DATA AUGMENTATION FOR RUMOR DETECTION USING CONTEXT-SENSITIVE NEURAL LANGUAGE MODEL WITH LARGE-SCALE CREDIBILITY CORPUS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we address the challenge of limited labeled data and class imbalance problem for machine-learning-based rumor detection in social media. We present an offline data augmentation method based on semantic relatedness for rumor detection. Unlabeled social media data is exploited to augment limited labeled data. Context-aware neural language model and a large credibility-focused Twitter corpus are employed to learn effective representations of rumor tweets for our semantic relatedness measurement method. A language model fine-tuned with the large domain-specific corpus shows a dramatic improvement on training data augmentation for rumor detection over pre-trained language models. We conduct experiments on 6 different real-world events based on 5 publicly available datasets and 1 augmented dataset. Our experiments show that the proposed method allows us to generate a large amount of training data with reasonable quality via weak supervision. We present preliminary results achieved using a state-of-the-art neural network model for rumor detection with augmented data for rumor detection.

1 INTRODUCTION

Research areas that have recently been received much attention in using Machine Learning (ML) and Natural Language Processing for automated rumor and fake news detection (Helmstetter & Paulheim, 2018; Kwon et al., 2017; Shu et al., 2017; 2018; Wong et al., 2018) and fact-checking (Boididou et al., 2018; Vosoughi et al., 2017; Kochkina et al., 2018b). One major bottleneck of state-of-the-art ML methods for rumor studies is that they require a vast amount of labeled data to be trained. However, the manual annotation of large-scale and noisy social media data for rumors is highly labor-intensive and time-consuming (Zubiaga et al., 2016). Due to limited labeled training data, existing neural networks (NNs) for rumor detection usually have shallow architecture (Chen et al., 2018; Ma et al., 2016). The scarcity of labeled data is a major challenge facing of studying rumors in social media (Aker et al., 2017). Another problem is that publicly available datasets for rumor-related tasks such as PHEME data (Kochkina et al., 2018b) suffer from imbalanced class distributions (Liu et al., 2017). Existing methods for handling the class imbalance problem (e.g., oversampling and the use of synthetic data (Xu & Chen, 2015)) may cause over-fitting and poor generalization performance. A methodology for rumor data augmentation with the minimum of human supervision is necessary. Previous studies presented that variations of rumors share similar propagation patterns (Maddock et al., 2015; Chen et al., 2018; Baxter et al., 2015). Based on this hypothesis, we argue that enriching existing labeled rumor data with duplicated tweets or corresponding variations is a promising attempt for rumor detection methods that rely on the structure of rumor propagation in social media. We propose a novel data augmentation method for automatic rumor detection based on semantic relatedness. We exploit a publicly available paraphrase identification corpus as well as context-sensitive embeddings of labeled references and unlabeled candidate source tweets. Pairwise similarity is used to guide the assignment of pseudo-labels to unlabeled tweets. ELMo (Peters et al., 2018), a state-of-the-art context-sensitive neural language model (NLM), is fine-tuned on a large credibility-focused social media corpus and used to encode tweets. Our results shows that data augmentation can contribute to rumor detection with deep learning with increased training data size and a reasonable level of quality. This has potential for further performance improvements using deeper NNs. We present the data augmentation results for three

real-world events and the performance of a state-of-the-art DNN model for rumor detection with augmented data in Section 5.

2 DATA

Four publicly available datasets covering a wide range of real-world events on social media as well as a Twitter paraphrase corpus are used for this project.
SemEval-2015 task 1 data (Xu et al., 2014) This data is built for paraphrase identification and semantic similarity measurement. This dataset is employed in our semantic relatedness method in order to fine-tune a optimum relatedness threshold through a pairwise comparison between labeled reference and unlabeled candidates source tweets embeddings (see details in Section 4).
PHEME dataset (6392078) (Kochkina et al., 2018a) This is an extension of the PHEME dataset of rumors and non-rumors and contains 9 manually labeled rumor event data set. This data is used as a reference dataset for data augmentation.
CrisisLexT26 (Olteanu et al., 2015) This dataset comprises tweets associated with 26 hazardous events happened between 2012 and 2013. A subset of data is manually labeled based on informativeness, information types, and information sources. This dataset is used as a reference dataset on which data augmentation is performed.
Twitter event datasets (2012-2016) (Zubiaga, 2018) This data consists of over 147 million tweets associated with 30 real-world events unfolded between February 2012 and May 2016. We use this data as a pool of candidate source tweets. Among the 30 events, we choose six events for which we can generate references using publicly available datasets and resources: 'Ferguson unrest', 'Sydney siege', 'Ottawa shootng', 'Charliehebdo', 'Germanwings plane crash', and 'Boston marathon bombings'. We refer to five events except the 'Boston marathon bombings' as **'PHEME5'**. Although the 'Boston bombings' event is not available in PHEME data, it is suitable for our task as references can be built using *CrisisLexT26* and a fact-checking website 'Snopes.com' [1] (refer to Section 3).
CREDBANK (Mitra & Gilbert, 2015) This data comprises more than 60M tweets grouped into 1049 real-world events, each of which were manually annotated with credibility ratings. This large corpus is leveraged to fine-tune ELMo model in order to provide better representations for rumor-related tasks (refer to Section 3).

