Propaganda Analysis

Vinay Damodaran
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
vdamodar@andrew.cmu.edu

Kinjal Jain
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
kinjalj@andrew.cmu.edu

Kavya Nerella
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
knerella@andrew.cmu.edu

Aiswarya Vinod Kumar
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
avinodku@andrew.cmu.edu

Abstract

We propose to work on a SemEval 2020 Task: Detection of Propaganda Techniques in News Articles. The motivation to work on this problem is based on the current state of affairs in world politics where politicians use mass propaganda techniques to create a divisive culture among people to garner more supporters. Due to the lack of a good automated censoring/ filtering system and the lack of resources to manually identify the propaganda that flows through social media these days, this propaganda flows unchecked. Having a system that is properly able to identify propagandist agendas by being able to segment the fine-grained instances aiming to spread malice, would help us tackle this issue prominent in everyday social media.

1 Task Description

The task is divided into 2 sub-tasks:

1. Span Identification: Given a plain-text document, identify those specific fragments which contain at least one propaganda technique.
2. Technique Classification: Given a text fragment identified as propaganda and its document context, identify the applied propaganda technique in the fragment. This can include overlapping spans corresponding to different techniques as well.

2 Dataset

The dataset is well-defined as part of the aforementioned task. It is a collection of 550 news articles split into 446 training articles (350k tokens) and 48 development articles. There are 18 different types of propaganda techniques identified in these News articles namely:

1. Presenting irrelevant data
2. Misrepresentation of someone's position
3. Whataboutism
4. Causal oversimplification
5. Obfuscation, intentional vagueness, confusion

6. Appeal to authority
7. Black-and-white fallacy, dictatorship
8. Name calling or labeling
9. Loaded language
10. Exaggeration or minimisation
11. Flag-waving
12. Doubt
13. Appeal to fear/prejudice
14. Slogans
15. Thought-terminating cliché
16. Bandwagon
17. Reductio ad hitlerum
18. Repetition

Out of these 18 categories some underrepresented classes are merged to finally have 14 classes. For each document, a plain text file and two tab separated annotation files are provided corresponding to each sub-task. For sub-task 1, multiple lines containing "Article ID", "Starting Offset of the Span" and "Ending offset (exclusive)" are provided. For sub-task 2, "Propaganda Class/ Technique" is also given. Evaluation is performed by calculating the micro-average F1 score.

3 Literature Survey

The dataset was published in EMNLP 2019.

The paper mentions tasks very similar to the current SemEval shared tasks with slight differences. As the competition does not mention any baselines specifically, the tasks and baselines mentioned in the paper are explained below.

1. Sentence Level classification(SLC) : which asks to predict whether a sentence contains at least one propaganda technique.
2. Fragment-level classification(FLC) : which asks to identify both the spans and the type of propaganda technique.

3.1 Baselines

- BERT
  - For FLC, each token's last hidden representation is passed through a 19 way linear classifier which classifies if this token belongs to one of the eighteen propaganda techniques or to none of them. The linear layer is fine-tuned during training data.
  - For SLC, the CLS token is passed through a linear layer which does binary classification
- BERT-Joint:
  - Both the linear layers for SLC and FLC are trained jointly.
- BERT-Granularity:
  - Here, SLC information is also used for FLC classification. The output of linear layer on top of CLS token is concatenated with the output of linear layer on top of BERT's hidden representation and this is inputted to a 19 way classifier.
- Multi-Granularity Network:
  - Suppose there are k tasks of increasing granularity, e.g., document-level, paragraph level, sentence-level, word-level, subword-level, character-level.
  - Each task has an output layer embedding o_k and a trainable function f. The gate f consists of a projection layer to one dimension and an activation function. The resulting value is multiplied by each element of the output of a layer of higher granularity o_{k+1}.

- As a result, examples strongly classified as negative in a lower granularity task would be ignored in the high granularity task.
- Having the lower-granularity as the main task means that higher-granularity information can be selectively used as additional information to improve the performance, but only if the example is not considered as highly negative.

Previously a similar task was organized at NLP4IF . The proceedings of the workshop have papers with insights to approach the problem. We consider this as a good starting point.

4 Timeline

Tentative Duration	Task(s)
Week 1	Literature Survey Finalise
Week 2-3	Baseline Implementation
Week 4	Analyze Baseline Determine Improvement Scopes
Week 5-7	Work in parallel over different improvements
Week 8	Analyze Improvements
Week 9-10	Experimentation Further Improvements Poster + Report Preparation

References

[1] Giovanni Da San Martino1, Seunghak Yu2, Alberto Barron-Cede, Rostislav Petrov4 , Preslav Nakov1 (2019) *Fine-Grained Analysis of Propaganda in News Articles*

[2] Tariq Alhindi, Jonas Pfeiffer, Smaranda Muresan (2019) *Fine-Tuned Neural Models for Propaganda Detection at the Sentence and Fragment levels*

[3] https://propaganda.qcri.org/nlp4if-shared-task/