Project Proposal

Team 8
Enze Yuan 1155139040
Xiaoyue Cai 1155137702
Jun Xiao 1155132409
Ruding Tan 1155137502

Executive Summary

The rate and volume at which fake news are spread overturn the possibility of efficient fact checking.Additionally,creating tools for automatic detection is as challenging due to lack of dataset containing articles which present fake or manipulated stories as compelling facts.At the same time,many
governments and technology companies begin to take actions and some related competitions were held to tackle this problem. In the last three years,most of the research in this area are developed along with the publishing of some new techniques and methods.In our project, we consider a classification task on the Fake News Corpus dataset,one of the open source datasets on github.Some algorithms or techniques such as SVM will be used in our project,and we will also deploy some related course content(e.g.clustering and classification, text processing and NLP) to our project.A machine learning model will also be built to aid us solve the problem.

Background and Motivation

Fake news is news, stories or hoaxes created to deliberately misinform or deceive readers, which was named Collins 2017's Word of the year.[1]

Usually, these stories are created to either influence people's views, push a political agenda or cause confusion and can often be a profitable business for online publishers.[2] Because of the high velocity of propagation through the Internet and lack of corresponding regulation, the phenomenon that a lot of disinformation mixed in other news actually affects our attention and decision. For instance, in the months leading up to the 2016 American election, the top 20 fake news stories had more shares, reactions, and comments on Facebook (8.7 million engagements) than the 20 top hard news stories (7.3 million engagements), according to a Buzzfeed analysis.[3] Under this circumstance, People find it's hard to recognize true news, they will lose confidence in the media and easily misled by the fake. As Ogilvy's Global Media Influence Survey mentioned, journalists believe that the public's trust in traditional media as a source for news has declined 22% since 2016.[4]

With more and more fake news on the Internet, governments and technology companies began to take actions. In May 8 2019, Singapore officially just passed its new anti-fake news law, called The Protection from Online Falsehoods and Manipulation Act in Parliament with an overwhelming majority from the House.[5]

According to naive realism and the confirmation bias theories,as a human, we have no natural expertise at distinguishing real from false news.And this defect of humanity may have great impacts on some real world issues such as politics and finance. So it is always interesting to look for a way to tell lies from truth.In our project,we just narrow the big problem down to a problem in a specific region:news. Our motivation for this problem is to design a method which focuses on automatically fact check..

Related Course Topics

We intend to use a classification-approach to detect fake news, and the method will be content-based, meaning that our analysis will mostly depend on textual content of news articles, e.g. body, title, URL. By formulating fake news detection as a two-class or multi-class classification problem, the main concern of this approach is to find effective features for training classifier ,hence we need to deploy techniques or algorithms that can be extracted features from textual content to our project. So two course topics are regarded to be closely related to our project, namely Clustering and Classification and Text Processing and NLP.

Besides,for what concerns the content-based classification method, many techniques/tools are used in previous research contributions, such as traditional machine learning (Logistic Regression, SVM), Matrix Factorization. These are also the contents we will learn in Big Data Analysis course, and we will try to use these techniques/tools in our project. Given that we roughly halfway through the first half the syllabus, we believe there may be more related course content that we can use in our project as the course proceeds.

Algorithms and Techniques

Statistics: We will select several statistic features for string pair matching based on common sets of words, stop words, tokens, characters and so on.This feature set contributes some predictive features in the top of most important features selected by tree based models.

NLP techniques: For textual inputs, we apply traditional NLP techniques such ngrams of characters or words to vectorize text into numerical space. The news title pair then is transformed using pairwise distances such as cosine, euclidean, city-block, jaccard or just simply using summation or subtraction.

SVM:We may use features extracted from the features that we pre-defined to train a SVM classifier.

Machine Learning:Some previous research contributions have adopted several machine learning methods,such as logistic regression,so it might be useful in our project.

BERT:It is a model created by Google to preprocess the data. We might use it in our pre-training round .This model can significantly improve the accuracy of other data mining models. Logistic Regression (LR) might be used as the reduced model to reduce all the outcomes to the final solution.

Deliverables and Demonstration

By the end of the project, we plan to test our method on collection of articles gathered from Sina Weibo and Twitter and compare the performance of our method with other models. We are inspired by Zhiwei Jin et a(2016)[7]. and consider two performance measures:

- **Accuracy** is the percentage of correctly identified fake and real news. It is an

overall measurement;
- **Precision and recall** for fake news and real news, respectively, represent a model's effectiveness on identifying each class.

So basically we will submit a test report as the deliverable. Hopefully, we could build an UI in form of a website or a Little Program based on Wechat to let others test questionable news on it. Thus we probably will demonstrate our analysis result by using a screen recorder to record a demo video of our UI.

Existing Works

Most of the research on false news has developed in the last 3 years as a helpful first step towards fake news detection was introduced during the 2017 Fake News Challenge Stage 1(FNC-1) [8]and BERT[9] was published by researchers at Google as a powerful tool in NLP tasks in 2018.

