2015

Spread and Skepticism: Metrics of Propagation on Twitter (Extended Abstract)

Samantha Finn
sfinn@wellesley.edu

P. Takis Metaxas
pmetaxas@wellesley.edu

Eni Mustafaraj
emustafa@wellesley.edu

Spread and Skepticism: Metrics of Propagation on Twitter (Extended Abstract)

Samantha Finn
Computer Science
Wellesley College
sfinn@wellesley.edu

Panagiotis Takis Metaxas*
Computer Science
Wellesley College
pmetaxas@wellesley.edu

Eni Mustafaraj
Computer Science
Wellesley College
emustafa@wellesley.edu

Abstract

Social media has become part of modern news reporting, used by journalists to spread information and find sources, or as a news source by individuals. The quest for prominence and recognition on social media sites like Twitter can sometimes eclipse accuracy and lead to the spread of false information. Could we use the so-called "wisdom of crowds" to detect the likelihood that some claim spreading may be true or false? This paper, part of an ongoing research, offers evidence that most of the time, false claims do not spread like true ones, and that the reaction of the audience following a story on Twitter is correlated with the validity of the story.

Using the `twittertrails.com` system, we have examined the spreading patterns of a number of claims that have been discussed in the news by investigative journalists. Here, we first introduce two new metrics for measuring the spreading and skepticism around the propagation of a claim. Then, employing a classification algorithm that takes into account the behavior of the crowd posting or retweeting about a story, our system leads us to observe that true and false rumors have different footprints in terms of how they propagate and invoke skepticism by their audience. In particular, we observe that, more often than not, false rumors are more likely to be negated if exposed to a large audience.

1 Introduction

The so-called "24 hour news cycle" has led to an increased sensationalism of news stories. Especially with the increase in cable news channels and online news media, the need to catch the attention of the public has led to faster and more hyped up reporting. As Twitter's role in breaking news or spreading it increases, so does the need for evaluating the information credibility of spreading rumors. This problem has been studied in different ways by the research community. Recently, we have developed TWITTERTRAILS, a system that allows its user monitor the spreading of rumors on Twitter (Finn et al. 2014). TWITTERTRAILS can help casual users who are not interested in a lengthy investigation, by providing an automatically calculated label as "likely false" or "likely true" for the credibility question, based on the behavior of the crowd that follows the story under investigation. This binary classification makes use of two novel features: the *spreading score* and the *skepticism score*. The spreading score reflects how far and wide the rumor is spreading. It is based on a re-purposing of the *h-index*(the well-known citation metric for publications (Small 1973)) for the set of tweets relevant to the credibility assessment task. The skepticism score is the ratio of h-indexes for tweets countering a rumor over those supporting it.

This paper, which is a work in progress, makes the following contributions:

We introduce two new metrics, the spreading score (how much the story spread) and the skepticism score (did other users show doubts about the story). These metrics leverage the "wisdom of the crowds" applied to sets of tweets related to the story being investigated.

We propose, train, and evaluate an observable algorithm for classifying as true or false claims (stories) seen on Twitter. This algorithm uses as features the two new metrics: the spreading score and the skepticism score, described below.

1.1 Data Collection

Our dataset consists of distinct "stories": collections of tweets related to a single subject. In this paper we focus on stories related to "claims," typically related to an event in real life which has happened or might happen, and which can be proven either true or false. To collect data to form a story, we use the Twitter Search API, which allows us to collect tweets from the last 6-9 days in reverse chronological order, based on search terms given to the API (it returns results similar to the search on Twitter's website). For example, to collect a dataset of tweets relevant to the claim that "welfare was giving out free cars", we used the search terms: *congress free cars*, *obama free cars*, *welfare cars*. In cases where the search results were not all relevant to the claim being investigated, we would also modify the dataset in order to remove irrelevant tweets.

In this paper, we study 134 claims which we have manually verified as being true or false. The datasets range in size from a few dozen tweets to over 60,000, and have been collected during 03/2014 - 03/2015.

