Witness Identification in Twitter

Rui Fang, Armineh Nourbakhsh, Xiaomo Liu, Sameena Shah, Quanzhi Li
Research & Development, Thomson Reuters
NYC, USA
{rui.fang, armineh.nourbakhsh, xiaomo.liu, sameena.shah, quanzhi.li}@thomsonreuters.com

Abstract

Identifying witness accounts is important for rumor debunking, crises management, and basically any task that involves on the ground eyes. The prevalence of social media has provided citizen journalism with scale and eye witnesses prominence. However, the amount of noise on social media also makes it likely that witness accounts get buried too deep in the noise and are never discovered. In this paper, we explore automatic witness identification in Twitter during emergency events. We attempt to create a generalizable system that not only detects witness reports for unseen events, but also on true out-of-sample "real time streaming set" that may or may not have witness accounts. We attempt to detect the presence or surge of witness accounts, which is the first step in developing a model for detecting crisis-related events. We collect and annotate witness tweets for different types of events (earthquake, car accident, fire, cyclone, etc.) explore the related features and build a classifier to identify witness tweets in real time. Our system is able to significantly outperform prior methods with an average F-score of 89.7% on previously unseen events.

1 Introduction

Citizen journalism or street journalism involves public citizens playing an active role in collecting, reporting, analyzing, and disseminating news and information. Apart from the fact that it allows bringing in a broader perspective, a key reason for its rise and influence is because of witness reports. Witnesses are able to share an eyewitness report, photo, or video of the event. Another reason is the presence of a common person's perspective, that may otherwise be intentionally or unintentionally hidden because of various reasons, including political affiliations of mass media. Also, for use cases involving time-sensitive requirements (for example, situational awareness, emergency response, and disaster management) knowing about people on the ground is crucial.

Some stories may call for identifying experts who can speak authoritatively to a topic or issue (also called cognitive authorities). However, in breaking-news situations that involve readily perceivable information (for example, fires, crimes) cognitive authorities are perhaps less useful than eyewitnesses. Since most of the use-cases that value citizen reports involve gaining access to information very quickly, it is important for the system to be real time and avoid extensive searches and manual screening of enormous volume of tweets.

Social media has provided citizen journalism with an unprecedented scale, and access to a real time platform, where once passive witnesses can become active and share their eyewitness testimony with the world, including with journalists who may choose to publicize their report. However, the same scalability is available to spam, advertisements, and mundane conversations that obscure these valuable citizen reports. It is clear that discovery of such witness accounts is important. However, presence of significant amount of noise, unrelated content, and mundane conversations about an event that may be not very useful for others, make such a task challenging.

In this paper, we address the problem of automated witness account detection from tweets. Our contributions include: (1) A method to automatically classify witness accounts on social media using only social media data. (2) A set of features (textual and numeric), spanning conversa-

Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, pages 65–73,
Austin, TX, November 1, 2016.

tions, natural language, and meta features suitable for witness identification. (3) A large scale study that evaluates the above methods on a diverse set of different event types such as accidents, natural disasters, and witnessable crimes. (4) Making available an annotated witness database. (5) A real time out-of-sample test on a stream of tweets. In many cases, the presence of witness reports may be the first indication of an event happening. We use the proposed method to determine if surge in witness accounts is related to potential witnessable events.

2 Related Work

A witness may be described as "a person who sees an event happening, especially a crime or an accident"[1]. WordNet defines a witness to be "someone who sees an event and reports what happens" (Miller, 1995), suggesting an expansion from being able to perceive an event to being able to provide a report. From a journalism perspective, witnesses may be defined as "people who see, hear, or know by personal experience and perception" (Diakopoulos et al., 2012).

The motivation behind our definition of witness accounts is that this paper is part of a bigger study on early identification of emergencies and crises through social media. The aim of the larger study is to detect such events prior to news media. In such cases, it is crucial to detect and verify witness accounts before the events are reported by news outlets, and therefore it is important to distinguish between first-hand accounts of the events, and those which are reflected by news reports. The latter type of messages would not be helpful to the study even if they conveyed situational awareness or provided insight into the event.

