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 assessedtwl2391There-m
w,poposw,collec,and vrify,gr,com-
of news articleswsrelitabys essgined
thatwsuldsusefulsbaingroba
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.

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

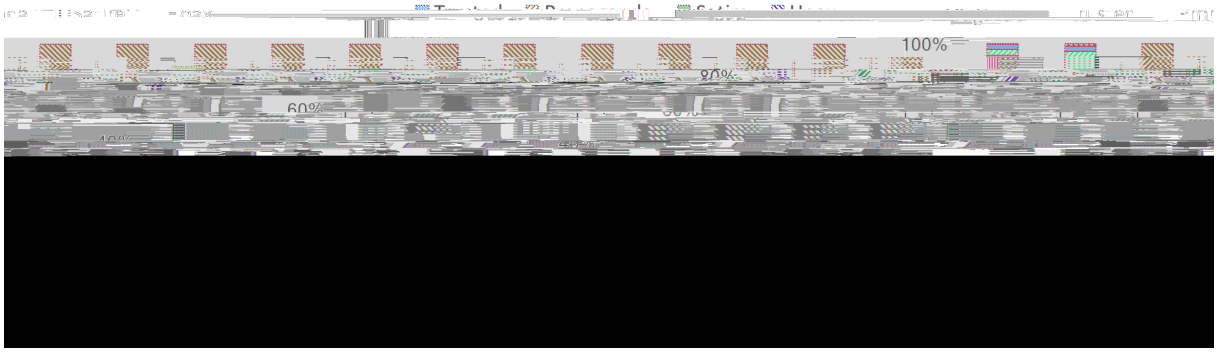

Figure 1: Classification of news articles from four test datasets by a model trained on Rashkin et al.'s training data. Labels assigned by the classifier are Capitalized (plot legend), actual labels of test items are in lowercase (x-axis).

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*

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 Ya-

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)*