Profiling Spreaders of Disinformation on Twitter: IKMLab and SoftBank Submission

Notebook for PAN at CLEF 2020

Timothy Niven, Hung-Yu Kao, and Hsin-Yang Wang

Intelligent Knowledge Management Lab
National Cheng Kung University, Tainan, Taiwan
and
AI Strategy Office, SoftBank Corp., Tokyo, Japan
tim.niven.public@gmail.com, hykao@mail.ncku.edu.tw, hsinyang.wang@g.softbank.co.jp

Abstract The problem we address is classifying whether a Twitter user has spread confirmed disinformation or not. We used two types of features that had validity in the training set: features that indicate thoughtfulness, and features reflecting emotional states. We attempted to capture thoughtfulness via the rate of function word usage and constituency tree features reflecting sentence complexity. We added features for sentiment in general and negative sentiment in particular to measure emotional arousal. We also experimented with custom lexicons for anger and distrust. Our classifier was an ensembled Support Vector Classifier, Random Forest, and Naive Bayes algorithms. We only considered the English data. Our cross-validated training set accuracy was 89.3%, but significantly overfit the training data, achieving 61.0% test set accuracy.

1 Introduction

The "Profiling spreaders of fake news" task at PAN 2020 [15] supplies 100 tweets per 300 users, where half of the users are confirmed to have spread deliberate disinformation.[1] The tweets are almost entirely retweets and shared news headlines. Task participants are required to develop techniques to separate the spreaders of disinformation from the controls. The task authors provide English and Spanish datasets to facilitate multilingual techniques. This is an exciting contribution, as solutions that work across languages are likely to reflect learning about the task, as opposed to "solving datasets" (i.e. overfitting the bias distribution shared across training and held-out testing sets, e.g. as a result of the data collection process) [9].

[1] Following Taiwan's Digital Minister, Audrey Tang, we prefer the term "disinformation," as "fake news" carries connotations that can undermine journalism [4].

2 Our Approach

Regrettably, for our submission we only had a chance to work on the English data. Our approach was to restrict ourselves to considering *how* the tweets discussed their topics, and not what they were talking about. Although the dataset comprises 30,000 tweets there are only 300 labels.[2] We were concerned that a solution such as BERT [6] might focus on keywords that have statistical validity across both train and test, but nevertheless overfit the dataset as a whole [11]. We were also interested to find features that help us understand the phenomenon. We therefore consider predictive accuracy as important but not all-encompassing, and favour solutions that are interpretable.[3]

The features we chose are designed to measure thoughtfulness and emotional arousal. Measurements were made from lexicons and constituency tree parses. For lexical categories, we counted the frequency of words for each tweet, and then took the mean over tweets per user. Formally, let the set of words in category c be $\mathcal{L}_c$. For each user, u, tokenize all 100 tweets yielding token vectors $\mathbf{t}_i^{(u)}, i \in [0, 100]$, and count the proportion of words belonging to category c. Then average over all user tweets to get the final frequency of usage per user:

$$f_c^{(u)} = \frac{1}{100} \sum_{i=1}^{100} \frac{\sum_{t \in \mathbf{t}_i^{(u)}} \mathbf{1}[t \in \mathcal{L}_c]}{|\mathbf{t}_i^{(u)}|}.$$

Constituency tree features were obtained with the Berkeley Neural Parser [10] and will be detailed below.

2.1 Thoughtfulness

We reasoned that people who share more thoughtful content are less likely to spread disinformation. We developed two sets of features to measure and investigate this hypothesis: (1) the frequency of function word usage; (2) constituency tree features.

Function Word Usage Rates Higher usage rates of function words indicate more complex sentences, and therefore may be seen to express thinking effort. The training data shows some support for the existence of a threshold of function word usage rate above which a user is unlikely to have spread disinformation.

We use the open-source English function word lists provided by [12], and use the categories: personal pronouns, adverbs, and all function words. Distributions of f_c for these categories are given in Figure 1. We included all three features in our final classifier as cross-validation indicated they were useful, however the adverbs and personal pronouns are questionable features. The average number of tokens in a tweet after our tokenization method is just over 13. Yet the baseline usage rate for adverbs is 0.06 for adverbs and 0.04 for personal pronouns. Therefore, the tweets are not long enough to yield reliable statistics for these word categories. The complete function words category

[2] The labels apply to the 300 authors, each of which contribute 100 tweets to the dataset.

[3] In this respect, our choice to ensemble with XGBoost was questionable.

