Rumor Identification and Belief Investigation on Twitter

Sardar Hamidian and Mona T Diab
Department of Computer Science
The George Washington University
sardar,mtdiab@gwu.edu

Abstract

Social media users spend several hours a day to read, post and search for news on microblogging platforms. Social media is becoming a key means for discovering news. However, verifying the trustworthiness of this information is becoming even more challenging. In this study, we attempt to address the problem of rumor detection and belief investigation on Twitter. Our definition of rumor is an unverifiable statement, which spreads misinformation or disinformation. We adopt a supervised rumors classification task using the standard dataset. By employing the Tweet Latent Vector (TLV) feature, which creates a 100-d vector representative of each tweet, we increased the rumor retrieval task precision up to 0.972. We also introduce the belief score and study the belief change among the rumor posters between 2010 and 2016.

1 Introduction

Traditionally television, radio channels, and newspapers were the only news sources available. They are still the top trusted news sources but there is a large new trend toward digital sources. A considerable ratio of newspaper readers now read them digitally and the number of people relying on social media as a news source doubled since 2010. Social media helps you post your news online by a single click, this feasibility leads novel breaking news to show up first on micro blogs. Twitter is one of the most popular microblogging platforms with more than 250 million users. Accessibility, speed and ease-of-use have made Twitter a valuable platform to read and share information. However, the same features which make Twitter or any microblogging platform a great resource, but combined with lack of supervision make them fertile grounds for malicious or accidental misinformation in social media. Accordingly, this can lead to harmful incidences especially in sensitive circumstances, which then could cause damaging effects on individuals and society. There are many information seekers who do not rely on a single source to get information, but this is not always a good solution since even other news outlets sometime rely on social media when it comes to novel breaking news. Smart phones enable everyone to capture and tweet every single moment hours before TV cameras arrive. Considering that, social media is an appealing option for those who crave novel tempting news but on the other hand, could deceive anyone by well-structured and formatted rumors. In this study we work on a standard dataset of rumors collected by Qazvinian et al. (Qazvinian et al., 2011). In their work, the definition of rumor is defined as a statement whose truth value is unverifiable or deliberately false. We are using the same definition and not investigating the stimulus behind rumors creation.

We investigate the problem of detecting rumors in Twitter data. We start with the motivation behind this research, and then the history of similar studies about rumors is overviewed. Then the overall pipeline is exposed, in which we adopt a supervised machine learning framework, and then we investigate the belief change for president Obama rumors in three years, and finally, we compare our results to the current state of the art performance on the task.

Proceedings of NAACL-HLT 2016, pages 3–8,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

We prove that our approach yields superior results in comparison to other works to date.

2 Related Work

There is an extension body of related works on trustworthiness and misinformation detection. In this section we only focus on closely related works on the Natural Language Processing field that concentrate on information propagation and trustworthiness on social media, and specially on Twitter.

2.1 Social media and Trustworthiness

After the earthquake and tsunami occurred in Japan on March 11th 2012, Takahashi and Igata, (Takahashi and Igata, 2012) targeted two sets of related rumor tweets about the earthquake. They create the model to detect other candidate rumor tweets relying on a sequence of processes. Takahashi and Igata detect the target rumor list using the entities and then the re-tweet ratio for target rumors is calculated, and finally the clue keywords get extracted by analyzing the scoring of each content word *w*, using the ratio of word occurrence in correction tweets *(num in correction(w))* over rumor tweets *(num in rumor(w))*. In a similar study, Soroush, (Vosoughi, 2015) proposes his two step rumor detection and verification model on the Boston Marathon bombing tweets. The Hierarchical-clustering model is applied for rumor detection, and after the feature engineering process, which contains linguistic, user identity, and pragmatic features, he adopts the Hidden Markov model to find the veracity of each rumor. Soroush also analyses the sentiment classification of tweets using the contextual Information, which shows how tweets in different spatial, temporal, and authorial contexts have, on average, different sentiments.

Sina is the popular Chinese microbloging platform like Twitter. Yang et al. (Yang et al., 2012) studied the rumors classification problem on both Twitter and Sina. He extended his primary features including content, client, account, location, and propagation by adding client-based features, which refers to a program that is being used to post on a microblog and also the location-based feature, which is a binary feature, that indicates being inside or outside of China. Yang et al. cover a significant range of meta-data features and fewer sentiment and contextual features in the aforementioned work. The most relevant related works to ours are Qazvinian et al. (Qazvinian et al., 2011)(V11) which use three sets of features, including content-based, network-based, and Twitter specific meme features. For content-based features, they extract lexical and part-of-speech patterns. For network-based features, they build two features to capture four types of network-based properties utilizing the log likelihood of retweet and reply properties in Tweets, and finally, the Twitter specific meme features include hashtags and URLs. In our previous work (Hamidiain and Diab, 2015)(S15) we used the V11 data set with a new set of features, more labels, different machine learning, and an experimental approach. We proposed Rumor Detection and Classification (RDC) within the context of microblogging social media and suggested Single-step and Two-step models (SRDC and TRDC) in a supervised manner and investigate the effectiveness of the proposed list of features and various preprocessing tasks.

