Heinz 95-845: Project Proposal

Alton Lu altonl/lualton@cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Sam Hobbs shobbs/shobbs@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Michael Pan zpan1/michaelpan@cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

1. Project Details

There is a rapidly growing interest in detecting misleading news information, whether it come from dubious media sources or from politician's statements on twitter. While misleading or false information is nothing new, social media allows misleading information to proliferation widely and quickly.

The creation of misleading or false news has been a big issue from the 2016 United States Presidential election, to the Brexit vote in the UK and then during the 2018 United States midterm elections. While governments have been working to find a legislative solution to these issues, much of the responsibility has fallen to the technology companies where these misleading stories spread.

Facebook and Twitter have attempted varying tactics at limiting the spread of fake news on their platforms. Twitter removed thousands of accounts related to discouraging voters in the 2018 midterm election. Facebook has also been seeing some success in reducing fake news.

However, none of these attempt are perfect. Misleading news stories can come in much more sophisticated forms. Continued research in this area is important. There are several attempts at producing a fake news classifier based on a variety of different context. This project will expand on these efforts. We have two possibilities and welcome thoughts about the scope and difficulty to fit within the rest of this course.

1. Our first proposed project is to combine two data sets and re-split around test and train. These two data sets include 1) full news stories and 2) single statements. One of the chief difficulties in detection of fake news is in the different domain and structures of the text. It's difficult to generalize one case to another because the data is far too different. We consider the possibility of mixing these two to create a more generalized model that can detect both fake news stories and false statements from politicians or twitter users.

2. Our second is to do a comparison study across different data sets. The two data sets we've found are the BuzzFeed and Getting Real data sets that both include full news

stories. Most studies only focus on a single corpus of data, but we would investigate the training and testing on one of the data sets and extend to predict across the second. This would allow us to test the efficacy of prediction to completely different news stories, different publishers, and different writing styles as the two data sets differ widely in publisher.

3. If the previous two are considered too large, our final objective is to take the statements data set and improve on the predictive analysis that Wang 2018 had. This would include the use of meta-data associated with politicians and adding credibility metrics included from PolitiFact (if someone lied a lot in the past, do they tend to lie again?).

1.1 Objectives

Our objective is to develop a predictive model that can assign a probabilistic score of veracity for a news article or statement. Articles and statements are not binary in truthfulness, so we want to represent the spectrum of truth. A working context for our objective is to have a 'veracity' score next to articles on Facebook or Twitter that users can see to determine the value of news.

As this is a classification task, we will use many of the similar machine learning techniques as referenced papers to ensure a consistent baseline. These include a regularized logistic regression classifier, a support vector machine classifier, a tree-based model, and a neural net.

We have a few potential datasets still being considered depending on the task. These data are outlined in section 2.4.

There are many possible limitations. The first is just in the difficulty of the task. It's fairly easy to distinguish from obvious fake news that include a lot of extreme language. But the type of fake news we'd ideally like to target are the half-truth that are difficult to distinguish. We foresee the possibility of getting some good accuracy score, yet having poor accuracy among 'hard-to-distinguish' classes, which defeats the purpose of our project. We are also aware of the possible mistakes in the data. Labeling these will naturally come with inherent biases and these could impact results.

This work will extend current research to consider more data. Most papers over the past year concentrate on the development of novel data and set an example of a predictive task, but none look across different data, structures, and domains. This paper will be a extension in considering the generalizability across data.

1.2 Parameters

We've stated previously multiple different possibilities. However, all of this is to implement the ideas present in Shu et al. 2017. Shu's paper described in great detail, possible methods for detecting fake news. Our analysis will actually be implementing those methods, although the specific data set used is not finalized yet. These analysis will be completely new, as Shu's paper only described the methodology.

Shu defines the problem in terms of News and Social features. The News includes the content of the article/statement and publisher meta-data. The social features include some measure of engagement or reaction, such as Twitter shares or Facebook reactions.

We are planning to run most processing in Python but will switch to R for some of the predictive analysis.

2. Proposal Details (10 points)

Please provide information for the following fields. Your proposal write-up should be no more than 2 pages.

2.1 What is your proposed analysis? What are the likely outcomes?

Our proposed analysis is to develop a classification score of misleading news using new and different linguistic and social features. The likely outcomes are that we can develop features, but accuracy remains quite low. Most benchmark papers are achieving 60 percent on normal text data sets and typically lower as the text get shorter.

2.2 Why is your proposed analysis important?

For any type of policy-making or to run a functioning democracy, the public requires a baseline of facts to learn and properly discuss. Information that is misleading, both unintentional and intention, limit public discourse. Moreover, the development of misleading information by countries outside the United States further threaten the public discourse.

