SENTIMENT AWARE FAKE NEWS DETECTION ON ONLINE SOCIAL NETWORKS

Oluwaseun Ajao[1], Deepayan Bhowmik[2] and Shahrzad Zargari[1]

[1] Department of Computing, Sheffield Hallam University, Sheffield, S1 1WB, UK
[2] Division of Computing Science and Mathematics, University of Stirling, Stirling, FK9 4LA, UK
oajao@acm.org, d.bhowmik@ieee.org, s.zargari@shu.ac.uk

ABSTRACT

Messages posted to online social networks (OSN) causes a recent stir due to the intended spread of fake news or rumor. This work aims to understand and analyse the characteristics of fake news especially in relation to sentiments, for the automatic detection of fake news and rumors. Based on empirical observations, we propose a hypothesis that there exists a relation between fake messages or rumours and sentiments of the texts posted online. We verify our hypothesis by comparing with the state-of-the-art baseline text-only fake news detection methods that do not consider sentiments. We performed experiments on standard Twitter fake news dataset and show good improvements in detecting fake news or rumor posts.

1. INTRODUCTION

In the task of detecting fake news in social media it is beneficial if all features associated with each message type are properly identified and utilised. Twitter posts with images offer more impression and influence over text only tweets. A Twitter message has been shown to have a lifespan of as little as less than one day and up to a 70 day span depending on the type of content and URL being shared [1]. This implies that except a message goes *viral* where it *infects* other users - leading to more engagements such as retweets, it normally tends to be short lived thus over-ridden by other posts before the end of the day. To create more engagements, often images are used which may not even be related to the post nor be true images of the event.

Previous work has shown that deception and false statements can be detected from the writing style of the authors or linguistics and sometimes be used to infer their personalities [2]. Some authors have shown that liars can even be detected as they tell complex stories, make fewer self-references to disassociate themselves from the story, and tend to have more frequent use of negative emotion words – as a sign of guilt [3]. Therefore, it is logical to consider emotions within the posted texts as a cue in relation to spreading fake news/rumour. We propose a hypothesis that there exists a relation between a fake message/rumour and the emotion/sentiment of the texts posted online. The proposed hypothesis is proven on a standard benchmark dataset by comparing with the state-of-the-art baseline text-only fake news detection methods that does not consider sentiments. An overall flow of the misinformation classification process is given in Fig. 1 The contributions of this work are as follows:

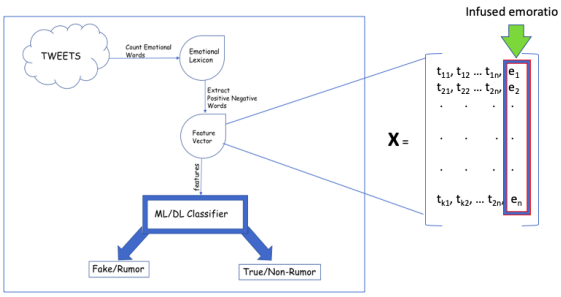

Fig. 1. Schematic Diagram of Text Rumor Classifier

- proposing a relationship that exists between fake news messages and emotional words used in the message text, and
- improvement in fake news detection and prediction following a sentiment-aware classification.

2. BACKGROUND

2.1. What is Fake News?

The Merriam Webster Online Dictionary [4] states Fake News as *'News reports that are intentionally false or misleading'*. We define Fake News in online social media as *'any story circulated, shared or propagated which cannot be authenticated.'* Thus, going by these definitions, we posit that Fake News can also include rumors, clickbait, propaganda, satire and parody as the truthfulness of the stories could often be unverifiable. Several methods have been aimed in the recent past to identify and tackle the problem of fake news. These could be broadly categorised into: *(a) Content-based:* Text (linguistics [5]); Media (images [6], GIFs and video) and URLs, *(b) User-based:* activity tracking (bots and spam [7]); bio information (registration age [8]); opposing views of other online users [9] and *(c) Metadata-based:* GPS Geotags, device source, Followers and Friends Network [10].

Authorized licensed use limited to: MIT Libraries. Downloaded on April 30,2024 at 20:05:07 UTC from IEEE Xplore. Restrictions apply.

2.2. Text-Based Fake News Detection

Ajao *et al.* [5] have shown that fake news can be detected using the text based only approach without prior knowledge of the topic domain. It is worth noting that fake and false information spreads much quicker and deeper than true information. [11] has so far created the largest rumour dataset of 126,000 messages spread by almost 3 million people and found that fake news diffused up to 100,000 people while the truth only reached 1,000 people. [12] identified that 'lone wolves' spread their message faster by creating fake accounts which express the same opinion in multiple ways to help propagate their message faster. A more effective way of achieving this by using social botnets - that retweet and share the same messages indiscriminately to gain popularity and achieve greater spread and coverage. In this work we aim to explore other semantic and multi-modal signal for misinformation in online social networks.

