The Workshops of the Tenth International AAAI Conference on Web and Social Media
Social Media in the Newroom: Technical Report WS-16-19

An Analysis of Event-Agnostic Features for Rumour Classification in Twitter

Laura Toloşi, Andrey Tagarev, Georgi Georgiev
Ontotext AD
Tsarigradsko Shosse 47A, Sofia, Bulgaria

Abstract

Recently, much attention has been given to models for identifying rumors in social media. Features that are helpful for automatic inference of credibility, veracity, reliability of information have been described. The ultimate goal is to train classification models that are able to recognize future high-impact rumors as early as possible, before the event unfolds. The generalization power of the models is greatly hindered by the domain-dependent distributions of the features, an issue insufficiently discussed. Here we study a large dataset consisting of rumor and non-rumor tweets commenting on nine breakingnews stories taking place in different locations of the world. We found that the distribution of most features are specific to the event and that this bias naturally affects the performance of the model. The analysis of the domain-specific feature distributions is insightful and hints to the distinct characteristics of the underlying social network for different countries, social groups, cultures and others.

Introduction

Social Media allows cheap and fast access to information and empowers the regular end-user to create and propagate news. The quick spread of unverified information is a consequence of this decentralized model. Twitter has become notorious for the speed of reactions and unreliability of information in emergency situations. Undoubtedly there's great value in the first hand, eye-witness reports by Twitter users on breakingnews events, but it is often hard to tell them apart from all misinformation and disinformation that comes along.

Automated rumor detection systems for microblogs have been proposed, to identify emerging stories, predict their potential of becoming viral, their veracity and credibility and assess people's support or denial. Most of the methods are based on supervised classifiers that rely on features like user profile, message characteristics, propagation patterns and topic-related features.

In this work we discuss the challenges of modeling an early rumor detector, applicable before a breakingnews story unfolds and thus not enough context on the underlying topic is available. Such model would assist journalists to routinely scan through a large body of Twitter content by sorting of the tweets by their rumor likelihood. The key challenge is the absence of topic-related features, which are usually of high predictive value. Expecting a low performance as compared to the models that do make use of topic features, we investigate the ability of the model to generalize to new, unseen topics. By using a dataset consisting of tweets about nine different breakingnews events, we show that even features expected to be topic-agnostic exhibit domain-specific distributions. We contend that for events that take place in different locations of the globe, in different countries from different continents, that raise the interest of particular audiences, probably the characteristics of the Twitter network of users that comment on those events are substantially different. Hence, due to domain-specific biases of the training set, real-world applications are expected to yield lower performance than that reported in most studies. The demographic differences of Twitter groups reacting (posting) to various high-impact events deserve further interdisciplinary investigation. In this paper, we do not propose a solution for dealing with the biases, but merely bring to attention an issue that is often ignored in papers.

Related work

(Qazvinian et al. 2011) define rumors as *'a statement whose true value is unverifiable'* and (Zubiaga et al. 2015b; 2015a) as *'a circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety.'*.

(Mendoza, Poblete, and Castillo 2010) presented an early study of Twitter rumor propagation, veracity and patterns of supporting and denying replies, tailored to one event: the great earthquake from 2010 in Chile. (Yang et al. 2012) propose a rumor detection model for Sina Weibo, the Chinese analogue of Twitter. Most features are also found in the previous studies by (Mendoza, Poblete, and Castillo 2010; Qazvinian et al. 2011), to which specific features available in Sina Weibo are added. This work is specifically focused on quantifying their added benefit and does not explain how are the 10 cross-validation folds created, whether they separate the topics or mix them.

(Castillo, Mendoza, and Poblete 2011) present a classifier for Twitter information credibility. The authors select a large number of annotated tweets and group them into topics. The model is a decision tree that uses message-based features (length, punctuation, presence of negative/positive words), user-based features (age, number of followers, number of followees, number of tweets authored), topic-based features (fraction of the tweets that contain URLs, fraction of the tweets that contain hashtags, sentiment aggregates) and propagation-based features (depth of the retweet tree, number of initial tweets in a topic). The performance of the model is 86%, but the model relies on topic-based features.

(Qazvinian et al. 2011) investigate tweets from five topics and propose a classifier for user belief: support or deny of a particular rumor story. The features are content-based (unigrams, bigrams, part of speech), network-based (tweeting user and retweeting user) and Twitter specific memes (hashtags and urls). The authors remark that the performance of the classifier on new stories is poor and observe that the accuracy is improving substantially with each added tweet from the domain, from 60% without any training example from the domain, to up to around 90%.

