News Verification by Exploiting Conflicting Social Viewpoints in Microblogs

Zhiwei Jin[1,2], **Juan Cao**[1], **Yongdong Zhang**[1], and **Jiebo Luo**[3]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]University of Rochester, Rochester, NY 14627, USA
{jinzhiwei, caojuan, zhyd }@ict.ac.cn; jluo@cs.rochester.edu

Abstract

Fake news spreading in social media severely jeopardizes the veracity of online content. Fortunately, with the interactive and open features of microblogs, skeptical and opposing voices against fake news always arise along with it. The conflicting information, ignored by existing studies, is crucial for news verification. In this paper, we take advantage of this "wisdom of crowds" information to improve news verification by mining conflicting viewpoints in microblogs. First, we discover conflicting viewpoints in news tweets with a topic model method. Based on identified tweets' viewpoints, we then build a credibility propagation network of tweets linked with supporting or opposing relations. Finally, with iterative deduction, the credibility propagation on the network generates the final evaluation result for news. Experiments conducted on a real-world data set show that the news verification performance of our approach significantly outperforms those of the baseline approaches.

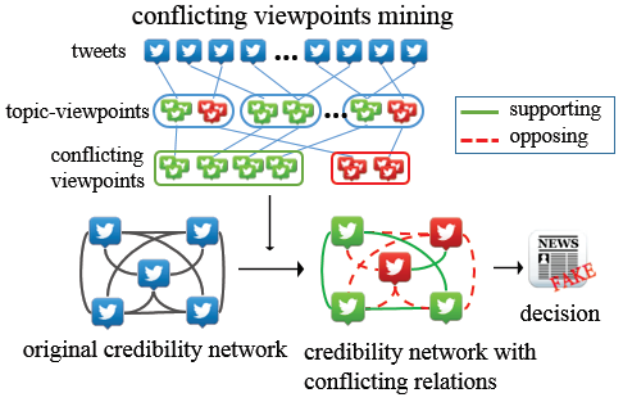

Figure 1: The framework of our proposed approach for news verification.

Introduction

With everyone serving as an information source, news in microblogs is substantial and valuable. Meanwhile, various fake news spreading on microblogs becomes a serious concern recently. For example, it was reported that 92 different rumors were widely spreading in Sina Weibo (a popular microblog service in China) during the first two days of the accident *"Malaysia Airlines Flight MH370 Lost Contact"* (Jin et al. 2014). It is a very challenging task to verify news automatically with the ever-increasing amounts of news in microblogs.

In this paper, we try to address this news verification issue upon deeper inspections, specifically for the conflicting viewpoints expressed by microblog users spontaneously regarding news. Generally, two kinds of relations exist among news tweets. One is supporting relation: tweets expressing the same viewpoint mutually support each other's credibility. The other is opposing relation, which is more subtle and exists among tweets expressing conflicting viewpoints. As microblogs are open media platforms, people can post their observations, opinions and feelings immediately when they read a piece of news. Thus, skeptical and even opposing voices would arise against the news along with original supportive voices in the case of fake news. These conflicting voices are very crucial for verifying the truthfulness of news.

Based on above observations, we propose to exploit the conflicting viewpoints in microblogs to detect relations among news tweets and construct a credibility network of tweets with these relations (Figure 1). First, conflicting viewpoints are mined through a topic model method. Based on this result, relations among tweets are revealed: tweets with the same viewpoint form supporting relations, while tweets with conflicting viewpoints form opposing relations. Rather than examining each tweet individually (Kwon et al. 2013; Gupta et al. 2013; Sun et al. 2013), we then construct a credibility network by linking tweets with detected relations to evaluate them as a whole. Credibility values of tweets would influence each other differently as indicated by these relations. Finally, credibility values of tweets propagate following these links on the network to produce a final decision. In the conflicting viewpoints mining process (Figure 2), news tweets are modeled as various topics. These topics are then clustered into conflicting viewpoints. Compared with traditional opinion mining methods(Kim and Hovy 2007; Park, Lee, and Song 2011; Thomas, Pang, and Lee 2006), our topic model is more focused on finding conflicting relations among tweets rather

