An Approach to Relevancy Detection: contributions to the automatic detection of relevance in social networks

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

Abstract. In this paper we analyze the information propagated through three social networks. Previous research has shown that most of the messages posted on Twitter are truthful, but the service is also used to spread misinformation and false rumors. In this paper we focus on the search for automatic methods for assessing the relevance of a given set of posts. We first retrieved from social networks, posts related to trending topics. Then, we categorize them as being news or as being conversational messages, and assessed their credibility. From the gained insights we used features to automatically assess whether a post is news or chat, and to level its credibility. Based on these two experiments we built an automatic classifier. The results from assessing our classifier, which categorizes posts as being relevant or not, lead to a high balanced accuracy, with the potential to be further enhanced.

Keywords: Automatic, Relevance Detection, Social Networks.

1 Introduction

Social Networks have an inherent capacity to spread information at a much higher pace than traditional media. They allow users to post and exchange messages almost instantaneously all over the world. This constitutes an ideal environment for the dissemination of news and of important information directly from their sources, or from the location of the events.

We have seen many cases of emergency situations [1] where some users disseminate information either by providing personal observations, or by sharing received messages from external sources, in their posts. From this pool of information, users generally combine and synthesize what they read, and then elaborate to produce their own interpretations, in a continuous cycle.

While this process can gather, filter, and propagate information very rapidly, it is not able to separate relevant information from simple false rumors. In 2010 we observed that immediately after a reported earthquake, many posts in Twitter did spread rumors which contributed to an increase of insecurity in the population [2]. Interestingly, we also observed that the spread of this information which turned out to be false was much more questioned than information which ended up being true.

Nevertheless, we also know that information disseminated from official and reputable sources is often considered more valuable, more shared/propagated, and generically understood as having a positive degree of relevance.

1.1 Research Focus and Outline of the Methodology

This introduction serves to focus our research which is to understand the spread of news information and its credibility over social media networks. In this paper we consider as "news" the information that is relevant to a large audience, as opposed to information which may be important, but only to a reduced set of people. We also use "credibility" in the sense of believability.

Our approach is based on a supervised learning methodology. We first identified a set of four relevant discussion topics. Then, each post on the topic was labeled by humans according to whether it corresponds to newsworthy information or to an informal conversation. After the data set is created, each item of the former class is assessed on its level of credibility by human judgement.

Our objective is to determine if we can automatically distinguish news from informal chat and, in the former case, to assess the level of credibility of content posted in social networks.

In the following section we focus on related work, namely why social networks are important sources for news and the way credibility relates with them. In a second stage, we extract some relevant features from each labeled topic and use them to build an automatic classifier that attempts to automatically determine if a social network message corresponds to newsworthy information. As a second step we try to, also, automatically assess its level of credibility (section 4). The base and rational of our automatic classifiers is described in section 5. Finally, in section 6 we present our conclusions and point out directions for future work.

2 Related Work

The literature regarding automatic assessment of newsworthy information has been increasing in the last couple of years and, regarding information credibility, it is even more extensive. Therefore, in this section our coverage is by no means complete nor extensive. Instead, we try to provide an outline of the research that is most closely related to the one used in this paper.

2.1 Social Networks as News Media

While most messages on social networks are conversational, people also use them to share relevant information and to report news [3,4,5]. Indeed, according to a 2010 study on the Twitter social network, the majority of trending topics can even be considered "headline news" [6]. For example, Twitter spreads news stories from traditional media like in the case of recent epidemics [7], it detects news events [8], also geolocating such events [9]. Another important feature is that it can find emergent and controversial topics [10]. Recently, it was described [11] an online monitoring system to perform trend detection over the twitter stream (although, many other systems currently exist). In the same line, we must also recall “Google Trends” as a similar system based on the user performed queries. Social networks have also been used during several emergency situations to share information [1,12].

2.2 Credibility

The perception of users with respect to the credibility of online news can be seen generically, as positive. In fact, apart from newspapers, people trust the Internet as a news source as much as other media [13]. A study conducted by Flanagin and Metzger [14] showed that in an absence of external information, the perceptions of online credibility is strongly influenced by style-related attributes, including the visual design, which is not directly related to the content itself. However, another study [15] showed that users may change their perception of credibility depending on the (supposed) gender of the author.

Meanwhile, we witness some search engines starting to display search results from social networks, particularly for trending topics. As a consequence, “spammers” are attracted to this mean of communication, which then leads the readers to a sense of distrust. The same reaction happens with some web pages that are heavily populated with offers of products or services [16].