(Liu et al. 2015) propose an early rumor debunking algorithm, that predicts veracity of events. Tweets are clustered into stories, features are defined, model is trained and predictions are carried out at story-level. The authors propose new features inspired by journalistic-verification instruments and train models on past topics, then test on other stories by adding increasing number of tweets form the event. The performance improves remarkably quickly after the first 100 tweets, showing that topic-dependent features carry significant predictive value.

Contribution and outline of the paper

Our goal was to train and evaluate an event-agnostic model for rumor classification. Having available a rich annotated set of tweets covering several distinct breakingnews events, we followed the steps of the state-of-the art rumor prediction models, limiting to features that are potentially event-independent. To our surprise, we found event-related biases in almost every group of features. We evaluate the models by a leave-one-topic-out cross-validation, a procedure that we believe is absolutely necessary for a realistic quantification of model performance in practice. In this paper, we describe the biases in detail.

The dataset, the features and the evaluation method are described in Section *Experimental setup*. Each feature group is discussed separately in the Section *Feature analysis*. The model trained on all features is described in Section *Classification tree model*. We comment on the results in Section *Conclusions*.

Experimental setup

Data

The data we are analyzing (Zubiaga et al. 2016) consists of tweets commenting on several breaking news stories and hoaxes from the recent past: a) *Putin missing* (in March 2015, due to a temporary absence from the media, social networks spread rumors that the Russian president Vladimir Putin might have been sick or dead); b) *Ferguson unrest* (in August 2014, an African-American young man was fatally shot by a white police officer in Ferguson, Missouri, provoking racist comments in the social media); c) *Prince concert in Toronto* (a false announcement that Prince would play in Toronto was spread in November 2014); d) *Gurlitt collection* (social media reacted to the news that a Swiss museum is about to accept works of art of a Nazi-era art collector); e) *Essien has ebola* (in the midst of an Ebola outburst in Africa, in October 2014 a hoax claiming that football player Michael Essien has contracted the virus was spread in the social media); f) *Ottawa shooting* (a Canadian soldier was killed in a shooting in front of the Parliament Hill in Ottawa in October 2014); g) *Germanwings crash* (in March 2015, the Germanwings Flight 9525 to Düsseldorf, Germany was hijacked and crashed in the French alps by the copilot, an act that was attributed later to mental illness but gave rise to many speculations); h) *Sidney siege* (a crisis that ended with two victims arose when a gunman took hostages in a cafeé in Sydney in December 2014); i) *Charlie Hebdo shooting* (in January 2015, two Islamist gunmen forced their way into the headquarters of Charlie Hebdo and killed 12 staff members).

Table 1: Counts of rumors/non-rumors by topic.

topic	**No. non-rumors**	**No. rumors**
charliehebdo	1695	474
ebola_essien	0	18
ferguson	892	291
germanwings_crash	690	332
gurlitt	196	190
ottawashooting	426	475
prince_toronto	4	237
putinmissing	123	143
sydneysiege	786	535

The procedure for data harvesting and rumor annotation is described in details in (Zubiaga et al. 2015a). Tweet threads consisting of a source tweet and its replies are annotated as either *rumor* or *non-rumor*, based on the definition given in the introduction of this article.

Table 1 shows statistics on the number of rumors and non rumors belonging to different topics. The non-rumor class is about twice as large as the rumor class.

Features

For an early rumor classifier we assume that topic-dependent features, as well as features like retweets, replies, propagation-based features are not available. We considered the following groups of features as likely to be topic-agnostic (most are discussed also in (Ma et al. 2015; Castillo, Mendoza, and Poblete 2011)) :

User-profile features. The number of Twitter *followers* of the user, the number of *followees* (or friends, other users that follow the activity of the user) and the number of *statuses* (tweets posted).

User id. The the unique identifier of a Twitter user.

Evidence. The list of *sources cited* to support the statement of the user, broken down to website domains, eg. *bbc.com*, *youtube.com*. None, one or several sources may be references in a tweet.

Table 2: Percentage of tweets in the left-out topic, the users of which have posted outside the topic, too.

topic	Percentage tweets
putinmissing	7.9%
ferguson	14.8%
gurlitt	12.4%
ottawashooting	44.7%
germanwings_crash	44.3%
sydneysiege	51.6%
charliehebdo	49.1%

Text style. A set of features describing the text style are the presence of certain *punctuation marks*, the presence of *capitalized words* and the *length of the message*.

