Detecting Journalistic Relevance on Social Media

A two-case study using automatic surrogate features

Álvaro Figueira * and Nuno Guimarães **
Department of Computer Science
CRACS / INESC TEC & University of Porto
Rua do Campo Alegre 1021/1055, Porto, Portugal
* arf@dcc.fc.up.pt
** nuno.r.guimaraes@inesctec.pt

***Abstract*— The expansion of social networks has contributed to the propagation of information relevant to general audiences. However, this is small percentage compared to all the data shared in such online platforms, which also includes private/personal information, simple chat messages and the recent called 'fake news'. In this paper, we make an exploratory analysis on two social networks to extract features that are indicators of relevant information in social network messages. Our goal is to build accurate machine learning models that are capable of detecting what is journalistically relevant. We conducted two experiments on CrowdFlower to build a solid ground truth for the models, by comparing the number of evaluations per post against the number of posts classified. The results show evidence that increasing the number of samples will result in a better performance on the relevancy classification task, even when relaxing in the number of evaluations per post. In addition, results show that there are significant correlations between the relevance of a post and its interest and whether is meaningfully for the majority of people. Finally, we achieve approximately 80% accuracy in the task of relevance detection using a small set of learning algorithms.**

Keywords—Relevance detection, Social networks, Text mining, CrowdFlower, Journalistic criteria, Surrogate features.

I. 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.

ASONAM '17, July 31-August 3, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07/$15.00
https://doi.org/10.1145/3110025.3122120

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 of a short circle of users. Users tend also to post private, personal, or just a very narrow scope information on 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].

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 just 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 'gossip analysis' [5] is based on the structure of the network, that propels information spreading, other researchers argue 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 [6]. Moreover, this research conducted on text virality indicates that common social network metrics alone (e.g. #likes, #retweets) are not sufficient for assessing such a complex phenomenon.

In addition, reinforcing the above-mentioned criteria, suggests that several virality components should be

considered, such as: appreciation, spreading, simple buzz, white buzz, black buzz, raising discussion and controversy. Similarly to these notions, 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 [1] and in [6], namely: 'controversy' 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 II.C.

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 "relevant" or "not-relevant" information. However, limitations identified by the authors 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.

II. THE RESEARCH METHODOLOGY

In order to detect relevance (or irrelevance) in text fragments, a methodology was proposed in another previous study [8]. The phases of this methodology 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.

A. Data Retrieval from Social Networks

The data was collected between the 1st and the 4th April 2016. It included Facebook posts and comments, and Twitter tweets. Data retrieval on Twitter was conducted by presenting the API with ten keywords ("refugees", "syria", "elections", "US", "Olympic Games", "terrorism", "daesh", "referendum", "UK" and "UE"), which were distributed by 100 queries.

In what Facebook is concerned, data retrieval was performed on the pages of fourteen international news providers ("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") and then only posts with the previous mention keywords were considered. The difference between the collection method among the two networks was enforced by the restrictions of their own API.

B. Data Preparation

Since the fragments were gathered for inclusion in a "CrowdFlower" task, it was important to guarantee that the text included a minimum of quality standards to be analyzed. The conditions that the text fragments had to fulfill to be included in the sample were: a number of words between 8 and 100 (to avoid absence of information or a considerable effort to classify), written in English (we used the [9] Java library), no profanity or slang words (to avoid compromising the seriousness of the task), no URLs, and no "retweets". Other pre-processing actions taken included the removal of emoticons and special characters.

Finally, a sample of 101 fragments was selected, assuring some quality control of the textual information and an equal representability of: each keyword, message type and social network.

C. CrowdFlower Classification

To classify the text fragments retrieved we rely on CrowdFlower since it offers more control over the experiment and the workers, when compared with Mechanical Turk (used in previous studies [7]). Therefore, each fragment of text was classified by 7 different workers from the UK or the USA (to control and diminish the cultural differences). In addition, none would be able to classify more than 10% of the total fragments, as was desirable to have as much as variability in the participants as possible. Furthermore, only the workers with the highest level of quality [10] would be allowed to complete this task and it was assured that each worker couldn't complete a job in less than 20 seconds (to help prevent random and unconsidered answers).

