Detecting Rumours in Disasters: An Imbalanced Learning Approach

Amir Ebrahimi Fard, Majid Mohammadi, and Bartel van de Walle

Delft University of Technology
A.EbrahimiFard@tudelft.nl, M.Mohammadi@tudelft.nl and
B.A.vandeWalle@tudelft.nl

Abstract. The online spread of rumours in disasters can create panic and anxiety and disrupt crisis operations. Hence, it is crucial to take measure against such a distressing phenomenon since it can turn into a crisis by itself. In this work, the automatic rumour detection in natural disasters is addressed from an imbalanced learning perspective due to the rumour dearth versus non-rumour abundance in social networks.
We first provide two datasets by collecting and annotating tweets regarding the Hurricane Florence and Kerala flood. We then capture the properties of rumours and non-rumours in those disasters using 83 theory-based and early-available features, 47 of which are proposed for the first time. The proposed features show a high discrimination power that help us distinguish rumours from non-rumours more reliably. Next, We build the rumour identification models using imbalanced learning to address the scarcity of rumours compared to non-rumour. Additionally, to replicate the rumour detection in the real-world situation, we practice cross-incident learning by training the classifier with the samples of one incident and test it with the other one. In the end we measure the impact of imbalanced learning using Bayesian Wilcoxon Signed-rank test and observe a significant improvement in the classifiers performance.

Keywords: Rumour detection, imbalanced learning, building dataset, feature engineering, Twitter.

1 Introduction

Rumours are unverified information circulating about the topics that people perceive important and are used as sense-making or risk management mechanism [4]. Rumours tend to thrive in situations that are ambiguous and/or pose a threat in which meanings are uncertain; questions are unsettled, information is missing, and/or lines of communications are absent. One of the contexts that satisfies all those conditions is the crisis. Due to the lack of information and mistrust towards the available sources of formal channels at the time of disaster being happening, people feel frustrated and seek information from informal channels. Eventually, if no information is available, people engage in affirmative rumouring, which means speculation based on whatever evidence and framework of understanding

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

they posses [4]. Traditionally, rumours were propagated by means of word of mouth. However, in recent years, the rapid growth of social networks escalated this problem and turned it into a serious issue by accelerating, widening, and deepening the circulation of misinformation among people [25,21]. Thus, it is of the essence to develop a solution for this problem, because it can otherwise turn into a potential threat to the main societal institutions such as peace and democracy.

One of the promising approaches to quell online rumours is artificial intelligence (AI) which can tackle misinformation at scale and across languages and time zones. Among the AI approaches, binary classification dominates the literature of rumour detection [33]. In this approach the classifiers are trained by historical samples form two classes of rumour and non-rumour. Recent studies have shown theoretically and empirically that compared to all different kinds of information circulating in social networks, rumours are in minority [5,6,15]. In other word, the fraction of non-rumours has the majority in the flow of information in social networks. From a data collection point of view, this leads to an imbalanced dataset containing uneven volume of rumour and non-rumour samples. This imbalance gives rise to the "class imbalance" problem or "curse of imbalanced dataset", which is the problem of learning a concept from a class with a small number of samples [16]. In machine learning, the problem of imbalanced data has been addressed by a learning paradigm called imbalanced learning.

In this study, we address the rumour dearth versus non-rumour abundance by imbalanced learning. As Figure 1 illustrates, we first collected more than 200,000 tweets regarding Kerala flood and Hurricane Florence. For the annotation, we used a large-scale labeling technique based on signal words of each incident.

We then extract 83 theory-based and early available features from every data point. Out of those features, 47 of which are proposed for the first time. Our feature selection method represents a high level of effectiveness for the newly proposed features. In the next step, we conduct a series of experiments using imbalanced learning. For the experiments, we use cross-topic learning in which a classifier is trained on one incident and is tested on the other one, instead of training and test on the same dataset. In the end, using Bayesian Wilcoxon Signed-rank test, we measure the effectiveness of imbalanced learning.

Given this, the main contributions of this article are summarised as follows:

1. Improving the models performance using imbalanced learning due to the imbalanced nature of rumour against non-rumour in social media.
2. Building two annotated datasets comprising more than 200,000 tweets regarding Kerala flood and the Hurricane Florence.
3. Proposing 47 features for computational rumour detection and evaluating them using five different learning methods.
4. Designing a set of cross-topic novel experiments which replicate the real-world situation in rumour detection through training with one dataset and testing with the other one.

