Ronnakorn Rattanakornphan

Motivation:

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, such as misleading thoughts, misinformation, propaganda, etc.. Therefore, people must be aware of and ready to face such danger.

A large part of social media can be described as a type of persuasion, to influence or guide other's 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.

Task:

There are 2 tasks in total.

1. Task 1: Multilabel classification: given a text, classify the (multiple) persuasion techniques presented in the text
2. 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

Data:

This project uses 2 datasets

The 1st dataset is from SemEval-2021 Task 6, consisting of

1. training data 688 text entries with 1184 category labels (not 688 labels since it's a multilabel dataset), 1497 spans
2. Validation data with 63 text entries with 123 category labels, and 182 spans.

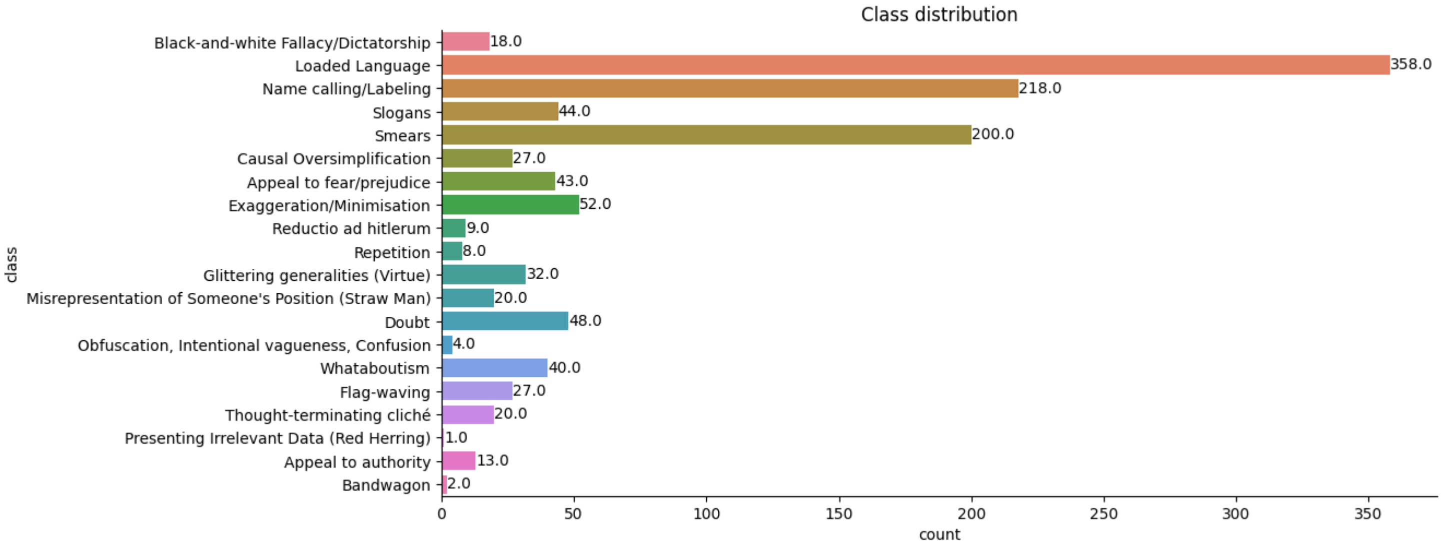
The 2nd dataset is the Twitter data from the IDEA cluster, consisting of 990,622 text records. No labels.

1. Only 10,000 entries are used in this project due to memory issues.
2. Each entry of Twitter data also has date and text sentiment information.