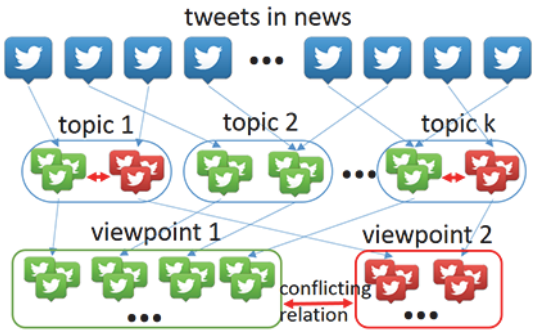

Figure 2: Conflicting viewpoints mining for tweets.

than their sentimental polarities.

In summary, our work makes several main contributions, including:

1. Mining conflicting viewpoints from tweets about news with a topic model method. The mining process is specifically designed to deal with viewpoint imbalance issues which frequently occur.
2. Constructing a credibility network over tweets with supporting and opposing relations established by conflicting viewpoints mining. Consequently, credibility propagation on this network is formulated as an optimization problem and an iterative solution is provided to solve it.
3. Lending itself to desirable early detection of fake news. Such early alerts of fake news can prevent further spreading of malicious content in social media.
4. Collecting a real-word data set from Sina Weibo. The data set is comparable in size with most of the recently released data sets and more importantly, contains ground truth from authoritative sources for a fair evaluation.

Related Work

Microblogs have gained huge popularity around the world, however, they also lead to a lot of false information spreading (Zhao, Resnick, and Mei 2015; Friggeri et al. 2014; Mendoza, Poblete, and Castillo 2010). Generally, existing studies on news verification could be categorized into two classes: classification-based approach and propagation-based approach.

Supervised classification is widely used to identify fake news. By formulating fake news detection as a two-class classification problem, the main concern of this approach is to find effective features for training classifiers. Castillo, Mendoza, and Poblete (2011) present the problem of false information detection on Twitter. Focusing on the newsworthy topics on Twitter, they provide thorough comparisons of various classification algorithms and interesting features for the task. Features are extracted from four aspects: the message, user, topic, and propagation features. Subsequently, similar studies (Yang et al. 2012; Sun et al. 2013) are performed to detect rumors on Sina Weibo with several new features. Most recently, Wu, Yang, and Zhu (2015) propose a hybrid SVM classifier which combines a random walk graph kernel with normal RBF kernel to detect rumors on Sina Weibo. The graph kernel is utilized to measure tweets' propagation graphs. With some intuitive rules, a tweet's propagation graph are pruned to represent its attributes concisely.

Rather than classifying each tweet individually, the propagation approach is presented recently to evaluate tweets credibility as a whole with inter-tweets links. Gupta, Zhao, and Han (2012) optimize the initial tweet classification results with a credibility propagation approach. Specifically, they evaluate an event's credibility by propagating credibility values between users, tweets and events. The entities' credibility values are initialized with results from classifier and optimized through propagation iterations. Entities are linked based on similarities. Recently, Jin et al. (2014) point out the user layer is misleading and they introduce a sub-event layer in the credibility network to capture deeper semantics within an event.

However, existing propagation-based methods ignore the crucial opposing relation among tweets. In this work, we exploit both supporting and opposing relations among tweets to potentially improve the performance of news verification.

Definitions

In microblogs, a piece of news consists of all tweets concerning it. Therefore, the credibility of tweets determines the news credibility altogether. Tweets in news are not isolated but linked with each other with supporting/opposing relations to different degrees. Theses relations are determined by the viewpoints tweets hold. Formal definitions for involved entities are listed as followings.

Definition 1 **Tweet**: *A tweet is a message posted by a user along with its social context.*

Compared with traditional news reports, tweets on social media like microblogs have several unique features: 1) content features (text contents, hash tag topics, external URL links), 2) social features (forward times, comment times), and 3) user features (user profiles, social relations). These features are useful for initializing tweet credibility value (Castillo, Mendoza, and Poblete 2011) and computing inter-tweet implications (Gupta, Zhao, and Han 2012).

