UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

JUEI HAO WENG

Measuring the Spreading of News on Twitter

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Engineering

Advisor: Prof. Dr. Leandro Krug Wives
Coadvisor: Dr. Vinicius Wolozyn

Porto Alegre
July 2019

“When the accomplishments and solid qualities
are equally blended, we then have the man of virtue.”
— CONFUCIUS

ACKNOWLEDGEMENTS

During the period the present work was being conducted, it reminded me of everyone who has helped me during this period of my life. Moreover, it reminds me of everyone who participated in my graduation path. For all professors who encouraged me to move forward, in both directly and indirectly way, all of my gratitude. For those professors who taught not only knowledge but also wisdom, my sincere thanks. I would like to thank all the colleagues who shared with me this learning process.

My special thanks to my advisor prof. Dr. Leandro Krug Wives for his motivation, patience and kind guidance throughout the present work, as well as my coadvisor Dr. Vinicius Wolozyn for his enthusiasm, always pushing me forward and sharing his knowledge whenever necessary.

Finally, I would like to thank my family and friends for their continuous support and help. Each one had an important contribution to the process that led to this work. Without their help, this work would not be possible.

ABSTRACT

Communication plays a large role in the development of societies over time as it enabled the exchange of information between people. Over time, new technologies have been developed to make communication more and more efficient. Nowadays, Twitter is one of the most popular social networking platforms where users exchange information all the time. There is a growing trend for Twitter to become one of the main sources of information regarding news and events. However, most content is still being produced by formal news agencies. In this sense, to measure and evaluate the coverage and the impact that traditional news has on Twitter is a relevant issue. In this work, we propose a novel method of retrieving news related information on Twitter and present a brief discussion about the possibility of measuring the dissemination of news through it. We conduct some experiments to evaluate and analyze the dissemination of specific news obtained from formal sources on Twitter, more specifically, we verify how to produce appropriate queries to find content related to the news produced by formal agencies. The results obtained in this study showed that simple word counting is a considerable accurate way for summarizing news content from traditional sources for Twitter content searches.

Keywords: Text Mining. Natural Language Processing. News Analysis. Text Classification.

Medindo a Divulgação de Notícias no Twitter

RESUMO

A comunicação exerceu um grande papel no desenvolvimento das sociedades ao longo dos tempos pois possibilitou a troca de informações entre as pessoas. Com o passar do tempo, novas tecnologias vêm sendo desenvolvidas para tornar a comunicação cada vez mais eficiente. Twitter é hoje uma das plataformas de redes sociais mais populares onde os usuários trocam informações a todo instante. Há uma tendência crescente do Twitter se tornar uma das principais fontes de informações referentes a notícias e eventos. Entretanto, grande parte do conteúdo de notícias ainda é produzido por agencias formais de notícias. Nesse sentido, avaliar e mensurar a cobertura e o impacto de notícias tradicionais no Twitter é um problema relevante. Neste trabalho, propomos um novo método de buscar informações referentes a notícias no Twitter e discutimos brevemente sobre possibilidade de se medir a disseminação de notícias através dele. Para tanto, executamos alguns experimentos que avaliam e analisam a disseminação de notícias específicas, obtidas em fontes formais, no Twitter. Mais especificamente, nós verificamos como produzir consultas adequadas para localizar conteúdo relacionado com as notícias produzidas em agências formais de notícias. Os resultados apresentados neste trabalho mostram a simples contagem de palavras como uma forma consideravelmente precisa na sumarização de textos de notícias de fontes tradicionais para buscas de conteúdo no Twitter.

Palavras-chave: Miniração de Textos, Processamento de Linguagem Natural, Análise de Notícias, Classificação de Textos.

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
NLP	Natural Language Processing
RSS	Rich Site Summary
URL	Universal Resource Locator

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The evolution of information and communication technologies has been changing people's lives in many ways. As a user of technology, our access to information is becoming more and more facilitated. Nowadays, traditional media sources are not able to follow the same pace of the Internet, when it comes to information spreading. An example of this is how Twitter is changing the way people get access to information, especially news.

Twitter[1] is a micro-blogging service that became very popular. People express themselves about a wide variety of topics using the Twitter platform, where they can post 280 characters text messages called *tweets* (Twitter had doubled the size of a tweet from 140 characters by 2017). Millions of tweets are generated every day. So, the way people can express ideas through tweets is limited by the size of the text messages. Taking this fact into account, the way tweets spread information also became an interesting research topic as information has to be summarized so that to fit into a tweet-size and at the same time does not fail to transmit the information itself. Many experiments have been conducted aimed to identify the topics of tweets, enabling the understanding of spreading of information and even the analysis of sentiment among Twitter users regarding a given topic.

Many experiments have been conducted aiming automatic trend discovery (MATHIOUDAKIS; KOUDAS, 2010) and sentiment analysis through Twitter streaming content (WANG et al., 2012; DELAVALD, 2018). The Twitter community has also been helping studies on real-time predictions and event detection.

Thus, studies related to news propagation throughout social network becomes popular; for instance, studies related to the spreading of fake news on social networks like Twitter and Facebook. However, the task of retrieving tweets related to a particular subject, or a piece of news published in traditional news sources like newspaper, is still an open problem.