A conditional random field (CRF) was used by [13] for text based rumor detection on the PHEME dataset. [5] employed a hybrid of recurrent neural networks and convolutional neural networks to show that fake news and rumors could be predicted achieving hight accuracy without prior knowledge of the topic domain and no feature engineering. [14] also used a text-based approach for fake news detection but considered the test, response and clustering of user features determined by support vector decomposition and integrated into a hybrid model.

2.3. Text Sentiment Analysis

Sentiment analysis also known as opinion mining seeks to understand the effective meaning of sentences and phrases. It assigns levels of classification to declarations made by the authors; also referred to as *polarity*. It could be as simple as binary levels such as positive and negative or sometimes neutral level of classification. Similar tools and methods were employed by [15] that used a weak supervised, semi-supervised and random-walk step to create lexicons and bag-of-words sentiments. Similarly, [16] using moving average of text sentiment scores over a period, established that negative and positive sentiments extracted from users on Twitter are true reflections of voters' confidence and approval ratings of the President. While sentiment analysis from text goes beyond polarity it could also include the determination of the emotional state of the authors such as *angry, anxious, depressed* and *excited*. Some sentiment dictionaries exist to help in the achievement of this task such as [17] and [18]. Sentiment analysis from text such as Twitter and blogs are well researched topic areas. However, to the best of our knowledge this is the first time it would be examined in the context of fake news detection in online social networks. For the scope of our current work we limit the sentiment analysis of our text to the negative and positive polarities of keywords from the text messages.

3. METHODOLOGY

3.1. Sentiment-Aware Misinformation

We hypothesize that there exists a relationship between a fake message or rumour and the sentiment of the texts posted online. Authors of misinformation posts have been found to conceal their emotions by use of negative emotional words as a sign of guilt in their communication [3]. Also could be that negative emotions tend to spread fast and thus become mechanisms with which these author convey their messages.

We also posit that sentiment may place a role in determination of the class of a tweet as a rumor or non-rumor. We observe such characteristics by analyzing the benchmark data [13] using world cloud visualization after text cleaning. Example of wordclouds from the Charlie Hebdo event is shown in Fig. 2. Therefore a sentiment analysis is proposed to be performed on each of the event corpus with a focus on the sentiment scoring function using Linguistic Word Count application's (LIWC) [19] psychological and linguistic analytic capabilities. Our sentiment analysis rely on an emotional ratio score as calculated in equation 1.

$$emoratio = \frac{\text{count of negative emotional words}}{\text{count of positive emotional words}} \quad (1)$$

In order to check if there was any level of significance between the two types of tweets (rumor and non-rumor), we calculate the t-statistic:

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\left(\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}, \quad (2)$$

and the Null Hypothesis:

$$H_0 : u_1 - u_2 = 0, \quad (3)$$

where u_1 is the mean of rumor corpus and u_2 is the mean of non-rumor corpus of the data. The initial assumption (Ho) is there's no difference between the average sentiment scores of the two populations i.e. rumors N_1 and non-rumors N_2 each having means $\bar{X_1}$ and $\bar{X_2}$ respectively.

In the analysis, we consider the Treatment 1 as the emoratio of rumor tweets of the 5 classes of events, $N_1 = 5$, average across the groups given as $\bar{X_1} = 3.74$, and variance of $s_1^2 = 3.15$. Similarly Treatment 2 is the emoratio vlaues of Non-Rumor events with $N_2 = 5$, $\bar{X_2} = 1.65$ and $s_2^2 = 0.48$. Thus the T-value calculation computed from Equation 2 is given as $t = 2.45058$ is greater than the p-value is 0.01995 (at 0.05 level of significance). It implies that we would reject the null hypothesis H_0, *i.e.*, there's significant difference in the mean of the sentiment scores of the two types of tweets.

3.2. The Algorithm

In the determination of the word relevance and usage within the corpus, we considered sentiments for the terms and

Authorized licensed use limited to: MIT Libraries. Downloaded on April 30,2024 at 20:05:07 UTC from IEEE Xplore. Restrictions apply.