(Morstatter et al., 2014) explore the problem of finding tweets that originate from within the region of the crisis. Their approach relies only on linguistic features to automatically classify tweets that are inside the region of the crisis versus tweets that are outside the crisis region. The tweets inside the region of the crisis are considered as witness tweets in their experiment setting. However, this is incompatible with our definition of a witness tweet. In our definition, a witness has to be in the crisis region *and* report on having witnessed the event. Thus, we do not consider all the tweets inside the crisis region as witness tweets.

(Cheng et al., 2010) explored the possibility of inferring user's locations based on their tweets. (Han et al., 2014) developed an approach that combines a tweet's text with its meta-data to estimate a user's location. The estimated user location, that is, if they are close to or within the crisis region is used as an indicator of witness tweets, but as discussed above, this is not sufficient for the purposes of our study.

There are few research studies that exclusively concentrate on *situational awareness*. (Verma et al., 2011) explore the automatic identification of tweets for situational awareness. They work on a related problem of finding potential witnesses by focusing on people who are in the proximity of an event. Such tweets may not contain content that demonstrates an awareness of the scope of the crisis and specific details about the situation. However, these tweets are not necessarily from a witness; they could be from a news report of the situation. Hence their problem is not equivalent to ours.

While computational models exist for situational awareness where all within region may be characterized as witness tweets but no real time system exists to identify eyewitness accounts; rather only characterizations of such accounts have been studied. For example, (Truelove et al., 2014) analyzed several characteristics of witness accounts in twitter from a journalistic perspective and developed a conceptual model of witness accounts. Their analysis is based on a case study event (a bushfire), without a computational model for witness identification. They found that witness accounts can be differentiated from non-witness accounts from many different dimensions, such as linguistic use and Twitter's meta data.

3 Data Collection and Annotation

We primarily concentrate on building a real-time system that is able to discover witness reports from tweets. To this purpose, we take a supervised classification approach. Preliminary data analysis revealed that different event types involved varied language specific to that event type, and varied temporal and spatial characteristics specific to the exact event. For example, words used in describing earthquakes might have phrases like 'tremors', 'shaking' but not 'saw suspect'. Also, witness characteristics depended on when and where an

[1] http://dictionary.cambridge.org/dictionary/american-english/witness

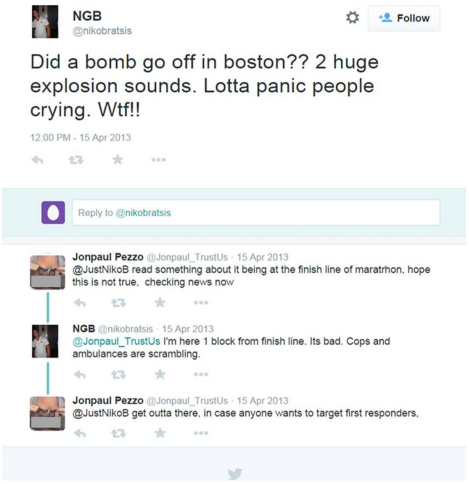

Figure 1: An example witness tweet

event took place. In the next section, we begin by describing our event types.

3.1 Selection of Events

As discussed before, eyewitness accounts are perhaps most useful to journalists and emergency responders during disasters and crises. Therefore we focus on these type of events in building our dataset. These include natural disasters such as floods and earthquakes, accidents such as flight crashes, and witnessable criminal events such as acts of terrorism.

We formed an events list by evaluating the disaster and accident categories in news agency websites, for example, Fox news disasters category[2]. We found the following events: cyclones, (grass)fires, floods, train crash, air crash, car accidents, volcano, earthquake, landslide, shooting, and bombing. Note that the events (within or cross category) may be distinct on several integral characteristics, like different witness/non-witness ratios. This is mainly due to the varying spatial and temporal characteristics of the events. For example, the Boston Marathon Bombing happened in a crowded place and at daytime. This led to a large number of eye-witnesses, who reported hearing the blast, and the ensuing chaos. Figure 1 shows an example witness tweet from Boston marathon bombing. On the other hand for the landslide that occurred 4 miles east of Oso, Washington, there were very few people near the landslide site. Thus, most of the tweets related to that landslide actually originated from some news agency report.