With this problem in mind, this study aims to address the following research questions:

RQ1 - *How to automatically retrieve tweets related to a given news article?*? Sometimes people use to search on Twitter what exactly is happening when the rumor of something starts to get spread on the Web, or even looking after comments related to news that is highly commented on the Web. Is there any pattern to be considered when it

[1] https://twitter.com/

comes to searches within Twitter social media and efficiently getting tweets related to the news?

RQ2 - *How to measure (and predict) the spreading of news on Twitter*? The spreading of news is usually related to the relevance of people (people who exert great influence among Twitter users), place, and many other factors involved in the event taken into consideration. However, is there any way to measure the spreading of news by analyzing the number of tweets within a certain period of time? If the number of topic-related tweets within a time window could determine the relevance of their related topic, then it would be possible to predict a topic spreading on the Web as well.

To evaluate those research questions, this work is structured as follows. The next chapter describes works related to Twitter content analysis developed in recent years for a better understanding of the context of the present work. Then, Chapter 2 presents background concepts, steps, and techniques used to design and run experiments to analyze the research questions. With the theoretical background presented, in Chapter 4, we describe the main idea and approach of this work and the steps we propose to conduct this experiment, from analysis of news articles content and respective extraction of information to the retrieval of related Twitter content for the measuring the impact of news on Twitter. Finally, in Chapter 5, we present the obtained results achieved by our proposal from the perspective of the research questions. In Chapter 6, we discuss the overall outcome of the present work and some future works insights.

2 THEORETICAL FOUNDATION

This chapter describes the conceptual background used in this work. It also addresses the main steps and techniques involved in the experiments designed and conducted to evaluate the research questions proposed.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a subfield of computer science that comprises studies on several artificial intelligence techniques to enable human language to be converted into formal language understandable by computers (COLLOBERT; WESTON, 2008). The NLP techniques and concepts used in this work are described in the following subsections.

For the correct processing of text written in natural language, we need to apply tokenization. Then, as some tokens (words) may not be relevant in terms of information, stop-words removal is usually applied. Finally, as orthographic variations may come into play, stemming can be used to minimize vocabulary differences. Those techniques are explained below. After, a quick overview of string similarity determination methods for content retrieval tasks is presented.

2.1.1 Tokenization

As described by Cheatham and Hitzler (2013), tokenization is an NLP technique commonly used in text mining at the content pre-processing stage. It consists of splitting a string into separated words. Standard tokenization uses white spaces characters. It is also possible to use different delimiters depending on the application where the tokenization is applied. For instance, one can consider hyphen character as a word delimiter when it comes to separate hyphen joined words. Different languages could have specific delimiters characters and rules to tokenize strings. As part of our methodology, we applied tokenization during pre-processing of input data, preparing it for later topic summarization. The input text is split into words considering blank space character as a delimiter.

2.1.2 Stop-words removal

As described by Wilbur and Sirotkin (1992), stop-words are words that occur in topic related contents as well as in not related contents with the same likelihood. That is a word that presents no semantics relevance in topic description or representation useful as a search query. For instance, in English grammar, words like "*the*", "*if*", "*but*", "*and*", etc. are considered stop-words. We extract all stopwords from our pre-processed input text to only consider meaningful words in the later summarization process

2.1.3 Stemming

Stemming is a technique of extracting the root part of a word. (LOVINS, 1968) defines a stemming algorithm as a procedure in which words with the same root are reduced to a common form. It is a common technique applied in text labeling algorithms using word frequency approach. This technique is used as part of our topic summarization method to be compared to the raw word counting summarization method regarding content search accuracy metrics.

2.1.4 String similarity

Determining string similarity is part of text pre-processing in many textual content retrieval systems. It helps to identify similar contents to be retrieved for a search query. There are four methods of similarity determination between texts: word co-occurrence/vector-based document model methods, corpus-based methods, hybrid methods, and descriptive feature-based methods (ISLAM; INKPEN, 2008). The commonly used methods are the vector-based methods, where the input text is represented as a word vector. In this experiment, we applied this technique over the retrieved search content to filter possible duplication.

2.2 Text Classifiers

Text Classification or Text Categorization is a technique of classifying text data into one of pre-defined categories or classes. Text classifiers are used in applications like

sentiment analysis and topic detection. In this work, we propose to use a text classifier for topic detection and summarization. The classifier must act in classifying retrieved search content regarding its relationship to the input text data topic.

2.3 Text Summarization

Text summarization is a technique in which one or more text data is converted to a reduced version with the more representative content. Text Summarization has been used to reduce features within the text classification tasks. There are several techniques developed in recent years, such as SumBasic (VANDERWENDE et al., 2007) and LexRank (ERKAN; RADEV, 2004) among others. Inouye et al. (2011) presented Hybrid TF-IDF and demonstrated its efficiency on summarizing Twitter posts. They analyzed the Twitter post domain for text processing, comparing different text summarization algorithms. An interesting observation on the outcome is that due to tweets unstructured content and the short length, simple word frequency, and redundancy reduction seems to be the best techniques for summarizing Twitter topics.

3 RELATED WORK

Regarding the spreading of news on Twitter, Vosoughi et al. (2018) investigated how rumors are disseminated on Twitter. They analyzed more than 126,000 stories tweeted by about 3 million people, and those tweets were checked by undergraduate students using an automated rumor-detection algorithm. They found that false stories inspired fear, disgust, and surprise. On the other hand, true stories inspired anticipation, sadness, joy, and trust. Another interesting finding is that fake news was more novel than true ones, which suggests that people were more likely to share novel information. They conclude that falsehood diffused significantly farther, faster, and more broadly than truth.

