Revisiting Rumour Stance Classification: Dealing with Imbalanced Data

Yue Li and **Carolina Scarton**
Department of Computer Science, University of Sheffield, UK
riona.yueli@gmail.com, c.scarton@sheffield.ac.uk

Abstract

Correctly classifying stances of replies can be significantly helpful for the automatic detection and classification of online rumours. One major challenge is that there are considerably more non-relevant replies (*comments*) than informative ones (*supports* and *denies*), making the task highly imbalanced. In this paper we revisit the task of rumour stance classification, aiming to improve the performance over the informative minority classes. We experiment with traditional methods for imbalanced data treatment with feature- and BERT-based classifiers. Our models outperform all systems in RumourEval 2017 shared task and rank second in RumourEval 2019.

1 Introduction

A key step in the task of automatically analysing rumour veracity is to analyse the view of other users on a particular rumour (Procter et al., 2013), i.e. the stance of its replies. RumourEval 2017 and 2019 (Derczynski et al., 2017; Gorrell et al., 2019) are shared tasks that provide tree-structured conversation threads consisting of tweets directly or indirectly replying to a rumourous tweet and aim to label the stance of these replies towards the rumour (task A). Specifically, it is framed as a four-class classification problem: *support*, *deny*, *query*, and *comment* (SDQC). *Supports* and *denies* are arguably the most informative stances for rumour verification (Mendoza et al., 2010), while *comments* are considered the least useful. However, the data for this task is highly imbalanced with *supports* and *denies* corresponding, respectively, to 18% and 7% of instances in the RumourEval 2017, while *comments* are 66%.[1]

Systems submitted for RumourEval 2017 task A, evaluated in terms of accuracy, have a high performance for the majority class (*comments*), whilst the minority classes are under-performed. `Mama Edha` (García Lozano et al., 2017) adjusts the weights of the four labels, while only correctly classifying 1% *denies* and 37% *supports*. `ECNU` proposes a two-step classifier, however, only 1% of *denies* and 28% of *supports* are accurately predicted (Wang et al., 2017). `IITP` over-samples the underrepresented classes (Singh et al., 2017), although only 12% *denies* and 44% *supports* are recognised. The winner, `Turing` (Kochkina et al., 2017), is unable to identify any *denies* in the test data. In RumourEval 2019, with macro-$F1$ as evaluation metric, `eventAI` (Li et al., 2019) (third place) achieves 55% and 79% of correct *supports* and *denies*, respectively. Ranked first, `BLCU NLP` (Yang et al., 2019) increases the *supports* and *denies* with external similar datasets. However, the expanded data is still skewed towards *comments*, and 38% *supports* and 51% *denies* are accurately predicted. `UPV` (Ghanem et al., 2019) set different weights for each class, but 72% *supports* and 91% *denies* are mis-classified. Despite `GWU` (Hamidian and Diab, 2019) designing a rule-based model to help predict the instances of minority classes, their system does not correctly classify any *denies*. Other work (Zubiaga et al., 2018; Akhtar et al., 2018; Ma et al., 2018; Xuan and Xia, 2019) that also experiments with RumourEval or PHEME datasets does not consider imbalanced data approaches, and under-performs in *support* and *deny* classes.

In this paper, we experiment with well-known techniques for dealing with data imbalanced problems (e.g. SMOTE (Chawla et al., 2002)). We create feature-based models and a BERT-based model (Devlin

[1] RumourEval 2019 dataset has a similar distribution: *supports* = 14%, *denies* = 7% and *comments* = 72%.

Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), pages 38–44
Barcelona, Spain (Online), December 12, 2020.

et al., 2019). Results show that our models not only have higher performance for the minority classes but also have higher overall performance (in terms of macro-$F1$ and geometric mean recall – GMR) than all systems submitted to RumourEval 2017 and all but one system submitted to RumourEval 2019.[2]

2 Resampling mechanisms and threshold-moving

Random under- or over-sampling (RUS or ROS) RUS randomly discards samples in the majority class so that the class proportions can be balanced. Generally, it is computationally more efficient than over-sampling because it reduces the training data, although it may lead to under-fitting. ROS re-balances the class proportions through randomly replicating samples in the minority classes. However, these replications can increase the possibility of over-fitting (Prati et al., 2009).