Traditionally, fake news detection problem has been formulated as a classification problem. Previous classification-based approaches can be categorized into 3 types according to the features they use, namely content-based, context-based and content and context-based.

As for content-based ones, machine learning methods and deep learning approaches (Convolutional and Recurrent Neural Networks) are commonly used. Neural networks are employed by three top-performing systems[10,11,12] in FNC-1 to classify the instance of an entire news article relative to its headline and also used by Popat et al.(2018)[13]. to build a framework to classify true and false claims. Researchers such as Horne et al.(2017)[14], Potthast et al.(2018)[15] and Fairbanks et al.(2018) [16]solved the classification problem by considering several machine learning methods. Other techniques like tensor decomposition, loopy belief propagation are adopted by some researchers to tackle the problem.

As for context-based ones, they focus on information derived from social interactions between users, e.g. likes, comments and (re)tweet, as to detect fake content. Based on different datasets, researchers proposed several methods, but machine learning and deep learning are the most commonly used techniques in those approaches.

Other researchers consider both news content and the associated context interactions as to detect malicious information items(2019)[17].

In terms of the datasets, there are very limited training datasets nowadays as social media impose limitations on the collection of public data. Among the previously released datasets for fake news, most notably BuzzFeed News dataset contains news articles in English during 2016 US presidential elections, however it only consists of 26 known fake articles and 679 articles , meaning that dataset might not be big enough for our project. Some open source datasets like BS detector contains articles annotated by news veracity detector tool and hence, cannot be trusted, as the labels are not verified manually and there are many anomalies like movies or food reviews.

As for the dataset we used in project, Fake News Corpus, we haven't spotted any related search paper yet.So it might lead to the difference between our project and existing work,because some methods or approaches may not be practical to examine our dataset given that those approaches or methods are trained with dataset with different characteristics.

Dataset

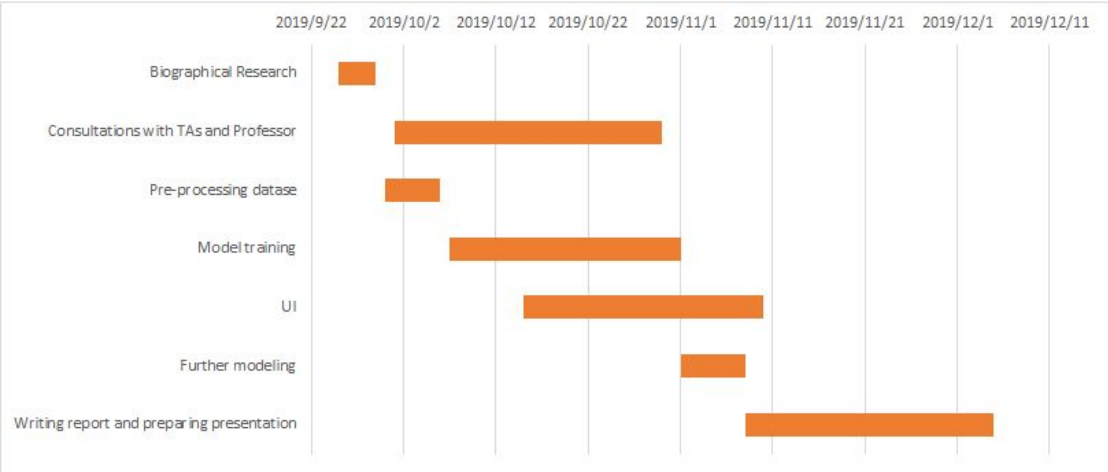
We will consider a problem about identifying false news on an open source datasets called Fake News Corpus, which contains millions articles scraped from a curated list of 1001 domains from http://www.opensources.co/.

This dataset contains 9,408,908 articles which are categorized into 11 categories (e.g. Fake news, Satire, Extreme Bias, Credible) and the volume of it is about 9.1GB.The dataset is formatted as csv and contains 16 fields(e.g.id, domain, URL, content, title).As the diagram shows,the dataset is unbalanced since the number of credible news is about two times of that of fake news.Before we build our model,we will produce a balanced dataset by using Hadoop to pre-process the original one.

Type	Tag	Count (so far)	Description
Fake News	fake	928,083	Sources that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports
Satire	satire	146,080	Sources that use humor, irony, exaggeration, ridicule, and false information to comment on current events.
Extreme Bias	bias	1,300,444	Sources that come from a particular point of view and may rely on propaganda, decontextualized information, and opinions distorted as facts.
Conspiracy Theory	conspiracy	905,981	Sources that are well-known promoters of kooky conspiracy theories.
State News	state	0	Sources in repressive states operating under government sanction.
Junk Science	junksci	144,939	Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
Hate News	hate	117,374	Sources that actively promote racism, misogyny, homophobia, and other forms of discrimination.
Clickbait	clickbait	292,201	Sources that provide generally credible content, but use exaggerated, misleading, or questionable headlines, social media descriptions, and/or images.
Proceed With Caution	unreliable	319,830	Sources that may be reliable but whose contents require further verification.
Political	political	2,435,471	Sources that provide generally verifiable information in support of certain points of view or political orientations.
Credible	reliable	1,920,139	Sources that circulate news and information in a manner consistent with traditional and ethical practices in journalism (Remember: even credible sources sometimes rely on clickbait-style headlines or occasionally make mistakes. No news organization is perfect, which is why a healthy news diet consists of multiple sources of information).