This paper is organised as follows. Section 2, provides an overview on the related studies in computational rumour detection. Section 3 explains data col-

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

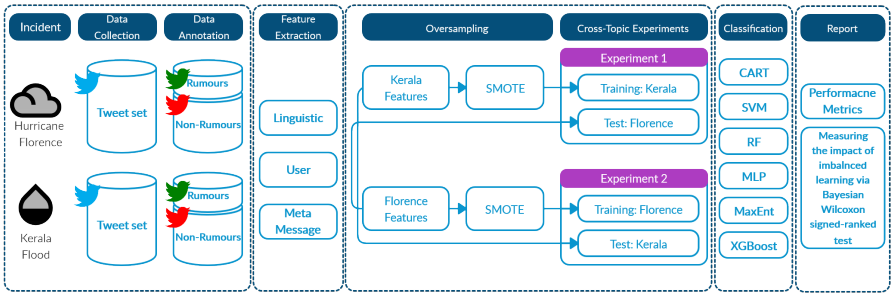


Fig. 1: Research flow of rumour detection with imbalanced learning.

lection, annotation, and feature extraction. In Section 4 we report and discuss the experiments results and features importance. Finally in Section 5 we conclude this study and give some future research directions.

2 Related Work

Computational rumour detection is a classification problem that aims to distinguish rumours from non-rumours. Similar to other classification problem, data collection, feature extraction, and model training constitute the pillars of a rumour detection model. In fact a dataset comprising rumours and non-rumours, a set of relevant features, and algorithm for the context of rumour spreading are prerequisite conditions for building a rumour detection model.

Since almost a decade ago that first articles in rumour detection with the computational approach were published, most of the studies meet this set of conditions and contribute in at least one of those areas. Several studies provided annotated datasets. They are mostly collected from Twitter [12,19,24,32] and Sina Weibo [29,17] as they provide a fairly easy API access. Some other studies, proposed new features according to specification of rumour propagation in predefined categories of content, user, network, and temporal features [12,2]. For instance due to the importance of rumour identification as early as possible, some of scholars proposed features to capture early appearance of rumours. For instance, Zhao et al. [31] propose a novel technique for the early rumour detection using signature text phrases. In another research, Wang et al. [26] show a feature pattern consisting of both user's attitude and information diffusion for the early detection of rumours in social networks. In the same vein, Kwon et al. [13] showed that some features are more informative in the early stages of rumour diffusion. Furthermore, some other studies proposed new algorithms for identification of rumours. For instance, Ma et al. [18] used the recurrent neural network (RNN) for the rumour detection, Lozano et al. [8] used the combination of convolutional neural networks (CNN) with both automatic rule mining and manually written rules, Chen et al. [3] utilized a CNN for short text categorization using multiple filter sizes, or Zubiaga et al. [34] model the rumour tweets

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

as sequences by extracting their features in the course of time and applying the conditional random field (CRF) for classification. New techniques do not necessarily mean designing a completely new algorithm, but it mostly means applying an algorithm which is developed before to the context of rumour detection. For example, LSTM (as an RNN technique) and CRF techniques were introduced in 1997 [10] and 2001 [14], respectively, but they were used later for rumour detection [34,18], and were truly counted as significant contributions.

The other area of importance is rumour detection in different subject domains. As we discussed earlier, any domain that falls into 3C's categories can be a hotbed for rumour emergence. There are several studies [2,12] that highlight the importance of domain in rumour detection, but they do not analyse the impact of domain in rumour detection [20]. To the best of our knowledge, there is one single research on computational rumour detection in a specific domain which is recently published by Sicilia et al [20]. They study the rumour detection in the health domain, in particular, the case of Zika virus. In this work, they take the context into account by collecting data and proposing novel features. However for the third element namely algorithm, they leave it without contribution by using classifiers from different learning paradigms.

3 Data and Features

For this study, we prepared two datasets regarding the Kerala flood and Hurricane Florence. The 2018 Kerala flood was the worst monsoon flooding in a century in Southern India with 400 fatalities and $2.7bn worth of damages. The Hurricane Florence was a category four hurricane hit Carolinas in the south-east of the United States. The hurricane caused more than 50 fatalities and up to $22bn damages. We set up a streaming API of Twitter to collect all tweets, retweets, replies, and mentions that included hashtags, text strings, and/or handles related to Kerala flood and Hurricane Florence. We could collect 100,880 tweets regarding Kerala flood and 101,844 tweets for Hurricane florence.