Synthetic minority over-sampling technique (SMOTE) is one of the most popular methods to over sample the minority class (Chawla et al., 2002). The mechanism is to artificially generate new samples based on k-nearest neighbours of each observation in the minority class. Although SMOTE also has other variants, such as Borderline-SMOTE (Han et al., 2005), we only experiment with the original SMOTE.

Adaptive synthetic (ADASYN) sampling approach (He et al., 2008) is another over-sampling method similar to SMOTE, where the new synthetic samples are also interpolated based on each observation's k-nearest neighbours in the minority class. The main difference between SMOTE and ADASYN is the number of synthetic samples generated for each observation in the minority class. For SMOTE, it only depends on the required ratio of over-sampling, while in ADASYN, the number depends on the level of hardness of learning the data observation. ADASYN may focus too much on outliers, while SMOTE may associate outliers with inliers. Therefore, both of them could result in a sub-optimal decision.

SMOTE + Edited Nearest Neighbors (ENN) (SMOTEENN) is a hybrid resampling method that combines over-sampling and under-sampling (Batista et al., 2004). Generally, it can achieve better performance than solely using SMOTE. In this method, firstly, the minority class is over-sampled by SMOTE. Then ENN will examine both majority and minority class and remove the data samples that are mis-classified by their three-nearest neighbours, which works as a data cleaning method.

Threshold-moving (TM) (Maloof, 2003; Sheng and Ling, 2006) usually does not change the original class proportions. The classifier is trained with the imbalanced data, but the decision threshold that transforms the output probability into class label is changed. For example, we usually set 0.5 for a balanced binary classification. As there is no closed-form expression for a threshold that can maximise macro-$F1$ (Lipton et al., 2014), we set the threshold according to the class proportions, which has been proved to maximise macro accuracy based on two assumptions: (1) the class proportion of the test set is similar to that of the training set, and (2) the prior of a class is equivalent to its proportion in the training set (Collell et al., 2018). Therefore, our process for threshold moving is: (1) compute the output probability P_k for class k and (2) assign the class with highest P_k/a_k, $a_k = num_k/num_{total}$, in which num_k is the number of class k in the training set, and num_{total} is the total number of the training set.

3 Experiments and Results

3.1 Experimental setup

Techniques presented in Section 2 are explored in two types of classification models. We use implementations of resampling methods from the `imbalanced-learn` python toolkit (Lemaître et al., 2017).

Feature-based classifiers Our feature-based approach is an adaptation of (Aker et al., 2017). We use Twitter-based features like number of re-tweets, presence of URLs and hashtags, number of followers for the user, among others. These features are then concatenated with a word vector representation for the tweets, using a pre-trained Twitter GloVe embedding model (Pennington et al., 2014). We train Random Forest (`RF`), Multi-Layer Perceptron (`MLP`) and Logistic Regression with stochastic gradient descent (`LR-SGD`) models using the `scikit-learn` python toolkit (Pedregosa et al., 2011).

[2] Our implementation is available at `https://github.com/YLi999/RumorStanceClassification`

Average GMR and standard deviations

	NT	**RUS**	**ROS**	**SMOTE**	**ADASYN**	**SMOTEEN**	**TM**
RF	0.000 ± 0.000	0.513 ± 0.025	$\underline{0.453 \pm 0.039}$	0.229 ± 0.242	0.000 ± 0.000	0.037 ± 0.079	0.457 ± 0.040
MLP	0.357 ± 0.139	0.541 ± 0.046	0.428 ± 0.154	$\underline{0.442 \pm 0.078}$	$\underline{0.494 \pm 0.057}$	$\underline{0.477 \pm 0.035}$	0.508 ± 0.036
LR-SGD	0.000 ± 0.000	0.519 ± 0.076	0.149 ± 0.195	0.234 ± 0.162	0.110 ± 0.178	0.230 ± 0.178	0.409 ± 0.067
BERT	$\underline{0.482 \pm 0.057}$	$\underline{0.622 \pm 0.027}$	0.442 ± 0.056	-	-	-	$\mathbf{0.626 \pm 0.028}$