[2]http://www.foxnews.com/us/disasters/index.html

3.2 Data Collection

In order to study the identification of eyewitnesses, we needed to identify some events and collect all related tweets for each event. Some previous studies(Yang et al., 2012; Castillo et al., 2011) used TwitterMonitor(Mathioudakis and Koudas, 2010) that detected sudden bursts of activity on Twitter and came up with an automatically generated Boolean query to describe those trends. The query could then be applied to Twitter's search interface to capture more relevant tweets about the topic. However, TwitterMonitor is no longer active. We formulated the required search queries manually, by following a similar approach.

3.3 Query Construction

Each query was a boolean string consisting of a subject, a predicate, and possibly an object. These components were connected using the AND operator. For instance,"2014 California Earthquake" was transformed to "(California) AND (Earthquake)". Each component was then replaced with a series of possible synonyms and replacements, all connected via the OR operator. For instance, the query may further be expanded to "(California OR C.A. OR CA OR Napa) AND (earthquake OR quake OR earthshock OR seism OR tremors OR shaking)". Finally, we added popular hashtags to the search query, as long as they didn't exceed Twitter's limit of 500 characters. For instance, the query would be expanded by hashtags such as "NapaEarthquake". As we read the retrieved tweets, more synonyms and replacements were discovered which we added them back to the query and searched in Twitter again. We repeat this process several times until the number of retrieved tweets is relatively stable. This process can help us find a good coverage of event tweets and witness tweets. However, we believe it is very hard to evaluate the accurate recall of our query results since we have to (1) have the complete twitter data of a specific time period and (2) label a huge amount of tweets.

3.4 Search

Each query was applied to Twitter to collect relevant tweets. Twitter offers a search API that provides a convenient platform for data collection. However, the search results are limited to one week. Since some of the items in our data-set

Table 1: Descriptive statistics of the events in the collected data-set

Event	# Witness tweets	# Total tweets
Cyclone	37	13,261
Grass fire	5	6,739
River flood	27	6,671
Flight crash	17	7,955
Train crash	5	7,287
Car accident	32	19,058
Volcano	2	3,096
Tornado	7	6,066
Earthquake	127	40,035
Landslide	1	3,318
Shooting	3	5,615
Bombing	138	31,313

spanned beyond a week's time, we could not rely on the search API to perform data collection. Instead, we decided to use Twitter's search interface, which offers a more comprehensive result set. We used an automated script to submit each query to the search interface, scroll through the pages, and download the resulting tweets.

For our event categories, we found 28 events with a total of 119,101 related tweets. If there were multiple events of either category then they were merged into their respective category. For example, tweets from 6 distinct grass fire events were merged into a single grass fire event. Similarly 3 train crashes, 3 cyclones, 3 flight crashes, 3 earthquakes, 2 river floods, 2 car accidents, and 2 tornadoes were merged. Table 1 provides further details on the different events.

3.5 Witness annotation

We first applied the following two filters to automatically label non-witness tweets.

1. If tweet text mentions a news agency's name or contains a news agency's url, it is not a witness tweet. For example, "Breaking: Injuries unknown after Asiana Airlines flight crash lands at San Francisco Airport - @AP"

2. If it is a retweet (since by definition it is not from a witness even if its a retweet of a witness account).

After the above filtering step, 46,249 tweets were labeled as non-witness tweets, while 72,852 tweets were left for manual annotation. Two annotators were assigned to manually label a tweet as a witness tweet in case it qualified as either of the following three categories(Truelove et al., 2014):

- Witness Account: Witness provides a direct observation of the event or its effects. Example: "Today I experienced an earthquake and a blind man trying to flag a taxi. I will never take my health for granted."

- Impact Account: Witness describes being impacted directly or taking direct action due to the event. Example: "Had to cancel my last home visit of the day due to a bushfire.".

