Journalistic Relevance Classification in Social Network Messages: an Exploratory Approach

Miguel Sandim, Paula Fortuna, Alvaro Figueira and Luciana Oliveira

Abstract Social networks are becoming a wide repository of information, some of which may be of interest for general audiences. In this study we investigate which features may be extracted from single posts propagated throughout a social network, and that are indicative of its relevance, from a journalistic perspective. We then test these features with a set of supervised learning algorithms in order to evaluate our hypothesis. The main results indicate that if a text fragment is pointed out as being interesting, meaningful for the majority of people, reliable and with a wide scope, then it is more likely to be considered as relevant. This approach also presents promising results when validated with several well-known learning algorithms.

1 Introduction

Nowadays social networks have become popular systems for sharing and exchanging messages between users. This high rate of information has also turned into a great source of potential, and interesting knowledge, that could be used for the creation of valuable information for a wider audience. In fact, much of the available information scattered among different "discussion groups" in social media, might actually be used in news, or in news creation, since thriving topics on most social networks many times reflect important current events which may be of interest for a more generic audience. On the other hand, we also know that more than usually, information in social media is not relevant outside a short circle of users. Users tend also to post private, personal, or just of a very narrow scope information on

Miguel Sandim, Paula Fortuna and Alvaro Figueira
CRACS / INESC TEC and University of Porto, Rua do Campo Alegre, 1021/1055, 4169-007 Porto, Portugal, e-mail: miguel.sandim@fe.up.pt, paula.fortuna@fe.up.pt, arf@dcc.fc.up.pt

Luciana Oliveira
CICE / ISCAP & INESC TEC, Polytechnic of Porto, Rua Jaime Lopes Amorim, Porto, Portugal
e-mail: lgo@eu.ipp.pt

their “pages”. In this panorama it is important to have systems capable of aiding in the identification of what might be interesting information to a wider audience. The goal of the present study is, therefore, to develop a classification model that can automatically identify relevant information in text messages on social networks.

The process of deciding if a particular text has relevant information is neither easy, nor objective, but it is, by far, the most important concern in handling information overload and retrieval [1]: what is relevant for one person, might not be relevant for another; what is not relevant now, might be in a few days or even in a few minutes from now; what is not relevant, can gain relevance just by the inclusion of some context. The combination of possibilities is endless. Moreover, the identification of reasons for personal relevancy diverge from person to person, thus consists on a psychological process by which relevance judgments are made [1] and are computationally difficult to be imitated.

Our approach to the detection of relevance is based on a generalized consensus about which information is relevant to be considered a ‘news’ from a journalist perspective. Although, each journalist may have its own writing style, and personal opinion about any subject, there are a set of guidelines which can help him within this process. Different authors ([2, 3, 4]) suggest some criteria to use: negativity, recency, proximity, consonance, unambiguity, superlativeness, personalization, eliteness, attribution, facticity, continuity, competition, cooption, composition and predictability, to name a few.

Research related to information spread was also found to be either based on the structure of the network it is introduced to or generated on, or on the nature of the content in itself. In fact, while [5] ‘gossip’ analysis is based on the structure of the network, that propels information spreading, [6] argues that virality is strictly connected to the nature of the content, and not to the types of edges linking nodes in specific co-occurrence or social pattern networks.

Moreover, research conducted on text virality [6] indicates that common social network metrics alone (e.g. #likes, #retweets) are not sufficient for assessing such a complex phenomenon and, reinforcing the above mentioned criteria, suggest that several virality components should be considered, such as: appreciation, spreading, simple buzz, white buzz, black buzz, raising discussion and controversiality.

Similarly, our system builds on a set of filters capable of detecting a set of unique characteristics that will enable to create a score for each social media post, allowing to discover “information with potential to be relevant”. Some of these unique characteristics have commonalities to research presented in [6] and in [1], namely: ‘controversiality’ and ‘positivess’, with the later having the same common ground as ‘white buzz’ and ‘reliability’ (or credibility) and ‘recency’, as mentioned in [1]. Other proposed content features add to research being conducted on the field, such as ‘interest’, ‘meaningfulness’ and ‘scope length’, which are further detailed in section 2.3.

In order to build a classification model it is fundamental to have annotated data with instances to train and test. In a previous study [7] workers from Mechanical Turk classified social network messages as “relevant” or “irrelevant”. The proposed system consisted of a social media crawler and respective classification into “rele-

vant" or "not-relevant" information. However, limitations identified in this preliminary stage of research led to the development of a more robust and comprehensive methodology. Instead of only asking the workers to answer a binary question about relevance, the workers were asked to give other information that could enlighten the process of journalistic relevance detection, namely by extending the text classification process, in order to include the above mentioned relevance cues. The increase of text classification comprehensives and complexity also allowed us to assure a higher level of trust on the gathered human classification. In the next section we describe this method; we present an analysis based on our results, and draw our conclusions about its efficiency.