*Corresponding author.

2 New Metrics: Spread and Skepticism

2.1 Spread

In order to measure the impact and visibility of a story on Twitter, we define a metric called *spread*. To measure spread, we consider a tweet as a "publication" and its verbatim retweets as citations and evidence of its visibility in the network. This is inspired by the relevant theory of Library Science of which academics are usually well aware (Small 1973; Hirsch 2005): The *h-index* of a collection of N publications is defined to be h when there are h publications in the collection that each have at least h citations and the remaining $(N - h)$ publications have less than h citations each.

In a similar fashion, we define the *spread* of a collection of tweets retrieved as relevant to a story as the h-index of the collection. We say that a story has a spread of h if there are h tweets that have received at least h retweets in the collection. We then define the *level of spreading* of a story on a discrete logarithmic scale of spreading scores: A story with score of at most 16, 32, 64, 128 is said to have spreading level *insignificant, low, moderate, high* (resp.). There are relatively very few stories that achieve spreading score above 128, and they are assigned spreading level *extensive*.

To help understand this concept, and what it means (and does not mean) in use, consider the following examples:

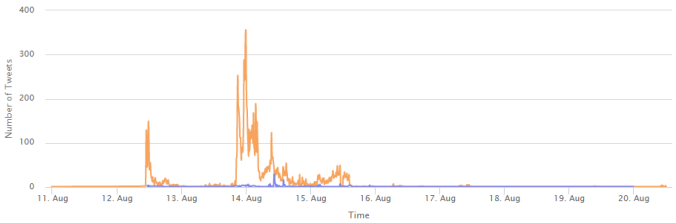

Figure 1: (This image is better observed in color.) From the claim that the wife of the police chief in Ferguson wrote a racist post on her Facebook account, the activity in 10 minute periods of tweets that support the claim (orange) vs those express doubt (purple).

1. **The claim that people on welfare will receive free cars**[1]: this story has a spread of 2, (on our scale, "insignificant level"). This means that, in our data, only 2 tweets received at least 2 retweets. One of those tweets received 5 retweets, while the other received 2. The spread captures not only the number of retweets received, but how many tweets received that number, so the fact that one tweet received 5 retweets does not increase the value. In addition, there are 15 more tweets which received only 1 retweet. In total, there are over 100 relevant tweets, the majority of which received no retweets at all. The spreading value of 2 does not represent how many tweets received less than 2 retweets.
2. **The claim that the wife of the police chief in Ferguson wrote a racist post on her Facebook account**[2]: this story has a spread of 33 (on our scale, "moderate level"). This means that, in our data, 33 tweets received at least 33 retweets. There are tweets receiving more than 33 retweets: for instance, one which has 256 retweets. The spread does not count the number of tweets receiving less than 33 retweets; there are almost 1000 tweets which received at least 1 retweet, and of those over 100 of them received at least 10 retweets. It also does not count the total number of tweets collected; in this case, over 10,000, the majority of which have no retweets, or are retweets themselves (the spreading value only counts the original post).
3. **The claim that Robin Williams died on August 11th, 2014**[3]: this is one of the most widely spread stories which we have collected, with a spread of 444 (on our scale, "extensive level"). This means at least 444 tweets were retweeted 444 or more times. There are tweets in our data which have received thousands of retweets, some even tens of thousands. And there are over 10,000 tweets with at least 1 retweet; of those, almost 4,000 have at least 10 retweets; and of those, almost 1,700 have over 100 retweets. In total, we collected over 60,000 relevant tweets for this story, about two thirds of which are themselves retweets.

From these examples one gets a better idea of what the spread represents (and does not represent): it measures both the propagation of highest reaching tweets and the number of high reaching tweets, each of these two things constrained by the other. It does not reflect the one or two tweets with the most retweets, even those that have exponentially more retweets than the value of the spread. Nor does it measure how many tweets were collected. Although these numbers are interesting and meaningful, the spread is meant to give an overall picture of the impact of a story: how visible it was, as well as how many people were engaged in it.