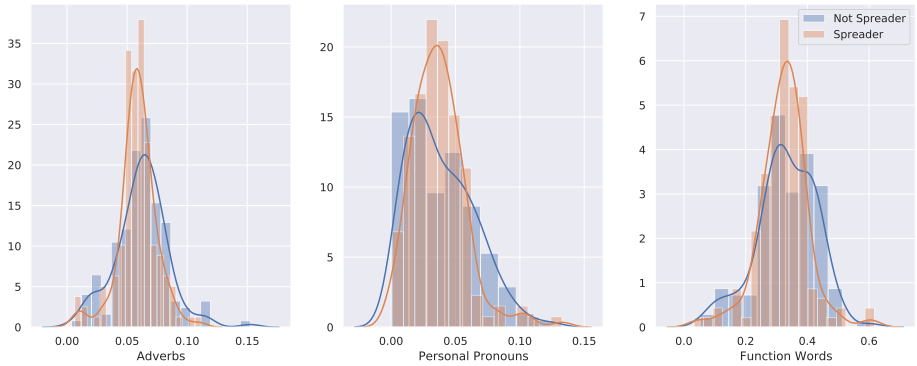

Figure 1. Distributions of usage rates of function word categories.

is more reliable, with baseline frequency of 0.33. There appears to be a threshold for function word usage rates above which the probability of spreading disinformation is low in this dataset. This threshold appears to be quite high compared to the rest of the distribution, suggesting this is not a feature with great power to solve this task on its own.

Constituency Tree Features Following the same intuition that sentence complexity indicates thoughtfulness, we investigated related features from constituency tree parses. Specifically, we calculated the average branching factor, and highest noun and verb phrases in each constituency tree. We consider a tweet as a single sentence, which is reasonable given that most tweets are headlines of news articles. From the constituency trees we extract the average branching factor, and average maximum noun and verb phrases as follows. Let $\mathcal{N}_i^{(u)}$ be the set of internal non-leaf nodes in the constituency tree for tweet i from user u. Let $\phi_{\text{bf}}(\cdot)$ count the number of children of a node. Then,

$$\mu_{\text{bf}}^{(u)} = \frac{1}{100}\sum_{i=1}^{100}\left(\frac{1}{|\mathcal{N}_i^{(u)}|}\sum_{n\in\mathcal{N}_i^{(u)}}\phi_{\text{bf}}(n)\right).$$

For noun and verb phrases, let $\mathcal{P}_i^{(u)}$ be the set of sub-trees representing either type of phrase, and let $\phi_{h_P}(\cdot)$ select the height of the tallest sub-tree in $\mathcal{P}_i^{(u)}$ of phrase type P. Then,

$$\mu_{h_{\text{NP}}}^{(u)} = \frac{1}{100}\sum_{i=1}^{100}\phi_{h_{\text{NP}}}(\mathcal{P}_i^{(u)})$$

$$\mu_{h_{\text{VP}}}^{(u)} = \frac{1}{100}\sum_{i=1}^{100}\phi_{h_{\text{VP}}}(\mathcal{P}_i^{(u)}).$$

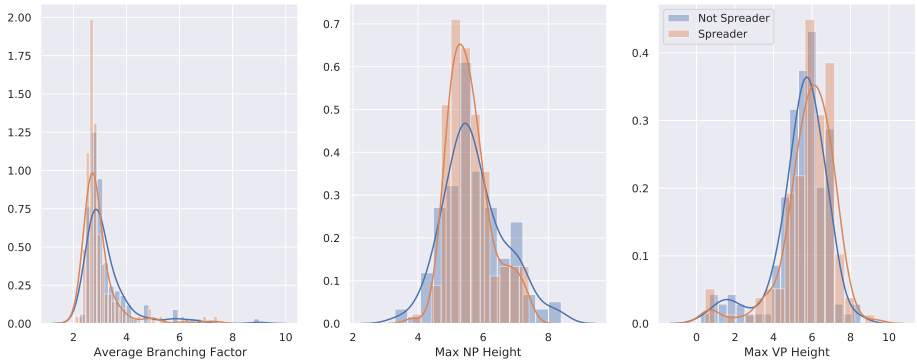

Figure 2. Distributions of constituency tree features.

The distributions over both labels are given in Figure 2. Once again the distributions are not very well separated, but indicate some weak support for our hypothesis that higher average branching factor and higher average noun phrases are more associated with the control group above a (fairly high) threshold. Interestingly, the pattern is opposite to what we expected for verb phrases, requiring further investigation.

2.2 Emotional Arousal

We initially investigated emotional arousal using the Linguistic Inquiry and Word Count (LIWC) dictionary [14]. However, due to being proprietary, we could not package it into our final software solution. We are also advocates of open science and would like to investigate alternatives to drive scientific progress, and provide resources equally available to less privileged researchers. Nevertheless, the value of the LIWC dictionary is beyond doubt, and still appears stronger than the alternatives we tried.[4]

Our LIWC analysis indicated the usefulness of: anger, negative emotion, and anxiety. The finding that anger is useful is consistent with previous work [8]. Using these three features, a support vector classifier could achieve 70% cross-validated accuracy on the training set. We therefore set ourselves the task of finding an equally good open source lexicon. We did not find one for anxiety, and so skipped that category.