Validation method

Due to rumor/non-rumor class imbalance in our dataset, we report AUC (Fawcett 2006; Spackman 1989; Sing et al. 2005) (area under the ROC curve) and not accuracy. We also report the F-measure when necessary.

The most realistic validation of a topic-agnostic classifier is to leave one topic aside for testing and to train on the remaining topics, in a cross-validation fashion. We call this *leave-one-topic-out validation*(LOTO). Unlike usual cross validation, the topics (i.e. folds) are unequal in size, therefore we report a weighted average of the AUCs with weights proportional to the sizes of the folds.

Finally, topics *ebola_essien* and *prince_toronto* were used only for training but not for testing, because they consist mostly of non-rumors (see Table 1) and ROC curves need samples from two classes to be evaluated.

Feature analysis

User id

The *user id* can be a good predictor of rumor if previous activity from the user has been recorded. News agencies, famous people and very active users are likely to have posted messages on previous rumor topics and we can estimate their probability for spreading rumors. Clearly, only a small fraction of the users in the test set are expected to appear in the train set, so for those instances we assign a NA value.

For a particular user, we estimate the rumor probability based on a train set and test on a test set, according to the LOTO procedure. Table 2 shows the percentage of the tweets in the left-out topic that have users that had posted in the train topics as well. *chaliehebdo* and *sydneysiege* stand out with half of the tweets being posted by users that have been active in other topics, too. Figure 1 shows pair-wise topic overlap computed as the fraction of *unique* common users. Our findings show that a large percentage of the users from *sydneysiege*, *germanwings_crash*, *ottawashooting* and *ferguson* are also active in *charliehebdo*. Table 2 shows that *putinmissing* has very few users active in other topics.

The probability of rumor given the user id is estimated using the relative frequency as:

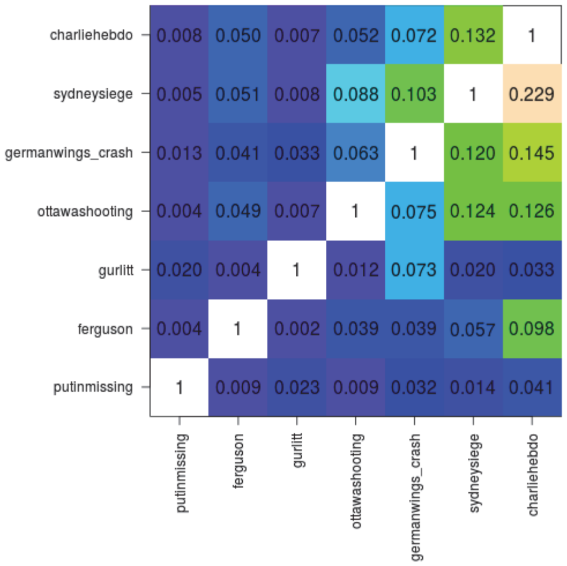


Figure 1: Pairwise similarity between topics w.r.t. users. Each cell represents the number of common users normalized by the total number of unique users of the row topic. Eg. 22.9% of the *sydneysiege* users are also active in *charliehebdo*.

$$P(\text{rumor}|\text{user id}) = \frac{\#\text{rumors of user = user id} + 1}{\#\text{tweets of user = user id} + 2},$$

where we included a Laplace correction (Lewis and Sauro 2006) for avoiding unlikely values of 0 and 1.

The LOTO validation shows better than random prediction AUC on all left-out topics, apart from the *putinmissing* (Figure 2). Specifically, *putinmissing* 19%, *ferguson* 59%, *gurlitt* 50%, *ottawashooting* 58%, *germanwings_crash* 53%, *sydneysiege* 66%, *charliehebdo* 64%. The weighted AUC is 59%.

User profile

We trained logistic regression models using the log-transform of number of *followers*, *followees* and *statuses*, since they follow heavy-tailed exponential distributions. Figure 3 shows the ROC curves by left-out-topic. The weighted average AUC is 62% but the variance among topics is large, about 7%.