To select the rumour cases we searched several credible news outlets and fact-checking websites (e.g. Snopes, The Washington Post, The Hindu, and Indian Express). We identified in total three rumours in Kerala flood and four rumours in Hurricane Florence with a high level of consistency among different news outlets and fact-checking sources. Then we extracted the rumour-related tweets corresponding to these events if the tweet contains the keyword relevant to the rumour [13]. The tweets without explicit keywords were assigned a non-rumour label. After the data annotation the Kerala dataset consists of 2,000 rumour-related and 98,880 non-rumour-related tweets. Florence dataset also comprises 2,382 rumours and 99,462 non-rumours[1]. The imbalance between the number of rumours and non-rumours in an incident aligns with the previous findings [6,15]

For feature extraction, we use 83 features to represent rumour and non-rumour tweets. The features are either taken from the literature of computational

[1] The datasets are publicly available: https://bit.ly/2WxVhY0

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

rumour detection or introduced in this work[2]. Tweet features are classified into three groups: linguistic & content, user, and meta-message. The linguistic & content category contains all the features about syntactic and semantic aspects of tweets. The features related to the account holders and their social networks fall into the user category, and all the features about tweets metadata belong to the meta-message class. In this study, we want to detect rumours in the disasters as early as possible, we can thus rely solely on the features that are available during the initial phases of rumour diffusion. In this regard, we deliberately ignore propagation and temporal features, as they are not available in the early diffusion phases [13]. We also skip features related to likes, retweets, and comments since they have a high volatility and varying values in the initial stages of rumour diffusion. Table 1 demonstrates all the features in three categories of linguistic & content, user, and meta-message. In this table, dagger (†) and diamond ($\diamond$) symbols specify features inspired by social bot detection literature and proposed ones, respectively.

4 Experimental Results

In this section we first report the models performance and measure the impact of imbalanced learning in the identification of rumours, then we evaluate the discrimination power of proposed features.

4.1 Models performance

In this section, we evaluate the features weights, and report the results of our models and compare their performance in cross-incidents experiments subsequently. Then we measure the impact of imbalanced learning on rumour identification. We also conduct a baseline analysis by comparing the performance of our models with state of the art.

For feature evaluation, our goal is to assign a weight to each feature. This allows us to rank features ordinaly as well as knowing to what degree each feature is informative for classification algorithms. The conventional methods such as χ^2 test and recursive feature elimination are either time-consuming or not suitable for our datasets due to their size[3]. Therefore, we define a score using random forest, XGBoost, adaptive boosting, regression tree, and extremely randomised trees as classification algorithms with an embedded feature selection mechanism. By summing up the features weights we obtain the degree of significance for each feature. To find the significant features we determine a threshold for the features score. If the score of a feature in each dataset is less than the given threshold, it is assumed to be *insignificant*. To determine the threshold, we choose

[2] The newly introduced features are explained and their relevance are discussed in the first section of the supplementary materials (available at: https://bit.ly/2PJ3FmR)

[3] The further explanations regarding the issues of the conventional feature selection methods can be found in the second section of the supplementary materials (available at: https://bit.ly/2PJ3FmR)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

Table 1: List of all the features in three categories: linguistic & content, user and meta-message. New features are marked with ⋄. The features inspired by social bot detection literature that, to the best of our knowledge, have not been used in rumour detection literature yet are marked with †.

Feature class	**Features**
Linguistic & Content	Number of exclamation marks in a tweet [2,32] Number of question marks in a tweet [2,32] Number of characters in a tweet [2] Number of words in a tweet [2] Number of uppercase letters in a tweet [2,32] Number of lowercase letters in a tweet [2] Number of first person pronoun in a tweet [2] Number of second person pronoun in a tweet [2] Number of third person pronoun in a tweet [2] Number of capital words in a tweet [2] Average word complexity in a tweet [24] Number of vulgar words in a tweet [24] Number of abbreviations in a tweet [24] Number of emojis in a tweet [24] Polarity of a tweet [30] Subjectivity of a tweet [27] Tone of a tweet [30] Positive words score of a tweet [29] Negative words score of a tweet [29] † Frequency of Part of Speech (POS) tags in a tweet (19 features) [23] ⋄ Frequency of Name Entity Recognition (NER) tags in a tweet (17 features) Opinion and insight score [24] Anxiety score [22] Tentativeness score [24] ⋄ Certainty score Sentence complexity [24]
User	Profile description (binary) [2,29,30,28] Verified account (binary) [2,29,28,32] Number of Statuses [29,28,2,13] Influence [29,28,2,30,13] Number of following [29,28,30,13] User role [24] ⋄ Attention Account age (day) [2] ⋄ Openness (binary) Profile location (binary) [29] † Profile picture (binary) [23] Profile URL (binary) [30] ⋄ Average follow speed ⋄ Average being followed speed ⋄ Average like speed ⋄ Average tweet speed † Screen name length [23] † Number of digits in screen name [23]
Meta-message	Number of hashtags in a tweet [19,2] Number of mentions in a tweet [2] Tweet URL (Binary) [2] Number of multimedia in a tweet [29,28,30] ⋄ Location sharing (binary)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