Data exploration

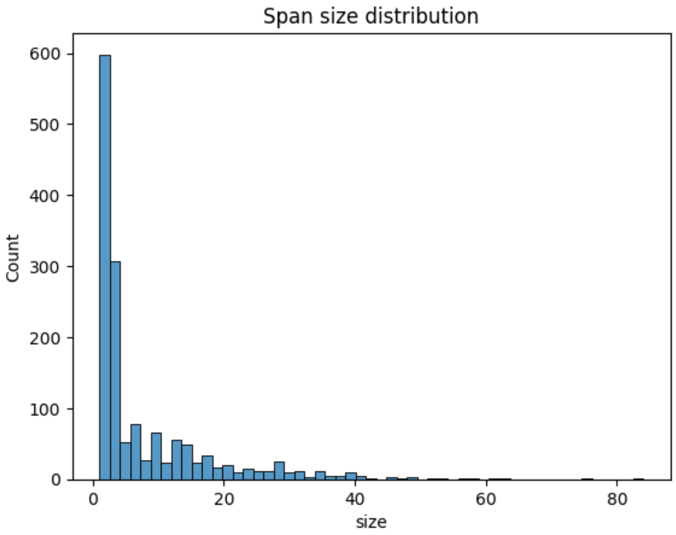
<u>Class distribution</u>

There are 20 classes in total.



There is a big imbalance among the class with the highest occurring class ("Loaded Language") having 358 entries but the lowest occurring class (Red Herring) having 1 entry.

