Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

#Unconfirmed: Classifying Rumor Stance in Crisis-Related Social Media Messages

Li Zeng
University of Washington
Mary Gates Hall, Suite 370
Seattle, Washington 98195

Kate Starbird
University of Washington
428 Sieg Hall
Seattle, Washington 98195

Emma S. Spiro
University of Washington
Mary Gates Hall, Suite 370
Seattle, Washington 98195

Abstract

It is well-established that within crisis-related communications, *rumors* are likely to emerge. False rumors, i.e. *misinformation*, can be detrimental to crisis communication and response; it is therefore important not only to be able to identify messages that propagate rumors, but also corrections or denials of rumor content. In this work, we explore the task of automatically classifying rumor stances expressed in crisis-related content posted on social media. Utilizing a dataset of over 4,300 manually coded tweets, we build a supervised machine learning model for this task, achieving an accuracy over 88% across a diverse set of rumors of different types.

Introduction

During crisis events, emergency responders, media, and members of the public utilize social media to disseminate, search for and curate crisis-related information, making sense of occurring uncertain events (Vieweg et al. 2010; Sutton et al. 2014). On social media, those behaviors result in a complex, dynamic event-related communication environment contained within a larger stream of informal communication on these platforms (Mendoza, Poblete, and Castillo 2010). Such crisis-related communication environments often naturally lead to the emergence of *rumors*. Generally defined, a rumor is a statement of information unverified during the time of communication (Shibutani 1966). Social media platforms have likely made it possible for rumors to spread faster and reach a wider audience than previously possible (Qazvinian et al. 2011; Spiro et al. 2012).

Detecting emerging misinformation and communicating such cases to both the public and emergency responders could alleviate concerns about the use of social media for crisis and risk communication (Hiltz, Kushma, and Plotnick 2014). Prior work has focused on understanding the spread of misinformation (Starbird et al. 2014), as well as introduced preliminary methods to automatically identify misinformation within a larger social media context (Qazvinian et al. 2011). While this later work is the closest to the study presented herein, it was not specifically focused on rumors during crisis, but rather rumors about high-profile, public figures (a much different case). Prior research on detecting misinformation has not yet been extended to a general framework that applies across rumors; current approaches have focused almost exclusively on false rumors and misinformation.

Recent exploratory work indicates that comparing the temporal and volume of affirmations and denials of rumor content could provide signals for both true and false rumors (Starbird et al. 2014), however more work is needed to quantify these patterns across rumors and crisis events. In addition, prior work has been limited in its scope as it relies on manually coded social media posts to determine whether content affirms or denies rumor claims. In short, methods are needed that allow scholars to automatically classify social media messages as rumor affirming or rumor denying.

We aim to fill this gap, building a classification model for rumor stance. These methods will not only allow large-scale investigations of rumoring behavior on social media because they will deliver large sets of labeled data, but they will also offer researchers the opportunity to gain a more nuanced view of rumoring and collective sense-making by indicating which features of rumor-related content are indicative of the posters stance towards rumor claims.

Data

Data for this project come from the microblogging platform Twitter. Data consist a labeled set of rumor-related Twitter posts (i.e. tweets) from a hostage crisis – termed the *Sydney Siege* by media – that occurred in Sydney in December, 2014. Custom Python scripts are used to access tweets via the Twitter Streaming API. Data collection relies on a curated set of event-related keywords and hashtags. Our research team used a mixed-methods approach to identify five notable rumors (named as *Flag*, *Airspace*, *Suicide*, *Hadley*, *Lakemba*, respectively) in the data. For each rumor, a set of rumor-related tweets are extracted from the larger dataset using a rumor-specific, text-based query; queries are designed to produce a comprehensive, low-noise sample of tweets related to a particular rumor story.

Data are labeled by human coders according to rumor stance. Two trained coders manually code every distinct tweet; [1] disagreements are arbitrated by a third coder. Un-

[1] We performed an inter-rater reliability analysis using the Fleiss' Kappa statistic (Fleiss 1971) to evaluate agreement among raters and obtained the result of Kappa as 0.892 ($p < 0.001$).

codable and rumor-unrelated tweets are removed. Remaining tweets are classified as one of three mutually exclusive categories: *affirm, deny* or *neutral*. *Affirming* tweets affirm the ongoing rumor story, serving to pass on or propagate the rumor. *Denial* tweets attempt to deny the rumor story, correcting misinformation. *Neutral* tweets do not take a stance, and are ignored in the subsequent analysis where we focus on the affirm/deny distinction. Detailed elaboration of the dataset and data processing processes can be found in our related work (Arif et al. 2016; Zeng, Starbird, and Spiro 2016).