3 METHOD

An overview of data augmentation method is presented in Figure 1. We exploit a limited amount of labeled data as weak supervision (i.e., references). The leftmost box shows how semantic similarity between a given pair of reference and candidate tweets is computed. Firstly, the contextual embedding model (ELMo) is fine-tuned with domain-specific corpus in order to learn representations of rumors. Given corpora that contain pairs of tweets, we apply language-based filtering and perform linguistic pre-processing. The pre-processing includes converting words in lowercase, removing 'rt @', URLs, and non-alphabetic characters, and tokenization. Tweets with a minimum of 4 tokens are considered to reduce noise (Ifrim et al., 2014). Then, we compute ELMo embeddings of tweets for subsequent semantic relatedness measurement. Cosine similarity is used to compare each tweet pair. We use *SemEval-2015 task 1 data* set as a benchmark for fine-tuning relatedness thresholds (see Section 4). Having an optimum threshold, we follow the same relatedness measurement procedure to compute semantic similarity for each pair of reference and candidate tweets in the *Twitter event datasets (2012-2016)* data. Candidate tweets refer to any tweets that report about an event of interest. Schemes for constructing references are varying between data sets. For *PHEME5*, we use annotations in the *PHEME* data. A reference dataset of "Boston marathon bombings" event is generated separately as mentioned in Section 2. Due to space constraints, we omit the detailed process here. The next step is to select rumors and non-rumors from candidate tweets based on the optimum relatedness threshold. In the final step, data collection [2] is performed to retrieve social-temporal context data (typically retweets and replies) for selected candidate tweets. Source tweets without context data are filtered out. We download source tweets for six selected events in *Twitter event (2012-2016)* and *CREDBANK* using Twitter API [3]. For *CREDBANK*, we downloaded *77,954,446*

[1] https://www.snopes.com/fact-check/boston-marathon-bombing-rumors/
[2] We will make our source code and augmented data publicly available.
[3] An open source tweet collector, available via https://github.com/socialsensor/twitter-dataset-collector

tweets (i.e., 97.1% of the original data). After deduplication, the train corpus contains *6,157,180* tweets with *146,340,647* tokens and *2,235,075* vocabularies.

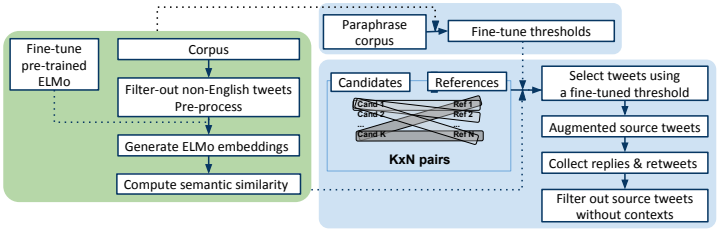

Figure 1: Overview of Data Augmentation.

Rumor-specific Embedding (ELMo) Previous research shows that fine-tuning NLMs with in-domain data allows them to learn more meaningful word representations and provides a performance gain (Kim, 2014; Peters et al., 2018). To fine-tune pre-trained ELMo, we generate a dataset using *CREDBANK*. Sentences are shuffled and spitted into training and hold-out sets (with proportion around 0.02%). We also generate test set that consists of 6,162 tweets in total using the *PHEME* data. Table 1 shows the number of tweets, tokens and vocabularies in the *CREDBANK* after deduplication. Following the practice in (Perone et al., 2018), a linear combination of the states of each LSTM layer and the token embeddings is adopted to encode tweets. Since the *CREDBANK* training set is still a relatively small dataset for NLM, we only fine-tune pre-trained ELMo with 1 epoch to avoid over-fitting. The result shows a large improvement in perplexity on both hold-out set and test set (See Table 2).