- Relay Account: Micro-blogger relays a Witness Account or Impact Account of another person. Example: "my brother just witnessed a head on head car crash".

If neither of the above three, then the tweet was labeled as a non witness account. After the annotation (The kappa score for the inter-annotator agreement is 0.77), we obtained in 401 witness tweets and 118,700 non-witness tweets.

4 Methodology

In this section, we outline our methodology for automatically finding witness tweets using linguistic features and meta-data. We first discuss the features, and then the models used.

4.1 Linguistic Features

Linguistic features depend on the language of Twitter users. Currently we concentrate only on English. Previous related works have also shown the utility of a few linguistic features (Morstatter et al., 2014; Verma et al., 2011) such as N-grams of tweets, Part-of-Speech and syntactic constituent based features. The following describes our new features:

Crisis-sensitive features. Parts-of-speech sequences and preposition phrase patterns (e.g., "near me").

Expression: Personal/Impersonal. If the tweet is a description of personal observations it is more likely to be a witness report. We explore several features to identify personal experiences and perceptions. (1) If the tweet is expressed as a first person account (e.g., contains first personal pronoun such as "I") or (2) If the tweet contains words that are from LIWC[3] categories such as "see" and

[3]http://www.liwc.net/

"hear", it is indicative of a personal experience; (3) If the tweet mentions news agency names or references a news agency source, it is not about a personal experience and thus not a witness report.

Time-awareness. Many witness accounts frame their message in a time-sensitive manner, for example, "Was having lunch at union station when *all of a sudden* chaos!" We use a manually created list of terms that indicate time-related concepts of immediacy.

Conversational/Reply feature. Based on analysis of the collected witness and non-witness tweets, we observe that the responses to a tweet and the further description of the situation from that original user helps confirm a witness account. We extract the following features: (1) If the reply tweet is personal in expression; (2) If the reply tweet contains journalism-related users; (3) If the reply tweet is from friends/non-friends of the original user; (4) If the reply tweet is a direct reply (to the original tweet).

Word Embedding The recent breakthrough in NLP is the incorporation of deep learning techniques to enhance rudimentary NLP problems, such as language modeling (Bengio et al., 2003) and name entity recognition(Collobert et al., 2011). Word embeddings are distributed representations of words which are usually generated from a large text corpus. The word embeddings are proved to be able to capture nuanced meanings of words. That is why word embeddings are very powerful in NLP related applications. In this study, the word embedding for each word is computed using neural network and generated from billions of words from tweets, without any supervision.(more details in Section 4.4)

4.2 Meta features

In addition to linguistic features, there are a few other indicators which might help identify witness accounts. (1) **Client application**. We hypothesize that witness accounts are likelier to be posted using a cellphone than a desktop application or the standard web interface; (2) **Length of tweet**. The urgent expression of witness tweets might require more concise use of language. We measure the length of a tweet in terms of individual words used; (3) **Mentions or hashtags**. Another indication of urgency can be the absence of more casual features such as mentions or hashtags.

Table 2: Description of features

contains first-person pronoun, i.e. "I","we"
contains LIWC keywords,i.e."see","hear" ?
contains news agency URL or name?
is a retweet?
contains media (picture or video)?
contains time-aware keywords?
journalist account involved in conversation?
situated awareness keywords in conversation?
contains reply from friend/non-friend
contains direct/indirect reply
type of client application used to post the tweet
length of tweet in words
contains mentions or hashtags?
similarity to witnessable emergency topics
word embeddings

4.3 Topic as a feature

As mentioned previously, witness accounts are most relevant for reporting on *witnessable* events. These include accidents, crimes and disasters. Thus, we hypothesize that features that help identify the topic of the tweets may help measure their relevance. Therefore we incorporate topic as a feature. We use OpenCalais'[4] topic schema to identify witnessable events. The following sections describe how we use these categories to generate topic features.

Table 2 shows the set of new features we proposed in witness identification.

