeira").

A utilização de apenas uma palavra de interesse para filtragem não é aconselhada, a não ser que sejam eventos muito específicos, como é o caso do furacão Leslie em que o próprio nome do evento pode ser utilizado como método de filtragem.

Exclui-se também a possibilidade de utilizar "palavra1" e "palavra2" devido ao seu valor muito baixo de filtragens, como foi possível observar nos casos de estudo.

Respetivamente à utilização de um topónimo para a micro validação, no estado atual desta metodologia, é aconselhável apenas que se utilize em casos concretos para a filtrar uma localização em concreto, já que o número de publicações com topónimos é muito reduzido para obter um grande número de contributos sobre o evento.

Fica comprovado que a utilização da combinação ("palavra1" ou "palavra2") tem melhores resultados, como é de esperar, do que somente a uma palavra, pois abrange um maior número de publicações relacionadas com o evento pretendido.

Esta metodologia baseia-se numa simples inquirição SQL à base de dados, sem estar suportada necessariamente por um algoritmo mais complexo - como é o caso da metodologia de Laylavi que esteve na base da motivação deste trabalho. Com a metodologia aqui desenvolvida, conseguem-se, de facto, resultados globalmente positivos, ainda assim com uma amostra de treino que acabou por se revelar insuficiente. Na verdade, e após os diferentes testes, percebemos que o desempenho desta metodologia pode ainda ser melhorada com uma amostra de treino mais bem construída, de uma forma estratificada.

É uma metodologia com flexibilidade de plataformas total (Facebook, Twitter, flickr, etc), apenas necessita de base de dados com os dados recolhidos e armazenados disponível.

No seu estado atual, inviabiliza-se a utilização deste método para uma utilização em tempo real porque cada evento obriga uma amostra de referência afinada para cada evento, assim como uma escolha de palavras de interesse pré-definidas para outros eventos para além de incêndios.

Tendo em conta os eventos catastróficos que acontecem todos anos em Portugal, este projeto é sem dúvida um projeto com potencial, visto que a fonte de informação não se rege apenas por uma fonte de informação, mas sim por mais redes sociais, assim como notícias de jornais online e também sensores físicos, sendo a multipolaridade deste conceito uma mais valia.

Concluindo, a exploração deste tema foi um fator de enriquecimento intelectual, não só pelo tema em si, mas também pelo seu objetivo, a mitigação de situações de emergência. Uma vez que, infelizmente, também vivenciei a situação de perigo dos incêndios de 2017, sem a possibilidade de contactar fontes exteriores, tendo um sistema desta dimensão disponível poderia, de certa forma, ajudar as entidades de proteção civil a mitigar uma situação de emergência e a obter a melhor tomada de ação possível antes de chegarem ao local, podendo-se deste modo evitar situações mais drásticas.

5.1. Trabalho Futuro

Apesar de se provar que o conceito é capaz de funcionar, ainda existe muito que pode ser melhorado e afinado: a adição de mais palavras de interesse, a adição de palavras de desinteresse para evitar que publicações não desejadas sejam filtradas (por exemplo publicações com a palavra fogo-de-artificio, sendo que à partida se referirá a uma festa popular, não terá importância para uma situação de catástrofe, reduzindo o número de verdadeiros negativos).

Incorporar um serviço de deteção manual por parte da proteção civil, permitindo ao utilizador escolher uma ou mais palavras de interesse, assim como a escolha de topónimos, se assim for necessário para localizar um evento. Deste modo, a entidade utilizadora deste serviço terá um controlo mais abrangente em relação à informação que pretende filtrar.

Um dos pontos de interesse futuros passa pela construção de uma Geotimeline (linha do tempo) com filtragem das publicações em tempo real à medida que são recolhidas (em caso de acontecimento de uma catástrofe) e utilização de componente espacial para a localização dos eventos (mapas), possibilitando também a análise de eventos acontecidos no passado.

Um outro ponto de interesse destaca-se pela possibilidade de automatizar o conceito de modo a que este consiga perceber quando um evento começou (por exemplo a partir de várias

palavras de interesse repetidas na recolha de publicações), começando assim um evento chave, começando a filtrar publicações de interesse relevantes ao evento que esteja a acontecer.