Another relevant work is the one by Zhao et al. (2011), who used topic modeling to analyze Twitter content of a period of three months. They compared the content of topics on Twitter with that of the New York Times, which is a traditional news vehicle. As a result, they found that Twitter users have lower interests in world news events, but users spread important events' news actively. In this work, we propose to find a simple and efficient way to address the relationship between published news from traditional news vehicles and their related content generated in the Twitter community.

Kwak et al. (2010) researched the topological characteristics of Twitter and had some findings regarding the dissemination of information through the platform. Their work provides a better understanding of how the behavior of information is spreading when it comes to retweets. After analyzing more than 100 million tweets, they could then identify several behaviors of information spreading in the Twitter community. One of them is that once a tweet is retweeted by one, it reaches an average of a thousand users through the platform, showing Twitter as a rapid spreading information source and its potential to achieve out a massive number of people.

Zaman et al. (2010) conducted a study on how information spreading occurs on Twitter and how to predict it. They presented a methodology to predict information spreading by analyzing characteristics of retweets in the platform, showing an important role played by retweets in information spreading through Twitter. That shows the critical role played by retweets in the Twitter community context. Since only tweets (Twitter posts) from user accounts one follows are shown in the timeline[1], the retweets are not considered a way to disseminate information. As we propose to measure the spreading of news through the Twitter platform, retweets are not considered in this experiment.

[1] About Twitter Timeline: <http://help.twitter.com/en/using-twitter/twitter-timeline>

Also, there is the study of Wang (2012) about sentiment analysis considering the US Presidential Election of 2012. It showed that political news and events influence the public sentiment on Twitter as this information get spread through the network. It also mentioned features to be considered when it comes to Twitter posts, and that pre-processing the messages is essential as well as their normalization. A similar experiment was conducted by Delavald (2018) regarding Brazilian politics crisis. The pre-processing procedures used in this work will be described in subsection 4.1.3.

Shaozhi et al. (2013) conducted a measurement study regarding the spreading of a single breaking news message on Twitter. They searched on Twitter and analyzed over 58M tweets and found 550K tweets are related to the news under consideration. The main topic of the news they were considering was about Michael Jackson's death, and they used two different queries to perform searching: "Michael Jackson" and "MJ". Although the experiment was not about the search precision but the spreading of messages through the social network, the results showed there are improvements regarding the query used in Twitter content retrieval.

The idea we bring in the present work is to extract representative words after processing news texts and search-related content on Twitter using them as search queries. Many works have been done regarding text mining and in particular, extracting keywords through words frequency approach (JOSHI; MOTWANI, 2006; BENYAMIN; HALL, 2013; YI et al., 2013). They propose different techniques for adapting several data mining and Knowledge Discovery in Databases (KDD) techniques for document labeling. A more straightforward technique was chosen for our experiment to verify its applicability regarding content searches on Twitter.

There are several techniques to deal with small or big data systems when it comes to information retrieval. With different approaches and application specifications, there is no best approach for general text data retrieval. Blair (1984) describes: "That is, retrieval strategies which work well on small systems do not necessarily work well on larger systems[...]".

Although there are studies analyzing the effects of news spreading from social platforms, the relation between news published on traditional news vehicles and social platforms have not been much explored. We propose a comparison of different methods for retrieving Twitter platform content related to news gathered from news sources, as well as a brief discussion on how to measure the spreading of news on social platforms and if there is any way of predicting it.

4 APPROACH TO MEASURE THE IMPACT OF NEWS ON TWITTER

Given the popularity of Twitter and its active users' community spreading information through the platform every day, we decided to use it as our experimental environment for the analysis of users' reaction to a piece of given published news. As this work aims to address the relation of online social platform, in particular, Twitter, content to given news, we used a simple word counting technique over a pre-processed news text, extracting keywords to be used for search-related tweets.

As mentioned before (see Chapter 3), there are several works reporting experiments with the keyword frequency approach showing the efficiency of the method of labeling text contents. In this work, we chose to simplify the way of implementing this technique, by directly counting the frequency of the words of input text, having stopwords removed from it. In section 4.1 we explain the implementation with more details.

We compared the proposed approach of keyword extraction based on occurrences with a variation using the stemming technique. We want to understand if the effect of applying stemming will be expressively evident when it comes to Twitter content searches, or it will not make such a difference at all.

4.1 Experimental Design

Driven by our research questions as mentioned, we propose an experiment using a simple way to efficiently retrieve Twitter content related to news of a given topic from traditional news vehicles (i.e., newspapers websites). Besides that, we discuss briefly how the obtained results could be used to measure, and even predicting the given news dissemination range through the Twitter community.

In this section, we discuss the details of the methodology used to conduct the present experiment. The steps proposed in our methodology will then be justified as we present the results in Chapter 5 aiming to answer the research questions mentioned previously.

For the analysis of news from traditional news sources, we defined a way to gather the news' feeds through conventional media websites, i.e., newspaper websites. Several publishers offer the syndication of their content through Rich Site Summary (RSS) technology, a way to post news in real-time to its subscribers since RSS is used for real-time distribution of Web content. Through its subscription, we can get access to the content of

the articles (i.e., text), as we want to make use of their content to search related posts on Twitter. The key idea is to process news' textual content, extracting keywords that better describe the main topic.