SentiWordNet For negative emotions, we settled on SentiWordNet [1]. Figure 3 shows the distributions for sentiment (of both kinds), and negative sentiment. An unexpected result is that more sentiment is associated with the control group, however negative sentiment shows the expected pattern. Again, these features appear to have relatively weak discriminative power, especially compared to the LIWC lexicon.

[4] One exception is the function word categories, for which the lexicon from [12] appears to yield very similar distributions. This is consistent with their findings, obtained on a different task and dataset. Our findings in this respect therefore give extra weight to the claim that this is a good open source alternative for the LIWC function word lexicon.

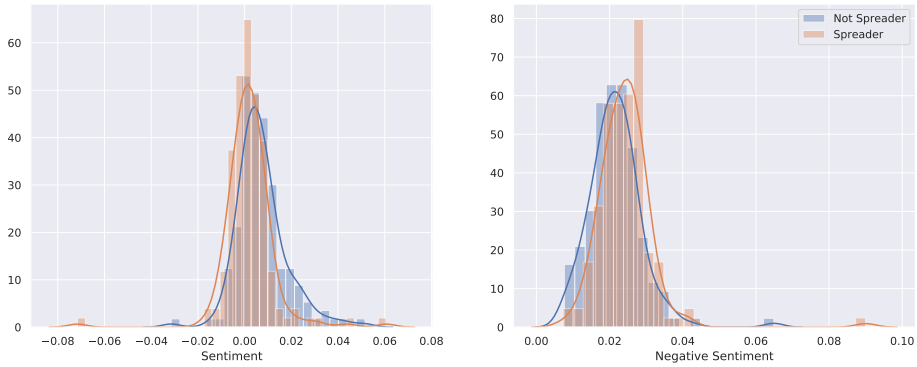

Figure 3. Distributions of sentiment and negative sentiment features.

Anger Of all the LIWC features, anger provided the clearest separation between the two classes, consistent with our understanding of how much disinformation works. Anger is known as an emotion that leads to action (REF). Therefore, if disinformation can bring someone to feel anger and outrage, it can promote its spread.[5] We did not find an open-source offering that could separate the two classes like LIWC does.[6] Given the importance of this feature, our broader goal of finding alternatives to LIWC, and general interest, we decided to experiment with making one. Our anger dictionary is far from a thorough or complete work. Our only criteria for success was a distribution that approximated the LIWC anger lexicon on this data. Since we used the training data to tune the lexicon, it also very likely overfits.

Inspired by previous work, we decided to exploit emoji on Twitter [2]. We used the twint library[7] to obtain tweets containing exclusively anger or joy emoticons in multiple 6-month time bins from the middle of 2020 to the end of 2014. We collection 20,000 tweets per emotion per time bin. We collected joy as a control, since in the average it is unlikely anger words will be present in joyous posts. The point of collecting over different times was to try and factor out the content words associated with anger at different times. We found "Trump" to be the most popular correlate of anger going back to before 2016, requiring a large number of timesteps to push him down our ranked list of anger words.

Utilizing a simple technique from social scientific research into media bias [7], we ranked words from this dataset by their χ^2 statistic comparing the frequency of each word's appearance in both anger and joy tweets. To account for variation over time, we scaled the χ^2 statistic by its entropy over time periods, promoting stable anger words, and demoting words whose correlation with each emotion is variable over time. The final results appeared quite reasonable, with the top five words being: angry, bloody, crap, disgusted, and unfair. We tuned the lexicon to the dataset by taking the top-50

[5] Taiwan's strategy for combating disinformation, "Humor over Rumor," is based on this understanding: humor being the best antidote to anger [16].

[6] Our search is unlikely to have been exhaustive.

[7] https://github.com/twintproject/twint.