The resulting labeled dataset is described in Table 1. Interestingly, Table 1 demonstrates that the proportions of rumor-affirming tweets and rumor-denying tweets across these rumors are unbalanced, with rumor-affirming tweets accounting for the large-majority of cases. Prior work suggests this may result from the tendency for misinformation or "bad news" to garner more attention.

Rumor name	Affirm tweets	Deny tweets	Prop. affirm tweets
Flag	1347	980	0.57
Airspace	636	356	0.64
Suicide	343	38	0.90
Hadley	516	25	0.95
Lakemba	64	70	0.47
Total	2906	1469	0.66

Table 1: Basic descriptive statistics for Sydney Siege data (duplicate tweets removed).

Methods

Our task in classifying rumor stance from tweets is to automatically apply rumor-affirming or rumor-denying labels to tweets with reasonable accuracy. Following standard techniques in statistical and machine learning, we construct a training dataset to fit and tune our model. So as not to bias our classification towards popular posts, we remove all exact duplicates. Evaluation of the model then occurs on held-out test data, allowing us to estimate out-of-sample performance. In practice we do this using cross-validation.

Due to the unbalanced nature of the raw data – the fact that we have far more tweets with an `Affirm` label than a `Deny` label – we consider both a balancing sampling strategy and proportional sampling strategy. For the balancing sampling case, we construct both the training set and the testing set with equal proportion of `Affirm` and `Deny` labels, whereas for the proportional sampling (which is closer to the real-world situation), both the training and testing sets have the same proportions of the binary labels as in the whole dataset. We take the majority-category label assignment as a baseline against which we compare our model.

Feature Extraction and Generation

- **Punctuation features**: Punctuation may be indicative of emotion and/or rumor stance. We extract the number of exclamation marks, question marks and '?!'.
- **Twitter-element features**: When people tend to confirm or challenge a story, they often refer to external resources (in the form of URLs), mention other users as evidence of information sources and add hashtags to make their tweets more easily seen by others.
- **LIWC features**: We extract other lexical features using the Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker 2010) dictionaries for negation words, swear words, negative emotion words, positive emotion words, emoticons, personal pronouns, adverbs, etc.
- **Tweet sentiment features**: We use an external sentiment classifier API provided by MetaMind[2] to perform a three-class sentiment classification on a tweet. The model returns one of the three exclusive sentiment labels - positive, neutral or negative for each tweet.
- **N-grams**: We do not include bag-of-words because the LIWC dictionaries capture most of the essential single words. We find that removing stopwords decreases accuracy. Therefore, we only apply word stemming. In addition, we apply reduction techniques to n-gram features by setting a minimum frequency threshold and lowercasing.
- **Part of speech features**: We apply a part-of-speech (POS) tagger using a Tweet NLP system (Owoputi et al. 2013). The system is used for POS tagging for informal online messages. By doing so, we obtain the part of speech for each word.

Classifier Training

For performing this classification task, we consider three different models - logistic regression, Gaussian Naïve Bayes and a random forest. For logistic regression, we experiment with both Lasso and Ridge regularization – both options do not seem to have substantial influence on performance and so we adopt L^2 regularization. We experiment tuning the random forest to choose the proper number of trees; considering both accuracy and computational complexity, we use 30 trees in the model.

We trained models for two cases: (1) a classifier trained on each of the five single rumor sets individually and (2) a classifier trained on the pooled rumor set that is obtained by aggregating all five rumors for the Sydney Siege event. Hypothetically, in the first case, the classifier would be able to pick up on rumor affirming and denying statements specific to the given rumor. For the pooled classifier, context of specific rumors might become blurred as all rumors related to a crisis event merge together. Therefore, in this later case, classifying rumor stances might become more challenging, because the classifier must be able to generalize across rumor cases, however, the task is closer to more general real-world, future applications of the model. For each model, we use a 10-fold cross-validation procedure. All of the results reported below are averaged across the 10-folds.

[2]Available at https://www.metamind.io; The claimed accuracy of this sentiment classifier is 81.73%.