Figure 4.1 describes our proposed framework, and its main steps are detailed below:

1. **Online news gathering through RSS**

   First, we start with the online news gathering step through RSS subscription on newspaper websites. We have chosen BBC and Reuters as our traditional primary sources of news for the experiment, as they offer free access to online contents. In section 4.1.1, we describe each of the sources with more details.

2. **News processing and extraction of keywords**

   As soon as we could get the main news for analysis, we process their content to extract keywords that could describe their topics for the later retrieval of related posts on Twitter. We used n most frequent keywords as our first extraction method for analysis and comparison. As mentioned, there are studies

3. **Assemble extracted keywords to form search query**

   With the keywords extracted, we managed to assemble them, forming search queries to be used on the Twitter search API, performing searches that will be detailed and discussed in section 4.1.2. We want to verify how the results behave for a different quantity of keywords used in search queries. A point to be considered is that the results could be more precise as far as we add more keywords in search queries. Until a certain number of keywords, the search could be able to retrieve no results, due to the query being too specific.

4. **Search on Twitter using previously assembled query**

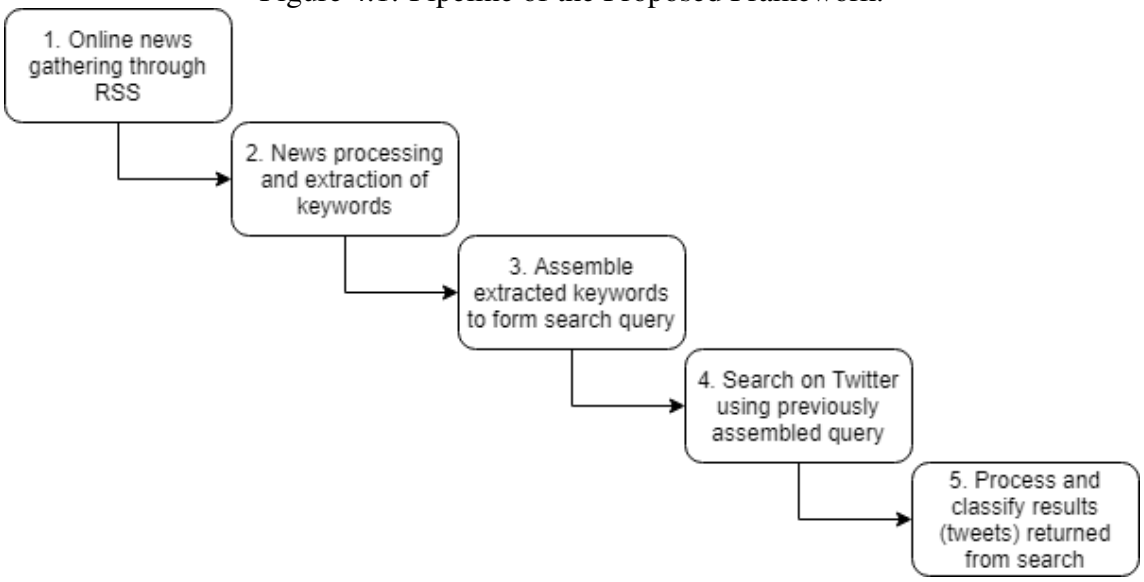
   Then, we search on Twitter with the search queries obtained on the previous step to extract tweets related to the content of the news we are processing.

   In this step, we add some filters to refine the searches, such as retweet and content filters. We will discuss these filters with more details in section 4.1.3.

5. **Process and classify results (tweets) returned from searches**

   Finally, the results were then processed and classified – a binary classification, as *related* or *not related* – according to their relation to the main topic in consideration, and we would have a measure of the precision of our framework. The evaluation metrics used in our experiment are detailed in section 4.2.

Figure 4.1: Pipeline of the Proposed Framework.



Twitter provides an Application Programming Interface (API) for the developer community that offers some functionalities such as posting tweets and searching tweets using queries. In order to interact with Twitter API, we use a Python framework called *Tweepy*[1]. Tweepy is a Python framework for extracting information through Twitter Development API. It offers several restful methods to access that Twitter platform.

4.1.1 News Sources

The experiment starts with news text processing as our algorithm input data. Through the RSS service provided by the news publishers, we can extract some valuable information from a piece of news, such as title, publication date and time, and so on, all in JSON format[2], which helps the next processing steps.

The traditional news vehicles chosen was BBC[3] news (and one of its division, BBC Brazil[4]) and Reuters[5]. The respective RSS sources URL from where the content was extracted for are the follows:

- **BBC:** <http://feeds.bbci.co.uk/news/rss.xml>
- **BBC Brazil:** <http://www.bbc.com/portuguese/topicos/brasil/index.xml>

[1]<http://www.tweepy.org/>
[2]JavaScript Object Notation (JSON) is a text format for structured data.
[3]<http://www.bbc.co.uk/news>
[4]<http://www.bbc.com/portuguese>
[5]<http://www.reuters.com/>

- **Reuters:** <http://www.reuters.com/tools/rss>

4.1.2 Query Generation Methods

For our experiment, we define a keyword as a word extracted from the input text that could be used to describe the main text topic. In this stage, the method of analysis we explore is using n most frequent words in the text. We treat the text as a list of words, eliminating stop-words and special characters that eventually occur. We also, as part of the tokenization process, convert every word to lower case to ensure that same words with different capitalization will not be counted as different words. By removing these words, we keep more representative words from which we choose the most recurring words to form a search query.