The paper is structured as follows: Section 2 defines the methodology that was followed throughout this study; Section 3 presents the results obtained from the exploratory analysis; Section 4 explains the transformation of the users answers on "Crowdflower" into a dataset, as well as the features extracted from each text fragment to potentially explain its relevance. Section 5 describes the experimentation process with several supervised learning algorithms and the results obtained; and finally, Section 6 offers an analysis over the developed work, its viability and envisioned future steps.

2 Related Work

3 Methodology

In order to detect relevance (or irrelevance) in text fragments, a methodology is proposed and described in this section. The phases of this methodology are summarised in fig. 1 and include: data crawling from social networks, data pre-processing, human classification with the use of the "Crowdflower" platform and the development of a classification model.

Each of the illustrated phases in fig. 1 is detailed in the next subsections.

3.1 Crawling from Social Networks

The first phase of this methodology consisted on data crawling from social networks. In this case, the text fragments analysed throughout this paper are posts and comments retrieved from two social networks - Twitter and Facebook - using each the corresponding official API. In order to do so, a Java program was developed to interface with the APIs and with a database built in PostgreSQL.

The data was collected between 1st and 4th April 2016 and included Facebook posts and comments and Twitter tweets. Facebook posts may take the form of status, link, image, video, offer or event. A Facebook post type status (mainly text) may be

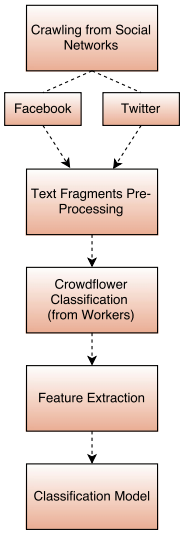

Fig. 1 Pipeline representing the methodology followed.

as long as 63206 characters. Facebook posts may receive comments, likes, shares and reactions (love, haha, wow, sad and angry). Post comments and post shares may also receive likes and replies. A Twitter tweet has a 140 character limit and may be marked as favourite and / or be retweeted (which would be the equivalent to a Facebook share).

Data retrieval on twitter was conducted by presenting the API with ten keywords (detailed in section 2.1.1), which were distributed by 100 queries. In what Facebook is concerned, data retrieval was performed on the pages of fourteen international news providers (detailed in section 2.1.2). A maximum of 1000 posts and of 20 comments per post was collected in each news provider page. These difference between the collection methods among the two networks were enforced by restrictions of their own API.

The initial retrieved dataset was composed of 11051 posts, 128673 comments and 76280 tweets.

3.1.1 Twitter

Regarding Twitter, tweets were gathered using the search method provided with one or more keywords from the following list:

- “Refugees” and “Syria”
- “Elections” and “US”
- “Olympic Games”
- “Terrorism”
- “Daesh”

- "Referendum" and "UK" and "EU"

These keywords were chosen based on their popularity in the initial gathering moment, since the probability of fetching a great quantity of tweets in current trending topics is higher. The search was conducted among the tweets from the previous seven days [8] from the collection moment.

3.1.2 Facebook

In regard to Facebook, the available API did not allow search of posts by keyword. In order to emulate this collection methodology, several posts and comments were collected from fourteen of the most popular international news providers' pages, namely: "Euronews", "CNN", "Washington Post", "Financial Times", "New York Post", "The New York Times", "BBC News", "The Telegraph", "The Guardian", "The Huffington Post", "Der Spiegel International", "Deutsche Welle News", "Pravda" and "Fox News". After the posts and comments collection a search by the keywords was conducted, using the ones specified in section 2.1.1, in order to obtain coherent subject distribution among both networks.

3.2 Text Fragment Pre-Processing

After the crawling from social networks, a control phase was conducted over the gathered text fragments. Since the fragments were extracted for inclusion in a "CrowdFlower" task, it was important to guarantee that the participants in the task had access to fragments with several quality standards. Therefore, only the text fragments with the following conditions were considered in the sample:

- Number of words between 8 and 100, since if the text fragment is too short in words there may not be enough information to answer the task questions. However if it is too long, it takes too much time and effort for the CrowdFlower's workers to complete the task.
- Written in the English language. A Naive Bayes classifier [9] was used to infer the text fragment's language, assuring homogeneity in the sample.
- With no profanity words, in order to avoid compromising the seriousness of the task.
- Containing all the words from at least one group (from section 3.1.1).
- Not a Twitter "retweet". This assures that all the text fragments are unique.