Definition 2 **News Event**: *In microblogs, a news event is considered as a set of tweets containing certain keywords during a certain period of time.*

With this definition, news credibility is computed as the average credibility of all tweets it contains. Some existing studies assume one tweet per news event (Kwon et al. 2013; Sun et al. 2013; Wu, Yang, and Zhu 2015). However, it is often difficult to find the appropriate single tweet for a certain news event. This simple assumption also ignores inter-tweet relations and other details.

Definition 3 **Viewpoint**: *A viewpoint is an implicitly expressed instance in response to the news.*

The original set of tweets representing news is complex, containing both positive arguments supporting the news and the negative ones questioning or refuting it. These two different sentiments towards news are defined as viewpoints in this paper. Compared with topic or sub-event (Jin et al.

2014), viewpoints, which contain sentimental information, is more informative to discern deeper semantics of news.

Mining Conflicting Viewpoints

Supporting and opposing relations are very critical in credibility propagation among tweets. On one hand, tweets with the same viewpoints form supporting relations to rise their credibility. On the other hand, tweets with contradicting viewpoints form opposing relations to weaken each other's credibility. Therefore, supporting and opposing relations will help us to judge news credibility from an overall perspective.

Traditional supervised opinion mining methods aim to detect sentimental polarities (negative/positive) of documents. However, in the context of news verification, we are more interested in the conflicting relations between tweets rather than their actual sentiments. Recently, Trabelsi and Zaiane (2014) enhance the LDA topic model (Blei, Ng, and Jordan 2003) for mining conflicting viewpoints in documents. But their assumption that conflicting viewpoints distributed evenly in each topic is impractical for news on social media, because public opinions in social media are often leaning to one side. To address these limitations, we propose to mine conflicting viewpoints from news tweets with unbalanced viewpoints.

First, each tweet is modeled as a pair of dependent mixtures: a mixture of topics and a mixture of viewpoints for each topic. Each topic-viewpoint pair is represented by a probability distribution over document terms. These topic-viewpoints are then clustered into conflicting viewpoints clusters.

The generative process of this model is:

1. for each topic-viewpoint pair kl, draw a multinomial distribution over the vocabulary:$\phi_{kl} \sim Dir(\beta)$.
2. for each tweet t
   (a) draw a topic mixture: $\theta_t \sim Dir(\alpha)$,
   (b) for each topic k, draw a viewpoint mixture: $\psi_{tk} \sim Dir(\gamma)$,
   (c) for each term w_{tn}: sample a topic assignment $z_{tn} \sim Mult(\theta_t)$; sample a viewpoint assignment $v_{tn} \sim Mult(\psi_{tz_{tn}})$; sample a term $w_{tn} \sim Mult(\phi_{z_{tn}v_{tn}})$.

Here, $Dir(\cdot)$ is a Dirichlet distribution, $Mult(\cdot)$ is a multinomial distribution, α, β, γ are fixed Dirichlet's parameters. The three hidden parameters $(\phi_{kl}, \psi_{tk}, \theta_t)$ can be inferred using the collapsed Gibbs Sampling algorithm (Griffiths and Steyvers 2004).

It has been proved that a topic-viewpoint pair is likely to be more similar to viewpoints from the same topic than viewpoints from different topics (Trabelsi and Zaiane 2014). If the distance between the topic-viewpoints of the same topic is large enough (larger than a given threshold h), they most probably belong to different viewpoints. Therefore, we set a threshold parameter to control the cannot-link constraints during clustering: topic-viewpoints with cannot-link constraint must be separated into different viewpoints after clustering.

The distance between two topic-viewpoints is computed as the Jensen-Shannon Distance (D_{JS}) (Heinrich 2005) of their probability distributions. Jensen-Shannon Distance is the symmetric version of the Kullback-Leibler Divergence (D_{KL}), which is a commonly used difference measure for probability distributions:

$$D_{JS}(\phi||\phi') = \frac{1}{2}[D_{KL}(\phi||\hat{\phi}) + D_{KL}(\phi'||\hat{\phi})] \quad (1)$$

$$D_{KL}(\phi||\hat{\phi}) = \sum_t \phi_t[\log_2\phi_t - \log_2\hat{\phi}_t] \quad (2)$$

here, ϕ and ϕ' are topic-viewpoint distributions and $\hat{\phi}$ is the average of ϕ and ϕ'.