Results

Rumor-Specific Prediction

We fit a classification model to predict whether tweets affirm or deny rumors in each specific rumor in the Sydney data. The first five rows of Table 2 show the results for each rumor-specific classification model. This model includes each of the features discussed previously. We show only the proportional sampling case here, however even sampling was also performed. Our results demonstrate that except for the *Hadley* rumor, which has a extremely high baseline, nearly all the rumor-specific classifiers are able to beat the baseline. For all five cases, the random forest has the best performance with strong precision, recall and accuracy.

Rumor Stance Prediction

Next, we use a pooled dataset – combining all tweets from each of the five rumors – to train the model. Results are shown in the last two rows of Table 2. We show the model evaluation for the case of both even and proportional sampling. The baseline in the case of an even sampling strategy is 0.5 due to the balance proportion of `Affirm` and `Deny` tweets in the data. All three models are able to beat the baseline and achieve accuracies around or over 0.8. The random forest classifier remains the most effective with the highest recall, accuracy and F-score.

Table 3 shows the confusion matrix for the best-performing classifier trained on the pooled data using proportional sampling strategy. The number of false negative (tweets predicted as denials, but actually are affirmations) is small, which indicates this model performs well in identifying denials. This result could be useful for our larger problem of interest - using crowd corrections as signal of misinformation - because rumor-denying tweets tend to be in a fewer number than rumor-affirming tweets in empirical cases suggesting they may get "lost" within the larger social media stream produced during the crisis.

Overall, the pooled model performs much better than the corresponding baselines, indicating the classification model is able to achieve around 83%-88% accuracy when predicting expressed rumor stances in crisis-related tweets. Compared to other classifiers, the random forest model tends to be more robust in both rumor-specific and pooled contexts using different sampling strategies.

Feature Importance

We examine feature importance in the trained random forest model to better understand which of the model features have higher importance in this classification task. Overall, features of note span a variety of feature categories, including n-grams, basic textual features, sentiment, part of speech, and lexical LIWC features, etc. Figure 1 presents the top 10 features for random forest model on the pooled data. Recall, this represents average results across 10-fold cross validation. Negation words seem to play an large role in classification in this case. This includes words and phrases such as `not`, `is not` (bi-gram), `not an` (bi-gram). We found that some combinations of part-of-speech are more significant, such as *adverb*, *verb + adverb*, *adverb + proper noun*, *adverb + determiner*, indicating particular "speech" patterns that users might use for expressing their stances toward a rumor. We also looked at the top 10 features for each rumor-specific model. We found out that there are differences in the top features for the classifier trained on different individual rumors, indicating different contexts for rumors within the same event. However, negation words and part-of-speech features remain significant for classifying rumor stances towards each rumor.

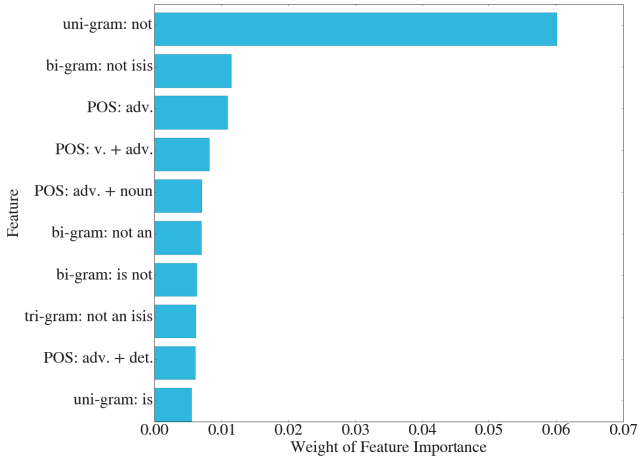

Figure 1: Feature importance for the trained random forest model using pooled data and proportional strategy

Discussion

This study was motivated by the recognition that manual coding of social media content can be very expensive, potentially prohibiting researchers from examining complete data about a particular social phenomena. As we build a large corpus of diverse rumors across multiple events, the trained classifier will improve in its generalizability and be extremely valuable for scaling analysis in future studies.

Building models to learn rumor stance from text allows for exploratory work that highlights particular features of expression online that might be indicative of the authors' stance towards the story discussed within. In the case presented here, we find that negation words and phrases such as *not* are strong indicators of rumor denying tweets. Noting such features could inform tools and systems designed to flag potential misinformation.