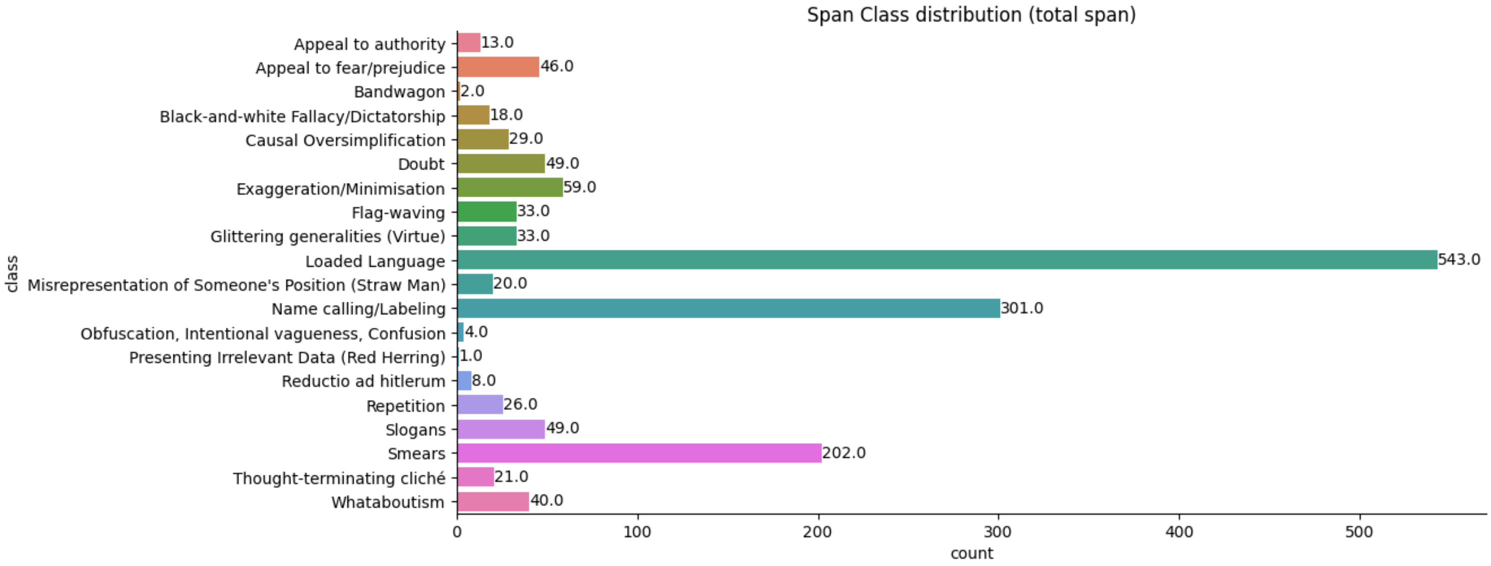
Span size distribution

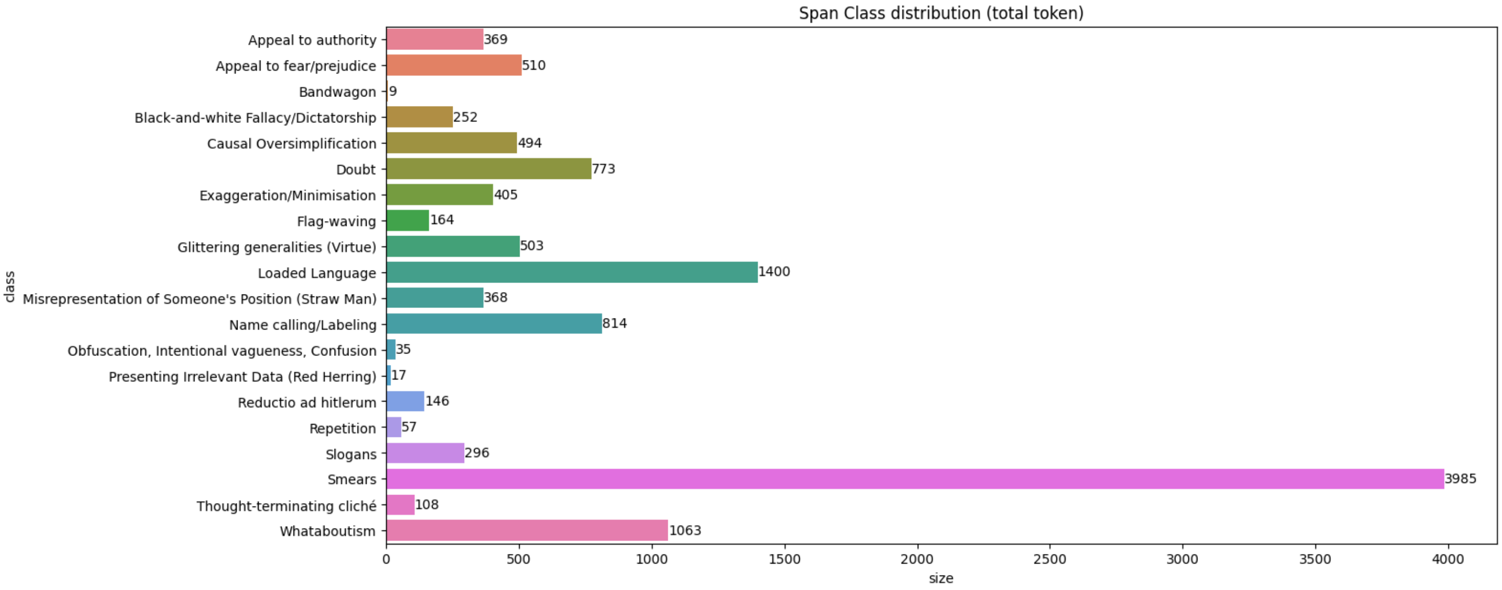


The size of the text span ranges from 1 token (1 word) to 84 tokens. The span with a longer size will be difficult to detect due to the low ratio of training data to all possible span (a longer span has a larger number of permutation and therefore require a large amount of sample to distinguish them). However, the number of long spans is quite low in the dataset and it'll be difficult for the model to recognize them.

The majority of the text span has a lower amount of tokens (usually less than 20 tokens). The number, however, is still a lot longer than the normal entity text span (ex. "Bank of America") which is usually less than 5 tokens. For this project, the longest possible span is limited to 16 tokens (instead of 84 tokens) to significantly reduce the memory and time to train the model.