Other pre-processing actions taken included the removal of links from the text. The complete dataset obtained after the control stage was composed of 1913 comments, 132 posts and 14860 tweets.

The text fragments, as specified in section 3.1.2, include official posts from news channel pages as well as comments in these pages, increasing the probability of having both relevant and irrelevant information in the collected fragments.

Finally, a sample of 101 text fragments was selected in order to assure a higher quality control of the fragments and an equal representativity of each keyword, message type and social network (see Table 2). Some statistics regarding the data selected include: posts from 10 distinct pages, comments from 28 unique users and tweets from 48 unique users. On average posts obtained 3247 likes, 741 shares and 573 comments; an average of 56 likes and 7 replies on comments; and an average of 2 favourites and 4 retweets on tweets. Facebook messages are composed, on average, by 22 words, while Twitter messages include an average of 17 words.

3.3 Crowdflower Classification

In order to perform the relevance classification of the dataset, the selected social network messages were incorporated in a classification task in the online platform "Crowdflower". This platform was chosen over other ones (e.g. Mechanical Turk) because it offers more control over the quality of the experiment and the users working on it.

The "CrowdFlower" task consisted in a list of eight questions that the users ("workers") had to answer about the journalistic relevance of a text fragment (see Table 1). The questions were compiled based on the journalistic criteria to find relevant information previously presented ([2, 3, 4]).

Table 1 Questions used in the "Crowdflower" experiment.

Relevance Criteria	**Question**
"Interesting"	Is the topic of the fragment "not interesting" or "interesting"?
"Controversial"	Is the topic of the fragment "not controversial" or "controversial"?
"Positive"	Is the fragment "negative" or "positive"?
"Meaningful"	Is the fragment "private/personal" or "meaningful for the majority of people"?
"New"	Is the information in the fragment "already known" (for the majority of people) or "new"?
"Reliable"	Is the information in the fragment "unreliable" or "reliable"?
"Wide Scope"	Has the information in the fragment a "narrow" or "wide" scope?
"Relevant"	Is the information in the fragment "irrelevant" or "relevant"?

Each of these questions allowed integer answers in a 5 point Likert scale.

One advantage of the "CrowdFlower" platform, as stated before, is the quality assurance among the "workers" in a task. In this study the following conditions were assured:

- Each fragment was classified by 7 different users, in order to analyze the consensus and subjectivity in the task.
- Each user classified at most 10% of the total fragments, because it was desirable to have as much as variability of participants as possible.
- Only Level 3 "CrowdFlower" users could complete this task. This is the best quality allowed in the platform and relates to the performance of the "workers" on test questions [10].
- All users were either from the UK or the USA, in order to control cultural differences.
- It was assured that each user took at least 20 seconds to complete the job, toward avoiding random and unconsidered answers.

After the experiment in "CrowdFlower" was concluded, a dataset was obtained with the text fragments and its classifications. A sample summary is presented in the next subsection.

3.4 Sample Summary

As a result of the previous phases, a total of 707 answers from 82 different users were collected. Regarding the characterization of this sample, 101 text fragments from 10 news providers' pages were included and the distribution of text fragments by keyword and message type is detailed in Table 2.

Table 2 Number of text fragments from each group of keywords and social network.

Keyword	FB Posts	FB Comments	TW Tweets
"Refugees" and "Syria"	5	5	8
"Elections" and "US"	5	5	8
"Olympic Games"	2	5	8
"Terrorism"	5	5	8
"Daesh"	2	5	8
"Referendum" and "UK" and "EU"	4	5	8

4 Exploratory Analysis

In order to better understand the process of relevance classification, an exploratory analysis was conducted using Pearson Correlation. The results of this analysis are presented in Table 3.

Table 3 Correlations between all the questions and the "Relevant" question for the 707 answers.

		"Relevant"
"Interesting"	*r*	0.61
	p	<0.001
"Controversial"	*r*	0.24
	p	<0.001
"Positive"	*r*	0.12
	p	<0.001
"Meaningful to the Majority"	*r*	0.60
	p	<0.001
"New"	r	0.15
	p	<0.001
"Reliable"	*r*	0.60
	p	<0.001
"Wide scope"	*r*	0.65
	p	<0.001

The correlations and *p* values indicate that the more the information is "interesting", "meaningful for the majority", "reliable" and with a "wide scope", the more it is perceived as being "relevant" by the evaluators.