Table 1: List of Annotated Rumors (Qazvinian et al, 2011)

Rumor	Rumor Reference	# of tweets
Obama	Is Barack Obama muslim?	4975
Michele	Michelle Obama hired many staff members?	299
Cellphone	Cell phone numbers going public?	215
Palin	Sarah Palin getting divorced?	4423
AirFrance	Air France mid-air crash photos?	505

3 Problem Definition and Approach

S15 and V11 results indicate that content features outperform other features in the Rumor Retrieval (RR) task. In this study we perform the rumor retrieval task with a new set of features. We employ content unigram feature, which lead to the highest results in among the content features. We employ the Tweet Latent Vector (TLV) to overcome the missing word and short length tweet issue. We extend the V11 data set to investigate the belief change for the specific rumor in different years.

3.1 Data

V11 published an annotated Twitter data set for the five different established rumors as listed in Table 1. The general annotation guidelines are presented in

Table 2: Rumor Detection Annotation Guidelines

0	If the tweet is not about the rumor
11	If the tweet endorses the rumor
12	If the tweet denies the rumor
13	If the tweet questions the rumor
14	If the tweet is neutral
2	If the annotator is undetermined

Table 2. The original data set as obtained from V11 did not contain the actual tweets for both the Obama and Cellphone rumors, but they only contained the tweet IDs. Hence, we used the Twitter search API for downloading the specific tweets using the tweet ID. Accordingly, the size of our data set is different from that of V11 amounting to 9000 tweets in total for our experimentation as it is shown in Table 3. The following examples are a sample of each of the annotation labels 0 (If the tweet is not about the rumor,) 11(If the tweet endorses the rumor,) and 12 (if the tweet denies the rumor) from the Obama rumor collection.

- **0**: 2010-09-24 15:12:32 , nina1236 , Obama:Muslims2̌019 Right To Build A Manhattan Mosque: While celebrating Ramadan with Muslims at the White House, Presi... http://bit.ly/c0J2aI
- **11**: 2010-09-28 18:36:47 , Phanti , RT @IPlantSeeds: Obama Admits He Is A Muslim http://post.ly/10Sf7 - I thought he did that before he was elected.
- **12**: 2010-10-01 05:00:28 , secksaddict , barack obama was raised a christian he attended a church with jeremiah wright yet people still beleive hes a muslim

3.2 Silver Data

V11 uses Twitter search API with regular expression queries, and collects data from the period of 2009 to 2010. We also run the same queries with the same keywords for the Obama rumor and collected more than 7000 tweets from 2014 and 2016. Collected tweets are labeled by applying the Rumor Retrieval (RR) pipeline. We named the new data as silver-data and use them to investigate how belief has changed toward the "Is Barak Obama Muslim?" rumor from 2010 to 2016. Table 3 shows statistics for the extracted tweets and silver-data. We labeled the silver-data as 0(Non-Rumor), 11(Believe), and merged 12(Deny-12, Doubtful-13, and Neutral-14). For tagging the silver data we used the original Obama data set as the train data set. Table 5 shows what labels are being used for the rumors retrieval and silver-data creation experiment.

Table 3: List Of Annotated Tweets Per Label Per Rumor

Rumor	**0**	**11**	**12**	**13**	**14**	**2**	**Total**
Obama	945	689	410	160	224	1232	3666
Michelle	83	191	24	1	0	0	299
Palin	86	1709	1895	639	94	0	4423
Cellphone	92	65	3	3	3	0	166
Air France	306	71	114	14	0	0	505
Mix	1512	2725	2452	817	321	1232	9059

Table 4: List of Tweets in Silver Data

	0 non-rumor	11 Believe	12 (Deny/ Doubtful/ Neutral)	Total
Obama2014	2940	3055	678	3738
Obama2016	1250	856	379	2485

3.3 Features

In designing the new set of features for the Rumor Retrieval (RR) task we considered two key points. First, addressing the missing words and length issue in Twitter (TLV) and second, extracting a feature that implies the user's belief about each rumor. We also present and conduct RR experiment applying S15 features as one of our baselines. We designed and employed a new set of features in S15 which are tagged by "*" in Table 6. Untagged features represent the features that are used in V11.