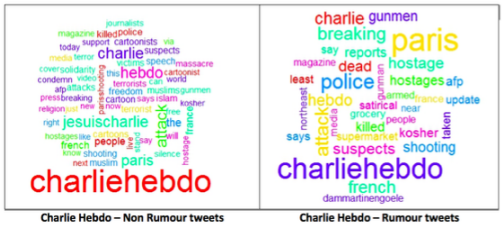

Fig. 2. Word Cloud of Charlie Hebdo Tweets

words. Topic models enable the identification of most relevant words and concepts within a text corpus [20]. We have used two models in extracting emotion scores: *a) Latent Semantic Analysis (LSA)* [21] and *b) Latent Dirichlet Allocation (LDA)* [22] as described below: **Latent Semantic Analysis:** Given an unobserved event topic t, a tweet corpus d containing a word w_i are conditionally independent.

$$p\left(d, w_i\right) = p(d) \sum_t p\left(w_i|t\right) p(t|d). \tag{4}$$

Latent Dirichlet Allocation: Given the parameters γ and η, a topic mixture Ψ with a set of M event topics t and words w will have joint distribution of:

$$p(\boldsymbol{\Psi}, \mathbf{t}, \mathbf{w}|\gamma, \boldsymbol{\eta}) = p(\boldsymbol{\Psi}|\gamma) \prod_{i=1}^{M} p\left(t_i|\boldsymbol{\Psi}\right) p\left(w_i|t_i, \boldsymbol{\eta}\right). \tag{5}$$

While both provided relevant sentiment score, our observation found the better relevance the results following the latter method. Therefore words extracted from the top 10 topics using LDA were supplied as input into our sentiment-aware rumor classifiers.

As sentiment analysis involves the identification of the positivity, negativity and neutrality of text and microposts such as tweets. By looking for keywords used in the posts, we are able to identify the words that are either good or bad portraying either positive or negative emotions. We posit that the consideration of the word sentiments and attaching appropriate weights to each of these identified words in the model building would further improve the performance of the fake news classifier.

Table 1 show the findings computed using Eq. (1) as part of the input features used in the classification. Overview of the proposed algorithm and description of the algorithm are shown in Fig. 1 and Algorithm 1, respectively. Results for various Machine learning and deep learning classifiers are also presented in Table 3. Given the proof that there is a strong significance and association between tweets spread as false rumors and Sentiment Analysis. The task is to develop a machine learning classifier that factors the sentiment score of each corpus containing n number of tweets in determining the weights used in the prediction model. This is achieved using the emotional ratio as described earlier.

Table 1. Emotion ratio in rumor and non-rumor Tweets

Corpus	Word Count	Positive Emotion	Negative Emotion	Emotion Ratio
Rumors				
Charlie	7054	0.82	4.34	5.29
Ferguson	5512	0.71	2.38	3.35
Germanwings	3895	0.41	2.31	5.63
Ottawashoot	7721	1.17	3.67	3.14
Sydneysiege	8250	0.81	1.03	1.27
Non Rumors				
Charlie	26004	2.52	5.78	2.29
Ferguson	14208	1.63	2.94	1.8
Germanwings	3689	0.73	1.68	2.3
Ottawashoot	6719	3.17	2.68	0.85
Sydneysiege	11874	2.7	2.73	1.01

Algorithm 1: Rumor Classifier Algorithm

Input: TweetCorpus, PosemoLexicon, NegemoLexicon;
Compute Latent Dirichlet Allocation;
Extract top k topics;
Extract relevant words for each k;
Extract negative emotion words;
Extract positive emotion words;
repeat
  Input: Receive next relevant tweets;
  Calculate *emoratio*;
  Extract word features from tweets into vector;
  Append the emoratio to the word feature vector;
  repeat
  until *all tweets have been appended*;
  Parse feature vector into classifier;
until *end of sequence*;
Output: y_1 *Predicted label of tweet - Rumor or Non-Rumor*;

3.3. Machine Learning and Deep Learning Classification

We compute the classification of the labeled dataset using a series of machine learning algorithms: logistic regression (LOGIT), support vector machines (SVM), decision trees, random forest and extreme gradient boosting (XG-Boost). We also implemented the long short term (LSTM) recurrent neural network implementation with hierarchical attention networks (HAN). We examine the benefits of using varied word embeddings as pre-trained language models for the input layer of the HAN model. We used the ones proposed by [23]. The pre-trained word vectors included the *Wikipedia 2014 Gigaword5 collection* which was pre-trained on six billion word tokens and the *Twitter collection* which was pre-trained on 2 billion tweets with 27 billion tokens; both in sizes of 100 dimensions. Both LSTM-HAN models were

Authorized licensed use limited to: MIT Libraries. Downloaded on April 30,2024 at 20:05:07 UTC from IEEE Xplore. Restrictions apply.

Table 2. Summary Statistics of Dataset

Name of Event	Event Date	Size	With Images
Charlie Hebdo	7th Mar 2015	2,058	1,087
Ferguson	9th Aug 2014	1,142	4390
Germanwings	24th Mar 2015	468	213
OttawaShoot	22nd Oct 2014	886	301
SydneySiege	15th Dec 2014	1,211	509
TOTAL		5,765	2,600

trained with an epoch size of 50, while a batch size of 64 was found to be optimal and learning rate was set at 10%. The pre-trained word embedding are optimised during the learning process as the deep neural network model benefits from this transfer learning approach.

