UNIVERSITY OF OTTAWA

FakeDetect – Fake News Detection

Project Proposal

Gurpreet Singh

(8908872)

Department of Electrical Engineering and Computer Science,
University of Ottawa, Ontario, Canada

gsing082@uottawa.ca

Submitted to Prof. Diana Inkpen

University of Ottawa

Major Project Proposal

FakeDetect - Fake News Detection

Problem Statement: - To detect the fake and biased news from the news articles

Motivation: - In this age of social media and the online news, the negative effects of the fake news spread has increased manifolds. We all have witnessed the negative political and the business impacts of the fake news many times. The 2016 US Presidential elections is one of the well know example of the same. Reading news online is too habitual these days that if the proportion of the fake news keeps on increasing, it is going to have a tremendous negative results on the world directly or indirectly.

Dataset: - There are multiple resources which publish information about the news domain like OpenSources (http://www.opensources.co/), PolitiFact and MediaBiasFactCheck.com. The data from the MediaBiasFactCheck.com contains more than 2000 media resources which are classified into categories such as –

questionable,

right – strongly biased toward conservative causes,

right-center – moderate conservative bias,

least-biased – minimal bias,

left – strongly biased toward liberal causes,

left- center – moderate liberal bias

conspiracy – unverifiable information that is not always supported by evidence

Firstly, I will try to aggregate the datasets from the three sources mentioned above. In case I am not able to do so or if it won't be much useful, then I will use either the dataset from the politifact.com or MediaBiasFactCheck.com independently.

Methodology: - I will crawl a good number of domains to compile the articles, large enough to provide accurate solutions further. Pre-processing of the articles like removing those containing the video files, removing comments etc. will be the next step. I am planning to train the model using non-linear SVM using k fold cross-validation but that will keep on varying depending upon the accuracy of the results we get.

Entering the input as the article will generate the list of domains whose contents have the overlap with the entered article. I am planning to use the n-gram search for this. Details like questionable, fake, bias etc can also be added along with the domains to give the highlight about the kind of article. Multiple features like POS tags, average sentence, punctuation marks etc. will be considered to train the model. I will work on more features which we can put to increase the accuracy of the results.

Dependencies:-

Scikit-learn for modelling

Submitted to Prof. Diana Inkpen

NLTK for pos tagging

Pattern for Polarity and Profanity features.

Again, all the tools and methodology might keep on varying as we proceed in the project.

Evaluation Measures: - Evaluation measures such as Precision, Recall and F-score will be used to give the classification report of the final model.

One more thing, we can do is to use different combinations of the features to train the model and look at the accuracy one by one to determine the best subset of the features. Let's say we have 5 features in total F1, F2, F3, F4 and F5, we can then use multiple combination of different sizes like (F1+ F2+ F3) as one subset, (F1+F3+F5) as another features subset and many more. It will help us in determining the most optimal subset of features to train the model.

Possible Outcome: - We will be able to classify the news snippet as fake, biased or genuine. We can use boxplots to visualize the similarity score/distance between the news snippet and the reliable source. The more the distance, the less is the similarity. Other visualizations can also be considered to display some user friendly results.

Evaluation Scheme

The evaluation will be based on the two deliverables:
1) Implementation and testing;
2) Technical report (problem description, methods, datasets, evaluation experiments and discussion of the results; conclusion and future work)

Evaluation Breakdown

The marking scheme will be:

- Technical report 50% of the mark.
- Implementation 50% of the mark.

Submitted by:

Name : Gurpreet Singh

Student No.: 8908872

School of Electrical Engineering and Computer Science

Submitted to:

Dr. Diana Inkpen

Dr. Diana Inkpen

Date: Sept 14, 2018

References :-

[1] https://github.com/clips/news-audit

[2] https://medium.com/@vishwanigupta/combating-fake-news-with-the-help-of-machine-learning-b56537609564

[3] https://medium.com/@vishwanigupta/kpca-skip-gram-model-improving-word-embedding-a6a0cb7aad49