3.3.1 Tweet Latent Vector (TLV)

The main intuition behind TLV is to create the latent vector representative of each tweet, since in most of the tweets, there are too few observed words

Table 5: Labels Used in Rumor Retrieval and Rumor Type Classification for Silver Data

1st Step		2nd Step
Method	Labels	Labels
(2-way, 2 step)	(0,2)(11-14)	(11)(12,13,14)

	ID	Value
	* Time	Binary
Twitter and Network Specific	* Hashtag	Binary
	Hashtag Content	String
	URL	Binary
	Re-Tweet	Binary
	*Reply	Binary
	User ID	Binary
Content	Content Unigram	String
	Content Bigram	String
	Pos Unigram	String
	Pos Bigram	String
Pragmatic	*NER	String
	*Event	String
	*Sentiment	String
	*Emoticon	Binary

Table 6: List of S15 features used for RR Experiment .

to tell us what the sentence is about. We assume that the semantic space of both the observed and missing words make up the complete semantic profile of a sentence. We propose the Tweet Latent Vector (TLV) feature by applying the Semantic Textual Similarity (STS) model proposed by (Guo and Diab, 2012) (Guo et al., 2014), which built on the WordNet+Wiktionary+Brown+training data set. STS preprocess each short text by tokenization and stemming, then changes the preprocessed data by removing infrequent words and TF-IDF weighting, and finally uses the model to extract the latent semantics, which is represented as a 100-dimension vector.

3.3.2 Committed Belief

For the belief feature we investigate the level of committed belief for each tweet, which is a modality in natural language, and indicates the author's belief in a proposition. We relied on the Werner et al. (Werner et al., 2015) belief tagger to tag the Committed Belief as(CB) where someone(SW) strongly believes in the proposition, Non-committed belief (NCB) where SW reflects a weak belief in the proposition, and Non Attributable Belief (NA) where SW is not (or could not be) expressing a belief in the proposition (e.g., desires, questions etc.) There is also the ROB tag where SW's intention is to report on SW else's stated belief, whether or not they themselves believe it. The feature values are set to a binary 0 or 1 for each CB, NCB, NA, and ROB corresponding to unseen or observed. The following example illustrates how the belief feature values are created.

Did yall <NA>know</NA> 1 in 5 people <CB>thought</CB> obama is a Muslim

Feature Values : CB:1 NCB:0 NA:1 ROB:0

3.3.3 Content Unigram

Similar to the content lexical features proposed in S15 and V11 we use the bag of word (BOW) feature set comprised of word unigrams. The feature values are set to a binary 0 or 1 for the word unigram vector representative of each tweet.

4 Experimental Design

All the experiments are conducted and evaluated based on various experimental settings. We utilized different data sets, features, and machine learning approaches, which are elaborated in this section.

4.1 Data

We conduct our experiments with two data sets: for the RR experiment we use the mixed data set (MIX) which comprises all the data from the five rumors. We split each of the three data sets into 80% train, 10% development, and 10% test. For the belief investigation experiment we only rely on the Obama dataset. After tagging the silver data by applying the RR model, we randomly select 400 rumors (200 believer-11 and 200 denier-12) from 2010(Gold Data), 2014, and 2016(Silver-data), and investigate how tweet writer's beliefs about the Obama rumors have changed in recent years.

4.2 Baseline

For the RR experiment we adopt three baselines: Majority, S15 features, and the V11 model. The Majority baseline assigns the majority label from the training data set to all the test data. In the S15 baseline we perform the RR experiment by relying on the features that are proposed in S15 and shown in Table 6. We performed the RR experiment with different models in Weka platform and chose the SMO, which yield to the highest result in this experiment. We also compared our results with V11, which reported the results as Mean Average Precision.

4.3 Machine Learning Tools

For the experiments we employ SVM Tree Kernel model, which was proposed by Alessandro Moschitti (Moschitti, 2004). In another experiment, we perform the RR task by applying S15 features, which are illustrated at Table 6 by hiring the SMO classifier on Weka (Hall et al., 2009).

4.4 Experiments and Evaluations

We implement two main experimental pipelines: Rumor Retrieval (RR) and Belief investigation. Content and TLV features are employed for the RR task and then we conduct our experiment in two different phases. In the development phase we utilized development data for tuning. Then the model, which could reach the highest performance, is used on the test data set. Evaluating the performance of the proposed technique in rumor detection should rely on both the number of relevant rumors that are selected (recall) and the number of selected rumors that are relevant (precision), since both of them are presented in this work. In another experiment we investigate the belief change in the Obama rumors. We define two scores for analyzing the belief for the rumor poster/ writer. T_{iCB} and T_{iNCB} are defined for each rumor in the Obama data set. Each of $T_{iBeliefTag}$ corresponds to a number of seen tag in each tweet. We calculate the belief scores for each Obama rumor dataset separately. We apply formula 1 on believer (11) and denier (12) rumor in the Obama data sets.

$$\frac{\#R_{11}BeliefTag}{\#R_{12}BeliefTag} \tag{1}$$

5 Results

In this section the impact of different experimental setups are discussed. We first elaborate on each experiments and then compare our methodology with the baselines.

