The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity

Fatemeh Torabi Asr
Discourse Processing Lab
Simon Fraser University
Burnaby, BC, Canada
ftorabia@sfu.ca

Maite Taboada
Discourse Processing Lab
Simon Fraser University
Burnaby, BC, Canada
mtaboada@sfu.ca

Abstract

Misinformation detection at the level of full news articles is a text classification problem. Reliably labeled data in this domain is rare. Previous work relied on news articles collected from so-called "reputable" and "suspicious" websites and labeled accordingly. We leverage fact-checking websites to collect individually-labeled news articles with regard to the veracity of their content and use this data to test the cross-domain generalization of a classifier trained on bigger text collections but labeled according to source reputation. Our results suggest that reputation-based classification is not sufficient for predicting the veracity level of the majority of news articles, and that the system performance on different test datasets depends on topic distribution. Therefore collecting well-balanced and carefully-assessed training data is a priority for developing robust misinformation detection systems.

1 Introduction

Automatic detection of fake from legitimate news in different formats such as headlines, tweets and full news articles has been approached in recent Natural Language Processing literature (Vlachos and Riedel, 2014; Vosoughi, 2015; Jin et al., 2016; Rashkin et al., 2017; Volkova et al., 2017; Wang, 2017; Pomerleau and Rao, 2017; Thorne et al., 2018). The most important challenge in automatic misinformation detection using modern NLP techniques, especially at the level of full news articles, is data. Most previous systems built to identify fake news articles rely on training data labeled with respect to the general reputation of the sources, i.e., domains/user accounts (Fogg et al., 2001; Lazer et al., 2017; Rashkin et al., 2017). Even though some of these studies try to identify fake news based on linguistic cues, the question is whether they learn **publishers' general writing style** (e.g., common writing features of a few clickbaity websites) or **deceptive style** (similarities among news articles that contain misinformation).

In this study, we collect two new datasets that include the full text of news articles and individually assigned veracity labels. We then address the above question, by conducting a set of cross-domain experiments: training a text classification system on data collected in a batch manner from suspicious and reputable websites and then testing the system on news articles that have been assessed in a one-by-one fashion. Our experiments reveal that the generalization power of a model trained on reputation-based labeled data is not impressive on individually assessed articles. Therefore, we propose to collect and verify larger collections of news articles with reliably assigned labels that would be useful for building more robust fake news detection systems.

2 Data Collection

Most studies on fake news detection have examined microblogs, headlines and claims in the form of short statements. A few recent studies have examined full articles (i.e., actual 'fake news') to extract discriminative linguistic features of misinformation (Yang et al., 2017; Rashkin et al., 2017; Horne and Adali, 2017). The issue with these studies is the data collection methodology. Texts are harvested from websites that are assumed to be fake news publishers (according to a list of suspicious websites), with no individual labeling of data. The so-called suspicious sources, however, sometimes do publish facts and valid information, and reputable websites sometimes publish inaccurate information (Mantzarlis, 2017). The key to collect more reliable data, then, is to not rely on the source but on the text of the article itself, and only after the text has been assessed by human

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 10–15
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

annotators and determined to contain false information. Currently, there exists only small collections of reliably-labeled news articles (Rubin et al., 2016; Allcott and Gentzkow, 2017; Zhang et al., 2018; Baly et al., 2018) because this type of annotation is laborious. The Liar dataset (Wang, 2017) is the first large dataset collected through reliable annotation, but it contains only short statements. Another recently published large dataset is FEVER (Thorne et al., 2018), which contains both claims and texts from Wikipedia pages that support or refute those claims. This dataset, however, has been built to serve the slightly different purpose of stance detection (Pomerleau and Rao, 2017; Mohtarami et al., 2018), the claims have been artificially generated, and texts are not news articles.

Our objective is to elaborate on the distinction between classifying **reputation-based** labeled news articles and **individually-assessed** news articles. We do so by collecting and using datasets of the second type in evaluation of a text classifier trained on the first type of data. In this section, we first introduce one large collection of news text from previous studies that has been labeled according to the list of suspicious websites, and one small collection that was labeled manually for each and every news article, but only contains satirical and legitimate instances. We then introduce two datasets that we have scraped from the web by leveraging links to news articles mentioned by fact-checking websites (Buzzfeed and Snopes). The distinguishing feature of these new collections is that they contain not only the full text of real news articles found online, but also individually assigned veracity labels indicative of their misinformative content.