a very small number of 0.001 in order to select features with minimum level of informativeness. The selection of this number as the significance threshold is inspired by the significance level in null hypothesis testing. Additionally, we call a feature *consistently significant* if its score is higher than the threshold in both datasets.

Figure 2 illustrates the feature significance in the Florence and Kerala datasets with blue and orange bars, respectively. In this figure, the length of each bar is proportionate to the importance of the corresponding feature. Figure 3 displays significant and consistently significant features based on their types and sources. In this figure, the abscissa represents the source or type, and the ordinate is the number of features in each category. Figure 3a illustrates the number of significant and consistently significant features as well as the total number of features for each feature category. According to this figure, 60% of linguistic & content features, 41% of user features, and 60% of meta-message features are significant. Among the significant features, 70% of linguistic & content features, 43% of user features, and 33% of meta-message features are considered as consistently significant. This means that linguistic & content, user, and meta-message features have the highest fraction of consistently significant features, respectively. Figure 3b shows the same measures for the proposed features and the ones which were extracted from the literature. Based on this figure, 70% of the literature features and 45% of the proposed features are significant. Out of the significant ones, 64% of literature and 62% of the proposed features are consistently significant.

In this study seven classifiers belonging to different learning paradigms were considered for the experiments [20]: multi-Layer perceptron (MLP) as a neural network, support vector machine (SVM) as a kernel machine, classification and regression trees (CART) as a decision tree, random forest (RF) as an ensemble of trees, XGBoost as a boosting approach, and maximum entropy (MaxEnt) as an exponential model. We tried to use entirely distinct algorithms in order to verify their effectiveness in the early rumour detection problem. We conduct four experiments to address two crucial challenges in computational rumour detection. The first challenge is the imbalanced nature of rumours to non-rumours [15,6]. This is what we are also observing in this study as the ratio of rumour to non-rumour in datasets is approximately 1:50. To tackle this issue, in addition to training the classifier with imbalanced dataset we use an oversampling technique to train the classifiers with the balanced form of the training sets. For oversampling, we use synthetic minority oversampling technique (SMOTE) as a powerful imbalanced learning technique that has shown a great deal of success in various applications. The SMOTE algorithm creates artificial data based on the feature space similarities between existing minority examples [9]. It is worth emphasising that we use this algorithm only for training data, in other words, the test dataset is still intact and preserves its imbalanced shape.

The second challenge is the common practice of the field namely training and testing on the same dataset. In a real rumour propagation case, there is almost no time for data collection, feature extraction, and model training; thus, such an approach for rumour detection cannot be operationalised. The other approach is