We also vary the number of keywords used in the query to perform searches. After the first tests, we limited the number of keywords between 3 to 5; this was chosen experimentally as by using more than five keywords to perform the searches, it became too specific, and the searches did not return any result. On the other hand, using under three keywords to form queries, the results topics became too random.

We propose to compare the accuracy of the results of two different keywords extraction methods while varying the number of keywords used. First, with ordinary content extracted from news syndication, we extract the n most frequent words of a text with all words counted and ordered by their frequency of occurrence in the text. Therefore, the top n most frequent words will compose the search query to be used on the social network search engine.

The second method is to apply stemming during word counting and, also, a similarity check during Twitter content retrieval, removing duplicated results, and verify the accuracy of applying these techniques.

4.1.3 Extraction on Twitter

After assembling the queries based on the techniques described in the previous subsection, they are used to retrieve all posts/tweets that the Tweepy API search mechanism can provide.

4.1.4 Processing and Classification of search results

From the search results, we separate the tweets by their relation to the input news text. We consider all comments regarding some main topics mentioned in the input news as related to the article. For instance, for the news entitled "Theresa May calls for mental health to be priority"[6] we consider as related all content regarding the announcement of Theresa May that mental health should be priority and teachers should have lessons in identifying children's mental illness included in their training.

In order to create a corpus for the classification task, we manually classified search results returned from processing two news articles. The tweets are classified by keywords, and we consider those as representatives for the news article in question.

4.2 Evaluation Metrics

We evaluate two different steps of our experiment. The first part of the evaluation is on Twitter content extraction. The precision is the main assessment metric for our extraction task. We define the precision of the extraction task as: $Precision = \frac{RT}{RC}$. Where RC is the number of total tweets retrieved from search, and RT is the quantity of tweet related to the news under consideration from RC.

With the tweets retrieved, the second part of the evaluation is on Twitter content classification. In order to evaluate the classification task, we adopted standard Information Retrieval metrics such as precision, recall, and F-1. These metrics can be briefly described as follows:

- *Precision:* the fraction of tweets classified as related that are really related to a news. $Precision = \frac{tp}{tp+fp}$
- *Recall* is the fraction of related tweets that were successfully identified. $Recall = \frac{tp}{tp+fn}$
- *F-1* corresponds to the harmonic mean between precision and recall. $f1 = 2 * \frac{precision*recall}{precision+recall}$

where tp is the number of positive instances correctly classified as positive, tn number of negative instances correctly classified as negative, fp negative instances wrongly classified as positive, and fn is the number of positive instances wrongly classified as negative.

[6] <http://www.bbc.com/news/education-48658151>

We defined positive instances as tweets that are related to news and negative instance as tweets that are not related to the news.

5 RESULTS AND ANALYSIS

With initial search tests, we could observe that the API retrieves both ordinary tweets and retweets. Since we are interested in analyzing the spreading of news through the Twitter community as described previously (Chapter 3), retweets are filtered from the search results. The Twitter API provides the retweet filter functionality by merely adding the string "-filter:retweets" to the search query. Thus, the retrieved results will all be ordinary posts that we would like to analyze.

Another issue from our first tests was that Tweepy returned some posts in the results that did not even contain the keywords used in the search query. Aiming to solve this problem, we added a content check component on the results. To be considered a related tweet, it must contain at least one of the keywords used to form the search query. This filter is implemented by simply performing string checking on each of the results, i.e., if it contains at least one keyword in its content or not.

The first results of tweets retrieval using the most frequent words in texts showed to be positively efficient regarding the similarity of topics between them. The following tables show the analysis of our preliminary results. In some cases, we achieved very high precision among retrieved tweets' content topic - with BBC News we obtained over 90% of precision using five keywords search queries (Table 5.1). It could be an issue as we could verify that some searches retrieved lots of very similar content tweets, even identical in some cases. Another issue verified in this approach is that the number of tweets retrieved in searches could sometimes be lower than the first limit of 20 tweets for each search we previously established - with Reuters News using five keywords search queries it only retrieved around half of the established quantity, in average.

The later could be related to too specific search queries used in the case of a higher number of keywords combined. A solution for it should be to avoid using too many keywords to form search queries. As it can be noted in Table 5.1, the more number of keywords in the search queries, the more specific is the search itself and, consequently, the more accurate are the retrieved content.

A fact to consider is that using a larger quantity of keywords in queries may eliminate the chances of also related tweets to be retrieved, once they may not contain all the keywords in its content. One better approach should consider the semantic meaning of words used to perform searches.

After the preliminary results, we chose two articles – one from BBC and one

Table 5.1: Preliminary search results analysis

Search results using n most frequent keywords				
BBC (2 articles)	Keywords	3	4	5
	Tweets	40	36	22
	Precision	50%	63.9%	90.9%
Reuters World (3 articles)	Keywords	3	4	5
	Tweets	48	47	21
	Precision	52.1%	61.7%	90.5%

Table 5.2: Total quantity of tweets retrieved during the period of 7 days of analysis

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Total
BBC Article	85	699	295	68	20	14	23	3211
BBC Brasil Article	190	143	54	12	4	4	0	434

from BBC Brazil, both published in June 16th – and redid the experiment analyzing now over a period of 7 days since the publication day of the input news article, applying the approaches presented in chapter 4. The precision of novel search results are shown in the figures 5.1 and 5.2.