4. RESULTS AND DISCUSSIONS

4.1. Dataset

We used the PHEME [13] labeled Twitter dataset, from which images were also retrieved for testing on the deep learning model. The corpora consists of 5800 tweets about 5 notable world events widely reported in the electronic, print and conventional news media. They occurred at various times between August 2014 and March 2015. We present the statistic about these news stories in Table 2. All items were hand labeled by journalists. About 45% of the dataset had images and only these were further selected for further enriching the feature set in terms of the embedded texts.

4.2. Results

The emotional ratio of negative to positive words is computed in Table 1. Our statistical test shows that the rumor dataset were significantly different in terms of being more negative sentiments and adverse emotional words from the emotional lexicon [18]. This is further proven in the fake news classifier models were our focus on using emotional words in the classification feature set gave better results over the state of the art which used the same dataset [5] and [13]. Specifically as shown from Table 3. SVM and HAN model with Twitter pre-trained word embedding performed best with 86% for sentiment-aware text only rumor detection. Also, Our results comprises of four variants of the classification feature set; the features from words within the text (TX), the emotional ratio (ER) and use of additional features (AD) including counts of uppercase words, exclamation marks, positive and negative emoticons, user mentions, hashtags and quotations. Table 4 gives summary results in terms of accuracy for these feature combination types. However, considering only the

Table 3. Range of Classifier Results after Emotional Analysis

Classifier	Accuracy	Precision	Recall	F-M
LOGIT	0.84	0.84	0.84	0.84
SVM-Linear	**0.86**	0.86	0.86	0.86
Decision Trees	0.77	0.77	0.77	0.77
Random Forest	0.85	0.85	0.85	0.85
XG-Boost	0.84	0.83	0.84	0.83
LSTM_HAN(Wiki)	0.85	0.86	0.81	0.84
LSTM_HAN(Twitt)	**0.86**	0.86	0.82	0.84
Baseline [5]	0.82	0.44	0.41	0.42
Baseline [13]	N/A	N/A	0.68	0.55

Table 4. Combined features (subset with image-only Tweets)

Classifier	ER+TX	AD+TX	ER+AD+TX
LOGIT	0.84	0.82	0.83
SVM	**0.89**	0.81	0.80
Decision Tree	0.77	0.81	0.81
Random Forest	0.85	0.86	0.85
Grad Boosting	0.85	0.85	0.85
XG-Boost	0.83	0.82	0.83

2600 tweets that had images in Table 4 i.e. column (ER+TX) we see that there's a further 3% improvement to 89% when there's a combination of the text with the emotional ratio if they contained an embedded image within the message. This further strengthens the impact of images in conveying rumors in online social networks. However, these additional features (AD) did not improve the performance of the models.

5. CONCLUSIONS

We proposed a new hypothesis that the use of emotional words is beneficial in sentiment-aware misinformation detection. We support the by proposing a novel sentiment-aware fake new detection algorithm and show improvement on a benchmark dataset over state-of-the-art algorithm that does not consider sentiment. The terrain of fake news and it's detection remains a actively researched topic because it continues to evolve rapidly and yet to be fully understood. This gap presents opportunities for progressive work to be done in the area. Additional sources of sentiment extracted from, *e.g.*, images, embedded text in the image and other visual media such as animations (GIFs) and videos may enhance model performance and is considered as future work.

Authorized licensed use limited to: MIT Libraries. Downloaded on April 30,2024 at 20:05:07 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 705–714.

[2] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference." vol. 77, no. 6, p. 1296, 1999.

[3] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[4] "Fake news - political scandal words." [Online]. Available: https://www.merriam-webster.com/words-at-play/political-scandal-words/fake-news

[5] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on twitter with hybrid cnn and rnn models," in *9th Int'l Conference on Social Media & Society. Copenhagen (July 18)*, no. Jul 2018, 2018.

[6] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 729–736.

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[8] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

[9] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs." in *AAAI*, 2016, pp. 2972–2978.

[10] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[11] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[12] S. Kumar, M. Jiang, T. Jung, R. J. Luo, and J. Leskovec, "Mis2: Misinformation and misbehavior mining on the web," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 799–800.

[13] A. Zubiaga, M. Liakata, and R. Procter, "Learning reporting dynamics during breaking news for rumour detection in social media," *arXiv preprint arXiv:1610.07363*, 2016.

[14] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 797–806.

[15] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[16] B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith *et al.*, "From tweets to polls: Linking text sentiment to public opinion time series." *Icwsm*, vol. 11, no. 122-129, pp. 1–2, 2010.

[17] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[18] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[19] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[20] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[21] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[23] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

Authorized licensed use limited to: MIT Libraries. Downloaded on April 30,2024 at 20:05:07 UTC from IEEE Xplore. Restrictions apply.