Alethiometer: a Framework for Assessing Trustworthiness and Content Validity in Social Media

Eva Jaho, Efstratios Tzoannos, Aris Papadopoulos, Nikos Sarris
Innovation Lab
Athens Technology Center
10, Rizariou str, 152 33 Halandri, Athens, Greece
{e.jaho, e.tzoannos, a.papadopoulos, n.sarris}@atc.gr

ABSTRACT

There are both positive and negative aspects in the use of social media in news and information dissemination. To deal with the negative aspects, such as the spread of rumours and fake news, the flow of information should implicitly be filtered and marked to specific criteria such as credibility, trustworthiness, reputation, popularity, influence, and authenticity. This paper proposes an approach that can enhance trustworthiness and content validity in the presence of information overload. We introduce Alethiometer, a framework for assessing truthfulness in social media that can be used by professional and general news users alike. We present different measures that delve into the detailed analysis of the content, the contributors of the content and the underlying context. We further propose an approach for deriving a single metric that considers, in a unified manner, the quality of a contributor and of the content provided by that contributor. Finally, we present some preliminary statistical results from the examination of a set of 10 million twitter users, that provide useful insights on the characteristics of social media data.

Keywords

social media; news; content validation; web application

1. INTRODUCTION

Social media refers to the interaction among people in virtual communities and networks, powered by Web 2.0 technologies, during which they exchange news, ideas, and generally information. However, it is not easy to digest the massive amounts of information that the community is offering. Hundreds of new blogs are appearing every day, hundreds of thousands of pictures and videos are uploaded and millions of tweets are posted every minute. Validation of content or presentation of it in an objective manner is a crucial challenge, in order to avoid manipulations and guarantee the democratic role of the media.

WWW'14 Companion, April 7-11, 2014, Seoul, Korea.
ACM978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579324.

In this paper, we present the framework of a platform for assessing truthfulness in one of the most popular social media, Twitter. It is a dynamic micro-blogging platform with an abundance of meta-data and text-based analysis capabilities, around three axes: *Contributor*, *Content* and *Context*. The analysis of the validity of *Contributor* concerns parameters such as trust, reputation and influence of an information source. *Content* validity is expressed through parameters such as the language used, the history and possible manipulations performed on the content. And finally, *Context* analysis examines whether the 'what', 'when' and 'where' of an online publication concur with each other. Joint analysis of the validity of *Contributor*, *Content* and *Context* provides a more thorough approach for revealing truthfulness.

There are recent results in the literature that justify the need to study a variety of parameters for the validation of information in social media. In [1], the authors find that a content-based analysis of author postings on a social networking platform is especially useful in identifying relevant and credible users to follow, apart from the connection characteristics of the authors in the social graph. They present a model for automatically identifying and ranking social media users according to their relevance and expertise for a given topic. Regarding the study of influence, the authors in [2] conclude that users who have high indegree are not necessarily influential in terms of spawning retweets or mentions, and therefore a variety of other measures must also be examined. Finally, a review of state-of-the-art technologies for automated deception detection in social media in [5] has identified multiple types of deceptive behavior: reputation fraud, opinion spam and social spam, each of which requires different detection methods.

Most of the previous research has focused on validating either the source of content or the content itself, but not these two aspects simultaneously. Furthermore, the analysis of the context of a post or article (publication date, place, etc.) and its coherence to the content itself can reveal mistakes that are often hidden in a well written text. Our approach also deals with the reputation fraud problem by taking into consideration, apart from contributor reputation, the validity and coherence of the content and context.

This position paper consists of two parts. In the first part, we present the framework of our metering platform, called "Alethiometer". In the second, we present some statistical results from a set of Twitter data, which further attest that different parameters describing a user (no. of followers, no. of tweets, account age) exhibit a different behavior and are

highly uncorrelated, so it is imperative to examine all these parameters in order to get the complete picture.

2. ALETHIOMETER

"Alethiometer" (from the Greek word αλήθεια, which means truth), analyses the validity of each tweet or author, based on a three 'C's framework: ***Content***, ***Contributor*** and ***Context*** analysis. For the analysis of each framework category, we define a set of related parameters, called *modalities*. In the following, we list these modalities, their meaning, and a short description of the assessment method.