Table 1: Statistics of the CREDBANK for fine-tuning ELMo.

	Train	Hold-out
# of tweets	6,155,948	1,232
# of tokens	146,313,349	27,298
# of vocabularies	2,234,861	6,517

Table 2: Perplexity before and after fine tuning before 1 epoch on hold-out set and test set. Reported values are the average of the forward and backward perplexity.

Dataset	Before tuning	After tuning
CREDBANK hold-out set	883.06	18.24
PHEME test set	475.06	32.02

4 Experimental Setup

Semantic Relatedness Fine-Tuning Table 3 compares different models for word representation on the SemEval-2015 data. We show the results based on the maximum F-score each model achieved. It shows the effectiveness of our *CREDBANK* fine-tuned ELMo over pre-trained model ("Original (5.5B)") and state-of-the-art word embedding models. Relatedness thresholds are fine-tuned based on precision achieved by the best-performing model.

Table 3: Comparison of the performance of different models for sentence representation.

Model	F-score	Precision	Recall	Threshold
ELMo+CREDBANK (average)	**0.6507**	**0.6088**	0.6986	0.652584
ELMo+CREDBANK (top)	0.6270	0.5660	0.7027	0.646989
ELMo Original 5.5B (average)	0.6281	0.5872	0.6752	0.630502
ELMo Original 5.5B (top)	0.6047	0.5554	0.6635	0.687503
GloVe (twitter.27B.200d)	0.5079	0.3417	**0.9890**	0.501713
Word2Vec (Google News)	0.4223	0.4796	0.3772	0.500273

Data Augmentation Table 4 shows the number of candidate source tweets and references after pre-processing and filtering out empty strings. We follow our data augmentation procedure described in Section 3. After pairwise similarity computation on all datasets, we adopt *0.8*, which achieves a precision of 0.9 in the benchmark task illustrated above, as relatedness threshold for selecting rumor source tweets from a pool of candidates. We randomly sample (3*(# of rumor source tweets)) non-rumor source tweets if the score of a candidate tweet is less than 0.3. Sampling more negative examples is an attempt to balance class distributions after source tweets without context data are removed. This is due to the fact that non-rumors generally draw less attention, and therefore, it is highly likely that they do not have replies or retweets.

Table 5: Rumor detection results for the PHEME5 and augmented datasets.

Data	Event	F-score	Precision	Recall	Accuracy
PHEME5		0.5354	0.5802	0.4970	0.7068
	charliehebdo	0.5413	0.4304	0.7293	0.7278
	germanwings	0.5972	0.6649	0.5420	0.6290
	fergusonunrest	0.1854	0.4583	0.1162	0.7463
	ottawashooting	0.6338	0.7795	0.5340	0.6742
	sydneysiege	0.5502	0.7169	0.4464	0.6880
PHEME5+Boston		0.4933	0.5800	0.4291	0.6958
	charliehebdo	0.5067	0.4521	0.5764	0.7528
	germanwings	0.5296	0.6821	0.4328	0.6098
	fergusonunrest	0.3181	0.4171	0.2570	0.7262
	ottawashooting	0.6209	0.7500	0.5298	0.6584
	sydneysiege	0.4392	0.7094	0.3180	0.6527
	bostonbombings	0.4593	0.5905	0.3758	0.6276

5 Results and Discussion

Data Augmentation Before filtering out source tweets without replies and retweets, *1,238* rumors and *3,714* non-rumors are collected for 'bostonbombings'. After filtering, *165* rumors and *228* non-rumors remain. Although the augmented data size is very limited for 'bostonbombings', experiments on 'sydneysiege' and 'ottawashooting' shows encouraging results. Total *25,486* rumors and *76,106* non-rumors are obtained additionally for 'sydneysiege', and total *21,519* rumors and *62,590* non-rumors are obtained additionally for 'ottawashooting'.

Table 4: Statistics of candidate source tweets and references after pre-processing.