4.4 Feature Extraction

In addition to the features introduced above, we experimented with several other potential features such as objectivity vs. emotion, user visibility and credibility, presence of multimedia in the message, and other linguistic and network features. They did not improve the performance of the classifier, and statistical analysis of their distributions across witness and non-witness messages failed to show any significant distinctions. Due to space limit, we provide the feature extraction details for two features.

Topic Features: Using OpenCalais' topic-classification api, we classified about 33,000 tweets collected via Twitter's streaming API in January-June 2015. We then separated those classified as WAR_CONFLICT, LAW_CRIME, or DISASTER_ACCIDENT. This resulted in 7,943

[4]http://www.opencalais.com/opencalais-api/

Table 3: Description of data set for training word embeddings

# of Tweets	198 million
# of words in training data	2.9 billion
# of unique tokens	1.9 million

tweets. Three researchers manually cross-checked the classification for accuracy. For each topic, 500 tweets on which all researchers agreed were chosen to represent that topic. We calculated TF-IDF metrics on these tweets and represented each topic as a vector of terms and their TF-IDF values. When applying these features to the training data, we calculated the cosine similarity between the term vector of each tweet and the term vector of each topic.

Word Embeddings: To extract word embeddings for each word in tweet, we use the word2vec toolkit[5]. word2vec is an implementation of word embeddings developed by Mikolov et al.(Mikolov et al., 2013). This model has two training options, continuous bag of words (CBOW) and the Skip-gram model. The Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. Based on previous studies the Skip-gram model produces better results and we adopt it for training.

We train the model on tweet data. The tweets used in this study span from October 2014 to September 2015. They were acquired through Twitter's public 1% streaming API and Twitter's Decahose data (10% of Twitter streaming data) granted to us by Twitter for research purposes. Table 3 shows the basic statistics of the data set used in this study. Only English tweets are used, and about 200 million tweets are used for building the word embedding model. Totally, 2.9 billion words are processed. With a term frequency threshold of 5 (tokens with fewer than 5 occurrences in the data set are discarded), the total number of unique tokens (hashtags and words) in this model is 1.9 million. The word embedding dimension is 300 for each word.

Each tweet is preprocessed to get a clean version, which is then processed by the model building process.

[5]Available at `https://code.google.com/p/word2vec/`

Table 4: A case study of transfer learning for witness identification

Models	Test on earthquake event
Model 1: trained on non-earthquake events	83.3%
Model 2: trained on earthquake event	87.0%

5 Experiments and Evaluation

To classify tweets as witness or non-witness automatically, we take a machine learning approach, employing several models such as decision tree classifier, maximum entropy classifier, random forest and Support Vector Machine (SVM) classifier to predict whether a tweet is a witness tweet or not. (SVM classifier performed the best for our method as well as on baselines, we only report results using SVM). As input to the classifier, we vectorized the tweet by extracting the features from the tweet's text and meta-data. Each of our features are represented as whether they occur within the tweet, i.e. Boolean features. The model then outputs its prediction of whether the tweet is a witness account.

5.1 Transfer learning

We first perform a case study of transfer learning. We trained one model on all event-types and tested on a specific type of event (e.g. earthquake). We then trained a second model for that specific type of event and compared the performance of these two paradigms. We choose earthquake events in our dataset for case study. We trained two models on 1000 tweets with witness and non-witness accounts and test on an event with 500 tweets. *Model 1* is trained on all other types of events, while *Model 2* is trained on another earthquake event. Table 4 shows the results. The F-1 score of *Model* 1 and 2 are 83.3%, 87.0% respectively. This suggests that event-based witness identifiers have better performance than general witness identifiers, but the model generalizes relatively well.

For the next experiment, we balanced the collected data by over-sampling the witness tweets by 10 times, and down-sampling the non-witness tweets to the same size accordingly. We then perform leave one out cross validation. For each event category, we use all tweets in other event cate-

gories to train the model. Once the training is done, we test the trained model on the tweets in the holdout event category. For example, for the cyclone category, we would use all tweets in all other 11 categories (grass fire, river flood, flight crash, train crash,...,) to train the model, and test the model on cyclone category tweets. This process was repeated for each event type.

