Project Proposal - Early Rumor Detection from Tweets
CS 585, UMass Amherst, Fall 2015

Anushree Ghosh
Rose Tharail John
Yamin Thuzar Tun

Motivation

Today Social Media is ubiquitous. Along with sharing of personal comments, it is an active avenue for spreading news about events in real time. This leads to the issue of the correctness of the information spread. Often there have been instances of gossip/rumor going viral without the actual event in question occurring. As the NLP research team from University of Michigan said, "a Rumor is a controversial statement that can be fact-checked".[1] With the world population being highly connected, unchecked outbreaks of rumors could lead to catastrophic effects on economy and society at large. For example, in April 2013, stock market collapsed dramatically and lost billions of dollars, after a false rumor spread that the White House was bombed causing President Obama serious injury. From this event, we can see how early detection of rumor could make resolving rumors possible and avoid catastrophic consequences of false rumor.

Problem Statement

We would like to explore some novel techniques employing Natural Language Processing to detect rumors before they lead to any adverse effects.

Related Work

Considerable research has been going on, in this topic, particularly in the realm of how they spread after natural calamities and disasters. By focusing on the March 2011 Japan earthquake and Tsunami, Takahashi and Igata [2] they analyze the timeline and volume of the spread of rumors and attempt to extract the characteristics of the tweets which are rumor candidates. Recent studies have tried to characterize tweets as rumors by analyzing the popularity of the post - the collective response behavior to the tweet[1]. It was noticed by researchers the truth behind the rumors were questioned severely, and therefore question phrases and question marks appeared several times in the tweets. [3]Castillo et al. therefore used the number (and ratio) of question marks as a feature to classify the credibility of a group of tweets.

Dataset

We could not find any publicly available labeled dataset of tweets about rumors. We will follow one of the following approaches to get the dataset:

1) Email the authors of the papers [1][2] we are referring to, for labelled rumor data sets.
2) HashKat is a dynamical network simulation tool designed to model the growth and information propagation within an online social network. It is an agent-based engine capable of simulating online networks. One possibility to obtain a dataset is to simulate the progression of rumors on twitter in HashKat.
3) Stream tweets from Twitter APIs with a tag of a known rumor in a particular date range and model based on those. The challenge is to label multiple thousands of tweets as rumors or non-rumors. This approach is the last option that we want to use since it requires tremendous amount of human labor for annotation.

Preliminary Analysis/Modes of Investigation

We will create two baseline models that consider the most intuitive characteristics of rumors, and a more sophisticated variant model that consider more generic characteristics of rumors. We will compare the performances of baselines against the variant model for evaluation.

Baseline Models

The major focus of the baseline model will be:

Model 1:

- **Question marks/words** as enquiry are generally indicative of rumors. In Table 1 of [1], a list of rumor mongering tweets clearly show a higher usage of skepticism and questions in the post. Thus this feature will have the count of question words and question marks used in the tweet.

Model 2:

- **Hashtag** This feature is the number of tweets that contain hashtags with keywords about the content of rumors.
- **Popularity of tweets** in terms of the number of retweets and replies - Rumors are highly re-tweeted and replied. Looking at the replies might give us a hint about its genuity. If the replies, question its authenticity, it might be a rumor. We could model the features of the replies themselves in terms of the degree of questions used in retweets and replies.

A simple logistic regression model can assign weights to the above features in each model, and classify if the tweet is a rumor or otherwise based on a certain threshold that we will be estimating in the preliminary analysis.

Variant Model

The following is the variant model that we would like to explore and compare its performance against the baseline model.

- **Enquiries and Corrections** - This model considers both enquiries and correction tweets as signal tweets for the purpose of recognizing rumor patterns. More detail about this model follows.

Approach

The baseline features mentioned above are not sufficient and robust enough to detect rumor tweets. For example, only a third of the tweets in the form of questions are actually enquiries about rumors. [1] We will describe more details about the pipeline of the variant model that considers enquiries and corrections as features.

The backbone idea of our variant model with enquires and correction features exploits the fact that rumor's nature of controversy evokes curiosity and desire for enquiry and debate. Rather than focusing on the content of rumor declaration tweet itself, we intend to look at the verification and correction tweets, which also belong in rumor tweet groups. We intend to apply UMichigan research team's techniques for implementing this model. [1] It involves designing of a pipeline to identify rumor correction tweets and not just the rumor tweets. Clustering and summarizing techniques can be applied to build a model that takes into consideration all the phases a rumor goes through right from origin to propagation and correction.

Rumor-Detection Pipeline

1. **Identify signal tweets** that are either verification tweets (eg. "Really?", "Is this true?") or correction tweets (eg. "That is a false rumor because …") using Chi-square scores as similarity scores.
2. **Cluster signal tweets** into different groups based on their common contents
3. **Summarize** the common content of each cluster in a single statement
4. **Select top-5 rumor clusters** using scoring system to indicate the likelihood of each summary statement in a cluster being a rumor. For the scoring system, we will train with two different classifiers- SVM and Decision Tree- using statistical features in Section 4.4 of [1] to obtain a better ranking function. [1]

Evaluation

We are planning to use a k-fold cross validation method to train our data on. For train-test splitting, we will use a 80-20% split, if we have fewer tweets and 90-10% split in case of a large number of tweets on the topic.

Final scope of project

We have mentioned several modes of investigations above to detect rumors. Based on the kind of dataset we are able to secure, we might take either of the approaches. For baseline calculations, we will include at least one of the given possible features.

Finally the scope of the project includes

1) Securing/ simulating labelled data
2) Feature extraction
3) Model training for baseline estimates
4) Techniques based on UMich paper [1]
5) Prediction & accuracy evaluation based on the clustering output

Softwares

For implementation, we will be using NLP toolkit, scikit-learn, Python or any other available machine learning libraries.

References

[1] Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts Zhe Zhao, Paul Resnick, Qiaozhu Mei. WWW '15 Proceedings of the 24th International Conference on World Wide Web. University of Michigan, MI

[2] Tetsuro Takahashi, Nobuyuki Igata. Rumor detection on twitter. SCIS-ISIS 2012, Kobe, Japan, November 20-24, 2012, IEEE 452

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web, pages 675–684. ACM, 2011.

[4] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, Qiaozhu Mei. Rumor has it: identifying misinformation in microblogs. EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing, University of Michigan, MI