2.1 Contributor modalities

- Reputation *(What do people think of this source of information?)*: Analyse comments in the course of time, to discover sentiments and opinions towards this source. Measured by the number of upvotes or likes.
- History *(What is the past activity of this source?)*: Information about how active a given source is on different social media platforms, combined with validity data. Measured by the update frequency of valid posts.
- Popularity *(Who cares about this source?)*: Information about how many and what kind of people are following the activity of this source, and how many are reading or are recommending this activity to others. Measured by the number of friends/followers, and the number of responses.
- Influence *(What happens because of this source?)*: Information about activities triggered by this source, such as re-posts, discussions or comments. Measured by number of retweets/shares, Klout influence score.
- Presence *(Where does this source appear?)*: Information about the type of source (individual, organization, officially verified account, fake identity, etc.) and its presence on multiple social media platforms. Measured by the number of accounts in different social media.

2.2 Content modalities

- Reputation *(What is the reputation of linked web content?)*: Do the linked web addresses lead to reputable sites? Measured in terms of domain reputation, page rank (GoogleRank or Alexa PageRank), or properties of the contributors to the content.
- Provenance *(What is the history of linked web content?)*: Finding the original occurrence of the content and its whole path across sources, places and time, and measuring the reputation of these sources.
- Popularity *(Who is interested in this content?)*: Information about how many people are following this content. Measured by the number of followers, and the number of responses.
- Influence *(What happens because of this content?)*: Analyse if this content is triggering discussions or other actions in the social sphere. Measured by number of retweets/shares.
- Originality *(Has the same content been used in the past?)*: Check whether the content or parts thereof have been used in the past (e.g., reused text or images that have appeared in the past).
- Authenticity *(Has the content been tampered with?)*: Check whether the content has been changed with respect to its original state (e.g., changed text or attached multimedia content).
- Objectivity and Diversity *(Are views presented from all sides?)*: Does an article (in case of linked web content) present views from all involved sides? Measured by the variation of opinions found for people, content, or general entities.

2.3 Context modalities

- Cross-checking *(Are there multiple similar reports?)*: Measured by the number of different and independent reports or mentions about the same thing.
- Coherence *(Is the content internally and externally coherent?)*: Measurement of text coherence (e.g., CohMetrix) and coherence between the content and tags, attached web-links, or attached multimedia.
- Proximity *(Has the report originated where and when it is claimed it originated / was produced?)*: Measurement of coherence between reference location/time and publication location/time.

2.4 Rating of modality parameters

Our approach for rating parameters is based on the Simple Aggregated Score proposed in [4], where normalized individual scores are added to provide an aggregated score.

Modality parameters are rated on a discrete 5-point scale, from 0 to 4. The rating is based on threshold values a_0, a_1, a_2, a_3 with the following mapping: $[0, a_0) \to 0$, $[a_0, a_1) \to 1$, $[a_1, a_2) \to 2$, $[a_2, a_3) \to 3$, $[a_3, \infty) \to 4$. The challenge is to select appropriate values for a_0, a_1, a_2, a_3. Ideally, we would like an interval-type scale, i.e. a scale that shows not only that a certain value is better than another, but also "how much" better it is. A simple method to achieve this is to adjust the scale so that it follows a uniform distribution. In this way, the same distribution mass is contained in each interval. To adjust the scale in this manner, we can take percentiles so that a_0 is the 20th percentile, a_1 the 40th, a_2 the 60th, and a_3 the 80th percentile.

We also need to somehow derive the significance of the parameters. Suppose that we need to compare the significance of an item i to an item j. Each item has a number of parameters. We can derive a score for the significance of each parameter by comparing its value with the value of parameters of all other similar items. Without loss of generality, suppose that we will compare the significance of items i, j, based on the parameters k, l, common to both items (e.g., the items are tweets, and k: no. of followers, l: no. of times item has been verified).

We first find the significance of each parameter, denoted by the function $S(\cdot)$. In order to do this, we collect the values of a sufficiently large number of items N for each parameter k, so that we have a set of values $\{k_1, \ldots, k_i, \ldots, k_j, \ldots, k_N\}$.

The significance of parameter k_i ($i \in \{1, \ldots, N\}$) can be calculated as:

$$S(k_i) = \frac{k_i - k_{min}}{k_{max} - k_{min}} \quad (1)$$

In the same manner we can calculate the values $S(k_j)$, $S(l_i)$, $S(l_j)$.

To derive a total significance value, we have to weight the significance of the values of parameters k, l between themselves. For example, is the no. of followers more important than the number of times a tweet has been verified? To answer this, we can calculate the dispersion of each parameter value around its mean. The idea is that the closest a value is to the mean, the more reliable it is, whereas farthest values are likely to be outliers, which should be weighted less.