Ensuring that misleading and fake news is properly flagged or removed is important to provide the baseline of facts for a functioning and educated public.

2.3 How will your analysis contribute to existing work?

There are a few papers that we will be extending. The primary work that we consider is from Conroy et al. 2015. This paper described a method of combining linguistic features and network methods for automatic detection of fake news. This was a survey paper that described a possibility of detection.

Further work was continued from Shu et al. 2017 which expanded heavily from Conroy to describe in more depth the possibility of detection fake news. The paper itself broke down the problem into content-based models and social-based models, arguing that the two needed to work together to develop the most accurate models. This paper also described the main datasets we've found.

Two papers that have attempted some of these predictive tasks are Wang 2018 and Perez-Rosas et al. 2017. Both papers described a dataset and the pipeline for prediction. Perez-Rosas attempted to develop more features that would increase predictive power while Wang attempted predicting with only surface-level linguistic representations. Wang achieved around 27 percent accuracy with Perez-Rosas had 65 percent accuracy. Both papers have large space for improvements.

2.4 Describe the data

We have four possible datasets.

1. BuzzFeed Fake News Corpus 2016

This is a corpus of news stories from nine publishers within a week of the US 2016 presidential election. It includes 3 known hyper-partisan left-wing publishers, 3 hyper-partisan right-wing publishers, and three main stream publishers. This was published in February 20, 2018.

2. BuzzFace 2018

This is a corpus created in Santia et al. 2018 that appends additional Facebook information to the BuzzFeed corpus. We are still debating whether to use this because our goal is to use linguistic cues to generalize to different types of text. The additional Facebook information may not be generalizable to other platforms like Twitter.

3. Liar

This corpus was created by Wang 2018 and utilizes data from PolitiFact. It contains 12.8k statements from politicians and classified on a scale of pants fire, false, barely true, half-true, mostly true, true. These are short statements that come from speeches, Facebook, Twitter, press releases, etc. and thus represent text from a wide-variety of contexts. This is particular valuable as it includes statements that are prepared (press releases) as well as potential off-the-cuff statements from interviews. We would like to combine this with another data set to generalize across different types of fake news.

4. Getting Real

This corpus is provided from Kaggle and is developed from a combination of the B.S. detector on Chrome and OpenSources. It contains 12,999 posts from 244 websites that have been labeled by the BS detector. It is more limited in scope as it only has stories labeled as false. Adding a normal new corpus, such as the Reuters dataset, might make this corpus more valuable.

2.5 What evaluation measures are appropriate for the analysis? Which measures will you use?

Our primary evaluation metric will be accuracy. A secondary goal is to identify the hard-to-classify text and measure accuracy based on those data points.

2.6 What study design, pre-processing, and machine learning methods do you intend to use? Justify that the analysis is of appropriate size for a course project.

The primary time of this project will be around pre-processing. We will replicate part of the processing from Perez-Rosas 2017 in linguistic features. These include calculating ngrams, punctuation, sentiment, readability, and syntactic POS tagging.

As the datasets are quite large and text data is unwieldy, we may be forced to work from a subset of the data. We think this project is challenging, but doable.

2.7 What are possible limitations of the study?

Limitations to the study include the difficulty of the task, generalizability, computing power, and our own skill sets. NLP is difficult challenge, even more so when predicting difficult content.

2.8 Who will your analytic pipeline? In one or two sentences, describe an example of its use.

Our working context is Facebook and Twitter feeds. Our imagined scenario is to develop a model that can take in any type of text and produce a truth score that is shown alongside

the news story or press release whenever it is shared on Facebook or Twitter. This would enable users to see and evaluate the story on their own.

References

[1] Amy Mitchell et al. "Americans Favor Protecting Information Freedoms Over Government Steps to Restrict False News Online," Pew Research.

[2] Mallory Locklear. "Researchers say Facebooks anti-fake news efforts might be working," Engadget.

[3] William Yang Wang. "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," Annual Meeting of the Association for Computational Linguistics 2017.

[4] Veronica Perez-Rosas et al. "Automatic Detection of Fake News," arXiv:1708.07104

[5] Giovanni C. Santia et al. "BuzzFace: A News Veracity Dataset with Facebook Users Commentary and Egos." Proceedings of the Twelfth International AAAI Conference on Web and Social Media.

[6] Niall. J. Conroy et al. "Automatic Deception Detection: Methods for Finding Fake News," ASIST 2015.

[7] Kai Shu et al. "Fake News Detection on Social Media: A Data Mining Perspective," Association for Computing Machinery's SIGKDD.