SI630 Project Proposal

Yunzhe Jiang

Abstract

The project is aimed at developing models for propaganda classification.

1 Introduction

Propaganda is used to promote or publicize a particular political cause or point of view,especially of a biased or misleading nature. Psychological and rhetorical techniques are applied in propaganda to make it work. Propaganda seems correct at the first sight. However, that the conclusion it makes sounds convincing is usually due to the misuse of logic or arousal of audience's emotion. All of those techniques are intended to go without being noticed to maximize its effect. Therefore it is important to detect propaganda and figure out what specific techniques that are being used.
By successfully detecting and classifying propaganda, people can look at information with more ration and logic. People or organizations in news and media industry could use this propaganda detector and classification to help them identify potential language traps. In addition, this classifier can be useful to educational institutes for teaching students about propaganda.

2 Problem Definition

The problem is to identify the what propaganda technique is used in the fragment, when given a text fragment considered as propaganda and its document context, Spans overlap for some fragments, so formally this is a a multi-label multi-class classification problem. However, here in this case, if multiple techniques are used in a span, the input file will be copied for multiple time of such fragments. Thus the problem can be algorithmically treated as a multi-class classification problem. The data has been originally annotated with 18 techniques. However, considering the relatively low frequency of some of the techniques, similar underpresented techniques are merged into one class:

- Bandwagon and Reductio ad Hitlerum into "Bandwagon,Reductio ad Hitlerum"
- Straw Men, Red Herring and Whataboutism into "Whataboutism,Straw Men,Red Herring"

Also techniques "Obfuscation,Intentional Vagueness,Confusion" are eliminated. Therefore this is a 14-classes classification task and the following techniques are considered:

- Loaded Language
- Name Calling, Labeling
- Repetition
- Exaggeration, Minimization
- Doubt
- Appeal to fear/prejudice
- Flag-Waving
- Causal Oversimplification
- Slogans
- Appeal to Authority
- Black-and-White Fallacy
- Thought-terminating Cliches
- Bandwagon, Reductio ad Hitlerum
- Straw Men, Whataboutism, Red Herring

3 Data

The data contains training articles, development articles and testing articles. Each article is shown in one .txt file in plain text format used as the input. In each .txt file which records the articles, the title of the article is in the first row, followed by an empty row. The content of the article starts from the third row, one sentence per line. Articles are retrieved from the newspaper3k library and has been automatically splited. Here is an example.

1. Manchin says Democrats acted like babies at the SOTU (video) Personal Liberty Poll Exercise your right to vote.
2. (empty line)
3. Democrat West Virginia Sen. Joe Manchin says his colleagues' refusal to stand or applaud during President Donald Trump's State of the Union speech was disrespectful and a signal that the party is more concerned with obstruction than it is with progress.
4. In a glaring sign of just how stupid and petty things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech not looking as though Trump killed his grandma.
5. When others in his party declined to applaud even for the most uncontroversial of the president's remarks, Manchin did.
6. He even stood for the president when Trump entered the room, a customary show of respect for the office in which his colleagues declined to participate.

The above example is an article containing four sentences. The first line is the title. However, the text is noisy in that some superscripts and subscripts are added, which makes the problem trickier. For example, "Personal Liberty Poll Exercise your right to vote." is actually not part of the title.
Some propaganda techniques are applied in this article. For instance, in the first line, "babies" uses both name calling and labeling. In the fourth line, ""stupid and petty" uses Loaded Language and "not looking as though Trump killed his grandma" is an instance of Exaggeration and Minimisation.

4 Related Work

The related works are as follows:

- Granik, Mykhailo and Mesyura, Volodymyr proposed an idea to detect and classify the propaganda based on Naive Bayes(Granik and Mesyura, 2017).
- Johnston, Andrew H and Weiss, Gary M proposed an idea to identify extremist propaganda with deep learning(Johnston and Weiss, 2017).
- Alhindi, Tariq and Pfeiffer, Jonas and Muresan, Smaranda proposed an idea of fine-tuned beural models for propaganda detection at the sentence and fragment levels(Alhindi et al., 2019).

5 Methodology

First, I plan to try feature extraction including word of bags and words embedding. Then I plan to use several baseline models first, which include support vector machine, Logistic regression, Naive Bayes classifier and random forest classifier. To add on, I will consider the relationship between words and phrases and try recurrent neural networks.

6 Evaluation and Results (1 point)

The baseline model I am going to compare against will be the random model. The baseline model will make classification predictions with the equal possibilities of each class.
The metric to evaluate the results will be accuracy.

7 Discussion

You can leave this section blank.

8 Work Plan

1. now to February 20th: finish the feature extraction, build the support vector machine, Logistic regression, Naive Bayes classifier and random forest classifier and evaluate their performance.
2. February 20th to March 8th: adjust feature extraction and try more advanced models such as RNN.
3. March 8th to April 15th: finish the project report.

9 Multi-person Team Justification

Acknowledgments

References

Tariq Alhindi, Jonas Pfeiffer, and Smaranda Muresan. 2019. Fine-tuned neural models for propaganda detection at the sentence and fragment levels. *arXiv preprint arXiv:1910.09702* .

Mykhailo Granik and Volodymyr Mesyura. 2017. Fake news detection using naive bayes classifier. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. IEEE, pages 900–903.

Andrew H Johnston and Gary M Weiss. 2017. Identifying sunni extremist propaganda with deep learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pages 1–6.

A Supplemental Material