Elsevier Editorial System(tm) for Applied Soft Computing

Manuscript Draft

Manuscript Number:

Title: Detecting relevant tweets and twitter user influence in very large tweet collections: the London Riots case study

Article Type: Full Length Article

Keywords: Tweet Topic Detection;
Fuzzy Fingerprints;
Text Mining;
Social Network Mining;
Page Rank User Influence

Corresponding Author: Prof. Joao Paulo Carvalho, Ph.D

Corresponding Author's Institution: INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa

First Author: Joao Paulo Carvalho, Ph.D

Order of Authors: Joao Paulo Carvalho, Ph.D; Hugo H Rosa, MsC; Fernando Batista, PhD

Abstract: In this paper we propose to approach the subject of detecting relevant tweets and Twitter user influence when in the presence of very large tweet collec- tions containing a large number of di↵erent trending topics. We use a large database of tweets collected during the 2011 London Riots as a case study to demonstrate the application of the proposed soft computing techniques. In order to extract relevant content, we extend, formalize and apply a recent technique, called Twitter Topic Fuzzy Fingerprints, which, in the scope of social media, outperforms other well known text based classification meth- ods, while being less computationally demanding, an essential feature when processing large volumes of streaming data. Afterwards we use Page Rank as a graph based centrality tool in order to identify who were the most in- fluential participants discussing the London Riots topic within the Twitter network.

Cover Letter

Lisboa, October 7th, 2015

Dear Editor-in-Chief

Please find attached the paper entitled "*Detecting relevant tweets and twitter user influence in very large tweet collections: the London Riots case study*", that we are submitting for possible publication in Applied Soft Computing.
The paper presents the theory and application of a soft computing technique, Fuzzy Fingerprints, for the detection and retrieval of relevant tweets in large tweet datasets. A real world case study is included as an application of the proposed techniques that allows finding influent users using the relevant retrieved tweets. As such, we believe that the work is of interest and within the scope of this journal despite the fact that hardly any works on the subject of soft computing applied to text mining and social network mining have been published in this forum.

Looking forward to receiving your news,

Sincerely,

João Paulo Carvalho
Hugo Rosa
Fernando Batista

Mailing address:
INESC-ID
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
E-mail: joao.carvalho@inesc-id.pt
Tel.: +351962148272

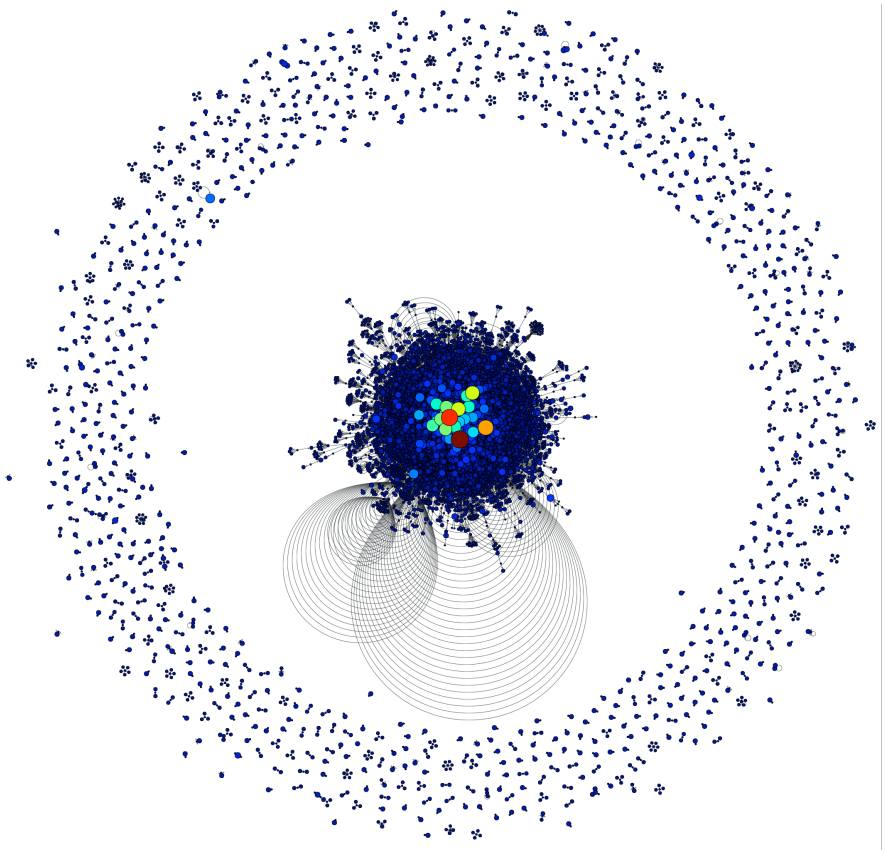
*Highlights (for review)

Detecting relevant tweets and twitter user influence in very large tweet collections: the London Riots case study

Highlights:

- We extend and formalize an efficient method for extracting relevant tweets from very large tweet collections based on text Fuzzy Fingerprints
- The proposed method is much faster and has better precision and recall than the most commonly used text based classification methods
- We show an application of the proposed method on a real world Twitter database collected during the London riots
- We apply a centrality based method on the relevant tweets in order find the most influential Twitter users during the London riots.

***Graphical abstract (for review)**



*Manuscript
Click here to view linked References

Detecting relevant tweets and Twitter user influence in very large tweet collections: the London Riots case study

Joao P. Carvalho, Hugo Rosa, Fernando Batista,

Abstract

In this paper we propose to approach the subject of detecting relevant tweets and Twitter user influence when in the presence of very large tweet collections containing a large number of different trending topics. We use a large database of tweets collected during the 2011 London Riots as a case study to demonstrate the application of the proposed soft computing techniques. In order to extract relevant content, we extend, formalize and apply a recent technique, called Twitter Topic Fuzzy Fingerprints, which, in the scope of social media, outperforms other well known text based classification methods, while being less computationally demanding, an essential feature when processing large volumes of streaming data. Afterwards we use Page Rank as a graph based centrality tool in order to identify who were the most influential participants discussing the London Riots topic within the Twitter network.