The limitation of the dataset is that the dataset was manually filtered, so some of the labels and URLs might be wrong, but it won't be a problem if we could use it to train the machine learning algorithm we build. We noticed that provider of the dataset claimed that he would not update it after the whole dataset was cleaned and published. We don't think it will not pose a practical issue given that we want to adopt a content-based algorithm to address the FNC(Fake News Detection)problem.

Project Timeline



Reference:

1. **Carson. J. (2019). Fake news: What exactly is it – and how can you spot it?. [online] The Telegraph. Available at: https://www.telegraph.co.uk/technology/0/fake-news-exactly-has-really-had-influence/ [Accessed 30 Sep. 2019].**
2. **Webwise.ie. (2019). Explained: What is Fake news? | Social Media and Filter Bubbles. [online] Available at: https://www.webwise.ie/teachers/what-is-fake-news/ [Accessed 30 Sep. 2019].**
3. **Kurtzleben. D. (2018). *NPR Choice page*. [online] Npr.org. Available at: https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election [Accessed 30 Sep. 2019].**
4. **McCarthy. J. (2018). *Journalists believe the public trusts traditional media 22% less than in 2016*. [online] The Drum. Available at: https://www.thedrum.com/news/2018/06/18/journalists-believe-the-public-trusts-traditional-media-22-less-2016 [Accessed 30 Sep. 2019].**
5. **Lim. S. (2019). *Singapore passes anti-fake news law after overwhelming majority of MPs vote in favour*. [online] The Drum. Available at: https://www.thedrum.com/news/2019/05/09/singapore-passes-anti-fake-news-law-after-overwhelming-majority-mps-vote-favour [Accessed 30 Sep. 2019].**
6. **Pham. L.(2019) *Transferring, Transforming, Ensembling: The Novel Formula of Identifying Fake News*.[online] Available at: https://people.eng.unimelb.edu.au/jianzhongq/wsdm19-cup-reports/reports/report1.pdf [Accessed 30 Sep. 2019]**
7. **Jin. Z, Cao. J, Zhang.Y and Luo.J.(2016) *News verification by exploiting conflicting social viewpoints in microblogs.* [online] Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12128/12049 [Accessed 30 Sep. 2019].**

8. **Pomerleau. D and Rao. D.(2017)** ***Fake news challenge.*** **[online] Available at: http://www.fakenewschallenge.org [Accessed 30 Sep. 2019].**
9. **Devlin. J, Chang. MW, Lee. K, Toutanova. K.(2018)** ***Bert: Pre-training of deep bidirectional transformers for language understanding.*** **[online] Available at: https://arxiv.org/pdf/1810.04805.pdf [Accessed 30 Sep. 2019].**
10. **Baird. S, Sibley. D, and Pan. Y.(2017)** ***Talos targets disinformation with fake news challenge victory.*** **[online] Available at: https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html [Accessed 30 Sep. 2019].**
11. **Hanselowski. A, Avinesh. P, Schiller. B, and Caspelherr.F. (2017)*****Description of the system developed by team athene in the fnc-1.*** **[Accessed 30 Sep. 2019].**
12. **Riedel. B. et al.(2017)** ***A simple but tough-to-beat baseline for the fake news challenge stance detection task.*** **[online] Available at: https://www.semanticscholar.org/paper/A-simple-but-tough-to-beat-baseline-for-the-Fake-Riedel-Augenstein/95a72b91db11f33384fd7b159561bf182c8c8d95 [Accessed 30 Sep. 2019].**
13. **Popat. K. et al.(2018)** ***Debunking fake news and false claims using evidence-aware deep learning.*** **[online] Available at: https://people.mpi-inf.mpg.de/~kpopat/publications/emnlp2018_Popat.pd [Accessed 30 Sep. 2019].**
14. **Horne. B. D and Adali. S.(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.*** **[online] Available at: http://homepages.rpi.edu/~horneb/FakenewsPoster2016_34inwide_banded.pdf [Accessed 30 Sep. 2019].**
15. **Potthast. M. et al.(2018)** ***A stylometric inquiry into hyperpartisan and fake news.*** **[online] Available at: https://www.aclweb.org/anthology/P18-1022 [Accessed 30 Sep. 2019].**
16. **Fairbanks. J. et al.(2018)** ***Credibility assessment in the news: Do we need to read?*** **[online] Available at: https://snap.stanford.edu/mis2/files/MIS2_paper_17.pdf [Accessed 30 Sep. 2019].**
17. **Pierri F, Ceri S.(2019)** ***False News On Social Media: A Data-Driven Survey.*** **[online] Available at: https://www.researchgate.net/publication/331246175_False_News_On_Social_Media_A_Data-Driven_Survey [Accessed 30 Sep. 2019].**