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

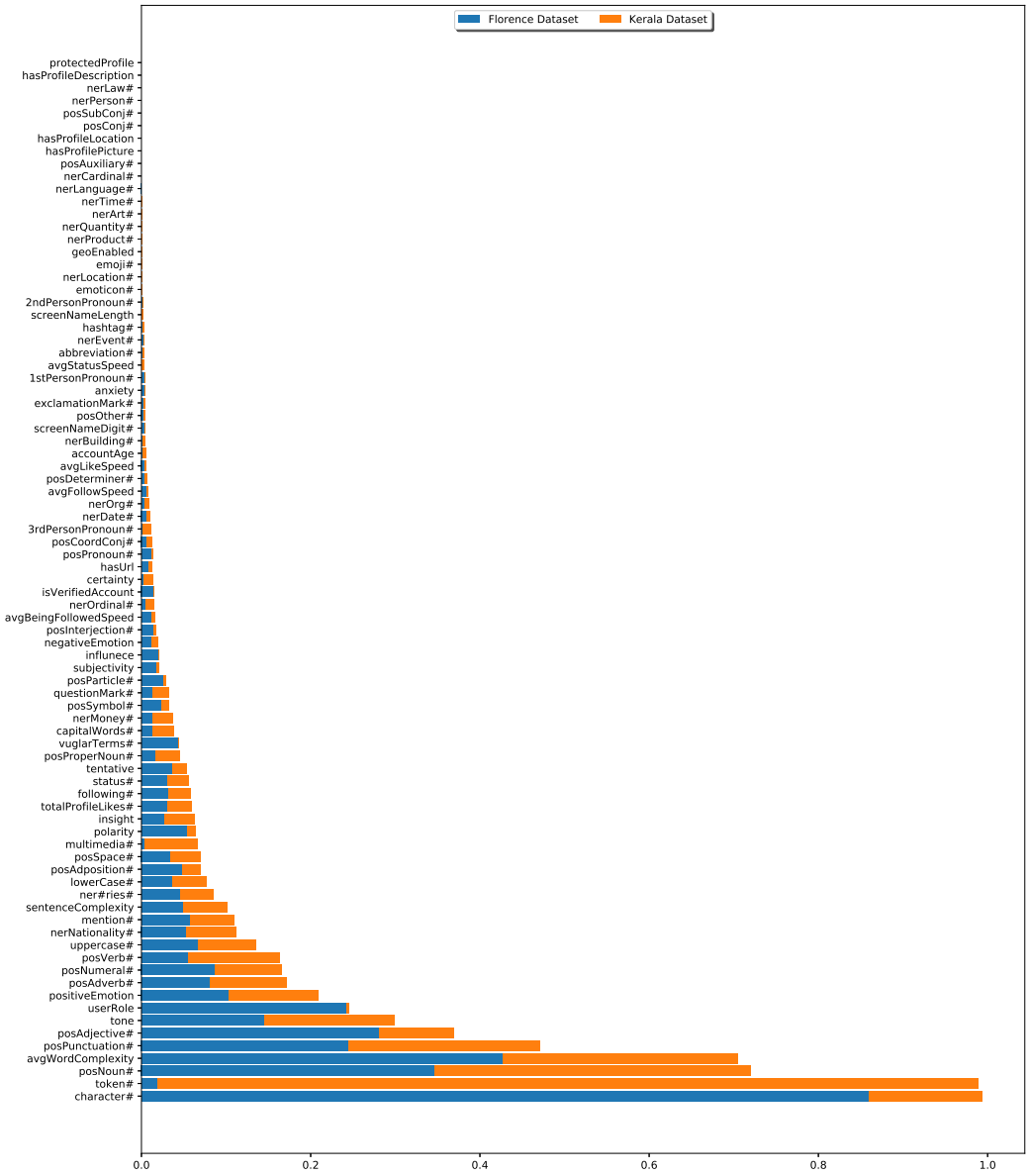
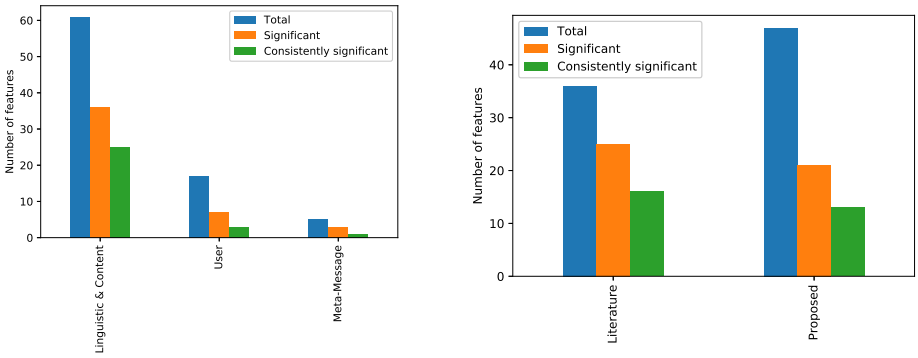


Fig. 2: Evaluation of the feature significance in the Florence and Kerala datasets. Blue and orange bars indicate the significance of the features in Florence and Kerala datasets, respectively. The longer a bar is, the more important the corresponding feature will be.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48



(a) Comparison of features based on their types.

(b) Comparison of features based on their source.

Fig. 3: Comparison of the feature significance. The blue, orange, and green bars denote the total, significant, and consistently significant number of features, respectively.

using historical rumours for identification of upcoming rumours. In this regard, we use one of our datasets for training and the other one for testing; then we switch the training and test set to assess the robustness of the proposed approach. In other words, once we use Kerala as training set and Florence as the test set (which is indicated by Kerala ⇒ Florence notation), then we switch the datasets and use Florence as training set and Kerala as test set (Florence ⇒ Kerala). To evaluate the classifiers performacne in different experimental settings we use precision-recall (PR) curve as it can provide an informative representation of performance assessment [9].