Span class distribution

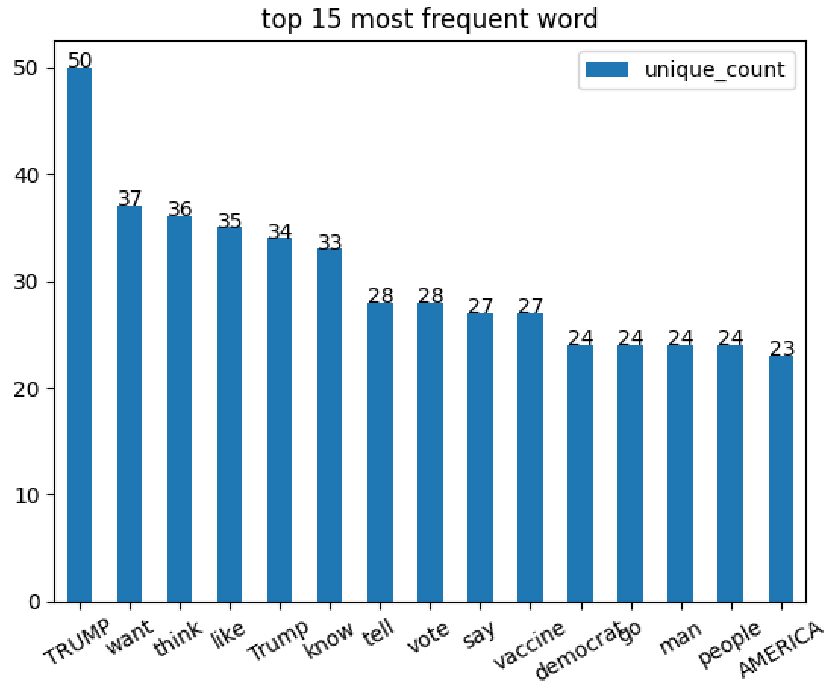




Similar to the text class distribution, the distribution of the text span class is also imbalanced, with the highest occurring class ("Loaded Language") having 543 text spans and the lowest occurring class ("Red Herring") having 1 text span.

The trend is different for the total number of tokens that belong to a specific class. Some classes (ex. "Smear") have roughly 20 tokens per span on average while some classes only have 3 tokens on average (ex. "Name calling")

Common word



The figure above shows the top 15 most occurring text in the training dataset. From the text shown, we can assume that the text revolves around political topics. The words "TRUMP" and "Trump" are treated as different tokens due to the need to distinguish the entity (ex. "Apple", the company vs "apple", a fruit) and possibly detect the different meanings that the capitalized word may contain.

Data preprocess

The preprocessing is as follows

1. The dataset (for both task 1 and task 2) is converted to the format (DocBin) required by the spaCy training pipeline.
2. The conversion process also involves tokenization, pruning stop words, and converting the character-based label into a token-based label (for task 2).
3. The dataset is then serialized into spacy object for feeding into the training pipeline.

Model (and parameter)

The model used in the project comes from spaCy, an open-source library for Natural Language Processing in Python. The particular model used in this project is the RoBERTa base model (transformer as embedder), spaCy's TextCategorizer (task 1), and SpanCategorizer (task 2).

In theory, continuing to run the training process would yield a more accurate model, but due to the constraint of time, a stopping criterion is defined. For task 1, the training process will stop if there's no improvement over 1600 iterations. For task 2, the threshold is raised to 4000 iterations. The score criterion for task 1 is the macro-averaged AUC while task 2 uses a combination of recall and F1-score.

All models are trained on the training set of SemEval-2021 Task 6, evaluated on the validation set of SemEval-2021 Task 6, and then used for inference on the Twitter dataset.

Result

Validation performance on SemEval-2021 Task 6 dataset

Task 1:

The model has

Metric	Value
macro average AUG	0.645
macro F1-score	0.181
macro precision	0.226
macro recall	0.169

The performance of the current model is quite low but could be easily improved by using a bigger model (the reported model uses the small model for efficiency).