For the set of values $\{k_1, \ldots, k_i, \ldots, k_j, \ldots, k_N\}$ we denote the sample average by $\bar{k}$ and the sample standard deviation by s_k. The dispersion of parameter k_i is

$$d(k_i) = \frac{|k_i - \bar{k}|}{s_k} \quad (2)$$

(we divide by s_k in order to have a normalization). In the same manner we calculate the dispersion for the other parameters $d(k_j)$, $d(l_i)$, $d(l_j)$. The weight of parameter k for item i will be $w(k_i) = (1 - d(k_i))/(d(k_i) + d(l_i))$, while that of parameter l for item i will be $w(l_i) = (1 - d(l_i))/(d(k_i) + d(l_i))$.

Finally, by combining the significance of the underlying parameters we can define the significance of the item. The total significance of item i will be $S_i = w(k_i) * S(k_i) + w(l_i) * S(l_i)$. In the same manner we calculate the total significance value for item j. The same method could also be used to evaluate the score of each modality (reputation. popularity, etc.) instead of using thresholds.

We note that a more accurate calculation of the significance of parameters should not take into account only the dispersion around the mean, but also other variables, e.g., the number of times an item has been verified. However, such refinements may lead to more subjectiveness in the ratings, and therefore we postpone their study for a later phase of our work.

3. PRELIMINARY RESULTS

In this section, we present a number of statistical results which we derived at a preliminary stage of our work, in order to derive more insights for the measured quantities. We have used a sample consisting of about 10 million users collected from a crawl we executed on twitter content from July 2013 for a period of three months.

The graphs in Fig.1(a) and 1(b) depict the empirical distribution for (a) the number of followers and the number of tweets, and (b) the user account age (in days) respectively. The curves in (a) are typical of heavy-tailed distributions, while graph (b) shows a multimodal heavy-tailed distribution with 3 different peaks. One at about 200 days (6.7 months), another at about 700 days (23.3 months), and the highest at about 1600 days (53.3 moths – 4.4 years). The heavy-tailed distribution is also evidenced from Table 1, where we show some important statistics. A very large standard deviation shows up, which is larger than the mean for all three parameters; this is characteristic of heavy-tailed distributions. For the first two curves (which are unimodal distributions), the fact that the median values are smaller than the mean shows that the distributions are positively skewed, which is also obvious from the graph (a). A comparison between Figs.1(a) and 1(b) shows that the number of followers/tweets is not highly correlated to the user account age, i.e. a 'new' user can have a large number of tweets/followers and vice-versa.

Table 1: Basic statistics for the number of followers, number of tweets and user account age.

Statistics	# of followers	# of tweets	User account age
Min	0	0	181
Mean	391	4794	963
Median	126	1228	902
Max	16385	73581	2239
Variance	1112878	82066572	238559

We further examine whether the distributions of the number of followers and the number of tweets follow a power law, i.e. a distribution of the form

$$p(x) = Pr(X = x) = Cx^{-\alpha} ,$$

where C is a normalization constant.

In order to fit the empirical data into power-law distributions, we follow the approach in [3]. By plotting the complementary CDFs on a doubly logarithmic plot, we notice that only a part of the tail of the distribution is a straight line, and therefore could follow a power law. Therefore, we aim to find the range of values $[x_{min}, x_{max}]$ (where x is the quantity of interest, either the number of tweets or the number of followers), in which the distribution exhibits a power-law behavior, and the corresponding values of the exponent α. Based on [3], the value of α is estimated as

$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min} - 1/2} \right]^{-1} , \quad (3)$$

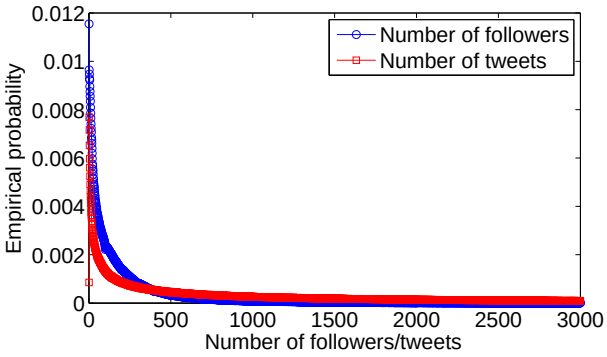
where n is the number of values between (and including) x_{min} and x_{max}, and x_i are the corresponding values of the quantity of interest.