In the upper panel of Table 2, the performance of each classifier is demonstrated in four figures using PR Curve. Each figure corresponds to an experiment, and each curve shows the classifier's performance regarding various threshold. By comparing Figure 4a to 4b and Figure 4c to 4d it is readily seen that oversampling had a positive impact and has improved classifiers performance in most cases. However, the degree of improvement is not the same for all classifiers. The improvement for classifiers with acceptable performance is insignificant, while classifiers with less impressive performance experience more salient enhancement. There are also classifiers which are insensitive to oversampling or the ones which receive slight negative influence from oversampling.

To be able to compare the classifiers' performance in a more concrete way, we have measured the area under the PR curve (AUPRC) for each classifier. The bottom panel of Table 2 shows the AUPRC regarding each classifier. As the table shows, for classifiers with the high score in Kerala ⇒ Florence such as SVM, MaxEnt, and XGBoost oversampling leads to slight improvement, but for classifiers with lower performance such as RF or MLP, oversampling results in a much higher improvement. Despite the poor performance of CART, oversampling

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

does not improve it that much. Similarly, in the Florence ⇒ Kerala experiment, CART has the lowest performance and oversampling cannot improve it significantly. MLP with very high performance receives marginal improvement. For the other two classifiers with high performance score oversampling shows a slight negative effect which is quite insignificant. For XGBoost and RF as classifiers with satisfactory AUPRC, oversampling fairly improves their performance. We use Bayesian Wilcoxon Singed-rank test [1] to verify the effectiveness of oversampling on the used classifiers. Based on this test, the probability of oversampling being effective is 0.99, which shows that oversampling significantly enhances the performance of classifiers.

For the baseline analysis, we use PHEME dataset [34] which is publicly available for computational rumour detection. We extract 83 features from each tweet and train the classifiers that we introduced before. We repeat the baseline experimental setup by using 5-fold cross-validation in the experiment. Table 3 demonstrates the precision, recall, and F-measure of classifiers along with the results of CRF in [34]. Based on this table, the proposed features and the canonical machine learning classifiers outperform significantly CRF in spite of the fact that the proposed experiment settings are much simpler than that in [34].

4.2 Features performance

In order to assess the impact of the proposed features on the rumour detection, we also eliminated the proposed features and conducted the same experiments. Then, we subtracted the scores corresponding to each classifier to obtain the discriminant power of the proposed significant features. Let f be the intersections of the proposed and significant features, then

$$DP(f) = PF_{Significant\,features} - PF_{leave\,f\,out\,set} \tag{1}$$

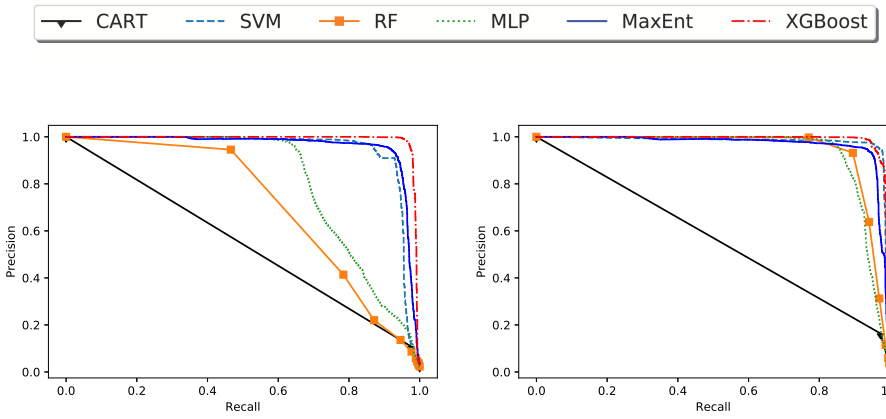
where DP is the discriminant power of the feature set f, $PF_{Significant\,features}$ is the performance of the classifier with all significant features, and $PF_{leave\,f\,out\,set}$ is the performance of the classifier after removing the feature set f. The higher positive values for the discriminant power means that the feature set f is more significant, and zero or negative values is an indicator that shows the feature set f is not effective in increasing the classifier performance.