Task 2: Span Categorization

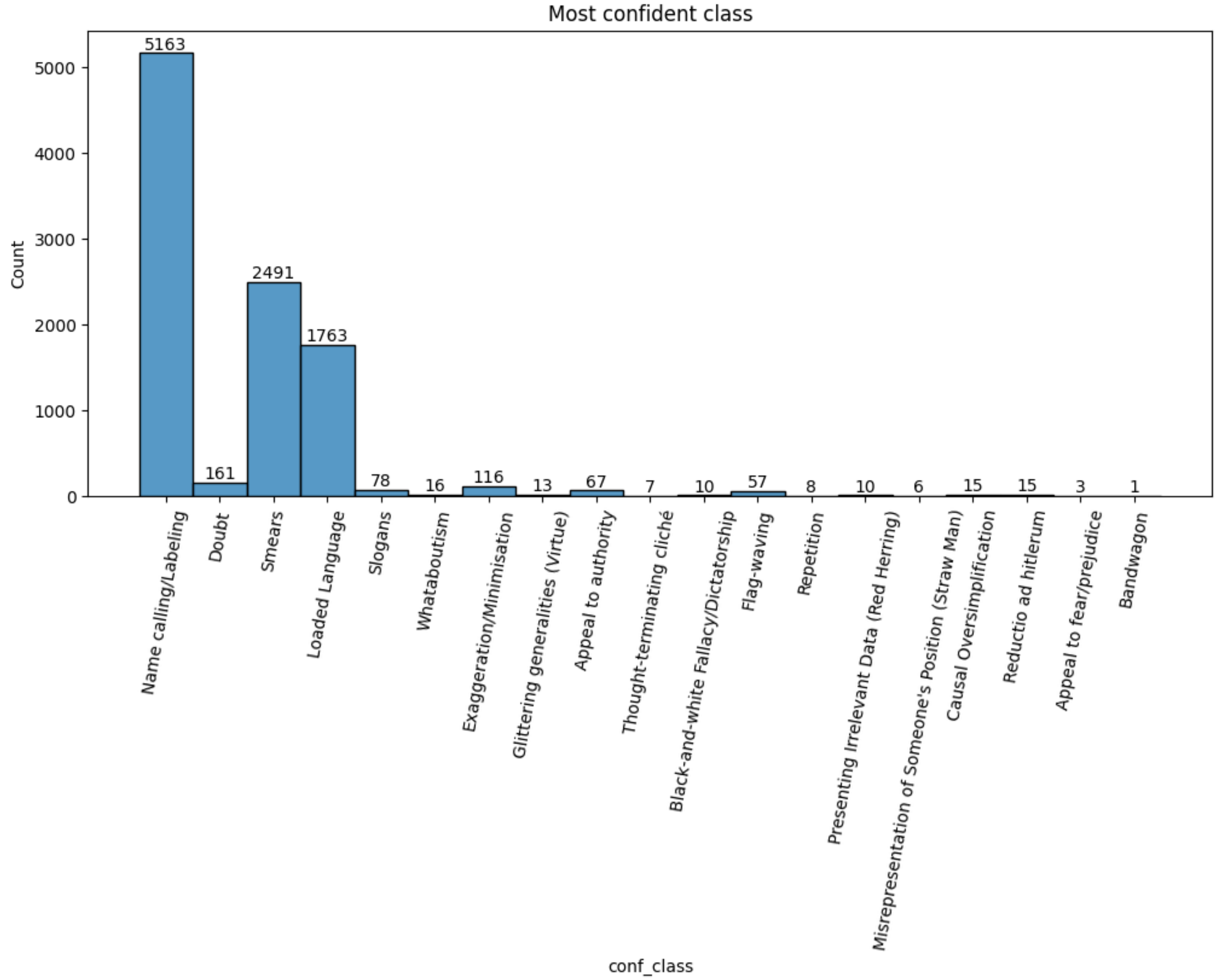
The model has

- F1-score: 0.167
- Precision: 0.351
- Recall: 0.110

The low recall implies that most of the persuasive span are not recognized. The cause of this might be the low amount of examples (only 1184 spans across 20 categories).