Bibliografia

BRUNS, A., & BURGESS, J. E. (2012). Local and global responses to disaster:# eqnz and the Christchurch earthquake. Paper presented at the Disaster and Emergency Management Conference.

BRUNS, A., DR KATRIN WELLER, D., BORRA, E., & RIEDER, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. Aslib Journal of Information Management, 66(3), 262-278.

FONTES, D.; FONTE, C.; CARDOSO, A. DE ALMEIDA J.-P.; ESTIMA, J. (2017): "A platform to integrate crowdsourced, pysical sensor and oficial geographic information to assist authorities in emergency response. 20th International Conference on Geo-Information Science (AGILE 2017). Wageningen, The Netherlands.

HAN, J., KAMBER, M., & PEI, J. (2011). Data mining: concepts and techniques: elsevier.

HANSEN, D., SHNEIDERMAN, B., & SMITH, M. A. (2010). Analyzing social media networks with nodexl: Insights from a connected world: Morgan Kaufmann.

INTERNET WORLD STATS (2019). Página de web, https://www.internetworldstats.com/stats.htm acedido a 20 de agosto de 2019.

LARSSON, A. O., & MOE, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New media & society, 14(5), 729-747.

LAYLAVI, F. (2016). A framework for adopting twitter data in emergency response (phd thesis) centre for disaster management and public safety, department of infrastructure engineering, school of engineering, the university of melbourne, victoria, australia. Doi: http://hdl.handle.net/11343/129095.

LAYLAVI, F., RAJABIFARD, A., & KALANTARI, M. (2017). Event relatedness assessment of twitter messages for emergency response. Information processing & management, 53(1). Doi: http://dx.doi.org/10.1016/j.ipm.2016.09.002.

MACHINE LEARNING MASTERY (2017) Página de Web, https://machinelearningmastery.com/clean-text-machine-learning-python/acedido a 10 de Agosto de 2019.

MATTSSON, M., & BOSCH, J. (1997). Framework composition: problems, causes and solutions. Paper presented at the proceedings of the international conference on technology of object oriented systems and languages.

MEDIUM (2017) Página de Web, https://medium.com/carwow-product-engineering/sql-vs-pandas-how-to-balance-tasks-between-server-and-client-side-9e2f6c95677 acedido a 22 de Agosto de 2019.

MORRIS, M. R.; TEEVAN, J. (2010): Collaborative Web Search: Who, What, Where, When and Why. Morgan & Claypool Publishers.

TRICE, M. (2015). Putting gamergate in context: how group documentation informs social media activity. Paper presented at the Proceedings of the 33rd Annual International Conference on the Design of Communication.

YEP, J., & SHULMAN, J. (2014). Analyzing the library's Twitter network Using nodexl to visualize impact. College & Research Libraries News, 75(4), 177-186.

ZHANG, S., ZHANG, C., & YANG, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17(5-6), 375-381.

Anexos

Anexo A: Código Desenvolvido em Python – Pré-processamento Inicial

Anexo B: Query, pré-processamento e filtragem SQL

Anexo C: Script de validação Excel (method_evaluation.ipynb)

Anexo D: CD com todo o conteúdo utilizado

Anexo A: Código Desenvolvido em Python – Pré-processamento Inicial

```
import unicodedata, string, nltk, re
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords

def preprocess(sentence):
    # letras minusculas + Tokenizar + remover stopwords
    sentence = sentence.lower()
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(sentence)
    filtered_words = [w for w in tokens if not w in stopwords.words('portuguese')]
    #sentence = re.sub(r'\s+', ' ', sentence, flags=re.I) #Remove espacos a mais
    return " ".join(filtered_words)

def remove_accents(sentence):
    # Remover acentuacao
    nfkd_form = unicodedata.normalize('NFKD', sentence).decode('utf8')
    only_ascii = nfkd_form.encode('ASCII', 'ignore')
    return only_ascii

def clean(df, column):
    # Usar os 2 metodos na coluna com info pretendida
    df[column] = df[column].str.replace('http\S+|www.\S+', '', case=False)
    df[column] = df[column].apply(preprocess).apply(remove_accents)
    return df

def Train(df,word):
    df = df.set_index('order_expression').filter(like = word, axis=0)
    df = df.set_index('fid')
    return df
```