Average macro-F1 and standard deviations

	NT	**RUS**	**ROS**	**SMOTE**	**ADASYN**	**SMOTEEN**	**TM**
RF	0.345 ± 0.012	$\underline{0.509 \pm 0.020}$	0.519 ± 0.011	$\underline{0.529 \pm 0.018}$	0.514 ± 0.002	0.333 ± 0.005	0.466 ± 0.014
MLP	0.466 ± 0.026	0.486 ± 0.038	0.507 ± 0.031	0.505 ± 0.044	$\underline{0.531 \pm 0.036}$	$\underline{0.461 \pm 0.024}$	0.502 ± 0.031
LR-SGD	0.402 ± 0.009	0.481 ± 0.065	0.384 ± 0.035	0.423 ± 0.024	0.418 ± 0.020	0.367 ± 0.020	0.534 ± 0.037
BERT	$\mathbf{0.584 \pm 0.029}$	0.502 ± 0.030	$\underline{0.529 \pm 0.025}$	-	-	-	$\underline{0.540 \pm 0.013}$

Table 1: GMR and macro-$F1$ on RumourEval 2017 development set. Best results overall are in bold. Best result for each approach are underlined.

BERT-based classifier (BERT) We employ the pre-trained *BERT-base-uncased* model (Devlin et al., 2018) with 12 transformer layers, hidden unit size of 768, 12 attention heads, and 110M parameters. The inputs are the texts of a rumourous tweet and a reply tweet, and we fine tune for three epochs with a batch size of 16, using the ktrain (Maiya, 2020) toolkit. During training, we apply the 1 *cycle policy* (Smith, 2018), and search the optimal learning rate among $5e^{-5}$, $4e^{-5}$, $3e^{-5}$, and $2e^{-5}$. Since ROS, RUS and TM can be directly applied to raw text, we only apply BERT with these three methods.

Evaluation For evaluation we use macro-$F1$ and GMR. Macro-$F1$ is the arithmetic mean between the $F1$-score $F_{1,c}$ of each class c: macro-$F1 = \frac{\sum_{i=1}^{C} F_{1,c}}{C}$ and is commonly applied in the evaluation of imbalanced binary classification. However, for multi-class problems, it is not robust to poor performance of the minority classes. GMR is denoted as $\sqrt[C]{\prod_{c=1}^{C} R_c}$, in which R_c is the recall of class c. False negatives may be more relevant than false positives in an imbalanced problem, therefore, it is important to assess models using recall-based metrics. Combining GMR with macro-$F1$ for evaluation can avoid choosing a model with high macro-$F1$ but actually with low recall for the minority classes.

3.2 Models assessment

For this experiment, we consider RumourEval 2017 data only.[3] We run each experiment 10 times to model variability and test on the development set. As baselines, we also train systems without any imbalanced data treatment (NT). Table 1 shows average and standard deviation of GMR and macro-$F1$. Although BERT in the NT case shows the highest macro-$F1 = 0.584$, BERT with TM (macro-$F1$ $= 0.540$) is still a better system, since it has the highest GMR $= 0.626$ and performs significantly better for *supports* and *denies*. When using feature-based classifiers, RUS leads to better models than the other approaches for both macro-$F1$ and GMR. The feature-based training data is high dimensional, which may harm the performance of resampling methods that are based on *k*-nearest neighbours. Systems with GMR $= 0$ are the worst case, since they could not correctly classify any *denies* in any of the 10 iterations (e.g. RF with ADASYN). Other systems, such as LR-SGD with SMOTE, fail to correctly predict any *denies* most of the time in 10 experiments, and consequently have high standard deviation of GMR (larger than 0.1). When using BERT, both RUS and TM result in a relatively good prediction on the minority classes, while TM perform better on the *comments* – the macro-$F1$ of BERT with TM is larger than that of BERT with RUS, although their GMRs are almost the same. TM works well with our neural models, BERT and MLP, which can provide good estimation of posterior probabilities. Finally, this analysis highlights the necessity of using both GMR and macro-$F1$ for evaluation. Some systems with high macro-$F1$ have low GMR, such as MLP with ADASYN (macro-$F1 = 0.531$, GMR $= 0.494$).