5 Surrogate Feature Extraction

In the previous section some characteristics of the information were presented as indicators of relevance in text fragments. However these variables were dependent on human classification and in order to classify a text fragment as "relevant" or "irrelevant" these features must be extracted automatically from the text or social network information. Therefore, several features were added aiming at replacing each question.

Table 4 Conversion between questions and automatic features.

Relevance Criteria	Goal	Surrogate Features	Description
Interesting	This group of metrics is based on the idea that people will react and share more information if it is interesting.	Number of user mentions	Number of "@" used in the text fragment to refer other users in the same social network
		Number of likes	Number of favorites in a tweet or number of likes in posts or comments from Facebook
		Number of shares	Number of "retweets" of a tweet or the number of shares of a Facebook post
		Comment count	Number of comments of a Facebook post and is not applicable to Twitter
Personal *vs.* Meaningful	Evaluate the subjectivity in the text fragment.	Sentiment Analysis [11]	Processed with the "polarity" function from the package QDAP [12] in R
		Number of Adjectives	Indicator for higher subjectivity [13]
		Number of pronouns (in first or second person)	Referred as an indicator for relevance [7]
Reliability	Use the credibility of the owner of the message.	Verification status	Status (verified or not) of the Facebook/Twitter profile that published the text fragment
		Number of followers	Number of followers of the Twitter profile or number of likes in a page from Facebook

5.1 Relation between relevance criteria and surrogate features

Aiming at evaluating the potential of automatic classification of relevance, a set of surrogate features matching the pre-established relevance criteria were extracted and developed, as represented in Table 4. In order to do so, social media metrics and additional methodologies were incorporated. At this stage, it was possible to correlate three of the relevance criteria with several automated processes. For instance, a set of surrogate social media metrics, such as number of user mentions, number of likes, shares and comments, can be indicative of 'interesting' content. Likely, performing sentiment analysis as well as adjective and pronoun counting can assist on evaluating the subjectivity of the messages. Finally, the verification status and the number of followers can be surrogate features for the relevance criteria 'reliability'.

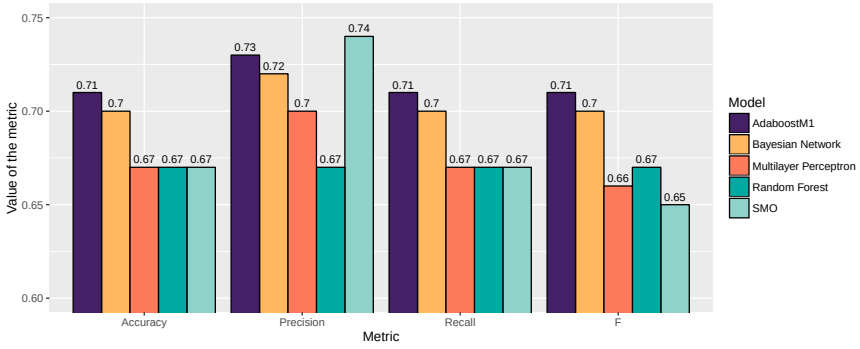

Fig. 2 Accuracy, precision, recall and F measure for each supervised learning algorithm.

5.2 Journalistic Relevance Class

Regarding the "Relevance" question, the numeric answer was converted into categorical. Each answer was transformed into a class according to the following rule: 1 or 2 became "Irrelevant", 3 became "Neutral" and 4 or 5 became "Relevant". Since each text fragment was classified by 7 users several agreement ratios were analysed (see fig. 3).

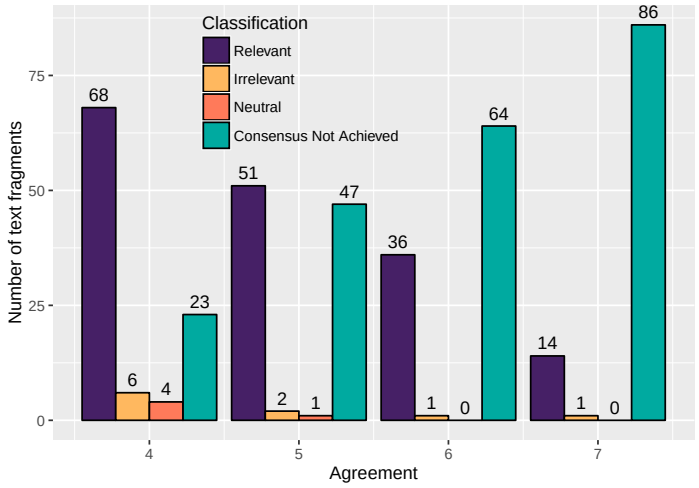

Fig. 3 Number of text fragments in each categorical answer ("Irrelevant", "Neutral" or "Relevant") with different agreement ratios.