The "CrowdFlower" task consisted in reading a text fragment and answering a list of eight questions about the 'journalistic relevance' of the information included in that fragment (interest, controversy, positiveness, meaningfulness, reliability, novelty, wide/narrow scope, and relevance). The created questions were based on the journalistic criteria to find relevant information previously presented ([2], [3], [4]).

After the experiment in "CrowdFlower" was concluded, a dataset was obtained with the text fragments and its classifications. A total of 707 answers from 82 different users were collected.

D. Exploratory Analysis

To better understand relation between variables in the process of relevance classification, an exploratory analysis was conducted using the Pearson Correlation. The correlations and values indicate that the more the information is "interesting" (r=0.61), "meaningful for the majority" (r=0.60), "reliable" (r=0.60) and with a "wide scope" (r=0.65), the more it is perceived as being "relevant" by the evaluators.

III. Relevance Classification Experiments

In this section, we present the criteria to establish the label in the training and testing dataset, the features extracted and their importance, and the results achieved in a small set of training models.

A. Journalistic Relevance Classification

Regarding the "Relevance" question, the numeric answer (using a 5-points Likert scale, to introduce some "confidence" of the worker) was converted into categorical (1 and 2 are not relevant, 3 is neutral, and 4 and 5 are relevant). Then an agreement of 5 out of 7 workers was established. Following these criteria, a balanced dataset (50% relevant, 49% not relevant) was achieved.

B. Surrogate Features

A set of surrogate features matching the pre-established relevance criteria were extracted and developed. 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 used as surrogate features for the relevance criteria 'reliability'.

C. Feature Importance

To understand the relative importance of each feature in this classification, we used the "Relief F" metric. Relief F is a statistical and weight based feature selection algorithm that is heuristic-independent [11]. 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 for the workers with ranking 0.15, 0.13, and 0.06 respectively.

D. Machine Learning Models

Now, equipped with a train set and relevant features to use, we conducted several experiments with different machine learning algorithms. From these experiences, we concluded that "AdaboostM1" and "Bayesian Networks" were the algorithms which achieved a higher accuracy (71% vs. 70%) and an F-score (71% vs. 70%).

E. Synthesis

This concludes the first stage of our research. Our current state is therefore: 1) we wanted to understand which characteristics are more likely to be used by people as indicators of relevance when relevancy is perceived. We followed a set of criteria commonly used by journalists to decide upon what is relevant. 2) We conducted an experiment in CrowdFlower which allowed us to focus on features taken from social media posts/tweets that potentially surrogate the journalistic criteria found to be important. 3) Using a ground truth created by the CrowdFlower workers we trained and tested several models achieving an accuracy and F-score around 70%.

IV. Increasing the number of samples

In the previous section, due to our CrowdFlower experiment on assessing a "ground truth" for the relevance classification task, we had high accuracy in several machine learning models. However, as the sample was quite small we assumed that it affected the strength on the confidence of the method proposed.

Therefore, we conducted a new CrowdFlower experiment, this time with 840 posts to be evaluated. In addition, we also relaxed the number of workers from 7 in the first experiment to 3 in this one. Although we recognize that a higher number of workers per post reinforces the confidence on the ground truth of our methods, we wanted to study how the number of evaluations affect the learning phase of the algorithms, especially in ambiguous tasks like relevancy classification. Our hypothesis is that with more learning cases, the overall performance of the machine learning algorithms will not be affected by a decrease on the number of workers.

The crawling methodology and text pre-processing were similar. However, due to the higher number of samples for classification and the reduce number of Level 3 CrowdFlower workers, we also had to include level 2 workers. All the other conditions regarding workers (mentioned in the previous experiment) were met.