The work presented herein provides a foundation for future studies of rumoring behavior on social media during crisis events. In particular, we hope researchers will be able to use these methods to augment manually coded data with machine labeled data to curate large-scale datasets of rumoring online.

Acknowledgements

This work is supported by, or in part by, the National Science Foundation under grants 1420255 and 1342252.

Rumor	Baseline	Training Sample	Model	Precision	Recall	Accuracy	F-Score
Flag	0.5789	Proportional	Logistic Reg.	0.875	0.895	0.836	0.871
			Naïve Bayes	0.890	0.872	0.843	0.881
			Random Forest	0.870	0.962	**0.879**	**0.913**
Hadley	0.9538	Proportional	Logistic Reg.	0.955	0.995	0.951	0.975
			Naïve Bayes	0.955	0.999	0.955	0.977
			Random Forest	0.958	1.0	**0.958**	**0.979**
Suicide	0.9003	Proportional	Logistic Reg.	0.906	1.0	0.907	0.950
			Naïve Bayes	0.911	0.981	0.897	0.945
			Random Forest	0.911	0.999	**0.911**	**0.953**
Airspace	0.6411	Proportional	Logistic Reg.	0.890	0.898	0.860	0.890
			Naïve Bayes	0.890	0.889	0.858	0.889
			Random Forest	0.874	0.967	**0.889**	**0.918**
Lakemba	0.5224	Proportional	Logistic Reg.	0.861	0.931	0.889	0.890
			Naïve Bayes	0.892	0.808	0.860	0.845
			Random Forest	0.858	0.939	**0.893**	**0.894**
Pooled	0.50	Even	Logistic Reg.	0.818	0.885	0.839	0.845
			Naïve Bayes	0.872	0.775	0.830	0.820
			Random Forest	0.851	0.870	**0.860**	**0.860**
Pooled	0.6642	Proportional	Logistic Reg.	0.856	0.950	0.857	0.898
			Naïve Bayes	0.898	0.875	0.851	0.886
			Random Forest	0.871	0.969	**0.884**	**0.917**

Table 2: Results of classification models rumor stance. Models for each rumor and pooled data are shown, along with model accuracy and F-scores.

		Predicted	
		Affirm	Deny
Observed	Affirm	563.4	17.6
	Deny	83.2	210.6

Table 3: Confusion matrix for the trained random forest model using the pooled data and proportional sampling strategy (average value for 10 runs).

References

Arif, A.; Shanahan, K.; Chou, F.-J.; Dosouto, Y.; Starbird, K.; and Spiro., E. S. 2016. How information snowballs: Exploring the role of exposure in online rumor propagation. In *In Proceedings of the ACM 2016 Computer Supported Cooperative Work (CSCW 2016)*. ACM.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378.

Hiltz, S. R.; Kushma, J.; and Plotnick, L. 2014. Use of social media by us public sector emergency managers: Barriers and wish lists. *In Proceedings of International Conference on Information Systems for Crisis Response and Management*.

Mendoza, M.; Poblete, B.; and Castillo, C. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the 1st Workshop on Social Media Analytics*, 71–79. ACM.

Owoputi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1589–1599. Association for Computational Linguistics.

Shibutani, T. 1966. *Improvised news*. Ardent Media.

Spiro, E. S.; Fitzhugh, S.; Sutton, J.; Pierski, N.; Greczek, M.; and Butts, C. T. 2012. Rumoring during extreme events: A case study of deepwater horizon 2010. In *Proceedings of the 4th Annual Web Science Conference*, 275–283. ACM.

Starbird, K.; Maddock, J.; Orand, M.; Achterman, P.; and Mason, R. M. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon mombing. In *Proceedings of the iConference*.

Sutton, J.; Spiro, E. S.; Johnson, B.; Fitzhugh, S.; Gibson, B.; and Butts, C. T. 2014. Warning tweets: serial transmission of messages during the warning phase of a disaster event. *Information, Communication & Society* 17(6):765–787.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24–54.

Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1079–1088. ACM.

Zeng, L.; Starbird, K.; and Spiro, E. S. 2016. Rumors at the speed of light? modeling the rate of rumor transmission during crisis. In *Proceedings of the Hawaii International Conference on Systems Science (HICSS)*.