In summary, conflicting viewpoints mining for tweets set with unbalanced viewpoints involves three main procedures:

1. Modeling a news corpus as a collection of topic-viewpoint pairs.
2. Computing the distances between topic-viewpoints of the same topic, and then comparing them with the pre-defined threshold h to form cannot-link constraints.
3. A constrained k-means clustering algorithm (Wagstaff et al. 2001) is applied under the computed constraints to cluster topic-viewpoints into two conflicting viewpoints.

With above steps, topic-viewpoints of the same viewpoints are clustered together and conflicting topic-viewpoints are separated.

Credibility Network with Conflicting Relations

To exploit inter-tweet relations for news verification, we link tweets to form a credibility network. In the defined credibility network, n tweets ($t_{1...n}$) are linked with each other. There are two kinds of links: supporting links between tweets taking the same viewpoint and opposing links between tweets taking different viewpoints. The tweet-tweet link is defined as a function $f(t_i, t_j)$. The function's polarity (positive/negative) defines the link type (supporting/opposing) and its value defines the link degree.

After all the tweets are linked and links are computed, initial credibility values are given to individual tweets, which are then propagated over the network until convergence. The credibility of a tweet is defined as a numeric value, and a threshold is used to determine it as real or fake. We define the credibility value $\in [-1, 1]$ and use 0 as the fixed threshold. The initial values are obtained from the predictions of a classifier trained at the tweet level with the extracted features and labeled training data.

Link Definition

The output of conflicting viewpoints mining is used to define the tweet-tweet link $f(t_i, t_j)$. The model distributions generated by the topic model in conflicting viewpoints mining is utilized to compute the distance between tweets; the viewpoints clustering result is used to determine the link type.

During conflicting viewpoints mining, a tweet t is modeled as a multinomial distribution θ_t over K topics, and a topic k is modeled as a multinomial distribution ψ_{tk} over L

viewpoints. So the probability of a tweet t over topic k along with L viewpoints is computed as $p\_tk = \theta_t \cdot \psi_{tk}$. The distance between two tweets t and t' are measured by using the Jensen-Shannon Distance: $Dis(t,t') = D_{JS}(p\_tk||p\_t'k)$. By taking the reciprocal of it, the distance is transformed to a similarity score.

Moreover, the type of link should be determined: supporting or opposing. It is reasonable to assume each tweet conveys only one major topic-viewpoint as tweets are very short messages. The largest proportion of p_tk is defined as the major topic-viewpoint of t. If the major topic-viewpoints of two tweets are clustered together (they take the same viewpoint), then they are mutually supporting; otherwise, they are mutually opposing. The final similarity/dissimilarity measure of two tweets is defined as:

$$f(t_i,t_j) = \frac{(-1)^a}{D_{JS}(p\_t_ik||p\_t_jk)+1} \tag{3}$$

where a is the link type indicator: $a = 0$ if t_i, t_j taking the same viewpoint; otherwise, $a = 1$.

Credibility Propagation

By formulating the credibility propagation on the credibility network as a graph optimization problem, entities' credibility values are propagated to obtain a final result (Yin, Han, and Yu 2008; Yin and Tan 2011; Vydiswaran, Zhai, and Roth 2011). However, negative(opposing) links existing in the network need specific attentions when defining the loss function.

Optimization Formulation

In the proposed credibility network there are one tweet credibility vector $\mathbf{T} = \{C(t_1), ..., C(t_n)\}$ ($C(t_i)$ denotes the credibility value of tweet t_i), and a $n \times n$ tweet-tweet link matrix $\mathbf{W} = [f(t_i,t_j)]$. With the link definition in the last section, $\mathbf{W}$ is symmetric but contains negative values.