A. Comparison of correlations

After the experiment finished, we proceeded to analyze the correlations between the "Relevance" question with the ones which were more correlated in the previous experiment (i.e. "Interesting", "Meaningful", "Reliable" and "Wide Scope") and compare them. The results reinforce and provide evidence that if the post is interesting and meaningfully for the majority, the more it is perceived as relevant. In fact, the correlation of both these questions increase in this second CrowdFlower experiment (r=0.72 and r=0.644 respectively). The "Wide Scope" feature slightly decreases (but not significantly) regarding the previous experiment. The most noticeable decline concerns the reliability question. However, despite being modest, 0.511 is still a correlation to consider in this analysis.

B. Feature Extraction and Classification Model

We advance to extract the surrogate features. At this point, we decided to extract the same features for comparison purposes. Regarding the agreement of the workers, we decide to classify a post as relevant if all workers classify it as such. Otherwise, the post is labelled as not relevant. After the dataset (with the automatically extracted features) was built, we proceed to compute the "Relief F" metric as we did in the previous experiment. We highlight some differences between this analysis the one conducted in the previous experiment: first, the ranking values are approximately 10

times smaller and the number of features with ranking values above 0 has increased. This is possibly due to the difference on number of entries (from 101 to 840) and, consequently, to the diversity of the sample. Second, the most influential attributes differ from the 101 posts experiment. The user mentions, the sentiment and the number of adjectives are the ones that achieve a higher-ranking value in this sample.

We then trained again the same machine learning algorithms (AdaboostM1, Bayesian Networks, Multilayer Perceptron, Random Forest, and Sequential Minimal Optimization) using the same training parameters as the previous experiment. The results provide evidence that despite relaxing on the number of evaluators, the machine learning models tested were not affected. In fact, the increase in the number of observations seems to compensate the lack of a higher number of evaluators per post, leading us to conclude that the model is robust. There is an overall increase on the values of the metrics used in the two experiments since all models increased in accuracy and only AdaBoost did not increase in F1-measure). The "Multilayer Perception" algorithm was the one reaching the highest difference in terms of F1-measure (6%) and the "SMO", the one in terms of accuracy (14%).

V. Conclusions

We presented an exploratory study about relevance classification in a journalistic perspective. We designed two different experiments but using the same methodology. 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); (3) submitting this final set for a classification job in CrowdFlower. The first experiment was conducted with 101 posts where each one was evaluated by 7 different CrowdFlower workers with a Level 3 performance. 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 of the experiment we applied five machine learning algorithms to our golden standard. In almost all computed metrics (accuracy, precision, recall and F-value) the "Bayesian Networks" and the "AdaboostM1" has 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. For our second experiment, we increased the number of posts to strengthen the learning of the several machine learning algorithms. We also relaxed on the number of workers to verify if the quality of our ground truth can be assured with a smaller number of evaluations per post. We started by comparing the correlations between both experiments, which reinforce what was already suggested on the first experiment. In addition, there is an increase on the correlation values between the question of "relevancy" and the questions of "interestingness" and "meaningful for the majority". The models also perform better in the second experiment achieving approximately 80% accuracy using the AdaBoostM1 algorithm. The significant correlations, the accuracy and the F-measure of both experiments showed that the quality control validated the proposed methodology to detect relevance in social network messages. In addition, it was also presented evidence that a higher number of samples, even when decreasing the number of workers, can achieve better results on the journalistic relevance classification task.

For future work, we propose to create two different workflows, depending on the source of the text (i.e. in which social network was extracted). Consequently, we will provide the models with automatically extracted features specific of each social network (for example, we can analyze the discrepancy of sentiment on the comments of a Facebook post or establishing links using hashtags in Twitter posts). Therefore, by having a designated model for each social network, our goal is to increase furthermore the overall performance on the journalistic relevance classification task.

Acknowledgment

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 – Fundação para a Ciência 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] Álvaro 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] Miguel Sandim, Paula Fortuna, Álvaro Figueira, and Luciana Oliveira. Journalistic Relevance Classification in Social Network Messages: an Exploratory Approach, pages 631–642. Springer International Publishing, Cham, 2017.

[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] Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In AAAI, volume 2, pages 129–134, 1992.