Table 4 tabulates the discriminant power of each classifier on two experiments. Interestingly, the average performance of all classifiers will increase if the proposed features are used. In particular, AUPRC of the first experiment has increased the most when the classification model is XGBoost. In the second experiment CART and XGBoost receive the highest improvement.

When we look at the classifiers in each experiment one by one, the performance often declines when we remove the proposed significant features. However, for MLP the performance improves after the removal of those features. This problem, namely performance improvement after feature reduction, is a non-trivial

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

Table 2: Assessing the performance of the classifiers using PR curve (upper panel) and AUPRC (bottom panel).



(a) Learning with imbalanced data when the training is with Kerala and test with Florence (Kerala ⇒ Florence)

(b) Learning with balanced data when the training is with Kerala and test with Florence (Kerala ⇒ Florence)

(c) Learning with imbalanced data when the training is with Florence and test with Kerala (Florence ⇒ Kerala)

(d) Learning with balanced data when the training is with Kerala and test with Florence (Kerala ⇒ Florence)

	Classifiers	Training with imbalanced data (AUPRC)	Training with balanced data (AUPRC)
Kerala ⇒ Florence	CART	54.2%	56.8%
	SVM	94.6%	97.8%
	RF	71.5%	94.3%
	MLP	80.9%	93.3%
	MaxEnt	95.9%	96.5%
	XGBoost	98.7%	98.8%
Florence ⇒ Kerala	CART	41.4%	44.6%
	SVM	97.7%	97%
	RF	81.3%	85.5%
	MLP	96.3%	96.4%
	MaxEnt	97.6%	97.2%
	XGBoost	83.9%	89.2%

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

Classifiers	PR	RE	F1-Score
CART	90.5%	91.7%	91.1%
SVM	92.6%	55.5%	69.4%
RF	93.9%	92.4%	93.1%
MLP	92.6%	90.6%	91.6%
MaxEnt	91.2%	79%	84.7%
XGBoost	95%	94.4%	94.7%
CRF [34]	66.7%	55.6%	60.7%

Table 3: Baseline analysis on PHEME dataset[34]

Classifiers	Florence ⇒ Kerala	Kerala ⇒ Florence
CART	10.4%	29.6%
SVM	0.3%	6%
RF	8.3%	3.1 %
MLP	0.2%	-3.4%
MaxEnt	7.6%	4.1%
XGBoost	16.3%	25%
Average	7.2%	8.4%

Table 4: The discriminant power of features for the classifiers with respect to AUPRC metric on two experiments

machine learning problem which has been revisited in the literature before [7,11]. One of the few insights about this problem that has been discussed in the literature is that classification performance achieved with different set of features is highly sensitive to the type of data and type of the classifier [11]; therefore getting the average of the classifiers performance may give a more reliable and broader picture of the discrimination power of chosen features.

5 Conclusion and Future Directions

In this work, we tackled the rumour propagation problem in disasters through imbalanced learning. We provided two datasets regarding Kerala flood and Hurricane Florence by collecting and annotating 100,880 and 101,844 tweets, respectively. We then used 83 features for creating a learning model by compiling existing features in the literature and introducing several new ones. In order to identify rumours in a timely manner, we focused solely on the early available features. We evaluated all features and observed that the proposed features could enhance the performance of the subsequent learning model for the rumour detection. For model building, we conducted a series of experiments using imbalanced learning. For the experiments, we used cross-topic learning which instead of training and test on the same dataset, it is trained on one event and is tested on the other one. Then we measure the impact of imbalanced learning using Bayesian Wilcoxon Signed-rank test. Our results show improvement in almost

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

all the classifiers when the oversampling technique is applied on the training sets.

One of the possible future direction in this domain is data collection and annotation.This field needs more publicly available datasets from different social media and in different kinds of disasters. Such datasets will enable scholars to validate their results in broader contexts and investigate the impacts of contextual factors on their results. Another avenue could be more domain specific studies. In order to have a universal system for rumour detection, we need to know the behaviour of rumours in different subject domains. Performing domain-specific rumour analysis on various subject domains would be a practical step toward discovering the behaviour of rumours.