For this period of analysis, we retrieved a total of 3211 tweets for BBC article and 434 tweets for BBC Brazil article. Table 5.2 shows the details of search results. Day 1 indicates the date of the article was published. All the tweets were manually classified and annotated with the tag "*_$related$*" for those who are related to the input news article and "*_$not_related$*" for those who are not related to.

The results showed a very high precision rate for our proposed search method. However, it has to be mentioned that the method provides us poorly in terms of volume of retrieved tweets. As the hypothesis we described in 4.1, with the query becoming over specified, we are actually filtering out a considerable part of related content as some of them may not contain all the keywords specified in the query.

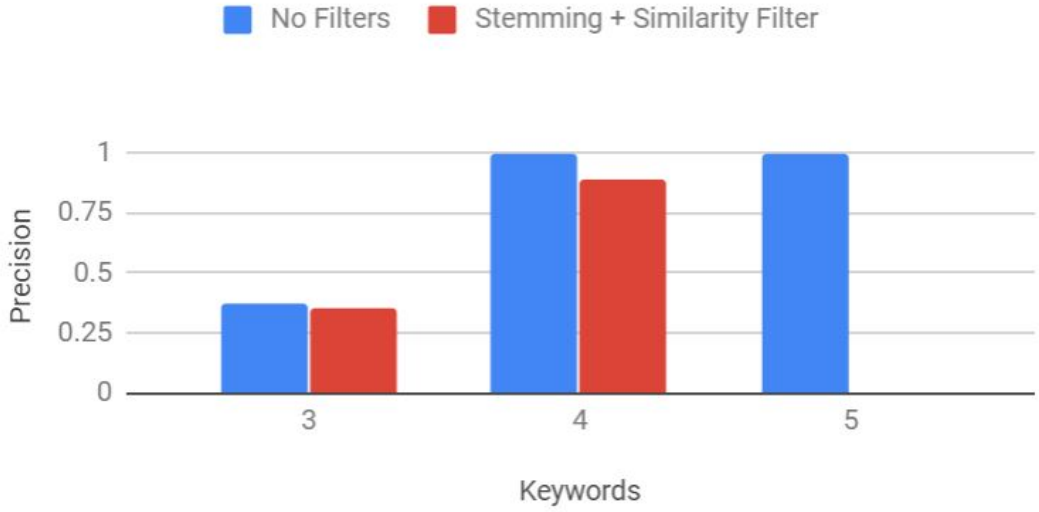
The stemming technique showed no big influence in the query generation process as the resulting query for some cases was identical in both applying stemming and not applying stemming. To picture better one of these cases, we describe below the resulting queries for the news articles analyzed:

1. *"Theresa May calls for mental health to be priority"* (from BBC News)

   Raw Counting (without applying stemming):

   - 3 Keywords: "mental health may"
   - 4 Keywords: "mental health may prime"
   - 5 Keywords: "mental health may prime minister"

Figure 5.1: Precision of tweets regarding news from BBC.



Source: Author

Stemming (applying stemming during word count):

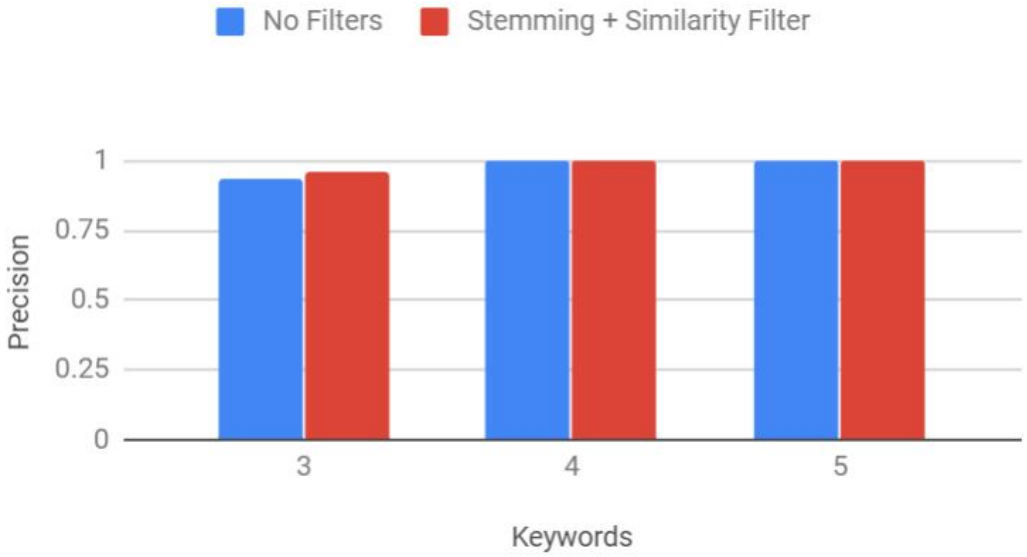
- 3 Keywords: "mental health may"
- 4 Keywords: "mental health may mrs"
- 5 Keywords: "mental health may mrs prime"

2. *"As declarações de Bolsonaro que levaram à demissão de Levy do BNDES"* (from BBC Brasil)
   Raw Counting (without applying stemming):
   - 3 Keywords: "bndes levy banco"
   - 4 Keywords: "bndes levy banco bolsonaro"
   - 5 Keywords: "bndes levy banco bolsonaro dados"

   Stemming (applying stemming during word count):
   - 3 Keywords: "bndes levy banco"
   - 4 Keywords: "bndes levy banco bolsonaro"
   - 5 Keywords: "bndes levy banco bolsonaro dados"

We can verify in figure 5.2 that the difference in precision between the two methods in comparison is only due to duplication of messages retrieved from searches since the generated queries are the identical between the two methods. It suggests that stemming technique may not be a good approach when it comes to single text summarization for Twitter content search query generation. An interesting investigation would be to verify the stemming technique applied in the retrieved content classification process.