Rashkin et al. dataset: Rashkin et al. (2017) published a collection of roughly 20k news articles from eight sources categorized into four classes: *propaganda* (The Natural News and Activist Report), *satire* (The Onion, The Borowitz Report, and Clickhole), *hoax* (American News and DC Gazette) and *trusted* (Gigaword News). This dataset is balanced across classes, and since the articles in their training and test splits come from different websites, the accuracy of the trained model on test data should be demonstrative of its understanding of the general writing style of each target class rather than author-specific cues. However, we suspect that the noisy strategy to label all articles of a publisher based on its reputation highly biases the classifier decisions and limits its power to distinguish individual misinformative from truthful news articles.

Rubin et al. dataset: As part of a study on satirical cues, Rubin et al. (2016) published a dataset of 360 news articles. This dataset contains balanced numbers of individually evaluated *satirical* and *legitimate* texts. Even though small, it is a clean data to test the generalization power of a system trained on noisy data such as the above explained dataset. We use this data to make our point about the need for careful annotation of news articles on a one-by-one fashion, rather than harvesting from websites generally knows as hoax, propaganda or satire publishers.

BuzzfeedUSE dataset: The first source of information that we used to harvest full news articles with veracity labels is from the Buzzfeed fact-checking company. Buzzfeed has published a collection of links to Facebook posts, originally compiled for a study around the 2016 US election (Silverman et al., 2016). Each URL in this dataset was given to human experts so they can rate the amount of false information contained in the linked article. The links were collected from nine Facebook pages (three right-wing, three left-wing and three mainstream publishers).[1] We had to follow the facebook URLs and then the link to the original news articles to obtain the news texts. We scraped the full text of each news article from its original source. The resulting dataset includes a total of 1,380 news articles on a focused topic (US election and candidates). Veracity labels come in a 4-way classification scheme including 1,090 *mostly true*, 170 *mixture of true and false*, 64 *mostly false* and 56 articles *containing no factual content*.

Snopes312 dataset: The second source of information that we used to harvest full news articles with veracity labels is Snopes, a well-known rumor debunking website run by a team of expert editors. We scraped the entire archive of fact-checking pages. On each page they talk about a claim, cite the sources (news articles, forums or social networks where the claim was distributed) and provide a veracity label for the claim. We automatically extracted all links mentioned on a Snopes page, followed the link to each original

[1] https://www.kaggle.com/mrisdal/fact-checking-facebook-politics-pages

Table 1: Results of the manual assessment of Snopes312 collection for items of each veracity label

Assessment / Veracity label	false	mixture	mostly false	mostly true	true	All
ambiguous	2	0	1	0	0	3
context	19	31	17	32	26	125
debunking	0	1	0	0	0	1
irrelevant	9	10	7	2	10	38
supporting	21	30	28	37	29	145
All	51	72	53	71	65	312

Table 2: Contingency table on disagreements between the first and second annotator in Snopes312 dataset

First annotator / Second annotator	ambiguous	context	debunking	irrelevant	supporting	All
ambiguous	0	0	0	0	0	0
context	1	0	1	8	71	81
debunking	0	0	0	0	1	1
irrelevant	0	36	0	0	16	52
supporting	0	11	1	0	0	12
All	1	47	2	8	88	146