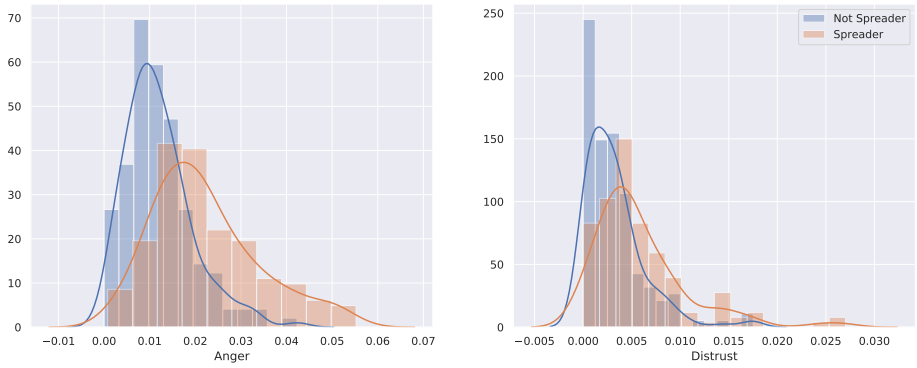

Figure 4. Distributions of anger and distrust features.

stems that were predictive of the "Spreader" class, as determined by average weight given to each word of a linear support vector classifer using 5-fold cross-validation. The resulting distribution of applying this lexicon is given in Figure 4. It does a much better job of distinguishing the two classes.

Distrust One recurrent theme in contemporary disinformation has been distrust in institutions. We therefore experimented with a creating a "distrust" lexicon. We used the χ^2 statistics of unigrams comparing their frequencies in the two class labels, and manually inspected the word list, pulling out words pertaining to distrust. The final lexicon includes 36 words such as: fake, suspect, unbelievable, doubt, question, lie, and scam. The resulting distributions are show in Figure 4. We generally see higher distrust scores for spreaders of disinformation, suggesting this could be a promising feature.

3 Results

We use NLTK [3] to tokenize the tweets, removing the special characters introduced by the dataset authors, such as "#URL#." The features described above are then calculated for each tweet, and averaged over users, as given in the equations above.

We separately trained support vector, random forest, and naive Bayes classifiers, using 5-fold cross-validation to find optimal hyperparameters, using scikit-learn [13]. The support vector classifier used an RBF kernel and $C = 1.$, achieving 69.7% accuracy on the training set. The random forest uses 40 estimators with a max depth of 5, achieving 68.3% accuracy. The naive Bayes achieved 64.3% accuracy. We then ensembled these classifiers with XGBoost [5], which achieved 5-fold cross-validated accuracy of 89.3% on the training set.

4 Discussion

It is very possible that the minor distributional differences in many of our features are an artifact of random sampling and fail to generalize, particularly those relating to thought-

fulness. However, cross-validation indicated their usefulness for this task, and the intuitions behind many of the features seem reasonable. It is likely our use of XGBoost has lead to overfitting due to the small number of labeled data points. We may also have overfit the training set when building our anger and distrust lexicons.

It would be much more interesting to compare these features across languages. Perhaps due in part to our unfamiliarity with Spanish, we were unable to easily find open-source resources to match what we have obtained for English.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010), http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf
2. Bandhakavi, A., Wiratunga, N., P, D., Massie, S.: Generating a word-emotion lexicon from #emotional tweets. In: Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014). pp. 12–21. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/S14-1002, https://www.aclweb.org/anthology/S14-1002
3. Bird, Steven, L.E., Klen, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
4. Butler, S., Hsu, I.: Q&A: Taiwan's digital minister on combatting disinformation without censorship. Committee to Protect Journalists (2019), https://cpj.org/2019/05/qa-taiwans-digital-minister-on-combatting-disinfor/
5. Chen, T., Guestrin, C.: Xgboost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Aug 2016). https://doi.org/10.1145/2939672.2939785, http://dx.doi.org/10.1145/2939672.2939785
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
7. Gentzkow, M., Shapiro, J.: What drives media slant? evidence from u.s. daily newspapers. Econometrica **78**, 35–71 (12 2006). https://doi.org/10.2139/ssrn.947640
8. Ghanem, B., Rosso, P., Rangel, F.: An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–18 (2020)
9. He, H., Zha, S., Wang, H.: Unlearn dataset bias in natural language inference by fitting the residual (2019)
10. Kitaev, N., Klein, D.: Constituency parsing with a self-attentive encoder. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia (July 2018)
11. Niven, T., Kao, H.Y.: Probing neural network comprehension of natural language arguments. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4658–4664. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1459, https://www.aclweb.org/anthology/P19-1459

12. Noble, B., Fernández, R.: Centre stage: How social network position shapes linguistic coordination. In: Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics. pp. 29–38. Association for Computational Linguistics, Denver, Colorado (Jun 2015). https://doi.org/10.3115/v1/W15-1104, https://www.aclweb.org/anthology/W15-1104
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
14. Pennebaker, J., Boyd, R., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015 (09 2015). https://doi.org/10.15781/T29G6Z
15. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (Sep 2020), CEUR-WS.org
16. Silva, S.: Coronavirus: How map hacks and buttocks helped taiwan fight covid-19. BBC (2020), https://www.bbc.com/news/technology-52883838