In our investigation, we discovered a poor calibration of the predictions among topics, meaning that the probabilistic output of the models is not comparable between folds, in the sense described in (Forman and Scholz 2010) (results not shown in this paper). Part of this is because of the class imbalance, but also due to the large differences in marginal distributions of the features among topics. Figure 4 shows that the mean *log(no.followees)* varies greatly from lowest in *gurlitt*, *prince_toronto* and *putinmissing* and highest in *ebola_essien* and *germanwings_crash*. Also, the standard deviation of the distributions is high, of about 2.5. The number of followers is more consistent among topics, with lower

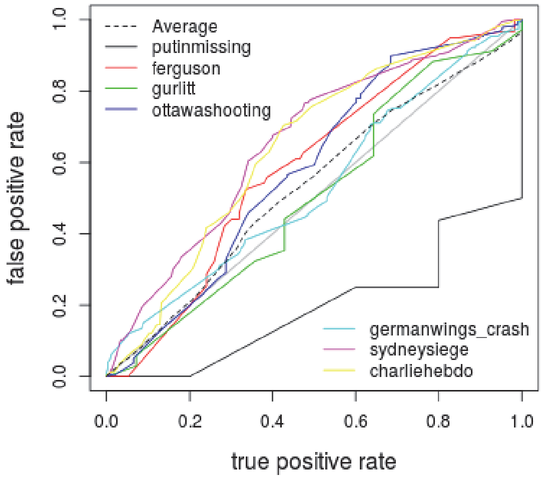


Figure 2: LOTO performance based on user ID only.

variance (1.7), with a notable outlier being *ebola_essien*, with users having a lower average number of followers. The number of statuses looks either bimodal or unimodal, with varying mean and a standard distribution of 1.75. These differences are biases that lead to poor prediction on some of the topics.

Evidence

We extract domain URLs cited in tweets and estimate the probability of a certain domain to be cited in a rumor, based on count statistics over the train data. We assume that this probability is somewhat correlated with the reliability of the sources. Figure 5a) shows the count of rumors versus count of nonrumors (logscale) for all URLs in our dataset. Observations which fall far away from the diagonal are sources referred mostly by rumors or nonrumors, respectively. *washingtonpost*, *independent.co.uk*, *telegraph.co.uk*, *theguardian.com* are from the notably nonrumorous sources. At the other end, *breakingnews.com*, *thegatewaypundit.com*, *blick.ch*, *globalnews.ca* are referred mostly in rumors.

For LOTO prediction we estimate rumor probabilities for each tweet, based on the list of one or more URLs referred. We use the following aggregation score:

$$P(\text{tweet is rumor}|\text{url}_1, \text{url}_2, ..., \text{url}_k) =$$
$$= \frac{\sum_{i=1}^{k} \text{No. of rumors containing url}_i \text{ in train set} + 1}{\sum_{i=1}^{k} \text{Number of tweets containing url}_i \text{in train set} + 2}$$

We have looked at the pairwise overlap of topics w.r.t the sources cited. Figure 6 shows that e.g. 45% of the sources cited in *sydneysiege* are also cited in *charliehebdo*. Conversely, 23% of the sources in *charliehebdo* are also cited in *sydneysiege*. Note that very few of the sources cited in *gurlitt* are also cited elsewhere (at most 15% overlap, with *germanwings_crash*).

The leave-one-topic-out prediction procedure for prediction performance returns the following AUCs (see also Figure 7 for ROC curves): *putinmissing* 70%, *ferguson* 36%, *gurlitt* 51%, *ottawashooting* 56%, *germanwings_crash* 45%, *sydneysiege* 56%, *charliehebdo* 53%. The weighted average AUC is 50%.

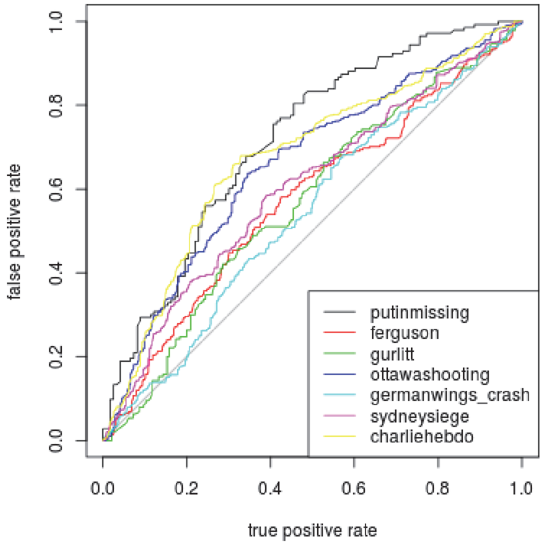


Figure 3: ROC curves for each left-out topic for logistic regression.

Emergent and News Trust We compared the rumor probabilities of URL domains computed based on our data to a veracity score from emergent.info and a credibility score form newstrust.net. We wish to know if there is consistency among the various reliability estimates for content providers.