news article, and extracted the text. The resulting datafile includes roughly 4,000 rows, each containing a claim discussed by Snopes annotators, the veracity label assigned to it, and the text of a news article related to the claim. The main challenge in using this data for training/testing a fake news detector is that some of the links on a Snopes page that we collect automatically do not actually point to the discussed news article, i.e., the source of the claim. Many links are to pages that provide contextual information for the fact-checking of the claim. Therefore, not all the texts in our automatically extracted dataset are reliable or simply the "supporting" source of the claim. To come up with a reliable set of veracity-labeled news articles, we randomly selected 312 items and assessed them manually. Two annotators performed independent assessments on the 312 items. A third annotator went through the entire list of items for a final check and resolving disagreements. Snopes has a fine-grained veracity labeling system. We selected *[fully] true*, *mostly true*, *mixture of true and false*, *mostly false*, and *[fully] false* stories. Table 1 shows the distribution of these labels in the manually assessed 312 items, and how many from each category of news articles were verified to be the "supporting" source (distributing the discussed claim), "context" (providing background or related information about the topic of the claim), "debunking" (against the claim), "irrelevant" (completely unrelated to the claim or distorted text) and ambiguous (not sure how it related to the claim). Table 2 provides information on the confusing choices: About 50% of the items received different category labels from the two first annotators. The first annotator had a more conservative bias, trying to avoid mistakes in the "supporting" category, whereas the second annotator often assigned either "supporting" or "context", and rarely "irrelevant". For the disagreed items, the third annotator (who had access to all outputs) chose the final category. Results in Table 1 are based on this final assessment. We use the "supporting" portion of the data (145 items) in the following experiments.

3 Experiments

In text classification, Convolutional Neural Networks (CNNs) have been competing with the TF-IDF model, a simple but strong baseline using scored n-grams (Le and Mikolov, 2014; Zhang et al., 2015; Conneau et al., 2017; Medvedeva et al., 2017). These methods have been used for fake news detection in previous work (Rashkin et al., 2017; Wang, 2017). For our experiments, we trained and tuned different architectures of CNN and several classic classifiers (Naive Bayes and Support Vector Machines) with TF-IDF features on Rashkin et al.'s dataset. The best results on the development data were obtained from a Support Vector Machine (SVM) classifier using unigram TF-IDF features with L2 regularization.[2] There-

[2] We used the same train/dev/test split as in Rashkin's paper. However, the performance of our SVM classi-

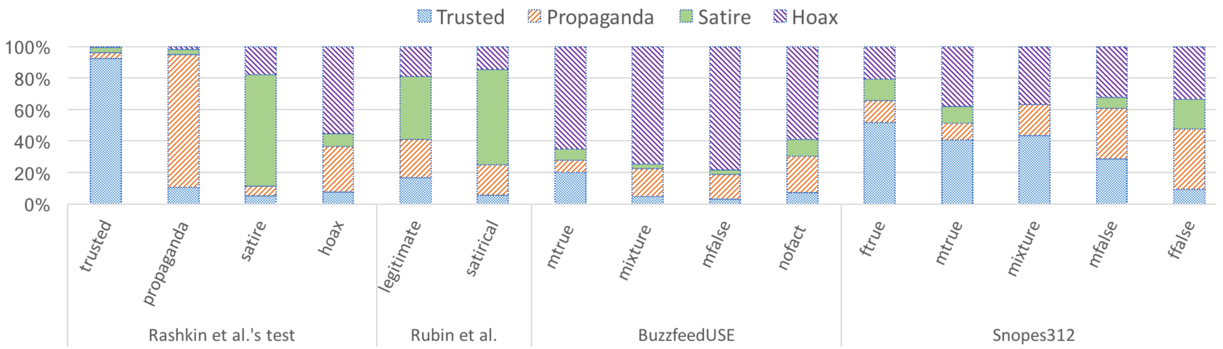

Figure 1: Classification of news articles from four test datasets by a model trained on Rashkin et al.'s training data. Labels assigned by the classifier are Capitalized (plot legend), actual labels of test items are in lowercase (x-axis).

fore, we use this model to demonstrate how a classifier trained on data labeled according to publisher's reputation would identify misinformative news articles.

It is evident in the first section of Figure 1, that the model performs well on similarly collected test items, i.e., *Hoax*, *Satire*, *Propaganda* and *Trusted* news articles within Rashkin et al.'s test dataset. However, when the model is applied to Rubin et al.'s data, which was carefully assessed for satirical cues in each and every article, the performance drops considerably (See the second section of the figure). Although the classifier detects more of the *satirical* texts in Rubin et al.'s data, the distribution of the given labels is not very different to that of *legitimate* texts. One important feature of Rubin et al.'s data is that topics of the legitimate instances were matched and balanced with topics of the satirical instances. The results here suggest that similarities captured by the classifier can be very dependent on the topics of the news articles.