2.2 The Skepticism Score

Skepticism measures the prominence of doubt and mistrust in a story. When calculating the spread, we measure the h-index over all the relevant data we collected for a story. Skepticism uses two subsets of the relevant data: those tweets which express doubt versus those that do not (we consider not actively questioning or debunking a claim, but still spreading the claim as an implicit show of support), in order to measure the presence and significance of the skepticism being expressed.

The first step to calculating the skepticism is to identify tweets in which the author expresses mistrust in the validity of a claim, whether they are wondering if the claim is false or expressing that it is an outright lie. This is a difficult task, as doubt, disbelief, and mistrust can be conveyed in many ways, including the use of sarcasm, which are hard to detect, as well as taking into account tweets in different languages. For now, we employ a simple and customizable negation algorithm, which we have found works fairly well for identifying tweets which express skepticism: we look for

[1] http://bit.ly/TT-Welfare-Free-Cars

[2] http://bit.ly/TT-Ferguson-Police-Chiefs-Wife

[3] http://bit.ly/TT-Robin-Willams-died

tweets using the following 10 commonly used keywords to express doubt or disbelief:

hoax, fake, doubt, false, scam, untrue, mistake, unreal, bogus, mislead

Consider the following tweet retrieved in a collection of tweets investigating the rumor that Betty White had died:

Guys, Betty White is not dead. It's fake news.

After separating our initial collection of relevant tweets into two subsets of doubting and supporting tweets, we define the *skepticism* of a story as the ratio of the h-index of doubting tweets over the h-index of supporting tweets. When the ratio is greater than 1.0, it indicates that doubt of the claim's validity has spread further than support for the claim. For examples of skepticism scores in practice, refer to the three example stories from Section 2.1:

1. **The claim that people on welfare will receive free cars**: this story has a skepticism of 0.5 (on our scale, dubious). The tweets which express doubt have an h-index of 1 in this story, and tweets that do not have an h-index of 2. So, the story has a skepticism of 0.5, which we have found to be a fairly high value. The tweets which express doubt spread about half as far as those that did not, though in the end neither spread very much (remember, this is a very low profile story, with an overall spread of only 2).
2. **The claim that the wife of the police chief in Ferguson wrote a racist post on her Facebook account**: this story has a skepticism of 0.27 (on our scale, dubious). This story has a much higher spread than the previous story (33 compared to 2). The tweets which express doubt have an h-index of 9, those that do not have an h-index of 33, which results in a ratio of 0.27. This is lower than the last story, and doubt that the claim is true still still outstripped by those who seem to believe it. The skepticism is still fairly high, though. In this case, the claim is false, but it still has some truth to it: the Facebook post that is being shared is real, but the claim that woman who wrote it was not the police chiefs wife is false.
3. **The claim that Robin Williams died on August 11th, 2014**: this story has a skepticism of 0.03 (on our scale, undisputed). This is one of the highest profile stories in our database, with a spread of 444. The h-index of tweets expressing doubt in this claim is 12, while those that do not have an h-index of 443. Although this has the highest h-index of tweets expressing doubt, it has the lowest ratio: 0.03. When a story has a greater spread, the tweets that express doubt on the claim have a greater audience and a greater chance to be retweeted. This claim was true, and many of the tweets expressing doubt express more of a hope that the story is false than strong conviction that it is (after all, there are many celebrity death hoaxes, including one about Robin Williams shortly before he died).

Two of our three examples are false; however in none of them do tweets expressing doubt spread further than those that do not. There are actually very few stories which we have collected where the negation of a claim spreads farther than support of the claim. It is usually after a false claim picks up steam that people begin to doubt it or realize that it is false. This doubt expressed en masse will slow down and eventually stop support of the claim, but very infrequently will it spread as far as the initial support (especially on a highly emotionally charged claim).

The graph in Figure 1 compares the activity of tweets expressing doubt to those that do not in the story investigating the claim that the wife of the police chief in Ferguson wrote a racist post on her Facebook account. The supporting tweets are in orange, and those express doubt are in purple. You can see that there is a much higher spike in activity for tweets that support the claim, but that activity starts to taper off after the doubting tweets begin to peak.