Recently, researchers from Indiana University created the “Truthy” service, which has started to collect, analyze and visualize the spread of tweets belonging to “trending topics”. The system uses features that are present in the tweets collected in order to compute a *truthiness score* for a set of tweets [17].

3 Data Retrieval

This section describes how we collected a set of messages related to pre-defined topics, from social networks

The topics, or *criteria* (using the social networks API terminology), were chosen according to what has populated, during the last weeks, the standard media headlines. Our intention was to have, as much as possible, topics that would lead to discussions, to the addition of extra/complementary information, to critics, as well as, eventually, to “passionate” arguments. All in all, we picked topics that supposedly would lead to a burst of information in the networks. The topics were:

- *Refugees*
- *Migrants*
- *Donald Trump*
- *Windows 10*

We collected the posts, the replies, the tweets, and the retweets (for the sake of simplicity we will call all these messages simply as ‘posts’), during almost 72 hours. We then labeled each post according to the topic used for the respective query to the social network. In the end of the collection period we had 15980 posts taken from the three social networks, respectively a shown in Table I:

Table 1. Number of collected post per social network.

Social Network	Number of collected posts
Google+	157
YouTube	8000
Twitter	7823

3.2 The News Assessment Procedure

Our first labeling round was intended to separate posts which spread information about “news” from cases concerning personal opinions, or out of topic posts, or even simple “chat”. To help us in this task we employed the Amazon Mechanical Turk service (MT), which uses workers to perform typically small and intelligent tasks (*Human Intelligent Tasks* – HITs). We presented to the evaluators the collected posts, each one labeled with the corresponding topic assigned during its retrieval. Then, we asked if the presented post was spreading news about a specific event, or if it was just conversation. The former case was labelled as ‘NEWS’ and the latest as ‘CHAT’.

The final set of posts was reduced due to incompatibilities between our initial set and the Amazon MT system. We discarded posts that had special characters not recognized by the MT system (mainly due to different encodings), and some other because they were too short (less than eight words), or even some other that were composed only by URLs. From this new data set, we randomly we selected 481 posts to present to evaluators. Each of these 481 posts was presented as a HIT to 5 different evaluators. We, then, assessed the panorama of agreement between evaluators: majority of three; of four; and, unanimity, concerning the number of labeled posts, as illustrated in Fig. 1.

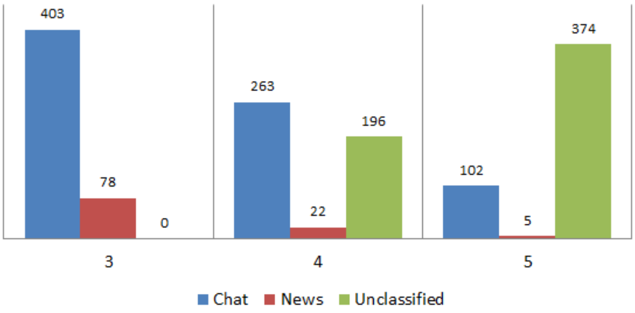

Fig. 1. Number of labeled messages according to majority levels.

We observed that unanimity for 'NEWS' was only possible in 5 posts (only 1% of the set), while at that scenario a large majority of posts (78%) was considered unclassified. For this reason, we decided to relax our level of agreement and to consider instead as an agreement, a majority of 3 evaluators. Therefore, a class label for each post was assigned if 3 out of 5 evaluators agreed on the label. In the other case, we label the posts as 'Unclassified'. Using this procedure 0% of the posts were left unclassified, being the whole sample distributed among two classes: 84% (403 cases) as 'CHAT', and 16% (78 cases) as 'NEWS'.

3.3 Credibility Assessment

After having a set of posts considered as NEWS, we focused on determining the credibility assessment of each post. We ran again a set of HITs over the collection of 481categorized posts. Using this collection of instances, we asked MT evaluators to indicate the credibility level for each message. We also provided the topic of the message in order to help them better understand the context.

In this evaluation we considered three levels of credibility: (i) "likely to be true", (ii) "likely to be false", and (iii) "I can't decide". As in the first round, we asked for 5 different assessments of each HIT. Labels for each topic were decided my majority, requiring an agreement of at least 3 evaluators. We illustrate the resulting panorama in figure 3. This result led us to pick a majority of three evaluators.