To formulate credibility propagation as a graph optimization problem, we make two assumptions over this network: entities with supporting relations should have similar credibility values; entities with opposing relations should have opposite credibility values. Thus, the following loss function is employed (Zhou et al. 2004).

$$Q'(\mathbf{T}) = \mu \sum_{i,j=1}^{n} \mathbf{W}^{i,j} \left(\frac{C(t_i)}{\sqrt{\mathbf{D}^{i,i}}} - \frac{C(t_j)}{\sqrt{\mathbf{D}^{j,j}}} \right)^2 + (1-\mu)\|\mathbf{T}-\mathbf{T}_0\|^2 \tag{4}$$

where, $\mathbf{D}$ is a diagonal matrix, because $\mathbf{D}^{i,j} = \sum_k \mathbf{W}^{i,k}$, and $0 < \mu < 1$ is a regularization parameter.

However, this loss function is not convex for $\mathbf{W}^{i,k}$ may be negative. A non-convex function may have many local minima, and is extremely difficult to optimize. To handle both similarity and dissimilarity, the loss function is redefined (Goldberg, Zhu, and Wright 2007):

$$Q(\mathbf{T}) = \mu \sum_{i,j=1}^{n} |\mathbf{W}^{i,j}| \left(\frac{C(t_i)}{\sqrt{\bar{\mathbf{D}}^{i,i}}} - s_{i,j} \frac{C(t_j)}{\sqrt{\bar{\mathbf{D}}^{j,j}}} \right)^2 + (1-\mu)\|\mathbf{T}-\mathbf{T}_0\|^2 \tag{5}$$

where $\bar{\mathbf{D}}^{\mathbf{i,j}} = \sum_{\mathbf{k}} |\mathbf{W}^{\mathbf{i,k}}|$ and $s_{i,j} = \begin{cases} 1, if\ \mathbf{W}^{i,j} \geq 0 \\ -1, if\ \mathbf{W}^{i,j} < 0 \end{cases}$.

The differences between the two loss functions are in the first term. In order to minimize $Q(\mathbf{T})$: when $\mathbf{W}^{\mathbf{i,j}} \geq \mathbf{0}$, t_i and t_j are mutually supporting, they should have similar credibility values; when $\mathbf{W}^{\mathbf{i,j}} < \mathbf{0}$, t_i and t_j are mutually opposing, they should have opposite credibility values or values both close to zero.

In our loss function (5), the first term is the smoothness constraint which guarantees the two assumptions of supporting and opposing relations; the second term is the fitting constraint to ensure variables not change too much from their initial values; and μ is the regularization parameter to trade off two constraints. Then the credibility propagation on proposed network is formulated as the minimization of this loss function:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}} Q(\mathbf{T}) \tag{6}$$

An Iterative Solution

With derivatives, the analytical solution to the objective function (6) can be easily derived as:

$$\mathbf{T}^* = (1-\mu)(\mathbf{I}-\mu\mathbf{H})^{-1}\mathbf{T}_0. \tag{7}$$

here, $\mathbf{H} = \bar{\mathbf{D}}^{-1/2}\mathbf{W}\bar{\mathbf{D}}^{-1/2}$.

The multiplication and inversion involved in this solution are time-consuming for a large matrix. It is very expensive or impractical to compute it given the large number of tweets in social media. Therefore, we provide an iterative solution to minimize the loss function. Moreover, we can prove it converges to the optimal solution.

The iterative process of the k-th iteration is defined as:

$$\mathbf{T}(k) = \mu\mathbf{H}\mathbf{T}(k-1) + (1-\mu)\mathbf{T}_0 \tag{8}$$

Now, we prove that it converges to the optimal solution. From (8), we have:

$$\mathbf{T}(k) = (\mu\mathbf{H})^{k-1}\mathbf{T}_0 + (1-\mu)\sum_{i=0}^{k-1}(\mu\mathbf{H})^i\mathbf{T}_0 \tag{9}$$

Since $0 < \mu < 1$ and $\mathbf{H}$ is similar to a stochastic matrix, $\lim_{k\to\infty}(\mu\mathbf{H})^{k-1} = 0$ and $\lim_{k\to\infty}\sum_{i=0}^{k-1}(\mu\mathbf{H})^i = (\mathbf{I}-\mu\mathbf{H}\mathbf{T})^{-1}$,

$$\mathbf{T}^* = \lim_{k\to\infty}\mathbf{T}(k) = (1-\mu)(\mathbf{I}-\mu\mathbf{H})^{-1}\mathbf{T}_0 \tag{10}$$

Therefore, the iterative solution (8) is proved to converge to the optimal solution for the optimization problem.