References

1. Benavoli, A., Corani, C., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. The Journal of Machine Learning Research **18**(1), 2653–2688 (2017)
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web - WWW '11. p. 675. ACM Press, New York, New York, USA (2011)
3. Chen, Y.C., Liu, Z.Y., Kao, H.Y.: IKM at SemEval-2017 Task 8: Convolutional neural networks for stance detection and rumor verification. In: The 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 465–469 (2017)
4. DiFonzo, N., Bordia, P.: Rumor psychology : social and organizational approaches. American Psychological Association (2007)
5. Dunbar, R.I.: Gossip in evolutionary perspective. Review of general psychology **8**(2), 100–110 (2004)
6. Fard, A.E., et al.: Rumour as an anomaly: Rumour detection with one-class classification. In: 2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC). pp. 1–9. IEEE (2019)
7. Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research **3**(Mar), 1289–1305 (2003)
8. García, L.M., Lilja, H., Tjörnhammar, E., Karasalo, M.: Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules. In: The 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 481–485 (2017)
9. Haibo He, Garcia, E.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (9 2009)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. Janecek, A.G.K., Gansterer, W.N., Demel, M.A., Ecker, G.F.: On the relationship between feature selection and classification accuracy. In: Proceedings of the 2008 International Conference on New Challenges for Feature Selection in Data Mining and Knowledge Discovery. pp. 90–105. JMLR.org (2008)
12. Kwon, S., Cha, M., Jung, K., On, W.C.: Prominent features of rumor propagation in online social media. In: International Conference on Data Mining. IEEE (2013)
13. Kwon, S., Cha, M., Jung, K.: Rumor Detection over Varying Time Windows. PLOS ONE **12**(1), e0168344 (1 2017)
14. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48

15. Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L.: The science of fake news. Science (New York, N.Y.) **359**(6380), 1094–1096 (3 2018)
16. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research **18**(1), 559–563 (2017)
17. Liang, G., He, W., Xu, C., Chen, L., Zeng, J.: Rumor Identification in Microblogging Systems Based on Users' Behavior. IEEE Transactions on Computational Social Systems **2**(3), 99–108 (9 2015)
18. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B., Wong, K.: Detecting Rumors from Microblogs with Recurrent Neural Networks. In: IJCAI. pp. 3818–3824 (2016)
19. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: identifying misinformation in microblogs. Proceedings of the Conference on Empirical Methods in Natural Language Processing pp. 1589–1599 (2011)
20. Sicilia, R., Lo Giudice, S., Pei, Y., Pechenizkiy, M.: Twitter rumour detection in the health domain. Expert Systems with Applications **110**, 33–40 (11 2018)
21. Starbird, K., Maddock, J., Orand, M., Achterman, P., Mason, R.: Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. In: IConference (2014)
22. Turenne, N.: The rumour spectrum. PLOS ONE **13**(1), e0189080 (1 2018)
23. Varol, O., Davis, C.A., Menczer, F., Flammini, A., Davis, C.A., Menczer, F., Flammini, A.: Feature Engineering for Social Bot Detection. In: Feature Engineering for Social Bot Detection, pp. 311–334. CRC Press (3 2018)
24. Vosoughi, S., Mohsenvand, M.N., Roy, D.: Rumor Gauge. ACM Transactions on Knowledge Discovery from Data **11**(4), 1–36 (7 2017)
25. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science (New York, N.Y.) **359**(6380), 1146–1151 (3 2018)
26. Wang, S., Moise, I., Helbing, D.: Early Signals of Trending Rumor Event in Streaming Social Media. In: Computer Software and Applications Conference (COMPSAC). pp. 654–659. IEEE (2017)
27. Wijeratne, S., Sheth, A., Bhatt, S., Balasuriya, L., Al-Olimat, H.S., Gaur Manas, Yazdavar, A.H.: Feature Engineering for Twitter-based Applications. In: Feature Engineering for Machine Learning and Data Analytics, pp. 359–384 (2017)
28. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 651–662. IEEE (4 2015)
29. Yang, F., Liu, Y., Yu, X., Yang, M.: Automatic detection of rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics (2012)
30. Zhang, Q., Zhang, S., Dong, J.: Automatic detection of rumor on social network. In: Natural Language Processing and Chinese Computing. pp. 113–122 (2015)
31. Zhao, Z., Resnick, P., Mei, Q.: Enquiring Minds. In: Proceedings of the 24th International Conference on World Wide Web - WWW '15. pp. 1395–1405. ACM Press, New York, New York, USA (2015)
32. Zubiaga, A., Liakata, M., Procter, R.: Exploiting context for rumour detection in social media. In: SocInfo. pp. 109–123 (2017)
33. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and Resolution of Rumours in Social Media. ACM Computing Surveys **51**(2), 1–36 (2 2018)
34. Zubiaga, A., Liakata, M., Procter, R.: Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media (10 2016)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_48