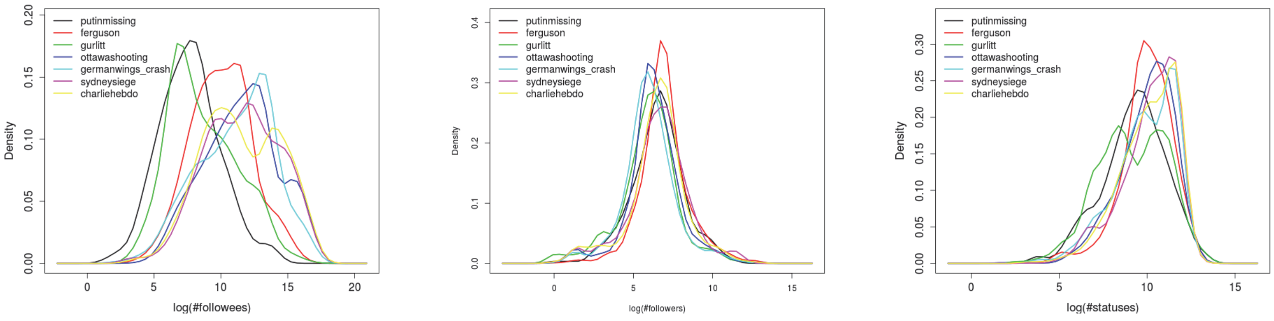
Emergent [1] is a real-time rumor tracking website that identifies and debunks rumors. With simple counting, one can estimate how often a source is *denying a fake story* or *supporting a true story* (source gets one positive mark) and how often a source is *supporting a fake story* or *denying a true story* (source gets a negative mark). We computed the smoothed frequency of negative marks and use it as a score for veracity.

NewsTrust[2] presents crowd-reported credibility of news sites. We collected scores for newspapers (http://newstrust.net/sources/list?medium=newspaper), magazines, blogs and online sources. We normalized the score to fall between 0 and 1, with higher values for less credible sources.

There's a significant number of common news sources between our tweets and Emergent (97) (Figure 8a)), compared to only 28 between out tweets dataset and NewsTrust (Figure 8b)). The comparison shows more consistency however between the rumor scores (based on the tweets) and the NewsTrust credibility scores; there is little correlation between the rumor scores and the veracity scores

[1] http://www.emergent.info

[2] www.newstrust.net



(a)

(b)

(c)

Figure 4: Marginal distributions by topic: a) No.followees; b) No.followers; c) No.statuses.

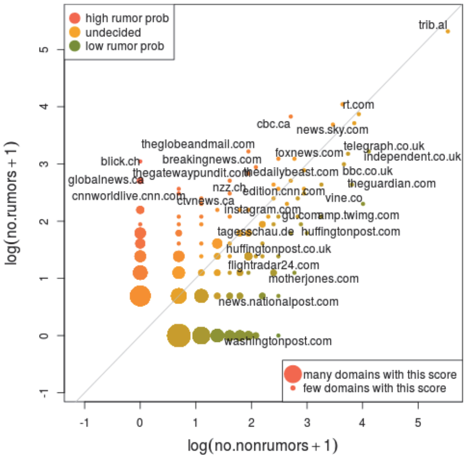


Figure 5: Rumor/non-rumor distributions by cited source.

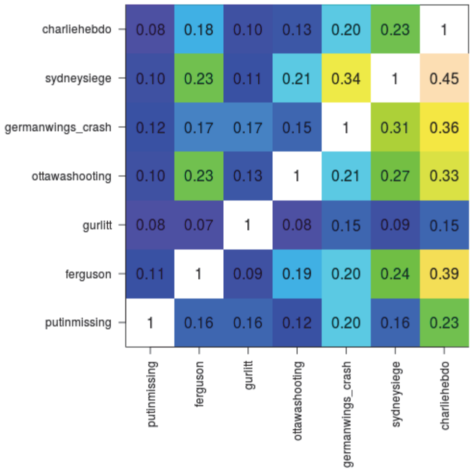


Figure 6: Pair-wise similarity of topics with respect to the sources cited.