5.2 Comparison of Prediction Models

We compared our proposed method with two baseline models from the literature(Diakopoulos et al., 2012; Morstatter et al., 2014).

- **Baseline 1**: A dictionary-based technique(Diakopoulos et al., 2012). The approach classifies potential witnesses based on 741 words from numerous LIWC categories including "percept", "see", "hear", and "feel". The approach applied one simple heuristic rule: If a tweet contained at least one keyword from the categories, then the tweet is classified as witness tweet.
- **Baseline 2**: A learning based approach(Morstatter et al., 2014). This method extracts linguistic features(as shown in Table 2) from each tweet and automatically classifies tweets that are inside the region of the crisis versus tweets that are outside the crisis region.

Table 5: Witness identification F-score for each event and model: Baseline

Testing Events	F-score	
	Baseline 1	Baseline 2
Cyclone	8.7%	75.1%
Grass fire	6.9%	**95.0%**
River flood	65.8%	83.3%
Flight crash	23.1%	77.2%
Train crash	39.9%	91.2%
Car accident	54.4%	86.1%
Volcano	46.0%	76.8%
Tornado	1.8%	83.9%
Earthquake	36.3%	77.3%
Landslide	46.1%	70.1%
Shooting	15.0%	80.9%
Bombing	34.5%	72.2%
Average	31.5%	80.8%

We experiment a set of models for witness identification:

- **Model i** (+Conversation) combines the new proposed 'conversational features' with all the features used in **Baseline 2**(Morstatter et al., 2014).
- **Model ii** (+Expression) combines the new proposed tweet 'expression features' with all features used in Baseline 2.
- **Model iii** (+Conversation+Expression) combines the new proposed conversational and tweet expression features with all features used in **Baseline 2**.
- **Model iv** (+Conversation+Expression+Meta) combines the previous classifier with meta features and topic-related features.
- **Model v** (**WE.**) uses only word embedding features which were obtained by an unsupervised learning process as described in subsection 4.4. As tweets are of various length, in order to get a fixed size feature vector representation of tweet to train the SVM, we explore min, average, and max convolution operators(Collobert et al., 2011). Specifically, we treat each tweet as a sequence of words $[w_1, ..., w_s]$. Each word is represented by a d-dimensional word vector $\mathbf{W} \in \mathfrak{R}^d$ (note that, $d = 300$ in our case). For each tweet s we build a sentence matrix $\mathbf{S} \in \mathfrak{R}^{d \times |s|}$, where each column k represents a word vector $\mathbf{W}_k$ in a sentence s. We can calculate the minimum, average, and max value of each row in the sentence matrix $\mathbf{S} \in \mathfrak{R}^{d \times |s|}$ and form a d x 1 vector, respectively. These d x 1 feature vector is used to train SVM classifier. Our empirical results shows that the max operator obtains the best results in a sample training data, so we only report this for the **WE.** model.
- **Model vi** (+Conversation+Expression+Meta+WE.) combines the handcrafted features used in **Model iv** with the word embedding features used in **Model v**.

For experiment and evaluation, we group similar events (for example, car accidents that happened in different times and locations) together, and perform a leave one out cross validation. More specifically, we used SVM classifier trained on

Table 6: Witness identification F-score for each event and model

Testing Events	F-score					
	Model i	**Model ii**	**Model iii**	**Model iv**	**Model v**	**Model vi**
Cyclone	75.5%	88.5%	**89.7%**	86.5%	87.0%	88.6%
Grass fire	94.7%	91.1%	93.6%	93.2%	93.1%	94.1%
River flood	83.3%	**91.5%**	91.4%	81.1%	82.2%	82.6%
Flight crash	77.5%	79.1%	81.5%	91.3%	85.7%	**91.5%**
Train crash	90.5%	90.9%	89.2%	92.8%	**92.9%**	**92.9%**
Car accident	88.1%	87.9%	88.5%	92.6%	90.7%	**92.7%**
Volcano	77.9%	81.0%	82.6%	**93.3%**	87.0%	93.1%
Tornado	85.9%	90.8%	**94.8%**	94.1%	93.8%	94.3%
Earthquake	78.8%	**80.8%**	80.7%	80.8%	80.5%	**80.9%**
Landslide	73.6%	80.7%	82.3%	85.7%	**85.9%**	85.5%
Shooting	82.8%	91.2%	92.2%	97.7%	93.0%	**97.8%**
Bombing	72.2%	72.8%	73.4%	82.0%	75.3%	**82.1%**
Average	81.7%	85.5%	86.7%	89.3%	87.2%	**89.7%**