[3] Since RumourEval 2019 has Reddit data, it is not possible to use the same level of metadata available for tweets in this dataset, which justify our focus only on RumourEval 2017 data for model selection.

	GMR	macro-F1
BERT-TM(ensemble)	**0.635**	**0.536**
`BERT-TM(single)`	0.626	0.513
`FBE-RUS`	0.618	0.484
`BERT-NT(single)`	0.403	0.516
`NileTMRG`	0.363	0.452
`ECNU`	0.214	0.467
`Turing`	0.000	0.434

Table 2: Comparison with selected systems from RumourEval 2017.

	GMR	macro-F1
eventAI	**0.726**	0.578
`BERT-TM(single)`	0.618	0.561
`BERT-TM(ensemble)`	0.605	0.571
`BLCU NLP`	0.571	**0.619**
`BUT-FIT`	0.519	0.607

Table 3: Comparison with selected systems from RumourEval 2019.

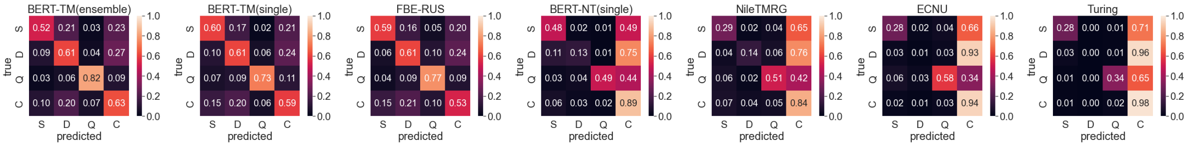

Figure 1: Confusion matrix for proposed models and selected systems for RumourEval 2017

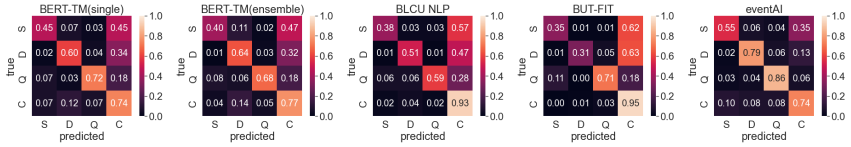

Figure 2: Confusion matrix for proposed models and selected systems for RumourEval 2019

3.3 Comparison with RumourEval submitted systems

As `BERT` with TM (`BERT-TM(single)`) is the best model on RumourEval 2017 development set, we further test it on both RumourEval 2017 and 2019 test sets, and compare it with the submitted systems. We also implement an ensemble of the three feature-based models (RF, MLP and LR-SGD) with RUS (`FBE-RUS`), and a bagging ensemble of `BERT-TM(single)` (Collell et al., 2018) – `BERT-TM(ensemble)`. The process of training `BERT-TM(ensemble)` is: (1) generate n training sets with simple bootstrap sampling; (2) fine-tune n `BERT` base classifiers; (3) compute the average of n probabilistic predictions for each class; and, (4) perform TM. Similar to `BERT-TM(single)`, the threshold is set according to the class proportion of training data. The optimal number of the base classifiers n is determined by the performance on development data ($n = 15$ in our case). We also present results for `BERT` without TM (`BERT-NT(single)`) on RumourEval 2017 test set.

For RumourEval 2017, we compare our models with `Turing`, `ECNU`, and `NileTMRG` (Enayet and El-Beltagy, 2017) (Table 2). `BERT-TM(single)`, `BERT-TM(ensemble)`, and `FBE-RUS` outperform other systems, showing similar performance for *supports* and *denies*. After applying TM on the output of `BERT-NT` (`BERT-TM(single)`), the performance on the minority classes is significantly enhanced (Figure 1). For RumourEval 2019, our models are compared with `BLCU NLP`, `BUT-FIT` (Fajcik et al., 2019), and `eventAI` (Table 3), outperforming `BLCU NLP` and `BUT-FIT` on *supports* and *denies* (Figure 2). Although `eventAI` performs better than our models, some details about its architecture are not provided in the paper and the code is not publicly available.