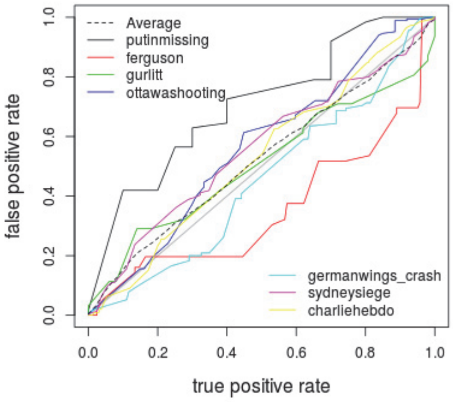


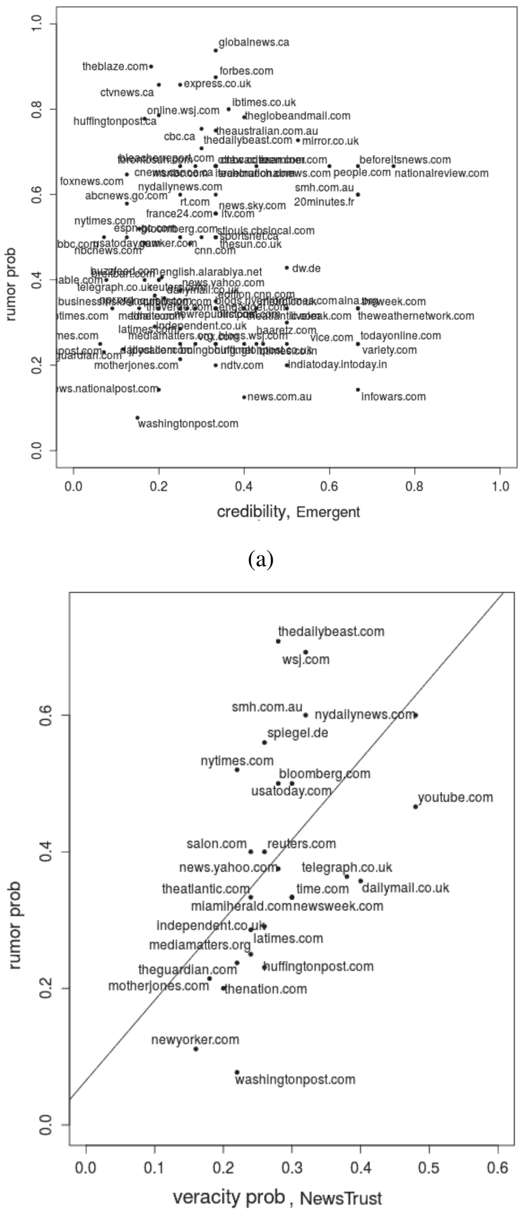
Figure 7: Performance based on sources cited, by left-out topic.

from Emergent. washingtonpost.com, newyorker.com, theguardian.com, motherjones.com are among the sources with low rumor probability and high credibility, whereas youtube.com, mydailynews.com, have higher rumor probability and low credibility.

Both Emergent and NewsTrust are valuable resources for rumor tracking. In practice however, their use is limited to the most popular (internationally) content providers, with bias towards news sites. Table 3 shows that the referred cited by the Twitter users are 57% .com domains, whereas .com domains represent 82% of the Emergent sources and 76% of the NewsTrust domains. Prior information is missing for .de, .fr, .ch domains, often cited in the events that took place in Europe for example (*germanwings_crash*, *charliehebdo*, etc).

Text style

Tweet length. We found no significant difference between the length of the message of rumors vs. non-rumors. We used t-test and Wilcoxon rank sum test and both showed large p-values. There are interesting differences in the length of tweets among topics (Figure 9a). The topics *gurlitt* and *prince_toronto* are significantly shorter than the rest, whereas the topic *charliehebdo* has longest tweets.



(a)

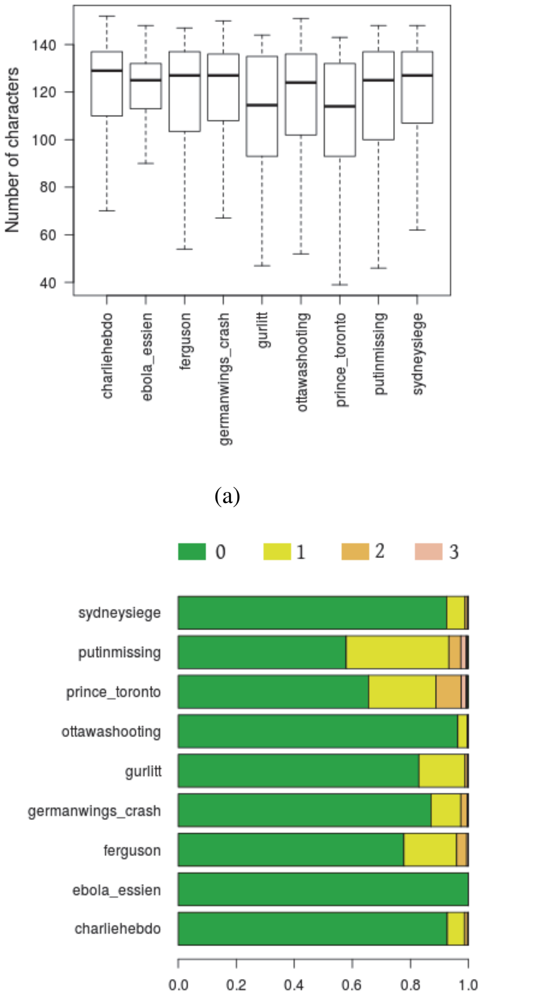
(b)