5.1 Rumors Retrieval

We perform the RR task by applying two sets of features and compare the results with the three baselines. We perform the RR experiment by employing the gold data set to detect Not-Rumor(0 and 2) and Rumor(11, 12, 13, and 14) in one-step two-way classification experiment. For the S15 baseline we applied all the 15 features listed in Table 6. We investigated the performance of different classifiers including J48(Decision Tree), Naive Base(NB,) and SMO and picked SMO which has outperformed the others. In similar experiment for the TLV task we employe TLV and Content features by applying the SVM Tree Kernel model, which lead to 0.972, 0.99 for the MIX and 0.971, 1.0 (precision and Recall) for the Obama gold data set. Table 7 shows how we outperform the other baselines (Majority, S15, and V11) by employing the proposed features.

Table 7: Precision and Recall Of RR task by Employing TLV+Content Unigram, S15, and V11 is reported as Mean Average Precision (MAP)

Data	Method	S15(pr,rec)	V11	TLV
	Majority	0.51,0.71	—	—
MIX	RR	0.94,0.94	**0.965**	**0.972,0.99**
	Majority	0.27,0.52	—	—
Obama	RR	**0.91,0.91**	—-	**0.971,1.0**

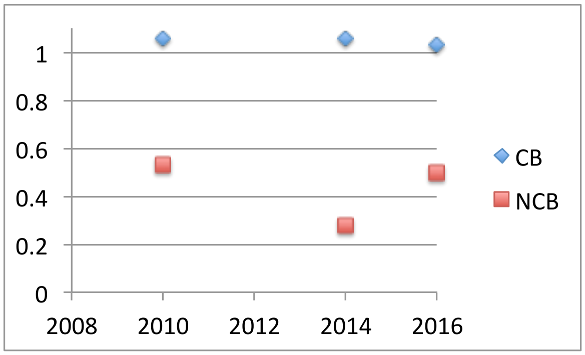

Figure 1: The CB and NCB score for the Obama Data set in 2010, 2014, and 2016

5.2 Belief Analysis

We propose formula 1 to measure the belief score for the Obama data set in different years. Then we investigate how the Committed Belief (CB) and Non Committed Belief (NCB) have changed among rumor believers as well as deniers from 2010 to 2016. Figure-1 shows the Committed Belief and Non-Committed Belief scores among the three data sets. Scores above one mean that the number of the committed belief words in rumor believers is more than in rumor deniers. It is interesting to see that the belief score for the all three years are higher than one. A simple interpretation of that would be, in all 2010, 2014, and 2016, people who were rumor believers in "Obama Being Muslim" show more belief than those who deny Obama being Muslim. On the other hand we see the NCB ratio, which is less than one for the same years. NCB means when SW presents a weak belief towards something. Having below one for NCB could be interpreted as deniers showing weak belief toward the fact that Obama is not a Muslim in 2010, 2014, and 2016. It is important to state that by receiving more data, we can attain more accurate behavior belief.

6 Conclusion and Future Work

In this paper, we proposed and studied the impact of Tweet Latent Vector and Belief on the problem of Rumor Detection in the context of twitter data. A new set of features are employed in our experiments to boost the overall performance of rumor retrieval and give better results in comparison to the other similar body work. We also proposed and analyzed the belief change model among rumor believers and deniers by defining the belief score. We are planning to expand the proposed methodology and investigate the trustworthiness problem from the belief and sentiment points of view and apply the model for streaming data on social media.

Acknowledgement

This paper is based upon work supported by the DARPA DEFT Program.

References

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 864–872. Association for Computational Linguistics.

Weiwei Guo, Wei Liu, and Mona T Diab. 2014. Fast tweet retrieval with compact binary codes. In *COLING*, pages 486–496. Citeseer.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Sardar Hamidiain and Mona Diab. 2015. Rumor detection and classification for twitter data. *The Fifth International Conference on Social Media Technologies, Communication, and Informatics, SOTICS, IARIA*, pages 71–77.

Alessandro Moschitti. 2004. A study on convolution kernels for shallow semantic parsing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 335. Association for Computational Linguistics.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics.

Tatsuro Takahashi and Nobuyuki Igata. 2012. Rumor detection on twitter. In *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*, pages 452–457. IEEE.

Soroush Vosoughi. 2015. *Automatic detection and verification of rumors on Twitter*. Ph.D. thesis, Massachusetts Institute of Technology.

Gregory J Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. Committed belief tagging on the factbank and lu corpora: A comparative study. *ExProM 2015*, page 32.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM.