code is available at
https://github.com/ronnakorn9/DataAnalytics2023_Ronnakorn/tree/main/proj

Detecting Persuasion Technique in Social Media

Ronnakorn Rattanakornphan[1] (rattar@rpi.edu)

([1]Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States)

Abstract

Social media allows people to quickly discuss and absorb vast amounts of information in a very short time. However, some information is undesirable or harmful, for example, misleading thoughts, misinformation, propaganda, etc..

In this project, we explore the possibility of using machine learning model to detect the persuasion technique (ex. “Name Calling”) in social media text. Particularly, we attempt to use transformer-based model (RoBERTa) to detect and classify the persuasion technique present in the X (formerly known s “Twitter”) text data.

We are able to use transformer-based model trained in SemEval-2021 Task 6 dataset to detect the persuasion technique presented in Twitter data. However, the recall rate of the detection is still low and there exist the issue of stability of the detected text. We show the detection example result and discuss the possible improvement in the end.

Motivation

A large part of social media can be described as a type of persuasion, to influence or guide others attitudes or behavior. For example, a person might defame an organization to convince other people in the group to turn away or become hostile toward that organization. Another person might be praising a product to convince other people to buy the item, regardless of the truth.

Many psychology studies report that notifying people of upcoming “persuasion” can increase their resistance towards said persuasion. Therefore, developing a machine learning model that could detect and warn the user in social media of such persuasion could help them make a more informed decision and be more attentive when traversing the vast social media in their daily life.

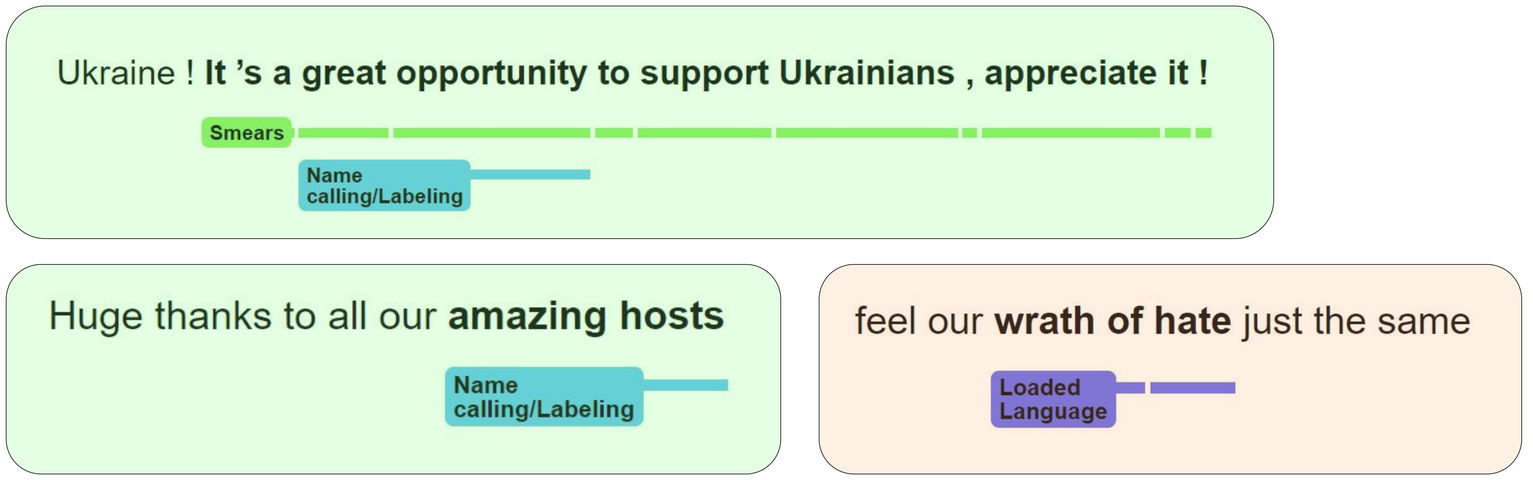

Fig.1 Visualization of persuasion span detection

Method

In this project, we use 2 differents models for 2 differents tasks.

Task 1: Multilabel classification: given a text, classify the (multiple) persuasion techniques presented in the text. We use transformer (RoBERTa) as the text encoding then use spaCy’s TextCategorizer for class detection.

Task 2: Span categorization: given a text, classify the (multiple) persuasion techniques presented in the text and also mark the section of the text the specified persuasion technique is presented. We also use transformer (RoBERTa) as the text encoding for this task (same model architecture but different instance) then use spaCy’s SpanCategorizer for class detection.

We train both model on SemEval Task 8 dataset. The training set has 688 text entries with 1184 categories label and 1497 text spans. The validation set has 63 text entries with 123 categories label and 182 text spans.

Both model is then run on X (Twitter) dataset. The result statistic is then collected and evaluated.

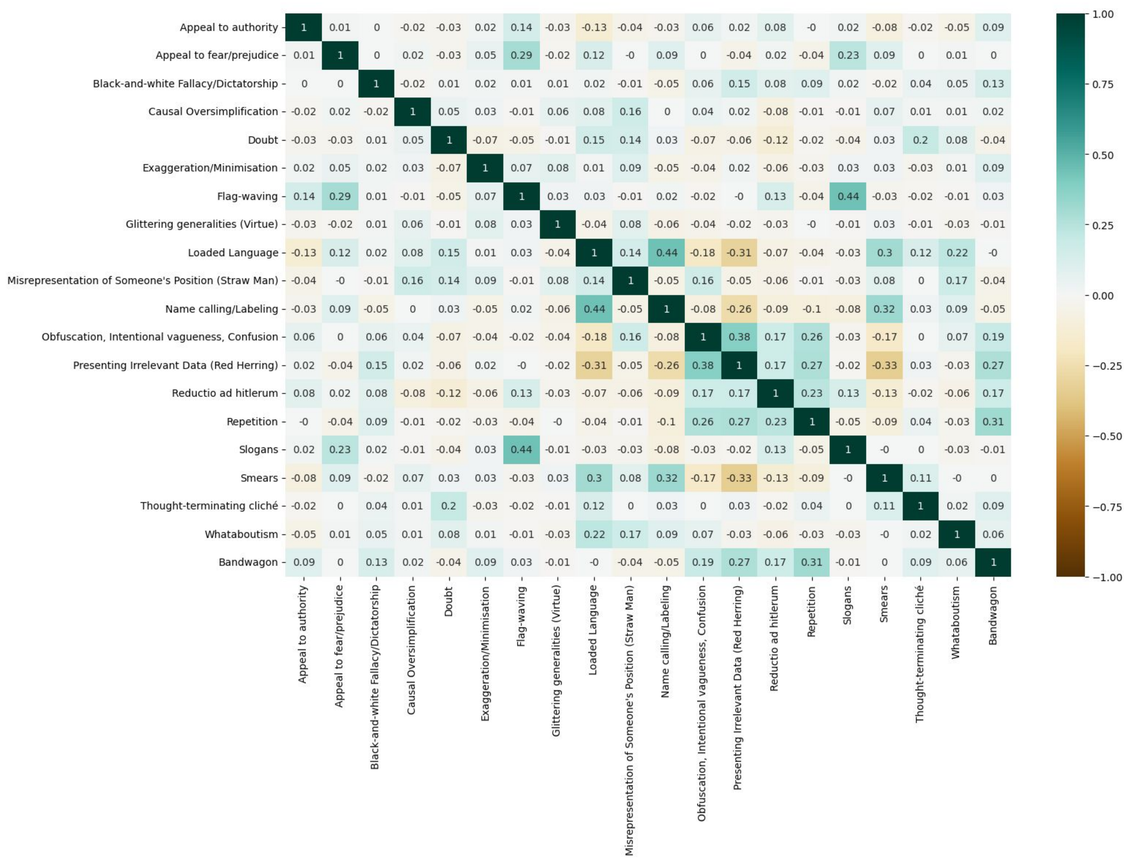

Fig. 2 Between-class correlation of the detected persuasion technique in X (Twitter) dataset.

Result

The both model has low recall in general. The large class imbalance mean that some smaller classes which has less than 5 samples could get ignored, pulling down the macro score significantly. A longer computation time, bigger dataset, and better data sampling are very likely to improve both model performance.

Metric (Task 1)	Value
macro average AUG	0.645
macro F1-score	0.181
macro precision	0.226
macro recall	0.169

Nevertheless, the 2nd model could still detect and label the persuasion technique presented in X (Twitter) text data (shown in Fig.1) although prediction is unstable at times (same word has different label).

```
{'Appeal to authority': 0.001771137351,
 'Appeal to fear/prejudice': 0.00932087376,
...
 'Loaded Language': 0.999467909,
...

 'Bandwagon': 0.000805786810}
```

Conclusion

We have developed 2 models for detecting the persuasion techniques in the text. The first model predicts the class probability (20 classes = 20 probabilities in total) while 2nd model detects the text span that has the persuasion techniques presented and also the class of the techniques.

Overall, it is possible to use a machine learning model to detect the persuasion technique in social media texts. We could improve the performance of model by extending the dataset further using the model output from the X (Twitter) dataset.