Next we examine the same model on our collected datasets, BuzzfeedUSE and Snopes312, as test material. The BuzzfeedUSE data comes with 4 categories (Figure 1). The classifier does seem to have some sensitivity to true vs. false information in this dataset, as more of the *mostly true* articles were labeled as *Trusted*. The difference with *mostly false* articles, however, is negligible. The most frequent label assigned by the classifier was *Hoax* in all four categories, which suggests that most BuzzfeedUSE articles looked like *Hoax* in Rashkin's data. Finally, the last section of 1 shows the results on the Snopes312 plotted along the 6-category distinction. A stronger correlation can be observed between the classifier decisions and the veracity labels in this data compared to BuzzfeedUSE. This suggests that distinguishing between news articles with true and false information is a more difficult task when topics are the same (BuzzfeedUSE data is all related to the US election). In Snopes312, news articles come from a variety of topics. The strong alignment between the classifier's *Propaganda* and *Hoax* labels with the *mostly false* and *[fully] false* categories in this dataset reveals that most misinformative news articles indeed discuss the topics or use the language of generally suspicious publishers. This is an encouraging result in the sense that, with surface features such as n-grams and approximate reputation-based training data, we already can detect some of the misinformative news articles. Observing classification errors across these experiments, however, indicates that the model performance varies a lot with the type of test material: In a focused topic situation, it fails to distinguish between categories (false vs. true, or satirical vs. legitimate articles). While a correlation is consistently observed between labels assigned by the classifier and the actual labels of target news articles,[3] reputation-based classification does not seem to be sufficient for predicting the veracity level of the majority of news articles.

4 Conclusion

We found that collecting reliable data for automatic misinformation detection at the level of full news articles is a challenging but necessary task for building robust models. If we want to benefit

fier was significantly better on both dev and test sets: 0.96 and 0.75 F1-score, respectively, compared to 0.91 and 0.65 reported in their paper. Source code will be made available at https://github.com/sfu-discourse-lab/Misinformation_detection

[3] A chi-square test indicates a significant correlation ($p < 0.001$) between assigned and actual labels in all four datasets.

from state-of-the-art text classification techniques, such as CNNs, we require larger datasets than what is currently available. We took the first steps, by scraping claims and veracity labels from fact-checking websites, extracting and cleaning of the original news articles' texts (resulting in roughly 4,000 items), and finally manual assessment of a subset of the data to provide reliable test material for misinformation detection. Our future plan is to crowd-source annotators for the remaining scraped texts and publish a large set of labeled news articles for training purposes.

Acknowledgement

This project was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. We would like to thank members of the Discourse Processing Lab at SFU, especially Yajie Zhou and Jerry Sun for their help checking the datasets.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 21–27.

Alexis Conneau, Holger Schwenk, Loc Barrault, and Yann LeCun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1107–1116, Valencia.

B. J. Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and Marissa Treinen. 2001. What makes web sites credible? A report on a large quantitative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 61–68, New York.

Benjamin D Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2972–2978, Phoenix.

David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy*, 2.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages II–1188–II–1196, Beijing.

Alexios Mantzarlis. 2017. Not fake news, just plain wrong: Top media corrections of 2017. *Poynter News*. Https://www.poynter.org/news/not-fake-news-just-plain-wrong-top-media-corrections-2017.

Maria Medvedeva, Martin Kroon, and Barbara Plank. 2017. When sparse traditional models outperform dense neural networks: The curious case of discriminating between similar languages. In *Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 156–163, Valencia.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Llus Mrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, LA.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge. Http://fakenewschallenge.org/.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2921–2927, Copenhagen.

Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of NAACL-HLT*, pages 7–17, San Diego.

Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. *BuzzFeed News*. Https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, LA.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 647–653, Vancouver.

Soroush Vosoughi. 2015. *Automatic detection and verification of rumors on Twitter*. Ph.D. thesis, Massachusetts Institute of Technology.

William Yang Wang. 2017. 'Liar, liar pants on fire': A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 422–426, Vancouver.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989.

Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 603–612. International World Wide Web Conferences Steering Committee.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 649–657, Montréal.