Event	# of cand.	# of ref.
bostonbombings	461,488	88
fergusonunrest	4,104,646	41
sydneysiege	1,287,812	61
ottawashooting	692,268	50
charliehebdo	9,699,630	60
germanwings	1,726,981	19

Rumor Detection We conduct rumor detection experiments using two different datasets: (1) *PHEME5*, (2) *PHEME5* with 'bostonbombings' data ('PHEME5+Boston'). We employ Kochkina et al.'s method as state-of-the-art baseline model of rumor detection with slight modifications (Kochkina et al., 2018b). For the sake of simplicity, we modify the implementation of 'MTL2 Veracity+Detection' for rumor detection only. We construct input by using a source tweets and the top (i.e., most recent) 24 replies in this task. We perform leave-one-out cross-validation on the PHEME5 and augmented datasets. The experimental results for rumor detection are presented in Table 5. It shows the performance of rumor detection for each event in the *PHEME5* and augmented datasets as well as for the entire datasets. When using the augmented data, we observe that overall performance decreases compared to when using the PHEME5 data. The 'ferguson' is the most difficult event for a rumor detection model as it has a unique class distribution distinguished from all other events (Kochkina et al., 2018b). It is worth noting that our data augmentation improves the performance of rumor detection on the 'ferguson'. The completion of data augmentation for events other than 'bostonbombings' has potential to boost overall and per event performance of rumor detection.

6 Conclusion and Future Work

We present a methodology of data augmentation for rumor detection that exploits semantic relatedness between limited labeled data and unlabeled data. This study is part of further research that aims to use a massive amount of publicly available unlabeled Twitter data and potentials of DNNs techniques in a wide range of tasks related to rumors in social media. Current study showed the potential efficiency and effectiveness of semantically augmented data for combating the labeled data scarcity and class imbalance problems of existing publicly available rumor datasets. In future work, we plan to augment data for more events to build comprehensive datasets for rumor detection, and conduct experiments on rumor detection via deep learning. We will evaluate the effectiveness of augmented data in alleviating over-fitting and its usefulness in facilitating deeper NNs for rumor detection. Further experiments will be conducted to examine the generalization of rumor detection models on unseen rumors.

REFERENCES

Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, Anna Kolliakou, Rob Procter, and Maria Liakata. Stance classification in out-of-domain rumours: A case study around mental health disorders. In *International Conference on Social Informatics*, pp. 53–64. Springer, 2017.

Claire Baxter, P Barratt, and Michael Thomson. Social media and the generation, propagation, and debunking of rumours. *Report on behalf of Department of National Defence, Canada. Ontario: Human Systems Incorporated*, 2015.

Christina Boididou, Stuart E Middleton, Zhiwei Jin, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, and Yiannis Kompatsiaris. Verifying information with multimedia content on twitter. *Multimedia Tools and Applications*, 77(12):15545–15571, 2018.

Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 40–52. Springer, 2018.

S. Helmstetter and H. Paulheim. Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, volume 00, pp. 274–277, Aug. 2018. doi: 10.1109/ASONAM.2018.8508520. URL doi.ieeecomputersociety.org/10.1109/ASONAM.2018.8508520.

Georgiana Ifrim, Bichen Shi, and Igor Brigadir. Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014*. ACM, 2014.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. PHEME dataset for Rumour Detection and Veracity Classification. 6 2018a. doi: 10.6084/m9.figshare.6392078.v1. URL https://figshare.com/articles/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3402–3413, 2018b.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. Rumor Detection over Varying Time Windows. *PLOS ONE*, 12(1):1–19, 2017. doi: 10.1371/journal.pone.0168344. URL https://doi.org/10.1371/journal.pone.0168344.

Yahui Liu, Xiaolong Jin, Huawei Shen, and Xueqi Cheng. Do rumors diffuse differently from non-rumors? a systematically empirical analysis in sina weibo for rumor identification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 407–420. Springer, 2017.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pp. 3818–3824. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3061053.3061153.

Jim Maddock, Kate Starbird, Haneen J Al-Hassani, Daniel E Sandoval, Mania Orand, and Robert M Mason. Characterizing online rumoring behavior using multi-dimensional signatures. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp. 228–241. ACM, 2015.

Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 994–1009. ACM, 2015.

Christian S Perone, Roberto Silveira, and Thomas S Paula. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*, 2018.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 430–435. IEEE, 2018.

Soroush Vosoughi, Mostafa'Neo' Mohsenvand, and Deb Roy. Rumor gauge: Predicting the veracity of rumors on twitter. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(4):50, 2017.

Kam-Fai Wong, Wei Gao, and Jing Ma. Rumor detection on twitter with tree-structured recursive neural networks. In *ACL*, 2018.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, 2014. URL `http://www.cis.upenn.edu/~xwe/files/tacl2014-extracting-paraphrases-from-twitter.pdf`.

Wen Xu and He Chen. Scalable rumor source detection under independent cascade model in online social networks. In *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, pp. 236–242. IEEE, 2015.

Arkaitz Zubiaga. A longitudinal assessment of the persistence of twitter datasets. *JASIST*, 69: 974–984, 2018.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. *CoRR*, abs/1610.07363, 2016.