Twitter data

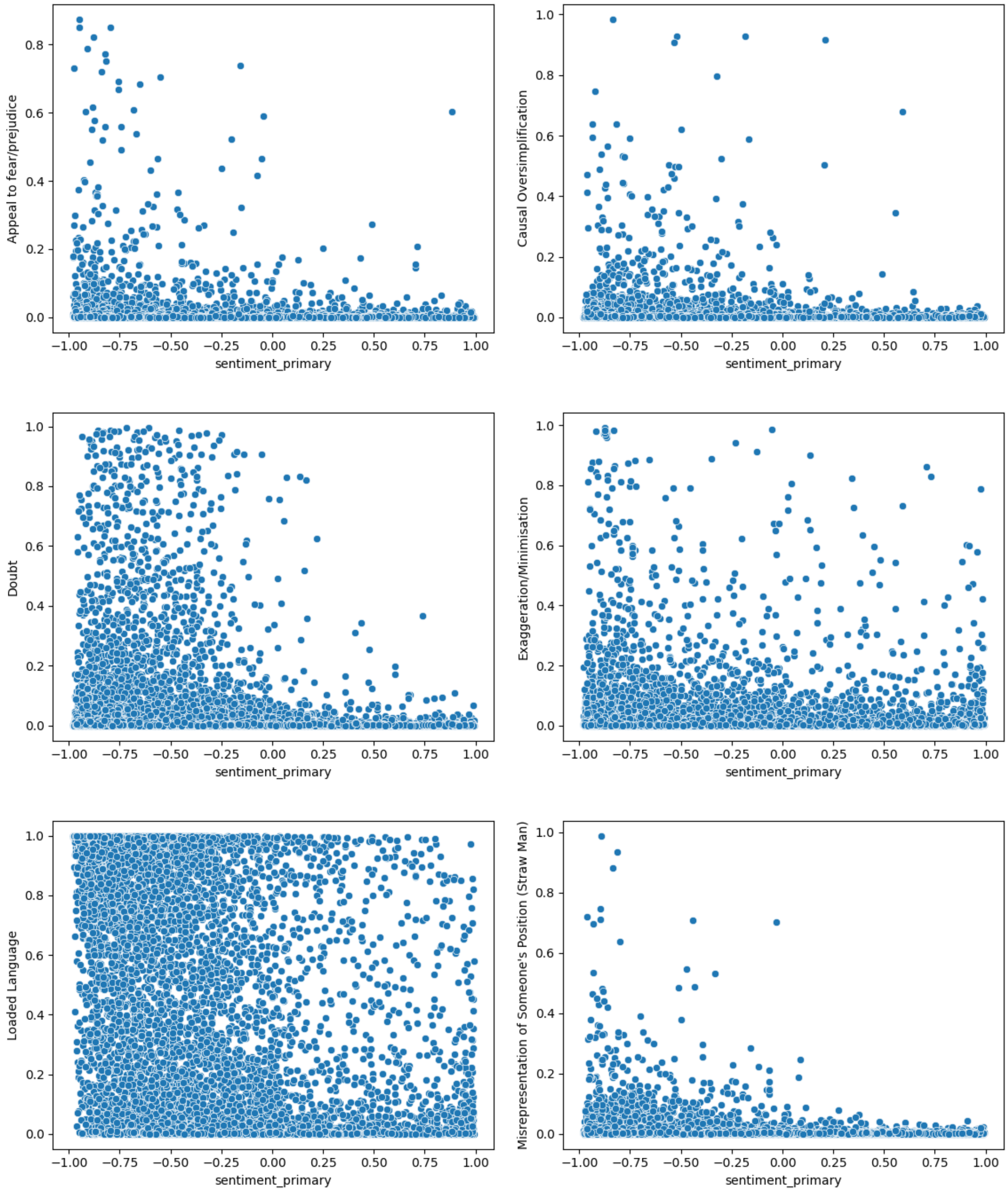
Class distribution

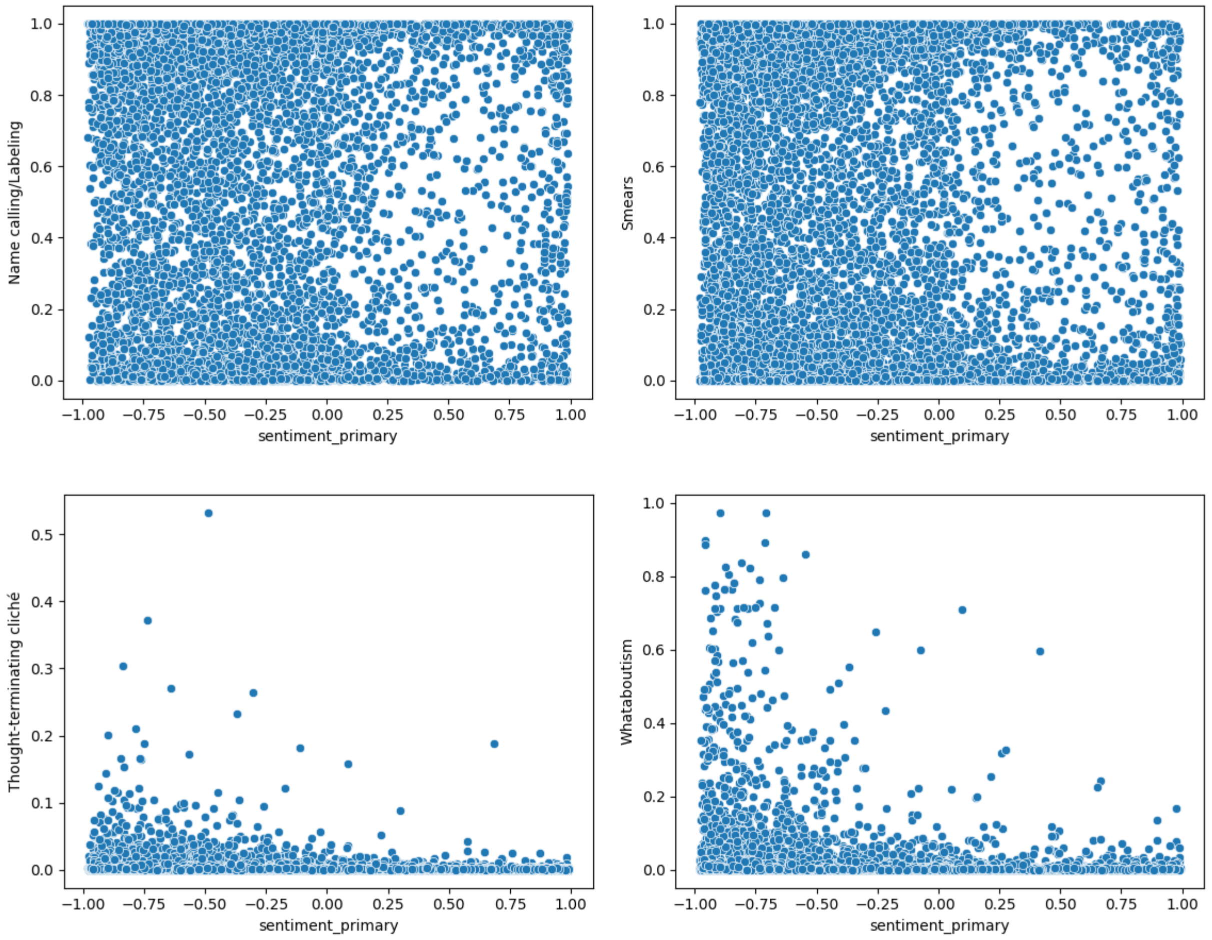


"Loaded Language", "Name Calling", and "Smears" are recognized the most. These are the 3 most common classes in the training data. It is uncertain whether the distribution is inherent in the data or the frequent training data makes the model more sensitive to/capable of detecting the aforementioned class.