Figure 5.2: Precision of news from BBC Brazil.



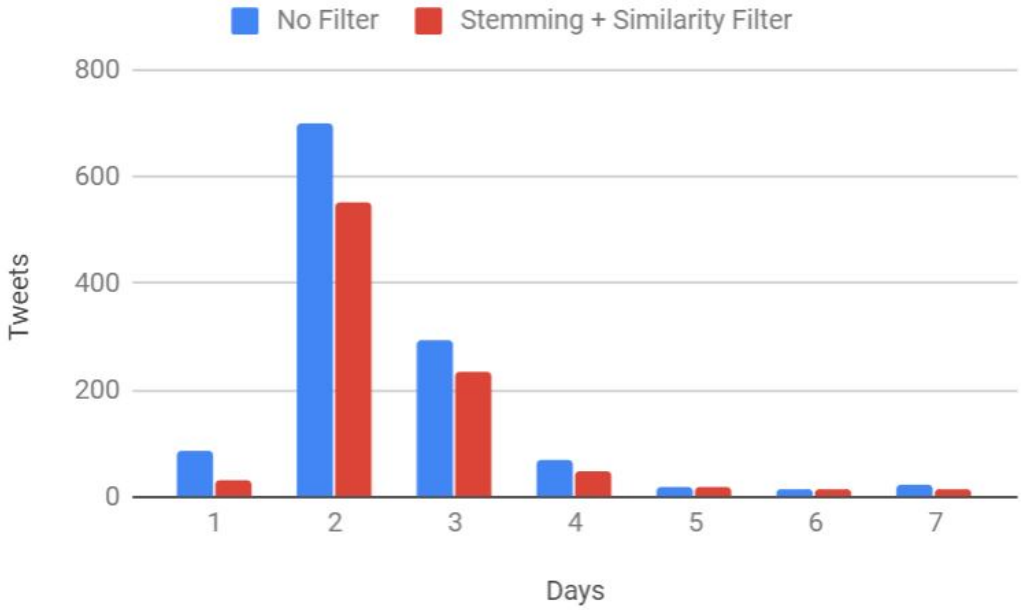
Source: Author

In order to answer one of our research question (RQ2) mentioned in Chapter 1, we want to see how the news articles we are considering spread through the period of 7 days. Thus, we separated the returning search content by the respective Twitter post creation date.

Figures 5.3 and 5.4 show the occurrences of the posts distributed over the period: an initial increase of people comments regarding the topic, and it decreases significantly after 2-3 days. The raw counting of related tweets quantity was associated with their respective dates of creation in the chart. The results shown are from using stemming technique and similarity check on retrieved results before saving to file (in red) and from not using such processing and filtering (in blue).

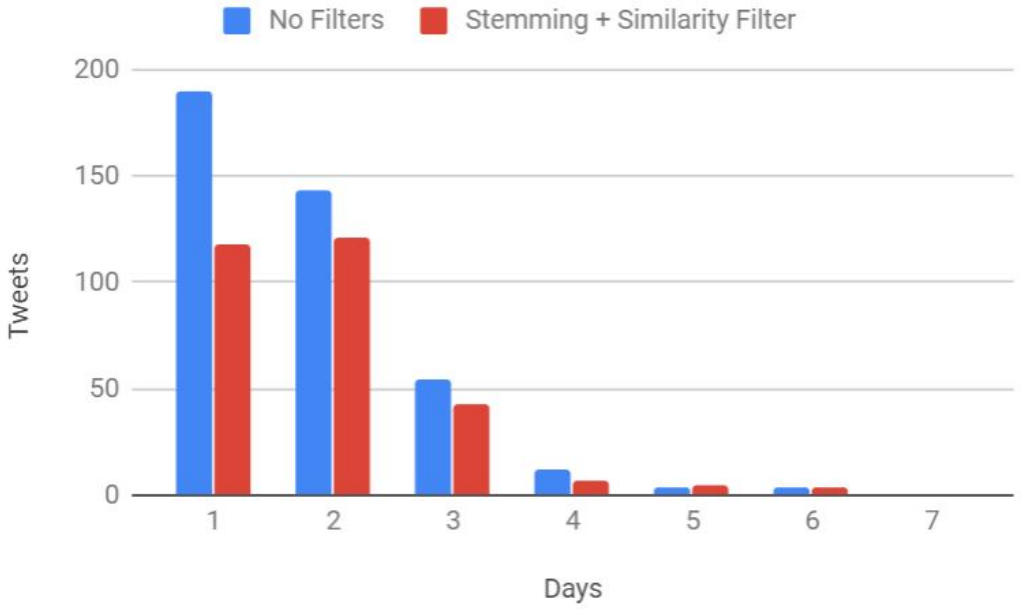
With the obtained results, our next step is to train a classification model in order to perform the classification task over the returned results. We want a classifier to provide real-time analysis on the retrieved content, separating relevant contents related to the input article being processed and then, based on the distribution mentioned before, one could experimentally infer – or even predicting – the spreading of the given article.