We examined a range of different x_{min} and x_{max} values, and chose the ones that produced power-law distributions which best fitted the data. The best fits were produced, in the case of the number of followers, for x_{min}=40, x_{max}=1000 and in the case of the number of tweets for x_{min}=200, x_{max}=3000, and the empirical and fitted CDFs are shown in Fig. 2. On the graph, the different behavior between the number of followers and the number of tweets appears in a straightforward manner.

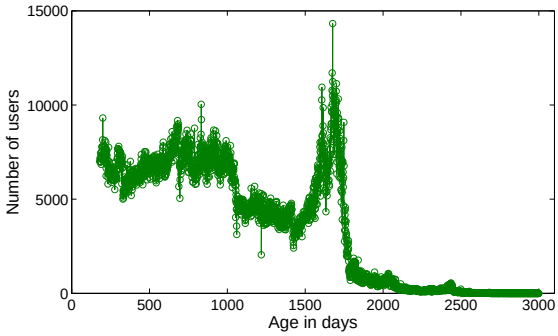
To further evidence this, we have calculated the correlation coefficient between the number of friends a user has (i.e. the number of users they are following), the number of followers and the number of tweets. For a series of n measurements of parameters X and Y written as x_i and y_i where $i = 1, 2, \ldots, n$, then the sample correlation coefficient is written as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and s_x and s_y are the sample standard deviations of X and Y.



(a)



(b)

Figure 1: Distribution of the number of followers, tweets and days a user has been active.

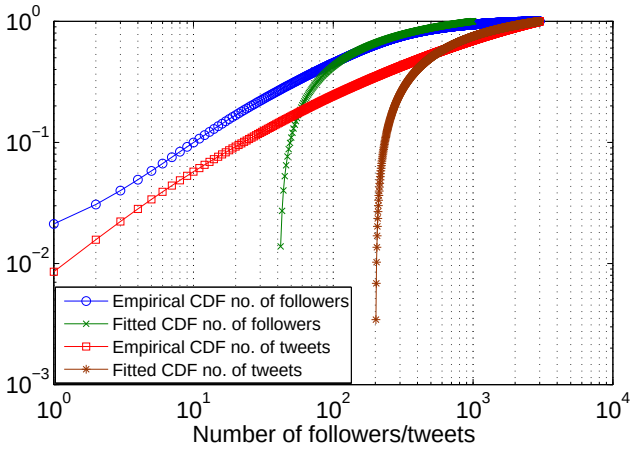


Figure 2: CDFs of the number of followers and tweets.

For the sample of 10^7, the values of the sample correlation coefficients between friends and followers was 0.1222, between friends and tweets 0.0800, and between followers and tweets 0.0197. The highest correlation therefore exists between friends and followers, whereas the lowest between followers and tweets. All the above correlation values are however quite small, which means that these parameters are relatively independent from one another and we have to consider each one individually, rather than calculate only one as representative of the others.

4. CONCLUSION AND FUTURE WORK

In this paper we have introduced Alethiometer and proposed an approach for providing a metric that will give a sense of truth to the news users. We conducted statistical analysis on a large corpus of Twitter data for some modalities, i.e., the number of followers, the number of tweets per user account, and the number of days an account is active, which attested the non-correlation between these modalities, and the need to examine them separately. In the next steps, we will extensively investigate the behaviour of other modalities and study the correlation not only between modalities of a contributor, but also between content, contributor and context modalities. We also defined an approach for deriving the significance of the modalities. Based on this approach, we will undertake a large-scale investigation in order to extract the most significant parameters for the validation of news in social media. Finally, the mapping of feature values on a qualitative scale is also an important research issue. To this end, we aim to examine appropriate methods for attributing qualitative characterizations to modalities according to the range of values they belong to.

5. ACKNOWLEDGEMENTS

This work has been supported by the European Commission under EU projects SocialSensor (FP7-ICT-2011-7-287975, `http://www.socialsensor.eu`) and REVEAL (FP7-ICT-2013-10-610928, `http://revealproject.eu`).

6. REFERENCES

[1] K. R. Canini, B. Suh, and P. Pirolli. Finding credible information sources in social networks based on content and social structure. In *SocialCom/PASSAT*, pages 1–8. IEEE, 2011.

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

[3] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, Nov. 2009.

[4] S. T. Moturu and H. Liu. Quantifying the trustworthiness of social media content. *Distrib. Parallel Databases*, 29:239–260, June 2011.

[5] L. Song, W. Zhang, S. S. Y. Liao, and R. C.-W. Kwok. A critical analysis of the state-of-the-art on automated detection of deceptive behavior in social media. In S. L. Pan and T. H. Cao, editors, *PACIS*, page 168, 2012.