As the iteration converges, each tweet receives a final credibility value, and the average of them is served as the final credibility evaluation result for the news.

Experiments

Data Set

News verification on social media is a fairly new problem, no standard data set is publicly available at the moment.

Table 1: Details of our data set.

	Fake News	Real News	All
Count	73	73	146
#Tweets	23456	26257	49713
#Distinct User	21136	22584	42310

Therefore, we build a data set collected from Sina Weibo for performance evaluation. This data set contains 73 fake news and 73 real news composed of 49,713 tweets, and involves 42,310 distinct users (Table 1).

To form a convincing ground truth, the fake news in this data set were collected from top fake news rank lists selected by authoritative sources, such as Xinhua New Agency; the real news are from a hot news discovery system of Xinhua News Agency. We find that even a "dummy" algorithm classifying all news events as real will achieve a very high accuracy if the number of fake news events is much smaller than the number of real ones and most existing studies also use balanced or near-balanced data sets. Therefore, we randomly sampled an equal number of real news. For each news, we extracted its keywords and duration time. With this information, we crawled tweets on Sina Weibo for this news. After removing duplicated news and news with less than 20 tweets, we finally build a balanced data set of 146 news.

We emphasize that our data set is comparable in size to those in existing studies in terms of news events while contains nearly 10 times more tweets. For example, in most recent studies, the two data sets in (Gupta, Zhao, and Han 2012) have 67 and 83 fake events respectively and the results in (Wu, Yang, and Zhu 2015) are reported on a set of about 5,000 tweets. Although Sina Weibo is the only data set source in this paper, our proposed method have little dependency on data platform or language type.

Performance Evaluation

We compare the performance of proposed approach to several baseline methods for the task of news verification.

Performance Measures To evaluate the performance quantitatively, we consider several performance measures: 1) Accuracy is the percentage of correctly identified fake and real news. It is an overall measurement; 2) Precision and recall for fake news and real news, respectively, represent a model's effectiveness on indentifying each class; 3) F_1 score is computed as the harmonic mean of precision and recall.

Experiment Setup We compare our approach with several existing methods in literatures. All these methods and experimental setup for them are listed here.

Castillo(2011) (Castillo, Mendoza, and Poblete 2011): A classification method at the news event level. To train the classifier, features of tweets in an event are aggregated to the event level. Reported result comes from a decision tree classifier through standard 4-fold cross validation procedure. 47 features from various aspects are extracted for training.

Kwon(2013) (Kwon et al. 2013; Wu, Yang, and Zhu 2015; Gupta et al. 2013): A classification method at tweet level. We trained a SVM classifier with 42 features extracted from three aspects: contents, users and social relations. The reported results are gained with a 4-fold cross validation. Credibility of news is generated as the average of credibility values of all tweets it contains.

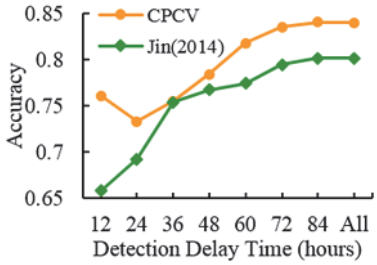

Figure 3: Fake news early detection.

The prediction outputs of this method serve as the initial credibility values for following two propagation-based methods.

Jin(2014) (Jin et al. 2014): A most recent propagation-based method. For parameters in this model, we take the same settings as (Jin et al. 2014).

CPCV: *Credibility Propagation with Conflicting Viewpoints*, our proposed method. For conflicting viewpoints mining, we take the same hyper-parameter settings in (Trabelsi and Zaiane 2014) for topic model, and we set the topics number $K = 10$, viewpoints number fixed as 2 and viewpoints clustering threshold $h = 0.95$. For regularization parameter in (5), we take the setting in (Zhou et al. 2004) as 0.99. The reported results are an average of 20 rounds, as the unsupervised topic model varies slightly in different runs.