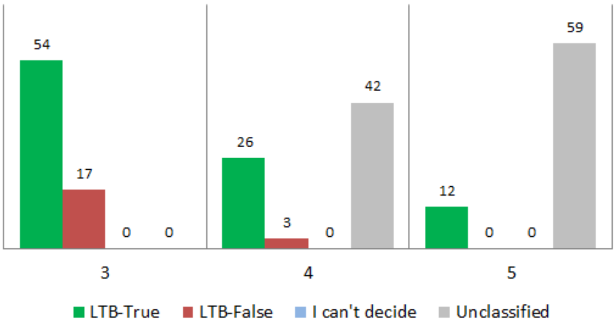

Fig. 2. Assessing credibility by majority.

In this round of evaluation, we also tried to distinguish the credibility of the messages according to the social network used. In the scenario of labeling messages by a majority of 3 we got that Google+ had 90% of "likely to be true" messages (and 10% of "likely to be false" messages), and YouTube had 73% of "likely to be true" messages (and 27% of "likely to be false" messages). Therefore, in this scenario there were no messages unclassified. This result is expressed in Table 2.

Table 2. Distribution of trustiness in the collected sample.

Social Network	Likely to be true	Likely to be false	Can't decide/Unclassified
Google+	90%	10%	0%
YouTube	73%	27%	0%

4 Creating an Automatic Analysis

4.1 Social Media Credibility

Our main hypothesis in this article is that the level of credibility of information disseminated through social media networks can be estimated automatically, up to a certain degree. We believe that there are several factors that can be observed in the social media platform itself and, in the message in particular, that are useful to assess the information credibility. However, we do also know that MT evaluators tend to fulfil every HIT as fast as they can, which usually leads to a reduced care when categorizing texts. We are aware that, as we get thousands of different evaluators, this behavior becomes common for some sub-set of the evaluators. Nevertheless, it should be taken into consideration and mitigated in future analyses.

From this experience, using our dataset, we propose, as hypothesis, that a reduced set of features take exclusively from the posts are sufficient to perform the automatic analysis with a satisfying degree of accuracy:

- The length of a post
- A set of words typically used in credible posts
- The number of occurrences of certain words
- The use of excessive punctuation
- The abundant use of smileys/emoticons

4.2 Automatically Detecting News

Consistently, the number of words in posts labeled as NEWS is bigger than those in posts labeled as CHAT. On the other hand, the use of pronouns is consistently bigger in CHAT posts. As a consequence, we created a "bag of words" that are more prone to appear in NEWS and that do not appear in CHAT posts. This set (117 words) was created using a traditional term frequency times the inverse document frequency metric (ie, the standard tf-idf). We also set thresholds for the length of a post according to the social network, and for a scoring function. This function scores when finding in a post, a word that is present in the bag, as well as the number of its occurrences. Finally, we fine-tuned the number of "symbols" for each social network setting thresholds for punctuation and smileys/emoticons.

4.3 Automatically Assessing Credibility

Similarly to the news detecting methodology, for assessing the credibility we took an approach based on the intrinsic characteristics of the post.

However, in this case it is not easy to use the bag of words because many of the words taken from the two categories, "Likely to Be True" (LBT), and "Likely to Be

False” (LBF), expressed using the tf-idf frequency are quite the same. I.e., there is much overlapping between the two sets. This situation is illustrated using word clouds (Fig. 3). Therefore, credibility cannot be assessed using only this type of criteria.

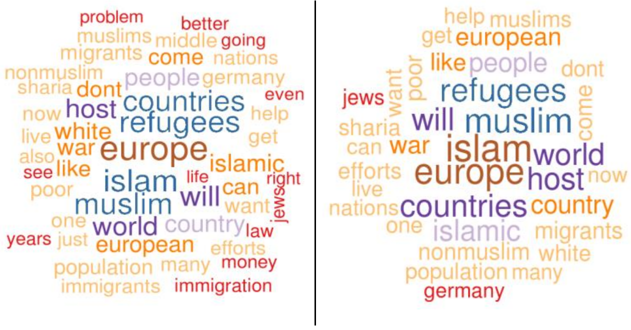

Fig. 3. Word clouds: “Likely to Be True” (left) and “Likely to Be False” (right).

Hence, we discarded the bag of words and created a function to promote/demote the resulting value according to: the post length, to the corresponding social network where the post was take from and, we set thresholds for the number of punctuation signs and smileys found. In the end of the automatic analysis, each post got a final score which was then used to categorize it.