Keywords: Tweet Topic Detection, Fuzzy Fingerprints, Text Mining, Social Network Mining, User Influence, Page Rank

1. Introduction

Twitter was originally created in 2006 as a public social networking service enabling users to send and read short 140-character messages. After the "Arab Spring" [1] and other protests and riots occurring between 2010 and

Email addresses: joao.carvalho@inesc-id.pt (Joao P. Carvalho), hugo.rosa@inesc-id.pt (Hugo Rosa), fernando.batista@inesc-id.pt (Fernando Batista)

2011, it became clear that important events are often commented in Twitter before they become "public news". Twitter rapidly became a major tool for spreading news, for dissemination of positions or ideas, and for the commenting and analysis of current world events. This has led to a change in how the public perceives the importance of social networks, and even news agencies and networks had to adapt and start using Twitter as a potential (and some times preferential) source of information.

However, using Twitter as a source of information involves many technical obstacles. As of mid 2015, more than 500 millions tweets covering thousands of different topics are published daily. Of these 500 million tweets, it is very unlikely that more than a few thousand, let us say in the range of 0.001%-0.01%, are relevant to a given discussion topic (even major topics). Therefore, filtering which content is relevant for a given discussion topic is far from trivial (section 3.3).

Twitter contributes to solve this problem by providing a list of top trends [2] and the hashtag # mechanism: when referring to a certain topic, users are encouraged to indicate it through the use of a hashtag. E.g., "#refugeeswelcome in Europe!" indicates the topic of the tweet is the current refugees crisis in Europe. Websites such as #hastags.org make good use of this information to present Twitter trends, e.g., https://www.hashtags.org/analytics/refugeeswelcome/. Other tools such as Twittermonitor [3] can also be used to obtain Twitter trends.

However, only roughly 16% of all tweets are hashtagged [4]. These numbers have been confirmed by our experiments, and can be partially explained by the fact that 140 characters is often not enough to communicate a thought, and including an #hashtag further aggravates the lack of available space. It is therefore clear that, in order to properly analyze a given discussion topic, it is essential to retrieve as much of the remaining 84% untagged information as possible. Since no other tagging mechanisms exist in Twitter, the process of retrieving tweets that are related to a given topic must use some kind of text classification process.

Assuming that it is possible to filter all the tweets related to a given topic, one still needs to find (among others) which tweets have more relevance. One important step to find relevant tweets within a topic, is to find who are its most important "actors". When big events occur, it is common behavior for users to post about it in such fashion, that it becomes a trending topic; users comment on the event, discuss it with their friends and followers, retweet what they feel is important, etc. Probably all these actions occur while being

unaware from where the event stemmed or who made it relevant. Much like real life, some users carry more influence and authority than others. Determining user relevance is vital to help determine trend setters [5], as well as separate important messages from spam and garbage. The determination of a user's relevance must take into account not only global metrics that include the user's level of activity within the social network, but also his impact in a given topic [6].

In this article, we will approach not only the issue of topic detection on Twitter, but also try to answer the question: "Which users were important in disseminating and discussing a given topic?" Topic detection will be based on an extension and formalization of the Fuzzy Fingerprints method previously presented in [7]. User influence will be computed using a centrality method, Page Rank [8], based on user mentions [9]. We will use a large database of tweets collected during the 2011 London Riots as a case study to show the application of the proposed techniques.

This paper is organized as follows. Firstly, an overview of other related work on the subjects of Topic Detection, User Influence and the London Riots is given. Secondly, we provide detailed explanation of the available dataset, the Twitter Topic Fuzzy Fingerprints method and the use of PageRank to determine the most important users. Finally, we experiment both methods with the London Riots dataset, present its results and discuss its merits.

2. Related Work

2.1. Topic Detection

The first goal of this work is essentially to automatically classify tweets into a set of trending topics. Tweet Topic Detection involves deciding if a given tweet is related to a given #hashtagged topic. Basically this can be categorized as a classification problem, albeit one with some particular characteristics that need to be addressed specifically: (1) it is a text classification problem, with an unknown and large number of categories, where the texts to be classified are very short texts (up to 140 characters); (2) it fits the Big Data paradigm due to the huge amounts of streaming data.

We distinguish between Topic Classification and Topic Detection. The former is broadly known in Natural Language Processing (NLP) as Text Categorization, and consists of finding the correct topic (or topics) for each document, given a restricted set of categories (subjects, topics) such as politics, sports, music, etc., and a collection of text documents [10], in this case,

tweets; the tweets will often belong to at least one of those categories and it is very rare that a tweet does not fit into any topic. The latter takes on a more detailed approach, where an attempt is made to determine the topic of the document, given a predetermined large set of possible topics, where the topics are so unique amongst themselves that there is a high probability that a tweet without a hashtag may very well not belong to any of the current trends.

When considering this difference, the most similar works on Topic Detection within Twitter are those related with emerging topics or trends, for example [3, 11, 12, 13]. In these works the authors use a wide variety of techniques regarding text analysis to find the most common related words and hence detect topics. In our work we already assume the existence of trending topics and we aim at efficiently detecting tweets that are related to them, despite not being explicitly marked as so.

It is also possible to find several works regarding Topic Classification. In [14], an attempt is made to classify Twitter Trending Topics into 18 broad categories, such as: sports, politics, technology, etc, and their experiments on a database of randomly selected 768 trending topics (over 18 classes) show that, using text-based and network-based classification modeling, a classification accuracy up to 65% and 70% can be achieved, respectively. Another interesting article, despite not on the theme of Topic Detection, demonstrates how to use Twitter to automatically obtain breaking news from the tweets posted by Twitter users [15]. In 2009, when Michael Jackson passed away, "the first tweet was posted 20 minutes after the 911 call, which was almost an hour before the conventional news media first reported on his condition". This further enforces the importance of automatically analysing the massive amount of information on Twitter.