Performance Comparison and Discussion From the performance comparison results in Table 2, we can observe that:

- Generally propagation-based approaches (Jin(2014) and CPCV) perform better than classification-based approaches (Castillo(2011) and Kwon(2013)). This is reasonable as propagation-based approaches exploit inter-tweet information.
- Our proposed method CPCV achieves the best result among all the methods for all evaluation measures. It reaches an overall news verification accuracy of 84%.
- CPCV significantly improves the accuracy performance by 4% compared with Jin(2014), because CPCV exploits sentimental conflicting relations among tweets in addition to common similarity relations.

The results validate the importance of supporting and opposing relations in news verification task and prove the effectiveness of conflicting viewpoints mining in our model.

Fake News Early Detection Another interesting experiment is fake news early detection, which aims to give early alerts of fake news. By setting a detection delay time, starting from the first tweet of news, only tweets posted no later than the delay time can be used for verification. It is a very challenging task to verify news with such limited information. But it is desirable for detecting fake news in a real-time

Table 2: Performance comparisons on our proposed data set.

	Accuracy	F precision	F recall	F F_1	R precision	R recall	R F_1
Castillo(2011)	0.74	0.746	0.726	0.736	0.733	0.753	0.743
Kwon(2013)	0.787	0.739	0.89	0.807	0.832	0.685	0.763
Jin(2014)	0.801	0.75	0.904	0.82	0.879	0.699	0.779
CPCV	**0.84**	**0.786**	**0.933**	**0.853**	**0.918**	**0.747**	**0.823**

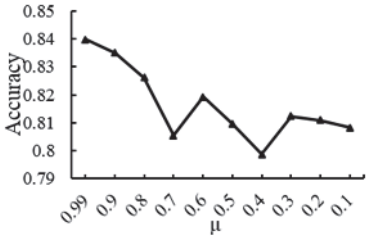

Figure 4: Accuracy of CPCV with varying μ.

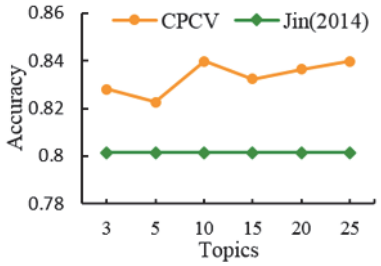

Figure 5: Accuracy comparison with varying topic numbers.

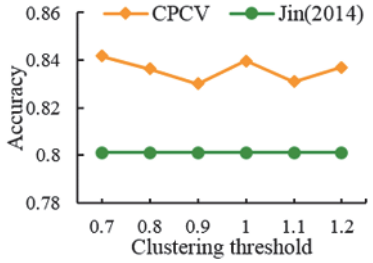

Figure 6: Accuracy comparison with varying clustering threshold.

situation: giving early alerts of fake news can prevent further spreading of malicious content on social media.

We compare the accuracy of news verification task by Jin (2014) and our proposed method (Figure 3):

- Accuracies of both methods increase along with the increase of detection delay time, because more information is available as time passed by.
- CPCV yields significantly better results than Jin(2014) on each time point except that their results are similar at 36 hours. Especially in the earlier stages, when the delay time is 12 hours, the verification accuracy of CPCV is 0.76 while that of Jin(2014) is 0.657. This result shows our method performs even better at early time stages of fake news.
- Another interesting observation is that the accuracy of CPCV at 60 hours (0.818) is already higher than that of Jin(2014) with all tweets used (0.801), which indicates our approach achieves better results than Jin(2014) even with only partial (about two-thirds) information available.

Impact of Regularization Parameter μ In the formulation of credibility propagation, the regularization parameter serves as a trade-off between tweets' links and tweets' initial values. To examine its impact on the performance, we plot the accuracy of CPCV with varying μ (Figure 4). The best performance is reached at $\mu = 0.99$. This means the link implications are very important in this model. Since links are formed and computed through conflicting viewpoints mining, which further validates the success of our exploitation of supporting and opposing relations among tweets.