Anexo B: Query, pré-processamento e filtragem SQL

```
#Parametros
Periodo = ['2017-06-01 00:00:00', '2018-10-31 23:59:59']
Palavras = ['incendi', 'fog']
out_file = '/Users/Foxxy/Desktop/file_res.xls'

queryfb = (
   "(SELECT *, to_tsvector('portuguese', lower("
        "unaccent("
         "regexp_replace(regexp_replace(message, 'http://[^:\s]+(\S+)',' ', 'g'), '[^\w]+',' ',
'g')"
          ")"
   ")) AS clean_message FROM ("
      "SELECT post_id, datahora, type, "
      "CASE "
      "WHEN type = 'link' "
      "THEN (unaccent(description)) "
      "ELSE lower(unaccent(message)) "
      "END AS message "
      "FROM facedata "
      "WHERE TO_TIMESTAMP(datahora, 'YYYY-MM-DD HH24:MI:SS') >
TO_TIMESTAMP('{}', "
      "'YYYY-MM-DD HH24:MI:SS') AND TO_TIMESTAMP(datahora, 'YYYY-MM-
DD HH24:MI:SS') "
      "< TO_TIMESTAMP('{}', 'YYYY-MM-DD HH24:MI:SS') "
   ") AS foo WHERE message IS NOT NULL AND ({whr})) AS stop_table"
).format(Periodo[0],Periodo[1], whr=" OR ".join([
   "message LIKE '%%{}%%'".format(w) for w in Palavras]))

queryfb2 = (
   "SELECT fid, stop_table.message, datahora, ARRAY_TO_STRING(array_agg("
      "word ORDER BY word_index), ' ', '*') AS order_expression, "
```

```
        "REPLACE(CAST(STRIP("
            "stop_table.clean_message) AS text), '', '') AS no_duplicated "
        "FROM ("
            "SELECT fid, word, CAST(UNNEST(word_index) AS integer) AS word_index
FROM ("
                "SELECT fid, SPLIT_PART(tst, ';', 1) AS word, "
                "STRING_TO_ARRAY(SPLIT_PART(tst, ';', 2), ',') AS word_index FROM ("
                    "SELECT post_id AS fid,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE("
                        "CAST(UNNEST(clean_message) AS text), "
                            "'',{{', ',\"{{'), ',\"{{', ';'), '}}\"', ''), "
                            "'('', ''), '}}', '') AS tst "
                    "FROM {tbl}"
                ") AS foo"
            ") AS foo2"
        ") AS foo3 INNER JOIN {tbl} ON foo3.fid = stop_table.post_id "
        "GROUP BY foo3.fid, stop_table.message, stop_table.clean_message, datahora"

    ).format(tbl=queryfb)
```

Anexo C: Script de validação Excel (method_evaluation.ipynb)

```
import numpy
import pandas
from gasp.fm import tbl_to_obj
from gasp.to import obj_to_tbl

out_file = r'exemplo.xlsx'
ref_data = r'…\file_ref.xlsx'
tst_data = r'…\file_res.xlsx'

ref_df = tbl_to_obj(ref_data)
tst_df = tbl_to_obj(tst_data)

df = ref_df.merge(tst_df, how='left', left_on='post_id', right_on='fid')
df['fid'].fillna('Nada', inplace=True)
df['cls_incendio'] = numpy.where(df.fid == 'Nada', 0, 1)
df['confusao'] = numpy.where(
   (df.is_incendio_v2 == 1) & (df.cls_incendio == 1), 'VP', numpy.where(
      (df.is_incendio_v2 == 0) & (df.cls_incendio == 0), 'FP', numpy.where(
         (df.is_incendio_v2 == 1) & (df.cls_incendio == 0), 'FN', 'VN'
      )
   ))
print obj_to_tbl(df, out_file)

#Accuracy
confusion_tbl = pandas.DataFrame()
confusion_tbl['number_of_rows'] = df.groupby(['confusao']).post_id.nunique()
confusion_tbl.reset_index(inplace=True)
confusion_tbl['percentage'] = (confusion_tbl.number_of_rows * 100.0) / df.shape[0]
confusion_tbl.head(10)
```

Anexo D: CD com todo o conteúdo utilizado

Este anexo físico contém o código integral em Python, assim como todos os ficheiros Excel utilizados nos testes de validação.