Figure 8: Comparison of rumor scores from tweets data and a) veracity scores from Emergent and b) credibility scores from NewsTrust.

Capitalization. At least one capitalized word was used in 28% of the tweets, almost 10% had more than 5 capitalized words. The most frequent capitalized words excluding stopwords were, in decreasing order of frequency: BREAKING, UPDATE, AFP, ISIS, NEWS, LIVE, US, PM, RIP, MORE, VIDEO, JUST, SYDNEY, SWAT, TV, DEVELOPING, CNN, RCMP, WATCH, CBD, PHOTO, STL, BBC, UK, AP, CBC, CEO, NHL, NOW , NSW, SHOOTING, CTV, OTTAWA, SYDNEYSIEGE, USA, EIL, PHOTOS, POLICE, STORY, ABC, CONVERT, FBI, MUSLIM, UPDATED, URGENT, UTC, ALERT, ATTACK, HOSTAGE, PK, TORONTO UPDATES, CAFE, GERMAN, LATEST, PEOPLE, PRINCE, URGENT, ARD, BILD, DETAILS,

Table 3: Domains of the sources cited.

domain	rumor data	Emergent	NewsTrust
.com	342	336	58
.de	47	2	1
.org	44	14	12
.ca	28	11	0
.ch	27	0	0
.co.uk	20	23	4
.fr	17	1	0
.net	16	6	0
.au	16	4	1
.in	10	6	0
.at	5	0	0
others	29	6	0



(a)

(b)

Figure 9: a) Lengths of tweets by topic. b) Punctuation by topic.

FRANCE, NEW, OMG, SIEGE, TERRORIST, TONIGHT, WILSON, WTF.

We selected a subset that we believe are topic-independent. For example, we removed 'OTTAWA', 'ISIS', 'FBI'. We chose the following subset of terms all occur in at least half of the topics: BREAKING, JUST, MORE, PHOTO, VIDEO, NEWS, UPDATE, DEVELOPING, LIVE, WATCH, NOW, DETAILS, LATEST, OMG, UPDATED, STORY and we investigated which of them have significantly different occurrence ratios in the rumor vs non-rumor classes. Fisher tests with multiple testing correction show that only BREAKING(p-value $9e{-}75$), MORE(p-value $1.4e{-}3$), NEWS(p-value $3.7e{-}12$), UPDATE(p-value $4.2e{-}15$), DEVELOPING(p-value $8.5e{-}4$) and STORY(p-value $7.3e-3$). However, the total number of occurrences is low for some of these words, of only 10 times.

The presence or absence of capitalization itself is also predictive of rumor, 40% of the rumors having at least one capitalized word, compared to 22% from the non-rumors.

Punctuation. The presence of '?', '!' or '...' is more often associated with non-rumors in our dataset, although the difference is not very large: 37% of the non-rumors contain at least one of the marks, as opposed to 32% of the rumors. The differences among topics are illustrated in Figure 9b.*putinmissing* and *prince_toronto* have highest proportion of tweets that contain at least one punctuation mark of the three considered.