For years, a wide range of methods has been applied to Text Classification problems, ranging from hand-coded rules to supervised and unsupervised machine learning. Some of the most well-known and commonly applied methods for text classification tasks include: K-Nearest Neighbours (kNN) and the Support Vector Machine (SVM).

The kNN is an example-based classifier. This means it will not "build explicit declarative representations of categories, but instead rely on computing the similarity between the document to be classified and the training documents" [10]. In this case, the training data is simply the "storing of the representations of the training documents together with their category labels". In order for kNN to "decide whether a document d belongs to a

category c, kNN checks whether the k training documents most similar to d belong to c. If the answer is positive for a sufficiently large proportion of them, a positive decision is made." The kNN is considered to be one of the simplest and best performing text classifiers, whose main drawback is "the relatively high computational cost of classification - that is, for each test document, its similarity to all of the training documents must be computed" [10]. In kNN, "the training is fast, but classification is slow. Computing all the similarities between a document that has not been categorized and a collection of documents, is slow" [16].

A support vector machine (SVM) is a very fast and effective binary classifier. According to [16] "every category has a separate classifier and documents are individually matched against each category". Given the vector space model in which this method operates, geometrically speaking, [10] describes SVM as a "hyperplane in the feature space, separating the points that represent the positive instances of the category from the points that represent the negative instances. The classifying hyperplane is chosen during training as the unique hyperplane that separates the known positive instances from the known negative instances with the maximal margin". In general, a larger margin means a lower classifier generalization error. SVMs can efficiently perform linear and non-linear classifications using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

In [17], Yang and Liu, performed several tests in a controlled study and reported that SVM and k NN are at least comparable to other well-known classification methods, including Neural Networks and Naive Bayes, and that significantly outperform the other methods when the number of positive training instances per category are small.

2.2. User Influence

The concept of influence is of much interest for several fields, such as sociology, marketing and politics. Empirically speaking, an influential person can be described as someone with the ability to change the opinion of many, in order to reflect his own. While [18] supports this statement, claiming that "a minority of users, called influentials, excel in persuading others", more modern approaches [19] seem to emphasize the importance of interpersonal relationships amongst ordinary users, reinforcing that people make choices based on the opinions of their peers. The point is that "influence" is an abstract concept, which makes it exceptionally hard to quantify.

Several studies have attempted to accomplish this goal. In [20], three measures of influence were taken into account, regarding Twitter: in-degree, re-tweets and mentions, where "in-degree is the number of people who follow a user; re-tweets mean the number of times others forward a user's tweet; and mentions mean the number of times others mention a user's name." It concluded that while in-degree measure is useful to identify users who get a lot of attention, it "is not related to other important notions of influence such as engaging audience". Instead "it is more influential to have an active audience who re-tweets or mentions the user".

In [21], the authors conclude that within Twitter, "news outlets, regardless of follower count, influence large amounts of followers to republish their content to other users", while "celebrities with higher follower totals foster more conversation than provide retweetable content".

InfluenceTracker [22] is a framework that rates the impact of a Twitter account taking into consideration an Influence Metric, based on the ratio between the number of followers of a user and the users it follows, and the amount of recent activity of a given account. It also calculates a Tweet Transmission rate where the "most important factor (...) is the followers' probability of re-tweeting". Much like [20], it also shows "that the number of followers a user has, is not sufficient to guarantee the maximum diffusion of information in Twitter (...) because, these followers should not only be active Twitter users, but also have impact on the network'.

In [23], with access to multiple extremists forums (Dark Web), the authors attempted to create an "effective way to identify the threat through social media" by "detecting the influential users automatically". By introducing "weights in forum social network to reflect the degree of influence", it was found that it makes a "make a substantial impact on the ranking result".

2.3. The London Riots

Between the 6th and 11th August 2011 thousands of people rioted in several boroughs of London with the resulting chaos generated looting, arson, and mass deployment of police. In the end five people died in what became known as the 2011 London Riots.

Although Twitter was said to be a communication tool for rioting groups to organize themselves, in [23], "researchers analyzed 600,000 tweets and retweets about the riots for evidence that Twitter was used as a central organizational tool to promote illegal group action" and concluded that "there is little overt evidence that Twitter was used to promote illegal activities at

the time, though it was useful for spreading word about subsequent events." According to The Guardian newspaper [24], Twitter did however play a big role spreading the news about what was happening and "was a valuable tool for mobilizing support for the post-riot clean-up and for organizing specific clean-up activities".

3. Material and Methods

3.1. Twitter Topic Fuzzy Fingerprints

Fingerprint identification is a well-known and widely documented technique in forensic sciences. In computer sciences a fingerprint is a procedure that maps an arbitrarily large data item (such as a computer file, or author set of texts) to a much compact information block, its fingerprint, that uniquely identifies the original data for all practical purposes, just as human fingerprints uniquely identify people. In order to serve for classification purposes, a fingerprint must be able to capture the identity of a given class. In other words, the probability of a collision, i.e., two classes yielding the same fingerprint, must be small.

Fuzzy Fingerprints were originally proposed for text classification in [26], where they were successfully used to detect authorship of newspaper articles (out of 73 different authors). For text classification purposes, a set of texts associated with a given class is used to build the class fingerprint. Each word in each text represents a distinctive event in the process of building the class fingerprint, and distinct word frequencies are used as a proxy for the class associated with a specific text. The set of the fuzzy fingerprints of all classes is known as the fingerprint library. Given a fingerprint library and a text to be classified, the text fingerprint is obtained using a process similar to the one used to create the fingerprint of each class, and then a similarity function is used to fit the text into the class that has the most similar fingerprint.

In order to use Fuzzy Fingerprints for tweet topic detection, several procedural changes were proposed in [7]. Here we formalize the process of creation of Twitter Fuzzy Fingerprints and Fingerprint Libraries based on a data set of #hashtagged tweets, and the respective process of tweet topic detection.

3.1.1. Twitter Fuzzy Fingerprint Creation and Twitter Fuzzy Fingerprint Libraries

The full set of properly classified tweets, i.e., tweets that are #hashtagged, are processed to compute the top-k word list for each of the #hashtags.