5 Automatically Detecting News and Assessing Relevance

Using our crawlers, we retrieved a new set of posts to be categorized by our automatic classifier and by the human evaluators at Mechanical Turk in order to compare the two categorizations. Our goal was to compare the precision of our system, tuned according to the two past experiments, against the human classification, and also to perform an assessment and analysis at the feature level. We retrieved 100 posts from the same social networks and according to the same criteria as in the previous study.

5.1 The MT Assessment

For this task we submitted to MT the whole set, such that each post was presented to 5 evaluators. Therefore, we had a total of 500 posts to be categorized.
This time instead of asking the workers to distinguish between ‘news’ and ‘chat’, we asked something slightly different: we asked them to classify the posts as “relevant” or “irrelevant” according to instructions described in **Table 1**.

Table 3. MT instructions for selection criteria.

Categories	Includes	Excludes
Relevant	Facts or data relevant for a journalist to write news	-----
Irrelevant	Facts or data only relevant to a reduced set of people. Information for which you can’t derive the context.	Information that could be incorporated in news.

Evaluators were also asked to not pay attention to: (1) misspelled words; (2) bad grammatical constructions; (3) verbs in the 1st person. Our rationale for these rules/instructions if that when people are asked to identify news, they are expecting to find a certain style of writing, which is characteristic from traditional media like television and radio. In our case, we don't want to be able to find a particular style; instead, we want to identify social network messages which may contain relevant information. Therefore, we ask the evaluators to check if, in the message, there is information capable of being adapted to be part of a news. Despite our intention to have 5 evaluators per post, we had some posts classified by the same person, which we immediately discarded form the resulting set to analyze, in the end we got 81% classified posts by majority and only 19, which we couldn't resolve, were left as unclassified, as illustrated in Fig. 4 (left chart).

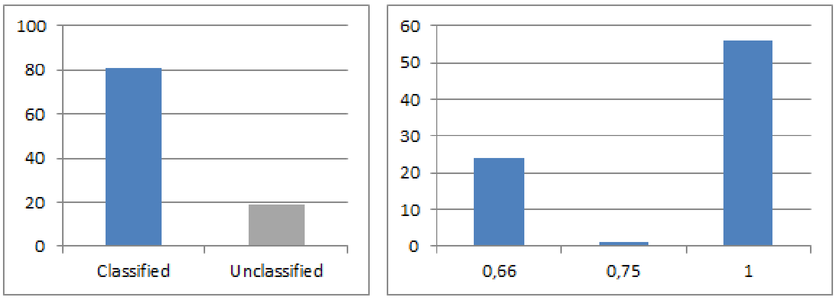

Fig. 4. Results from MT classification on the 100 posts.

We can also see (right chart of Fig. 4) that, regarding the classified posts (81 posts) more than 50 were classified in unanimity; 24 posts by majority of two-thirds (0.66) and only one by majority of three-quarters (0.75).

5.2 The Automatic Assessment

Our developed classifier took into account the features described in section 4.1 and in 4.2 of this article. In particular, we used the length of a post as a normalization process. Another important feature is the number of pronouns ('mine', 'ours', etc.) that are present in the post, usually together with sentences which include 'I', 'me' or 'you'. For example, from our results posts with more than 5 occurrences of pronouns tend to be categorized as non-relevant, or chat. We used thresholds for the amount of '!' and of smileys. The threshold is computed independently for each sentence in a post. Finally, we took into account the use of "bad language" or swear-words. In cases where these kind of words are used, the score of the relevancy for that post reduces in a percentage of the length of the post.

5.3 Assessment of the Results

Our methodology to assess our system's performance was to compare the predicted label computed for each post, against the label assigned to the post by the MT evaluator's. We recall that we tried to use up to 5 evaluators for each post (although in the end we had to discard some, as discussed in section 5.1). Therefore, in some cases we used the majority of the evaluator's opinions.

The metrics used for the evaluation were the 'precision', the 'accuracy' (or 'rand index'), and the 'F1', which are standard metrics for binary classifications. To compute these metrics we used the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In this context, we define TP as reporting a success in predicting some post to be relevant; a TN as reporting a success in predicting some post to be irrelevant; and, FP and FN as reporting errors predicting for each of the two cases, respectively.

Table 4. Computed metrics for the assessment.