Classification tree model

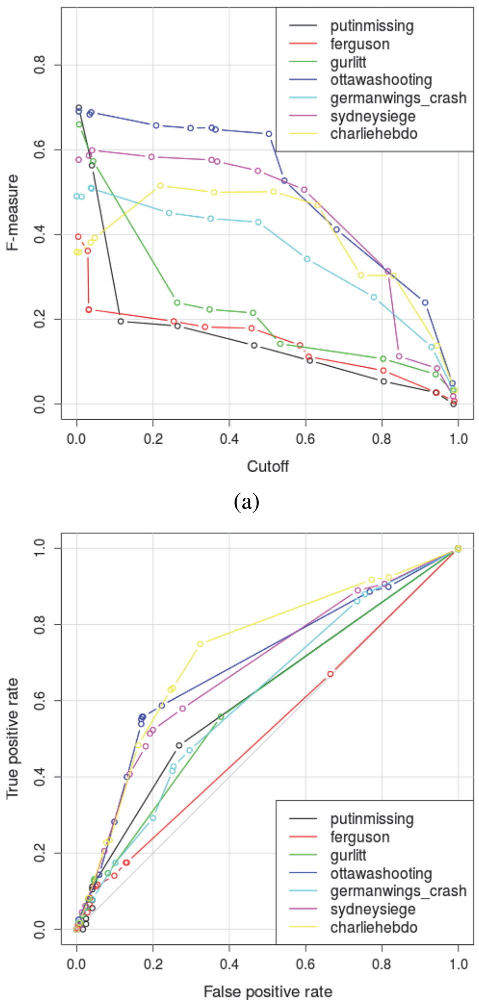
We trained a classification tree (Hothorn et al. 2006; Hothorn, Hornik, and Zeileis 2006) using the four groups of features described. Figure 10 shows the F-measure for each of the topics in the LOTO validation setting. The F measure is plotted against the score cutoff which is the response of the classification tree. Note the high variation of performance over the topics.

The weighted averaged AUC is 65% (see Figure 10b). The F-measure at different score cutoffs is shown in Figure 10a. Different cutoffs are optimal for the various topics, but in general very small values of around 0.1 result in better performance.

Conclusions

We analyzed the challenges of a topic-agnostic rumor classifier, which can be used to monitor the Twitter stream and flag the most likely tweets to become rumors. We used user id, user profile, text style and URL domains referred for evidence as a basis for prediction. The performance of the classifier is low as expected, around 65% F1 measure, but it is a realistic estimate for most application scenarios, where journalists do not have the resources to annotate tweets in real-time, as events unfold.

Most importantly, our analysis revealed the existence of biases among topics that, to our knowledge, haven't been discussed previously. Demographic differences of users that disseminate various topics can be caused by the location of the event, the public interested in the event and other complex factors. The users that take interest in speculating about the alleged death of Vladimir Putin and the users that comment on the shooting at the Charlie Hebdo office are likely different statistical populations of the Twitter network. These topic-specific characteristics deserve interdisciplinary investigation by sociologists, psychologists, statisticians. We believe it is critical that future rumor detection models should be tested on datasets that cover various stories, to ensure realistic evaluation.



(a)

(b)

Figure 10: a) F-measure. b) AUCs.

Acknowledgment

This work has been funded by the Pheme project (EU 7^{th} Framework Programme for research, technological development and demonstration under grant agreement no 611233).

References

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th*

International Conference on World Wide Web, WWW '11, 675–684. ACM.

Fawcett, T. 2006. An introduction to roc analysis. *Pattern Recogn. Lett.* 27(8):861–874.

Forman, G., and Scholz, M. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.* 12(1):49–57.

Hothorn, T.; Hornik, K.; van de Wiel, M.; and Zeileis, A. 2006. A lego system for conditional inference. *The American Statistician* 60(3):257–263.

Hothorn, T.; Hornik, K.; and Zeileis, A. 2006. Unbiased recursive partitioning: A conditional inference framework. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS* 15(3):651–674.

Lewis, J. R., and Sauro, J. 2006. When 100% really isn't 100%: Improving the accuracy of small-sample estimates of completion rates. *J. Usability Studies* 1(3):136–150.

Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; and Shah, S. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 1867–1870. New York, NY, USA: ACM.

Ma, J.; Gao, W.; Wei, Z.; Lu, Y.; and Wong, K.-F. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 1751–1754. New York, NY, USA: ACM.

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.

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*, EMNLP '11, 1589–1599. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sing, T.; Sander, O.; Beerenwinkel, N.; and Lengauer, T. 2005. Rocr: visualizing classifier performance in r. *Bioinformatics* 21(20):7881.

Spackman, K. A. 1989. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, 160–163.

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*, MDS '12, 13:1–13:7. New York, NY, USA: ACM.

Zubiaga, A.; Liakata, M.; Procter, R.; Bontcheva, K.; and Tolmie, P. 2015a. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, 347–353. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Zubiaga, A.; Liakata, M.; Procter, R.; Bontcheva, K.; and Tolmie, P. 2015b. Towards detecting rumours in social media. In *AAAI Workshop on AI for Cities*.

Zubiaga, A.; Liakata, M.; Wong Sak Hoi, G.; Procter, R.; and Tolmie, P. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS ONE* 11(3).