Considering F_j as the set of all words in the tweets containing #hashtag j, the result of processing F_j is a list of k tuples $\{v_i,\ n_i\}$ where v_i is the i-th most frequent word and n_i the corresponding count. I.e., we obtain an ordered k-sized list containing the most frequent distinct words for each topic.

Due to the small size of a single tweet, its features should be as unique as possible in order to make the fingerprints distinguishable amongst the various #hashtagged topics. Therefore we propose to also account for the Inverse Class Frequency (icf) of each word existing in all the computed k tuples $\{v_i,\ n_i\}$. Icf is an adaptation of the well-known Inverse Document Frequency (idf), where #hashtagged topics are used instead of documents to distinguish the occurrence of common words:

$$icf_v = \log \frac{J}{J_v} \tag{1}$$

In ()1), J is the #hashtag fingerprint library size (i.e., the total number of different #hashtags), and J_v is the number of #hashtagged topics where $v \in F$ (i.e., where word v is present).

The product of the frequency of word v with its inverse class frequency, $tficf_v = n_v \times icf$, is used to re-order the k-sized word list of each #hashtagged topic.

$$\mu_{ab}(i) = \begin{cases} 1 - (1-b)\frac{i}{kb} & i < a \\ \frac{a(1-\frac{i-a}{k-a})}{k} & i \geq a \end{cases} \tag{2}$$

The next step consists in fuzzifying each top-k list in order to obtain the #hashtag fingerprint. The choice of the fuzzifying function is critical: the chosen approach is to assign a membership value to each feature in the set based only on the order in the list. The reason for using the order instead of the frequency results from empirical experiments that show that the order of the frequency seems more relevant than the frequency actual value [26]. The more frequent features will have a higher membership value. We tested for several alternative membership functions, and all results presented in this work use a function $\mu_a b$ inspired in the Pareto rule, where roughly 80% of the membership value is assigned to first 20% elements in the ranking (2)

The fingerprint (Φ), which is based on the top-k list, consists on a size-k fuzzy vector where each position j contains an element v_i and a membership

value μ_i representing the fuzzified value of the rank of v_i (the membership of the rank).

An #hashtagged topic j will be represented by its fingerprint $\Phi_j = \Phi(F_j)$. Formally, fingerprint $\Phi_j = (v_j i, \mu_j i)|i = 1..k_j$ has length k_j, with $S_j = v_j i|i = 1..k_j$ representing the set of v in Φ_j. The set of all #hashtag fingerprints will constitute the fingerprint library.

3.1.2. Tweet Topic Detection using Twitter Fuzzy Fingerprints: Tweet to Topic Similarity Score

The original text fuzzy fingerprint detection method [26] consisted in creating a fingerprint for each text to be classified, and to compare its fingerprint with all fingerprints contained in the fingerprint library. That method is not applicable to very small texts, such as for example, tweets, since the word frequencies in a single tweet are not distinctive enough to create a fingerprint (within 140 characters very few relevant words, if any, are repeated). In order to address this issue we use a Tweet to Topic Similarity Score (T2S2) that tests how much a tweet fits to a given #hashtagged topic. The T2S2 score (3), does not take into account the size of the text to be classified (i.e., its number of words).

$$T2S2(T, \Phi_j) = \frac{\sum_{v} \mu_{\Phi_j}(v) : v \in (T \cap S_{\Phi_j})}{\sum_{i=0}^{j} \mu_{\Phi_j}(w_i)} \tag{3}$$

In (3), Φ_j is the #hashtagged topic of fingerprint j, T is the set of distinct words of the preprocessed tweet text, S_{Φ_j} is the set of word of the #hashtag j fingerprint and $\mu_{\Phi_j}(v)$ is the membership degree of word v in the #hashtag j fingerprint. Essentially, T2S2 divides the sum of the membership values of every word v that is common between the tweet and the #hashtag j fingerprint, by the sum of the top j membership values in $\mu_{\Phi_j}(w_i)$ where $w \in (\Phi)$.

T2S2 tends to 1 when most to all features of the tweet belong to the top words of the fingerprint, and approaches 0 when there are no common words between the tweet and the fingerprint, or the few common words are in the bottom of the fingerprint.

Tweets that have a T2S2 score to a given #hashtagged topic above a given threshold, are considered as being relevant to the topic and are retrieved from the database.

3.1.3. Parameter Optimization and Previous Results

According to [7], the Twitter Topic Fuzzy Fingerprints performed very well on a set of 2 millions English, Spanish and Portuguese tweets collected over a single day, beating other widely used text classification techniques. The training set consisted of 11000 tweets containing the 22 of the top daily trends (hashtagged topics). 350 unhashtagged test tweets were properly classified with an f-measure score of 0.844 (precision=0.804, recall=0.889).

Further work by the same authors [27], used a training set of 21000 tweets, from "21 impartially chosen topics of interest out of the top trends of the 18th of May, 2013". The test set was made of "585 tweets that do not contain any of the top trending hashtags" and "each tweet was impartially annotated to belong to one of the 21 chosen top trends". After extensive parameter optimization using a development set, the fuzzy fingerprint method scored an f-measure of 0.833 on the test set, when using k=20 fingerprints, words with less than 3 characters removed, no stopwords were removed and no stemming was performed. Any tweet with a T2S2 score above 0.10 was chosen for retrieval. This setup, proved to be not only more accurate than other well known classifying techniques (kNN and SVM), but also much faster (177 times faster than kNN and 419 times faster than SVM).

The described setup (fingerprint size, T2S2 threshold, and text pre-processing parameters) was chosen for the current London Riots case study.

3.2. PageRank User Influence

The second goal of this article is to assert user influence for a given Twitter topic. With the previous definitions of influence in mind, and the set of tweets regarding a given topic that the Twitter Topic Fuzzy Fingerprints method will provide, we propose a graph representation of user's influence based on "mentions". Whenever a user is mentioned in a tweet's text, using the @*user* tag, a link is made from the creator of the tweet, to the mentioned user, like so:

The tweet "*Do you think we can we get out of this financial crisis,* @$userB$*?*" from @$userA$, creates the link: @$userA \longrightarrow$ @$userB$.