Impact of Conflicting Viewpoints Clustering As CPCV is dependent on the performance of conflicting viewpoints mining, we evaluate it under different setting of two key parameters to inspect this influence. (These parameters are not used in Jin(2014), we draw its result as a contrast in the comparisons)

In Figure 5, we plot the accuracy performance of CPCV with respect to different choices of topic number K. With the topic number changing from 3 to 25, the verification accuracy varies between 0.82 an 0.84. Despite its variation, the performance of CPCV is always better than that of Jin(2014), even in the worst case (when topics number is 5).

In Figure 6, we plot the accuracy performance of CPCV with different setting of viewpoints clustering threshold h. The accuracy of CPCV varies in a narrow range from 0.83 to 0.84 while h changes from 0.7 to 1.2. So CPCV is quite stable with respect to clustering threshold h. Moreover, the performance of CPCV is still better than that of Jin(2014) under different settings of h.

Conclusion

In this paper, we exploit conflicting social viewpoints in a credibility propagation network for verifying news automatically in microblogs. The credibility network of tweets is constructed with both supporting and opposing relations computed from the viewpoints distributions of tweets. Conflicting viewpoints are discovered with an unsupervised topic model method. By formulating credibility propagation on this network as a graph optimization problem, we define

a sensible and effective loss function and provide an iterative optimal solution. Analysis of experimental results on a data set collected from Sina Weibo shows the effectiveness of our approach.

Acknowledgments

This work was supported in part by National High Technology Research and Development Program of China (2014AA015202), and National Nature Science Foundation of China (61428207,61571424, 61525206). Jiebo Luo would like to thank the support by the New York State CoE Institute for Data Science.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web (WWW)*, 675–684. ACM.

Friggeri, A.; Adamic, L. A.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Goldberg, A. B.; Zhu, X.; and Wright, S. J. 2007. Dissimilarity in graph-based semi-supervised classification. In *International Conference on Artificial Intelligence and Statistics*, 155–162.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web companion*, 729–736. International World Wide Web Conferences Steering Committee.

Gupta, M.; Zhao, P.; and Han, J. 2012. Evaluating event credibility on twitter. In *Proceedings of the SIAM International Conference on Data Mining*, 153. Society for Industrial and Applied Mathematics.

Heinrich, G. 2005. Parameter estimation for text analysis. Technical report, Technical report.

Jin, Z.; Cao, J.; Jiang, Y.-G.; and Zhang, Y. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining (ICDM)*, 230–239. IEEE.

Kim, S.-M., and Hovy, E. H. 2007. Crystal: Analyzing predictive opinions on the web. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1056–1064.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 1103–1108. IEEE.

Mendoza, M.; Poblete, B.; and Castillo, C. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, 71–79. New York, NY, USA: ACM.

Park, S.; Lee, K.; and Song, J. 2011. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 340–349. Association for Computational Linguistics.

Sun, S.; Liu, H.; He, J.; and Du, X. 2013. Detecting event rumors on sina weibo automatically. In *Web Technologies and Applications*. Springer. 120–131.

Thomas, M.; Pang, B.; and Lee, L. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 327–335. Association for Computational Linguistics.

Trabelsi, A., and Zaiane, O. R. 2014. Mining contentious documents using an unsupervised topic model based approach. In *Data Mining (ICDM), 2014 IEEE International Conference on*, 550–559. IEEE.

Vydiswaran, V.; Zhai, C.; and Roth, D. 2011. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 974–982. ACM.

Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S.; et al. 2001. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, 577–584.

Wu, K.; Yang, S.; and Zhu, K. Q. 2015. False rumors detection on sina weibo by propagation structures. In *IEEE International Conference on Data Engineering, ICDE*.

Yang, F.; Liu, Y.; Yu, X.; and Yang, M. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 13. ACM.

Yin, X., and Tan, W. 2011. Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web*, 217–226. ACM.

Yin, X.; Han, J.; and Yu, P. S. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering* 20(6):796–808.

Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, 1395–1405. International World Wide Web Conferences Steering Committee.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. *Advances in neural information processing systems* 16(16):321–328.