Metric	Value
TP	24
TN	35
FP	38
FN	3

Metric	Value
Precision	0,38709
Accuracy	0,59
F1	0,68417
Balanced Accuracy	0,68417

Although the 'precision' is low, it corresponds to the "closeness of agreement among a set of results" (ISO 5725 definition), which is not particularly important in this cases. On the other hand, 'accuracy' corresponds to "the closeness of a measurement to the true value" (ISO 5725), which is certainly important. The 'F1' metric, which combines the precision and the sensitivity, is also at a good level. We also computed the balanced accuracy, in order to avoid the effects derived from the unbalanced dataset (i.e., the unbalanced number of relevant vs. irrelevant posts). It is defined as:

$$\mathit{balanced\ accuracy} = \frac{0.5\ \times TP}{TP + FN} + \frac{0.5\ \times TN}{TN + FP}$$

Which, in the case of this experiment, results in a value of more than 68%.

6 Conclusions

Online users have at their disposal a constantly growing number of tools to spread their opinions or share information gathered from other sources. However, more than often the information is not relevant to most readers. Therefore, it is important to design systems that can help a reader to detect on his behalf what may be relevant information, in the sense of a set of data/facts that might interest to a broad audience.

In this article we presented an approach to create a system that is able to automatically detect relevant information in messages posted on most common social networks. During our research, despite believing that the categorization made by evaluators from Mechanical Turk may be not as accurate as a specialist would do (a journalist, for example), we found that a system based on very simple characteristics retrieved from the posts is able to achieve an accuracy of almost 70%. This result leads us to believe it is possible to achieve even higher accuracy using just the inherent characteristics of the posts.

Acknowledgments. 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. Vieweg. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. In Computer Supported Collaborative Work, February 2010.
2. Marcelo Mendoza , Barbara Poblete , Carlos Castillo, Twitter under crisis: can we trust what we RT?, Proceedings of the First Workshop on Social Media Analytics, p.71-79, July 25-28, 2010, Washington D.C., District of Columbia [doi:10.1145/1964858.1964869]
3. Akshay Java , Xiaodan Song , Tim Finin , Belle Tseng, Why we twitter: understanding microblogging usage and communities, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, p.56-65, August 12-12, 2007, San Jose, California [doi:10.1145/1348549.1348556]
4. Mor Naaman , Jeffrey Boase , Chih-Hui Lai, Is it really about me?: message content in social awareness streams, Proceedings of the 2010 ACM conference on Computer supported cooperative work, February 06-10, 2010, Savannah, Georgia, USA [doi:10.1145/1718918.1718953]
5. Pear Analytics. Twitter study. http://www.pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf, August 2009.
6. Haewoon Kwak , Changhyun Lee , Hosung Park , Sue Moon, What is Twitter, a social network or a news media?, Proceedings of the 19th international conference on World wide web, April 26-30, 2010, Raleigh, North Carolina, USA [doi:10.1145/1772690.1772751]
7. Vasileios Lampos , Tijl De Bie , Nello Cristianini, Flu detector: tracking epidemics on twitter, Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III, September 20-24, 2010, Barcelona, Spain
8. Jagan Sankaranarayanan , Hanan Samet , Benjamin E. Teitler , Michael D. Lieberman , Jon Sperling, TwitterStand: news in tweets, Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 04-06, 2009, Seattle, Washington [doi:10.1145/1653771.1653781]
9. Takeshi Sakaki , Makoto Okazaki , Yutaka Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th international conference on World wide web, April 26-30, 2010, Raleigh, North Carolina, USA [doi:10.1145/1772690.1772777]
10. Ana-Maria Popescu , Marco Pennacchiotti, Detecting controversial events from twitter, Proceedings of the 19th ACM international conference on Information and knowledge management, October 26-30, 2010, Toronto, ON, Canada [doi:10.1145/1871437.1871751]
11. Michael Mathioudakis , Nick Koudas, TwitterMonitor: trend detection over the twitter stream, Proceedings of the 2010 ACM SIGMOD International Conference on

Management of data, June 06-10, 2010, Indianapolis, Indiana, USA [doi:10.1145/1807167.1807306]

12. Bertrand De Longueville , Robin S. Smith , Gianluca Luraschi, "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires, Proceedings of the 2009 International Workshop on Location Based Social Networks, November 03-03, 2009, Seattle, Washington [doi:10.1145/1629890.1629907]
13. A. J. Flanagin and M. J. Metzger. Perceptions of internet information credibility. Journalism and Mass Communication Quarterly, 77(3):515--540, 2000.
14. A. J. Flanagin and M. J. Metzger. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. New Media Society, 9(2):319--342, April 2007.
15. C. L. Armstrong and M. J. Mcadams. Blogs of information: How gender cues and individual motivations influence perceptions of credibility. Journal of Computer-Mediated Communication, 14(3):435--456, 2009.
16. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), July 2010.
17. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. arXiv, Nov 2010.