This is also true for re-tweets:

The tweet "*RT @userC The crisis is everywhere!*" from @$userA$, creates the link: @$userA \longrightarrow$ @$userC$.

In graph theory and network analysis, the concept of centrality refers to the identification of the most important vertices's within a graph, i.e., most important users. We therefore define a graph $G(V, E)$ where V is the set of users and E is the set of directed links between them.

Arguably the most well known centrality algorithm is PageRank [8]. It is one of Google's methods to its search engine and uses web pages as nodes, while back-links form the edges of the graph (Figure 1). According to [27], "query-independent evaluation of web pages is the significant characteristic of this PageRank algorithm" as it calculates the value of each page offline "thus the ranking of web pages becomes static". By having each page with equal probability to be chosen as a starting point, "the Initial Probability Distribution is 1/N and the importance of any web page can be judged by looking at the pages that link to it".

It is defined by)4) as $PR(v_i)$ of a page v_i.

$$PR_{vi} = \frac{1-d}{N} + d \sum_{v_j \in M(v_i)} \frac{PR(v_j)}{L(v_j)} \tag{4}$$

In (4), v_j is the sum ranges over all pages that has a link to v_i, $L(v_j)$ is the number of outgoing links from v_j, N is the number of documents/nodes in the collection and d is the damping factor. The PageRank is considered to be a random walk model, because the weight of a page v_i is "the probability that a random walker (which continues to follow arbitrary links to move from page to page) will be at v_i at any given time". The damping factor corresponds to the "probability of the random walk to jump to an arbitrary page, rather than to follow a link, on the Web" and is required to "reduce the effects on the PageRank computation of loops and dangling links in the Web." [28]. The true value that Google uses for damping factor is unknown, but it has become common to use $d = 0.85$ in the literature. A lower value of d implies that the graph's structure is less respected, therefore making the "walker" more random and less strict.

In [9], it was shown that, in the context of Twitter User Influence, PageRank outperforms another well known network algorithm, Katz [29], which is why it is the method that is used in this work to determine the influence of users in disseminating a given topic.

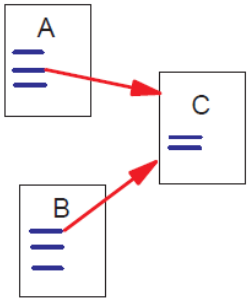


Figure 1: A and B are back-links of C

3.3. Data

The London Riots occurred between the 6th and 10th August 2011.

A large dataset, [25], known as TW-Master was created by The Guardian newspaper via the REST API during the riots, and then expanded using the users timeline. For each user, tweets created after August 1st 2011 were retrieved up to the 3,200 tweet limit from REST API statuses/user-timeline limitation. A total of 9,913,397 Tweets were collected from 8,819 Twitter users.

Following the event, The Guardian publicly released Twitter data which included a list of 200 influential twitter users based on re-tweets during the riot period. The released dataset contained a total of 1,132,938 tweets that were posted between August 1st 00:00:00am and August 31st 23:59:59. According to The Guardian, the dataset contains 17,795 tweets related with the London Riots. This data set was used for the case study presented in this article.

4. Calculation, Results and Discussion

In this section, we present the results we obtained by applying the proposed methods to the available London Riots dataset.

Using the Twitter Topic Fuzzy Fingerprints method, we created a "London Riots fingerprint" that will allowed us to retrieve from the London Riots database, tweets that are relevant but not contained in the 17795 tweets list made public by The Guardian (section 3.3). By obtaining a richer set of relevant tweets, it is possible to perform more detailed studies and analysis on the events that occurred in 2011. As an application example, we created a graph representation of the users in the extended set, and determined which

Table 1: Training Set Trend Distribution

Top Trend	Count
#londonriots	11490
#ukriots	2733
#riots	2332
#riotcleanup	1832
#lfc	1193
#london2012	93
#motogp	0
#eurovision	12
#libya	1517
#f1	898
#mariobrosep	20
#mcfc	628
#theparadigmshift	0
#projectallout	0
#seo	268
#ionlyhaveloveforgod	0
#architecture	0

users were most important in broadcasting the topic using the PageRank algorithm.

4.1. London Riots Fingerprint

As it was mentioned in section 3.3, the available data set consists of 1132938 tweets. The number of distinct hashtags in this set is huge (in the order of the thousands), but only 4 of those hashtags have enough occurrences and were considered relevant for the purpose of creating the London Riots Fuzzy Fingerprint: #londonriots; #ukriots; #riots; #riotcleanup.

13 additional #hashtagged topics were selected using REST API's deprecated method "GET trends/weekly", which returns the top trending topics for each day in a given week. They are used to perform the Inverse Topic Frequency step (see section 3.1.1). Despite being top trends for the days of the London Riots, some of them do not have any tweet occurrences in our database. Table 1 shows the list of topics.

The low (sometimes zero) value of tweets containing the top trending topics can be explained by Twitter's own view on what constitutes a trending topic. According to [2], "Twitter Trends are automatically generated by an algorithm that attempts to identify topics that are being talked about more right now than they were previously. The Trends list is designed to

help people discover the most breaking news from across the world, in real-time. The Trends list captures the hottest emerging topics, not just what is most popular. Put another way, Twitter favors novelty over popularity". This definition, alongside the information that only 1% of the tweets can be streamed, explains a seemingly low presence of the top trending topics, in contrast with a high presence of London Riots tweets, due to a possible bias in the data extraction performed by The Guardian.

The data set used for the creation of the fingerprint is composed of any tweet in the data set that contains at least one of the hashtags in Table 2. In order to make the most out of the London riots topic, the hashtags #londonriots, #ukriots, #riots and #riotscleanup were aggregated into a single #londonriots class. This set is composed of 23060 tweets, and is rather unbalanced, i.e., different classes classes/hashtags have different amount of tweets.