Because of this, the spread of a story tends to affect the skepticism somewhat, and vice versa. Some false claims never gain enough traction on Twitter for people to notice and debunk them. And in general, false claims do not spread as far as true claims. From our experience, skepticism values greater than 0.25 tend to be very strong indicators of a claim being false.

3 True vs. False Claims

To gain more insight into how claims behave on Twitter, based on their skepticism and spread, we can plot them on a graph. In our dataset of claims, we have manually determined whether they are true or false, utilizing traditional news media and websites such as snopes.com. We then plot the true and false claims based on their skepticism vs. spread, as seen in figure 2.

We observe in the graph that false claims tend to have low spread and skepticism that varying from very low to very high, while true claims have spread varying from very low to very high and low skepticism. There is also only a small area of overlap between true and false claims, those having low spread and low skepticism; in other words, there are no true claims with very high skepticism and low spread, and no false claims with very high spread and low skepticism. One final observation is that there are no claims with high spread and high skepticism.

It seems reasonable to conclude that, on Twitter, the more people see some false information, the more likely it is that they will either raise an objection or simply decide not to repeat it further.

On Twitter, when people are doubtful of the information in a tweet, they appear to be less likely to retweet it (though they may still discuss it in their own tweets), resulting in that information propagating less than information they believe. There are exceptions, where tweets that present false information receive many retweets, and even stories in our dataset that were false being widely discussed, but the general trend is that on Twitter, false claims have much less visibility (this is not true for all social media websites, Facebook being a good example (Friggeri et al. 2014)).

4 Labeling Claims

Based on the observation from the graph in Figure 2, we can see that spread and skepticism are meaningful metrics in which to attempt to differentiate true claims from false ones.