Figure 5.3: Distribution of tweets related to BBC news during the period of 7 days.



Source: Author

Figure 5.4: Distribution of tweets related to BBC Brazil news during the period of 7 days.



Source: Author

6 CONCLUSION

In this work, we presented an overview of recent works on analyzing Twitter content, from Twitter messages spreading dynamics analysis to event detection through Twitter streaming, passing by sentiment analysis and information retrieval approaches for online social networks. With the current scenario on information spreading seen, we presented a novel approach on measuring the spreading of news through Twitter social network and retrieval of its content. We bring two main research questions that drive our work: *How to automatically retrieve tweets related to a given news article?* and *How to measure (and predict) the spreading of news on Twitter?* Through raw counting of document words frequency, we showed that a good precision on tweets retrieval task could also be reached. However, there is a certain limit of context specification by using more keywords to perform the searches. A distribution of occurrences of Twitter posts could be visualized, and it showed an expected behavior through the period of analysis.

In the stage of processing search results, we noticed the need for a corpus in order to train our classifier model. Since it has to be the news article context-related dataset, we performed several manual analyses and successfully created a corpus for seven articles, with a total of 3,733 manually classified tweets against the respective article. It is worth mentioning that there are also 4,198 Twitter contents from searches of the other two articles not included in our result analysis as they were pending analysis.

We also proposed a classifier that can, in real-time, process a given news article as text data and scrapes tweets through Twitter API using the keyword frequency counting approach and identifies all related tweets for further analysis and processing. Due to limitations described below, we delimit the scope of the experiment as presented so far.

There were several technical issues during the experiment on using Tweepy framework that makes use of Twitter restful API, such as search rate limits and calls rate per time window. Another issue regards to the lack of dataset we could use without creating one on our own. The manual classification task demands too much effort.

Regarding the manual process of creating a corpus for the classifier, it was an issue that took us a long time. However, once with it prepared, it would be an excellent use for future experiments on similar approaches.

As our experiment covers some Twitter content retrieval techniques, the next step could be applying text processing techniques on tweets for classifying them over their relation to the input news article under consideration, as mentioned by Inouye et al. (2011).

The idea proposed in this work could be improved by experimenting different types of learning classifiers – that could benefit from the dataset we provided in this experiment – on the search result classification task.

REFERENCES

BENYAMIN, D.; HALL, M. A. **Search and retrieval methods and systems of short messages utilizing messaging context and keyword frequency**. [S.l.]: Google Patents, 2013. US Patent 8,380,697.

BLAIR, D. C. An evaluation of retrieval effectiveness for a full-text document retrieval system. 1984.

CHEATHAM, M.; HITZLER, P. String similarity metrics for ontology alignment. In: SPRINGER. **International Semantic Web Conference**. [S.l.], 2013. p. 294–309.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: ACM. **Proceedings of the 25th international conference on Machine learning**. [S.l.], 2008. p. 160–167.

DELAVALD, G. S. Uma análise de dados das reações à crise política brasileira no twitter. 2018.

ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. **Journal of artificial intelligence research**, v. 22, p. 457–479, 2004.

INOUYE, D.; KALITA, J. K. Comparing twitter summarization algorithms for multiple post summaries. In: IEEE. **2011 IEEE Third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing**. [S.l.], 2011. p. 298–306.

ISLAM, A.; INKPEN, D. Semantic text similarity using corpus-based word similarity and string similarity. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM, v. 2, n. 2, p. 10, 2008.

JOSHI, A.; MOTWANI, R. Keyword generation for search engine advertising. In: IEEE. **Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)**. [S.l.], 2006. p. 490–496.

KWAK, H. et al. What is twitter, a social network or a news media? p. 591–600, 2010.

LOVINS, J. B. Development of a stemming algorithm. **Mech. Translat. & Comp. Linguistics**, v. 11, n. 1-2, p. 22–31, 1968.

MATHIOUDAKIS, M.; KOUDAS, N. Twittermonitor: trend detection over the twitter stream. In: ACM. **Proceedings of the 2010 ACM SIGMOD International Conference on Management of data**. [S.l.], 2010. p. 1155–1158.

VANDERWENDE, L. et al. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. **Information Processing & Management**, Elsevier, v. 43, n. 6, p. 1606–1618, 2007.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, American Association for the Advancement of Science, v. 359, n. 6380, p. 1146–1151, 2018.

WANG, H. et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL 2012 System Demonstrations**. [S.l.], 2012. p. 115–120.

WILBUR, W. J.; SIROTKIN, K. The automatic identification of stop words. **Journal of information science**, Sage Publications Sage CA: Thousand Oaks, CA, v. 18, n. 1, p. 45–55, 1992.

YE, S.; WU, F. Measuring message propagation and social influence on twitter. com. **International Journal of Communication Networks and Distributed Systems**, Citeseer, v. 11, n. 1, p. 59–76, 2013.

YI, X. et al. Private searching on streaming data based on keyword frequency. **IEEE Transactions on Dependable and Secure Computing**, IEEE, v. 11, n. 2, p. 155–167, 2013.

ZAMAN, T. R. et al. Predicting information spreading in twitter. v. 104, n. 45, p. 17599–601, 2010.

ZHAO, J. J. W. X. Comparing twitter and traditional media using topic models. **Institutional Knowledge at Singapore Management University**, 2011.