The parameter setup used to execute the Twitter Topic Fuzzy Fingerprint method, was the same that studies [7, 30] have shown to be optimal in both performance and speed:

- threshold value for T2S2 = 0.10
- Size of the fingerprint, $k = 20$
- removing words with less than 3 characters from corpus
- not removing stopwords from the corpus
- not performing stemming operations

Table 3 shows the obtained London Riots fuzzy fingerprint.

Each of the remaining 1112938 tweets in the database was tested for similarity with the London Riots fingerprint. As a result, 25757 tweets were retrieved from the data set. This represents an increase of about 45% in the number of relevant tweets retrieved and made available by The Guardian.

An independent validation of the dataset was performed by an impartial third party. The analysis of the obtained 25757 London Riots tweets indicated a precision of 0.951 in the obtained results. The remaining 1107181 tweets, not considered relevant by the Fuzzy Fingerprints method, were also analyzed in order to detect False negatives, i.e., tweets that should have been identified as relevant, but were not considered relevant. Due to the cost

Table 2: The London Riots Fingerprint

Rank	Feature	μ
1	police	1.0
2	riot	0.8
3	rioters	0.6
4	cover	0.4
5	http://t.co/0hg1bhi	0.2
6	croydon	0.1875
7	clapham	0.175
8	@riotcleanup	0.1625
9	causes	0.15
10	cameron	0.1375
11	riots	0.125
12	shops	0.1125
13	hackney	0.1
14	#hackney	0.0875
15	#birminghamriots	0.075
16	boris	0.0625
17	birmingham	0.05
18	army	0.0375
19	#manchesterriots	0.025
20	rioting	0.0125

of analyzing such a huge number of tweets manually, the annotation was a combination of automatic, semi-automatic and manual procedures.

From the existing information, a table was produced containing relevant tweet meta-data. Such table was then provided to a human linguist annotator, whom manually annotated and validated a considerable number of tweets. The annotation process was conducted using the following predefined strategy: (1) check the text of individual tweets and validate or correct the initial annotation until finding a possible pattern, either related or not related with London Riots; (2) apply regular expressions to get a list of similar tweets, related with the pattern, and that can be easily checked altogether; (3) Check and mark the list of returned tweets and go back to step 1. The annotator used 3 different tags: "Y" (related with London Riots), "N" (not related) and "?" (not sure).

The previously described strategy is very efficient during the initial iterations, where a simple pattern returns big lists of similar tweets that can be check and marked altogether. However, as one proceeds with the annotation, patterns that return similar tweets that have not been previously checked are

much more difficult to discover. Simple heuristics, such as looking at the list of words triggered by the fingerprint, or sorting the list of the tweets by their T2S2 score, helped validating the most problematic tweets. It is also relevant to indicate that we have put more emphasis validating the tweets previously marked as Y in order to avoid populating the database with false positives. At the end of this process, 1328 annotation values were modified (0.12%), another 7858 (0.71%) were re-annotated with Y, and 1971 tweets (0.17%) were marked not sure. The initial automatic annotation was considered correct for all the remaining tweets.

4.2. London Riots User Influence

In this section we will present the results of PageRank's algorithm ranking for most influential users within the 25757 London Riots tweets obtained in the previous section. An empirical study of the users is made, in order to ascertain their degree of influence. The graphs and ranking were determined by Tiago Peixoto's "Graph-Tool" [31].

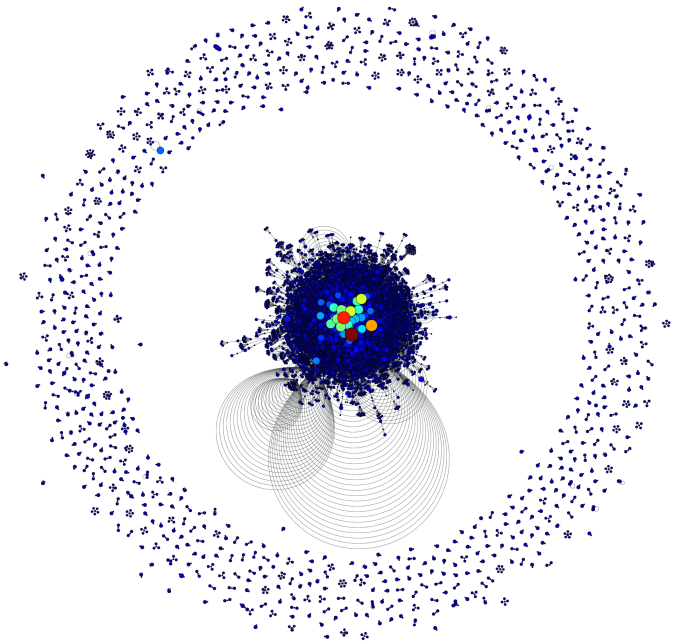


Figure 2: User influence Page Rank Graph - larger circles indicate larger user influence.

Table 3 shows how the PageRank algorithm behaves with our graph representation of user mentions in tweets. Figure 2 provides a visual tool to the

Table 3: London Riots Top 20 most influential users according to Page Rank

User	PageRank		Mentions	
	score	rank	#	rank
@guardian	0.0002854	1	160	2
@skynewsbreak	0.0002512	2	178	1
@gmpolice	0.0002128	3	122	4
@riotcleanup	0.0001767	4	107	6
@prodnose	0.0001761	5	67	14
@metpoliceuk	0.0001494	6	116	5
@marcreeves	0.0001476	7	69	11
@piersmorgan	0.0001465	8	78	8
@scdsoundsystem	0.0001442	9	69	12
@subedited	0.0001337	10	70	10
@youtube	0.0001257	11	48	20
@bbcnews	0.0001256	12	94	7
@mattkmoore	0.0001237	13	62	15
@richardpbacon	0.0001218	14	40	27
@lbc973	0.0001150	15	34	35
@skynews	0.0001113	16	74	9
@bengoldacre	0.0001055	17	61	17
@bbcnewsnight	0.0000988	18	68	13
@tom_watson	0.0000968	19	44	21
@paullewis	0.0000954	20	129	3
...				
@juliangbell	0.0000275	188	61	16

graph, as provided by PageRank. The bigger the size of the vertex, the more influential the algorithm deems it to be. The sum of posting and mentioned users is 13765 (vertices) and it has 19993 different user mentions (edges), achieving a network connectivity ratio of $\frac{edges}{vertices} = 1.46$.