In order to balance the number of instances in each class, the chosen agreement value was 5: a text fragment was considered "Relevant" if at least 5 workers answered "4" or "5" for the text relevance question. In any other case ("Irrelevant", "Neutral" or "Consensus Not Achieved") the text fragment was considered "Not Relevant". Therefore with this criteria the number of text fragments considered as "Relevant" and "Not Relevant" was 51 and 50 respectively.

6 Classification Model

In order to understand the importance of each feature, the "Relief F" metric [14] was computed. The results revealed that the message type (which distinguishes "FB Posts" from "FB Comments" and "Tweets"), the number of comments (if applicable) and the verified status of the author of the text fragment are the most influential attributes. The feature ranking obtained with this metric is presented in Table 5.

Table 5 Relief F attributes with value greater than "0".

Features	Ranking Value
message type	0.15
comment count	0.13
verified	0.06
followers count	0.01
shares	0.01

Several experiments with different models were also conducted, with "AdaboostM1" and "Bayesian Networks" being the algorithms which achieved higher accuracy (71% v.s. 70%) and F score (71% v.s. 70%). These results are summarised in fig. 2.

7 Conclusion

In this paper we presented an exploratory study about relevance classification in a journalistic perspective. The first stage of our methodology consisted of: (1) collecting posts from social networks (either from Facebook and Twitter) according to a set of popular, yet controversial, topics; (2) filtering the retrieved posts to gather a dataset with enhanced quality (e.g. with a reasonable quantity of words, written in English, etc); (3) submitting this final set for a classification job in "CrowdFlower".

Our analysis of the results pointed out that interesting, meaningful, reliable and wide scope information is more likely to be considered as relevant for a majority of 5/7 of workers. This exploratory analysis led us to identify surrogate features, which could be accessed/extracted, or computed, automatically to predict relevance.

In a second stage we applied five machine learning algorithms to our golden standard. In almost all metrics (accuracy, precision, recall and F-value) the "Bayesian Networks" and the "AdaboostM1" have the best performance for the available data. Regarding the features used, we found out that "message type" and "comment count" are the most important ones for this analysis. Besides, the significant correlations, the accuracy and the F-value showed that the quality control validated the proposed methodology to detect relevance in social network messages.

Finally, for the future work two different goals could be considered. Firstly it is important to increase the sample size of classified messages with the intent of strengthening the confidence in the methods used. Secondly new surrogate features should be researched (e.g. related with the wide scope of the information in the text fragments) to complete the automatic classification relevance model.

Acknowledgements This work is financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundao para a Ciłncia e a Tecnologia (Portuguese Foundation for Science and Technology) within project "Reminds/ UTAP-ICDT/EEI-CTP/0022/2014".

References

1. S Shyam Sundar, Silvia Knobloch-Westerwick, and Matthias R Hastall. News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology*, 58(3):366–378, 2007.
2. Allan Bell. *The language of news media*. Blackwell Oxford, 1991.
3. Johan Galtung and Mari Holmboe Ruge. The structure of foreign news the presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.
4. Herbert J Gans. *Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time*. Northwestern University Press, 1979.
5. Mursel Tasgin and Haluk O Bingol. Gossip on weighted networks. *Advances in Complex Systems*, 15(supp01):1250061, 2012.
6. Marco Guerini, Carlo Strapparava, and Gözde Özbal. Exploring text virality in social networks. In *ICWSM*, 2011.
7. Alvaro Figueira, Miguel Sandim, and Paula Fortuna. An approach to relevancy detection: Contributions to the automatic detection of relevance in social networks. In *New Advances in Information Systems and Technologies*, pages 89–99. Springer, 2016.
8. The search api — twitter developers. https://dev.twitter.com/rest/public/search. Accessed: 2016-04-21.
9. Shuyo Nakatani. Language detection library for java, 2010. Accessed: 2016-04-21.
10. Crowdflower. Crowdflower community - introducing contributor performance levels!, 2014. Accessed: 2016-04-28.
11. Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.
12. Tyler W. Rinker. *qdap: Quantitative Discourse Analysis Package*. University at Buffalo/SUNY, Buffalo, New York, 2013. 2.2.4.
13. Robert Rittman, Nina Wacholder, Paul Kantor, Kwong Bor Ng, Tomek Strzalkowski, and Ying Sun. Adjectives as indicators of subjectivity in documents. *Proceedings of the American Society for Information Science and Technology*, 41(1):349–359, 2004.
14. Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.