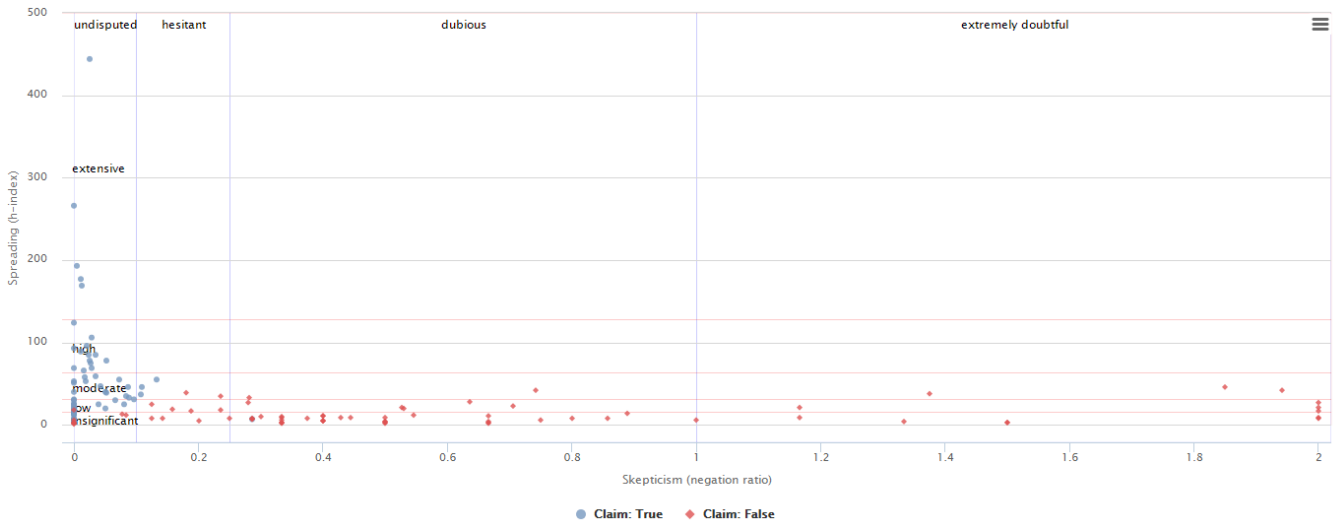

Figure 2: (This image is better observed in color.) This graph plots over 100 collected stories in our database (at the time of this writing). Spreading score is represented in the y-axis while skepticism score in the x-axis. Each story is represented by either a red diamond or a blue circle. Following investigation determining the ground truth of the rumor, rumors proven to be true are represented by a blue circle, while those shown to be false by a red diamond. Note that false rumors on Twitter exhibit low spreading and high skepticism scores.

We use this information to create two algorithms to automatically classify claims as true or false: the first, a simple algorithm which classifies claims based on whether their spread and skepticism values fall into a range of values, and a machine learning classifier which uses skepticism and spread as features.

4.1 Range Classifier

This classification of a story as true or false is based on the have also created an algorithm that will label a story as either likely true or likely false based on these two crowd metrics. Stories with low, insignificant or moderate spread and skepticism that is dubious or extremely doubtful (spread <32 and skepticism $>= .25$) are deemed likely false, while stories with moderate, high, or extensive spread and undisputed skepticism (spread $>= 16$ and skepticism <0.1) are considered likely true. Any claim with spread and skepticism outside these two ranges is undetermined.

These ranges are conservative, favoring accuracy over being able to label all stories. Though it correctly labels 94 of the 95 claims that fall into the two ranges (98.9%), of our 134 manually verified claims, it leaves 28.4% of them undetermined. Considering these undetermined claims as incorrectly labeled, this method has only a 70.1% accuracy.

4.2 Machine Learning Classifier

In order to automatically classify claims (or the stories) as either true or false, we trained two classifiers on the stories which we had manually labeled as either true or false claims. This subset of the collection contains 47% true claims, and 53% false (more or a less a balanced set that will not bias the classifiers). The supervised classification relies only on two input features: the spreading score and skepticism score (shown in Figure 2), and the binary label true or false. The two classifiers, Naive Bayes and Classification Tree, were trained with leave-one-out validation, displaying accuracy scores of, respectively, 0.924 and 0.940.

4.3 Discussion

One of the stated objectives of our research is to make the algorithmic choices of the tool observable to users. The algorithm for classifying a claim as true or false uses only two features allowing for a 2-D visualization where the separation of true and false stories can be easily inspected. While the accuracy of the classification is very high (around 92%) the two features on which the classification is based might suffer from variability introduced at previous levels in the calculation. For example, the spreading score can be influenced by the nature of the tweets collection returned by the Search API (Twitter APIs are constantly changing the rules of what they retrieve).

The skepticism score depends on the goodness of the classifier that labels the tweets as negation or non-negation tweets. Negation is easier to detect than polarity of sentiment, since in a 140-character tweet users have to mention both the original claim and their belief that it is not true. This limits the vocabulary used to express negation. However, given that no algorithm understands natural language and its subtleties, the classifier might err on both sides (miss negation or find negation where it doesn't exist). The only safe-guarding mechanism that we have for this problem is the inclusion or exclusion of certain negation words by the user through a dedicated interface. This is another way in which the user asserts its control over how the system works.

The current L-shape of the labeled stories in Figure 2, leads us to the following **observation**:

On Twitter, claims that receive higher skepticism and lower spreading scores are more likely to be false. On the other hand, claims that receive lower skepticism and higher spreading scores are more likely to be true.

We make this observation specific for Twitter because it may not hold for every social network. The Twitter's interface has the particular feature that both the claim and its negation will get the same amount of real estate on a user's stream (since they will be individual tweets). On the other hand, this apparently is not true for Facebook, where a claim gets much greater exposure than a comment, because a comment may be hidden quickly due to follow up comments. So, on Facebook most people may miss an objection to a claim. This may explain the result by (Friggeri et al. 2014) who find that false claims on Facebook live for a long time, even those claims that are verifiable by a quick search on Snopes.[4] By contrast, very few Snopes-included claims appear on Twitter.