data from all other types of events to classify tweet data from a new event. The F-score for each event as well as the average F-score are reported in Table 5, 6.

Table 5,6 show that our approaches were able to outperform previous two baseline approaches on categorizing witness tweets, with an average F-score of 81.0%, 85.5%, 86.7%, 87.2%, 89.3% and 89.7%, respectively.

The results indicate that our system is able to significantly outperform the two baseline approaches with an highest average F-score of 89.7% on previously unseen events.

It is interesting to observe that, the performance of **Model v** which uses only word embedding features obtained from unsupervised training on large tweet data-set, is comparable to the learning model (e.g. **Model iv**) that use hand-crafted features. Furthermore, when word embedding features are combined with handcrafted features (**Model vi**), the model's performance is further improved. One main reason is that the word embedding features explicitly encode many linguistic regularities and patterns which might not have been well captured by hand-made features. This result is in line with studies on other natural language processing task such as sentiment analysis (Tang et al., 2014).

We also observe that conversational features do not seem to improve performance to a considerable level (80.8% for Baseline 2 Versus 81.7% for **Model i**), we think that might be partially due to two reasons: (1) the fact that not all tweets lead to conversations (see statistics on Subsection 4.1); (2)the way we extract the conversational features is preliminary. In the future we will collect more data and explore more sophisticated features from conversations.

5.3 Witness identification on the real-time streaming Twitter data

In this section, we evaluate the hypothesis of whether detecting a witness accounts indicates that an event has taken place. We apply our witness identification model on streaming real-time Twitter data. For the time period that we tested in, the number of real-time tweets were 7,517,654 tweets. In the entire tweet collection, 47,254 tweets were identified as witness tweets. Based on a simple similarity measure, we clustered the tweets. If less than 3 tweets were found in a cluster, we eliminated that cluster. This led to 49,906 clusters or events. Of the 47,254 witness tweets, 1782 were from the clusters. Note that the proportion of witness tweets is 3.57% in the cluster events and only 0.63% in the streaming 1% sample. This suggests that there is a relationship between statistically finding more witness accounts and detection of events. In future, we aim to study this relationship in more detail.

6 Conclusion

We proposed a witness detection system for tweets. We studied characteristics of witness reports and proposed several diverse features. We show that the system is robust enough to work well on both in sample and true out of sample events.

References

[Bengio et al.2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

[Castillo et al.2011] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *WWW*, pages 675–684.

[Cheng et al.2010] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA. ACM.

[Collobert et al.2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

[Diakopoulos et al.2012] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2451–2460, New York, NY, USA. ACM.

[Han et al.2014] Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *J. Artif. Int. Res.*, 49(1):451–500, January.

[Mathioudakis and Koudas2010] Michael Mathioudakis and Nick Koudas. 2010. Twittermonitor: trend detection over the twitter stream. In *Proc. of the 2010 ACM SIGMOD International Conference on Management of Data*, pages 1155–1158. ACM.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[Miller1995] George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of ACM*, 38(11):39–41, November.

[Morstatter et al.2014] Fred Morstatter, Nichola Lubold, Heather Pon-Barry, Jrgen Pfeffer, and Huan Liu. 2014. Finding eyewitness tweets during crises. In *ACL Workshop on Language Technology and Computational Social Science*.

[Tang et al.2014] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1555–1565.

[Truelove et al.2014] Marie Truelove, Maria Vasardani, and Stephan Winter. 2014. Towards credibility of micro-blogs: characterising witness accounts. *GeoJournal*, 80:339–359.

[Verma et al.2011] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. 2011. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

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