5. Discussion

In this section, an empirical evaluation of the obtained results is performed.

There seems to exist a relation between the number of mentions and the ranking, since these users are some of the most mentioned users in our universe of tweets. According to PageRank, the following, are the top users:

- @guardian, Twitter account of the world famous newspaper "The Guardian".
- @skynewsbreak, Twitter account of the news team at Sky News TV channel.

This outcome agrees with [21] previous statement, that, "news outlets, regardless of follower count, influence large amounts of followers to republish their content to other users". This can be justified by the incredibly high London Riots news coverage.

Other users in our top 20, seem to fit the profile, namely @gmpolice, @bbcnews and @skynews. Most of the other users are either political figures, political commentators or jornalists (@marcreeves, @piersmorgan, @mattkmoore and @richardpbacon).

There are two interesting cases worth mentioning:

- @paullewis, shows up at 20th according to PageRank.
- @juliangbell, shows up at 188th according to PageRank.

The user @paullewis has 129 mentions, but PageRank penalizes it probably because it is mentioned by least important users, which means a less sum weight is being transfered to it in the iterative process. This logic also applies to users @bbcnewsnight, @skynews and @bbcnews. Additionally, @paullewis is also an active mentioning user, having mentioned other users a total of 14 tweets, while @skynewsbreak and @guardian have mentioned none. As a consequence, Paul Lewis transfers its influence across the network while The Guardian and SkyNews simply harvest it.

User @juliangbell, despite mentioned often (61 times), is down on the PageRank because of indirect gloating, i.e., he mentions himself in his own tweets. Looking at the data, we found this scenario, very often:

Tweet: *"@LabourLocalGov #Ealing Riot Mtg: @juliangbell speech http://t.co/3BNW0q6"* posted by @juliangbell himself.

The user is posting somebody else's re-tweet of one of his tweets. As a consequence a link/edge was created from @juliangbell to @LabourLocalGov, but also from @juliangbell to himself, since his username is mentioned in his own tweet. Julian Bell is a political figure, currently Labor Leader of Ealing Council and Chair at London Councils Transport and Environment Committee. It is acceptable to think that he would have a role in discussing the London Riots, since it was such a political event, but the self congratulatory behavior of re-tweeting other people's mentions of himself, is contradictory with the idea of disseminating the topic across the network.

6. Conclusion

In this work, we used Twitter Topic Fuzzy Fingerprints, a novel and efficient approach in tweet topic detection, to process The Guardians London Riots Twitter database.

This method allowed us to expand the number of tweets considered relevant for the events of the 2011 London Riots by 45% with a precision of 0.95, confirming the high effectiveness of text based fuzzy fingerprints when applied to text social network mining.

With the extended dataset, composed of 25757 tweets, we performed a study on user influence based on the most well known centrality network algorithm: Google's PageRank. Its results proved to be in direct agreement with other related works on the subject of influence, which claimed that news outlets were of vital importance for topic propagation on Twitter. It also allowed us to confirm that PageRank heavily penalizes Twitter users that try to manipulate the social network by indirectly citing themselves.

7. Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project PTDC/IVC-ESCT/4919/2012 and funds with reference UID/CEC/50021/2013.

8. Bibliography

[1] C. Huang, Facebook and twitter key to arab spring uprisings: report, http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report, accessed: 2014-05-02 (June 2011).

[2] Twitter, To trend or not to trend, https://blog.twitter.com/2010/trend-or-not-trend, accessed: 2014-03-28 (2010).

[3] M. Mathioudakis, N. Koudas, Twittermonitor: Trend detection over the twitter stream, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD '10, ACM, New York, NY, USA, 2010, pp. 1155–1158. doi:10.1145/1807167.1807306.
URL http://doi.acm.org/10.1145/1807167.1807306

[4] A. Mazzia, J. Juett, Suggesting hashtags on twitter, Master's thesis, University of Michigan (2010).

[5] R. Tinati, L. Carr, W. Hall, J. Bentwood, Identifying communicator roles in twitter, in: Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, ACM, New York, NY, USA, 2012, pp. 1161–1168. doi:10.1145/2187980.2188256.
URL http://doi.acm.org/10.1145/2187980.2188256

[6] J. Weng, E.-P. Lim, J. Jiang, Q. He, Twitterrank: Finding topic-sensitive influential twitterers, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, ACM, New York, NY, USA, 2010, pp. 261–270. doi:10.1145/1718487.1718520.
URL http://doi.acm.org/10.1145/1718487.1718520

[7] H. Rosa, F. Batista, J. P. Carvalho, Twitter topic fuzzy fingerprints, in: WCCI2014, FUZZ-IEEE, 2014 IEEE World Congress on Computational Intelligence, International Conference on Fuzzy Systems, IEEE Xplorer, Beijing, China, 2014, pp. 776–783.

[8] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bringing order to the web (1999).

[9] R. A. H. Rosa, J. P. Carvalho, F. Batista, Detecting user influenc in twitter: Pagerank vs katz, a case study, 7th European Symposium on Computational Intelligence and Mathematics.

[10] R. Feldman, J. Sanger, Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, New York, NY, USA, 2006.