5 Related Work

Monitoring and evaluating the propagation of a rumor has recently gotten a lot of attention. A new web service, `emergent.info` developed by journalist Craig Silverman is using journalists to evaluate online claims[5] and deem them as True/False/Unverified. They track the number of shares a rumor has on Facebook, Twitter and Google+ and report the numbers along with links to articles that supported and countered the rumor. Another application is RumorLens (Resnick et al. 2014). It analyzes the spread of rumors on Twitter and prompts user feedback to classify results as propagating, debunking or unrelated to the original rumor. It then uses a text classifier to garner more widespread results.

Rumor Cascades on Facebook have also been studied by a Facebook team (Friggeri et al. 2014). They are focusing on tracking the way that rumors propagate on Facebook, mainly those that have been verified by snopes.com. Unlike what we observe on Twitter, they find that rumors do not easily die on Facebook, but they may re-emerge long after they started.

The propagation of false rumors related to a Chilean earthquake was analyzed by (Castillo, Mendoza, and Poblete 2013) where they found that there are measurable differences in the way messages propagate, that can be used to classify them automatically as credible or not credible.

One of the earliest systems that focused on studying patterns of information propagation in online social networks like Twitter is Truthy (Ratkiewicz et al. 2011). Truthy is based on the concept of memes that spread in the network. Such memes are detected and followed over time to capture their diffusion patterns. Truthy is a more general-purpose system than the ones we mentioned previously in this section, which, despite its name, it doesn't provide explicit assessment of the veracity of the tracked memes. However, through visualizations of propagation patterns and other metrics (e.g., sentiment analysis), Truthy can enable a user to come to a certain conclusion on her own.

[4] At the time of this writing, Facebook is reportedly about to introduce a feature that will allow users mark stories as fake: http://cnnmon.ie/1CCznVQ

[5] "Why Rumors Outrace the Truth Online" by Brendan Nyhan, Sept. 29, 2014 http://nyti.ms/1pFXaAq

6 Conclusion and Future Work

TWITTERTRAILS was designed and implemented with the goal to provide a vital service to users who want to engage with Twitter as a source of reliable information, either for their own consumption, or as a source for journalism, both professional and amateur.

Through a simple classification algorithm that takes into account the spreading and skepticism levels by the crowd posting or retweeting about a story being investigated, our system leads us to observe that true and false rumors have different footprints in terms of how they propagate and invoke skepticism by their audience. False rumors are more likely to be negated if exposed to a larger audience.

There are, however, several limitations that we are currently dealing with which we acknowledge here. First, our system depends heavily on the availability of data from the Search API which places an upper limit on the number of tweets and on the date range we can retrieve. Second, the translation of the information need associated with a claim to the search query terms can be error prone. Despite our careful efforts, both of these can affect the recall rate of our system. Third, the set of claims we have investigated so far is not a random sample of all claims on Twitter. They are influenced by what journalists have been interested in examining and writing about. Often, they are claims that are sensational or can surprise us by their final outcome. This limitation, however, strengthens our findings since it focuses more on topics that we are more likely to surprise the journalists. Finally, our negation detection algorithm is rather naive and restricted to English, but this is a known problem that many researchers struggle with. Nevertheless, we are working in addressing each of the limitations.

References

Castillo, C.; Mendoza, M.; and Poblete, B. 2013. *Internet Research* 23(5):560–588.

Finn, S.; Metaxas, P.; Mustafaraj, E.; OKeefe, M.; Tang, L.; Tang, S.; and Zeng, L. 2014. Trails: A system for monitoring the propagation of rumors on twitter. Comp Journ Symp.

Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor Cascades. In *ICWSM*.

Hirsch, J. 2005. An index to quantify an individual's scientific research output. *PNAS* 102(46).

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *In WWW*, 249–252. ACM.

Resnick, P.; Carton, S.; Park, S.; Shen, Y.; and Zeffer, N. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. AJR.

Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS* 24:265–269.