4 Conclusion and future work

We experiment with traditional imbalanced data techniques for the task of rumour stance classification and show that: (i) our models are capable of outperforming all systems in RumourEval 2017 and all but one system in RumourEval 2019 in terms of both macro-$F1$ and GMR scores, and (ii) a more in-depth evaluation is needed in order to correctly assess this task. Further improvements may be achieved by employing model-based imbalanced data techniques (e.g. by setting different weights for each class during training), which is left as future work.

Acknowledgments

This work was funded by the WeVerify project (EU H2020, grant agreement: 825297).

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria, September. INCOMA Ltd.

Md Shad Akhtar, Asif Ekbal, Sunny Narayan, Vikram Singh, and Erik Cambria. 2018. No, that never happened!! investigating rumors on twitter. *IEEE Intelligent Systems*, 33(5):8–15.

Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Guillem Collell, Drazen Prelec, and Kaustubh R. Patil. 2018. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, 275:330–340.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Omar Enayet and Samhaa R El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474.

Martin Fajcik, Lukáš Burget, and Pavel Smrz. 2019. But-fit at semeval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. *arXiv preprint arXiv:1902.10126*.

Marianela García Lozano, Hanna Lilja, Edward Tjörnhammar, and Maja Karasalo. 2017. Mama edha at SemEval-2017 task 8: Stance classification with CNN and rules. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 481–485, Vancouver, Canada, August. Association for Computational Linguistics.

Bilal Ghanem, Alessandra Teresa Cignarella, Cristina Bosco, Paolo Rosso, and Francisco Manuel Rangel Pardo. 2019. Upv-28-unito at semeval-2019 task 7: Exploiting post's nesting and syntax information for rumor stance classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1125–1131.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Sardar Hamidian and Mona Diab. 2019. Gwu nlp at semeval-2019 task 7: Hybrid pipeline for rumour veracity and stance classification on social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1115–1119.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing (ICIC 2005)*, pages 878–887, Hefei, China. Lecture Notes in Computer Science.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328, Hong Kong, China. IEEE.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada, August. Association for Computational Linguistics.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize F1 measure. *Mach Learn Knowl Discov Databases*, 8725:225–239.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the The Web Conference 2018*, pages 585–593.

Arun S. Maiya. 2020. ktrain: A Low-Code Library for Augmented Machine Learning. *arXiv preprint arXiv:2004.10703*.

Marcus A Maloof. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, DC, USA.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: can we trust what we RT? In *Proceedings of the First Workshop on Social Media Analytics*, pages 71–79, Washington, DC, USA. Association for Computing Machinery.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Ronaldo C. Prati, Gustavo E. A. P. A. Batista, and Maria Carolina Monard. 2009. Data mining with imbalanced class distributions: concepts and methods. In *Proceedings of the 4th Indian International Conference on Artificial Intelligence (IICAI 2009)*, pages 359–376, Tumkur, Karnataka, India.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214.

Victor S. Sheng and Charles X. Ling. 2006. Thresholding for making classifiers cost-sensitive. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, pages 476–481, Boston, Massachusetts. American Association for Artificial Intelligence.

Vikram Singh, Sunny Narayan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2017. IITP at SemEval-2017 task 8 : A supervised approach for rumour evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 497–501, Vancouver, Canada, August. Association for Computational Linguistics.

Leslie N. Smith. 2018. *A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay*. US Naval Research Laboratory Technical Report 5510-026.

Feixiang Wang, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 8: Rumour evaluation using effective features and supervised ensemble models. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 491–496, Vancouver, Canada, August. Association for Computational Linguistics.

Kaizhou Xuan and Rui Xia. 2019. Rumor stance classification via machine learning with text, user and propagation features. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. Blcu_nlp at semeval-2019 task 7: An inference chain-based gpt model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.