[11] M. Cataldi, L. Di Caro, C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, in: Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD '10, ACM, New York, NY, USA, 2010, pp. 4:1–4:10. doi:10.1145/1814245.1814249.
URL http://doi.acm.org/10.1145/1814245.1814249

[12] S. P. Kasiviswanathan, P. Melville, A. Banerjee, V. Sindhwani, Emerging topic detection using dictionary learning, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, ACM, New York, NY, USA, 2011, pp. 745–754. doi:10.1145/2063576.2063686.
URL http://doi.acm.org/10.1145/2063576.2063686

[13] A. Saha, V. Sindhwani, Learning evolving and emerging topics in social media: A dynamic nmf approach with temporal regularization, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, ACM, New York, NY, USA, 2012, pp. 693–702. doi:10.1145/2124295.2124376.
URL http://doi.acm.org/10.1145/2124295.2124376

[14] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, A. Choudhary, Twitter trending topic classification, in: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDMW '11, IEEE Computer Society, Washington, DC, USA, 2011, pp. 251–258. doi:10.1109/ICDMW.2011.171.
URL http://dx.doi.org/10.1109/ICDMW.2011.171

[15] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, J. Sperling, Twitterstand: News in tweets, in: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09, ACM, New York, NY, USA, 2009, pp. 42–51. doi:10.1145/1653771.1653781.
URL http://doi.acm.org/10.1145/1653771.1653781

[16] M. Konchady, Text Mining Application Programming, Charles River Media, 2006.

[17] Y. Yang, X. Liu, A re-examination of text categorization methods, in: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, ACM, New York, NY, USA, 1999, pp. 42–49. doi:10.1145/312624.312647.
URL http://doi.acm.org/10.1145/312624.312647

[18] E. M. Rogers, Diffusion of innovations (1962).

[19] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01, ACM, New York, NY, USA, 2001, pp. 57–66. doi:10.1145/502512.502525.
URL http://doi.acm.org/10.1145/502512.502525

[20] M. Cha, H. Haddadi, F. Benevenuto, K. P. Gummadi, Measuring user influence in twitter: The million follower fallacy, in: in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social, 2010.

[21] D. F. S. G. Alex Leavitt, Evan Burchard, The influentials: New approaches for analyzing influence on twitter (2009).

[22] G. Razis, I. Anagnostopoulos, Influencetracker: Rating the impact of a twitter account, CoRR.
URL http://arxiv.org/abs/1404.5239

[23] C. C. Yang, X. Tang, B. M. Thuraisingham, An analysis of user influence ranking algorithms on dark web forums, in: ACM SIGKDD Workshop on Intelligence and Security Informatics, ISI-KDD '10, ACM, New York, NY, USA, 2010, pp. 10:1–10:7. doi:10.1145/1938606.1938616.
URL http://doi.acm.org/10.1145/1938606.1938616

[24] T. Guardian, http://www.theguardian.com/uk/2011/dec/07/twitter-riots-how-news-spread.

[25] S. R. Crockett, K, Twitter riot dataset (2011).

[26] N. Homem, J. P. Carvalho, Authorship identification and author fuzzy fingerprints, in: 30th Annual Conference of the North American Fuzzy Information Processing Society, NAFIPS2011, 2011.

[27] V. Lakshmi Praba, T. Vasantha, Efficient hyperlink analysis using robust proportionate prestige score in pagerank algorithm, Applied Soft Computing 24 (Complete) (2014) 86–94. doi:10.1016/j.asoc.2014.07.012.

[28] N. Q. Phuoc, S.-R. Kim, H.-K. Lee, H. Kim, Pagerank vs. katz status index, a theoretical approach, in: Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT '09, IEEE Computer Society, Washington, DC, USA, 2009, pp. 1276–1279. doi:10.1109/ICCIT.2009.272.
URL http://dx.doi.org/10.1109/ICCIT.2009.272

[29] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
URL http://ideas.repec.org/a/spr/psycho/v18y1953i1p39-43.html

[30] H. Rosa, Topic detection within social networks, Master's thesis, Instituto Superior Tcnico (2014).

[31] T. Peixoto, https://graph-tool.skewed.de/.

9. Vitae

Prof. João Paulo Carvalho has a PhD (2002) and MsC (1996) degrees from Instituto Superior Técnico, University of Lisbon, Portugal, where he is currently a Professor at the Department of Electrical Engineering and Computation. He has taught courses on Computational Intelligence, Distributed Systems, Computer Architectures and Digital Circuits since 1998. He is also a senior researcher at L2F Spoken Language Systems Laboratory, INESC-ID Lisboa, where he has been a researcher since 1991, and has coordinated 6 nationally funded research projects (3 as the main project leader) and has been involved as a member in over a dozen national and European projects. His current main research interest involves applying Computational

Intelligence techniques to social systems and to speech and natural language processing. He has authored over 100 papers in international scientific Journals, book chapters and peer-reviewed conferences. He was program co-chair and organizer of IFSA-EUSFLAT 2009, webchair for IEEE-WCCI 2010, Publicity Chair of Fuzz-IEEE2015, Program Chair of IPMU2016 and program committee member of several conferences in the area of soft computing and computational intelligence.

Prof. Fernando Batista has a PhD in Computer Science and Engineering (2011) from Instituto Superior Técnico (IST), University of Lisbon, Portugal. He has a Masters degree in Electrical and Computer Engineering (2003) also from IST, and the bachelors degree (1997) in Mathematics and Computer Sciences from Universidade da Beira Interior (UBI). Since 2000, he has been lecturer at the Lisbon University Institute (ISCTE-IUL). Since 2001, he is also a researcher at the Spoken Language Systems Laboratory (L2F), INESC-ID, participating in several European and National projects. From 1996 to 2000 he was a researcher at the Natural Language Processing Group, INESC. His current research interests include Text mining, Machine Learning, Rich Transcription, and the integration of different fields of Natural Language Processing. He was recently part of the local organization committee of EMNLP 2015 and has been involved in the LxMLS summer school since 2011. He is currently a member of the Institute of Electrical and Electronics Engineers (IEEE), and of the International Speech Communication Association (ISCA).

Hugo Rosa has a Masters Degree in Electrical Engineering and Computation from Instituto Superior Técnico, University of Lisbon, Portugal. He is currently a Junior Researcher at L2F Spoken Language Systems Laboratory, INESC-ID Lisboa, where he has been a researcher since 2013. In the last two years, he has authored 4 papers for his work with Prof. João Paulo Carvalho and Prof. Fernando Batista in applying Computational Intelligence techniques to social systems and to speech and natural language processing, namely Twitter.