Sentiment-to-class correlation

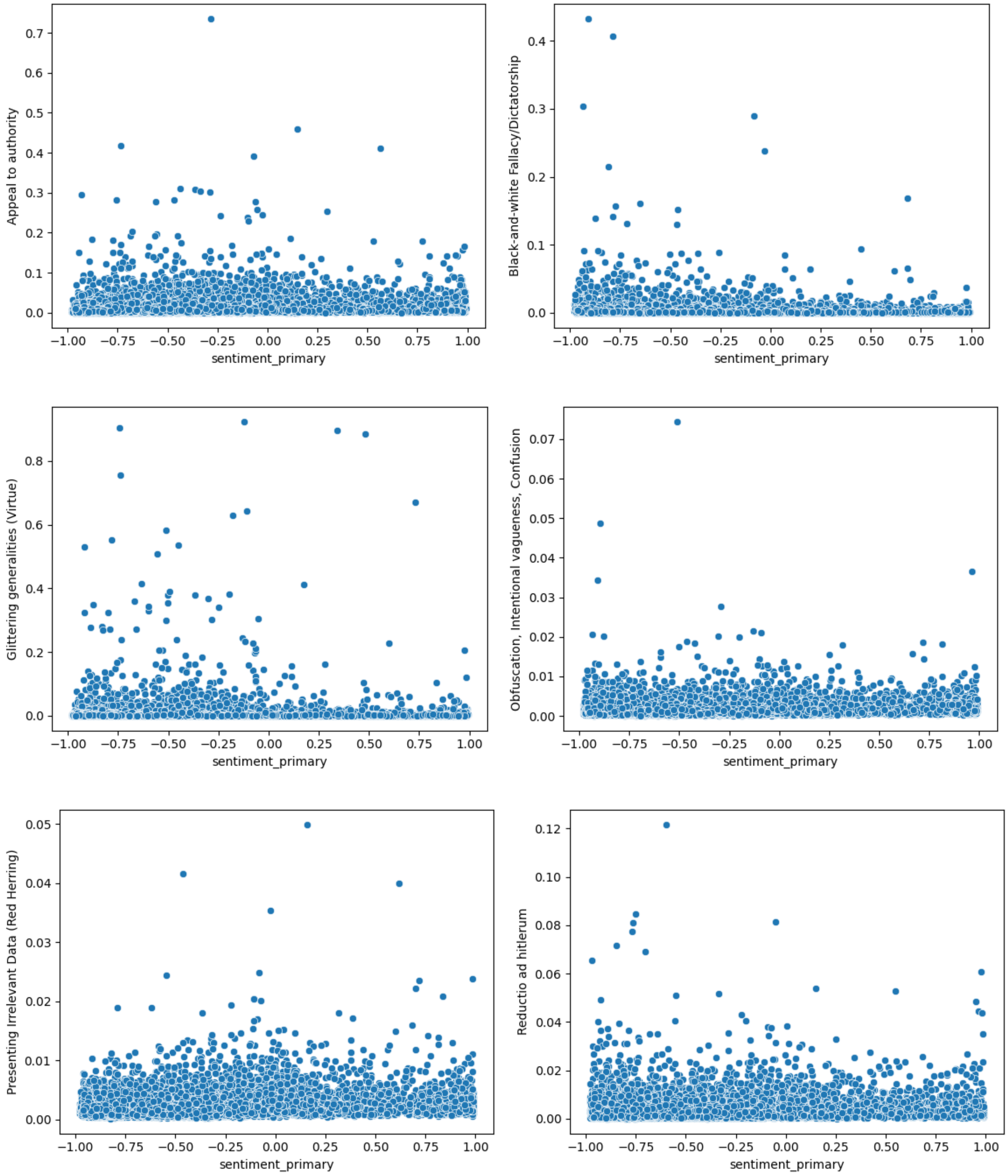
(More observable relation)

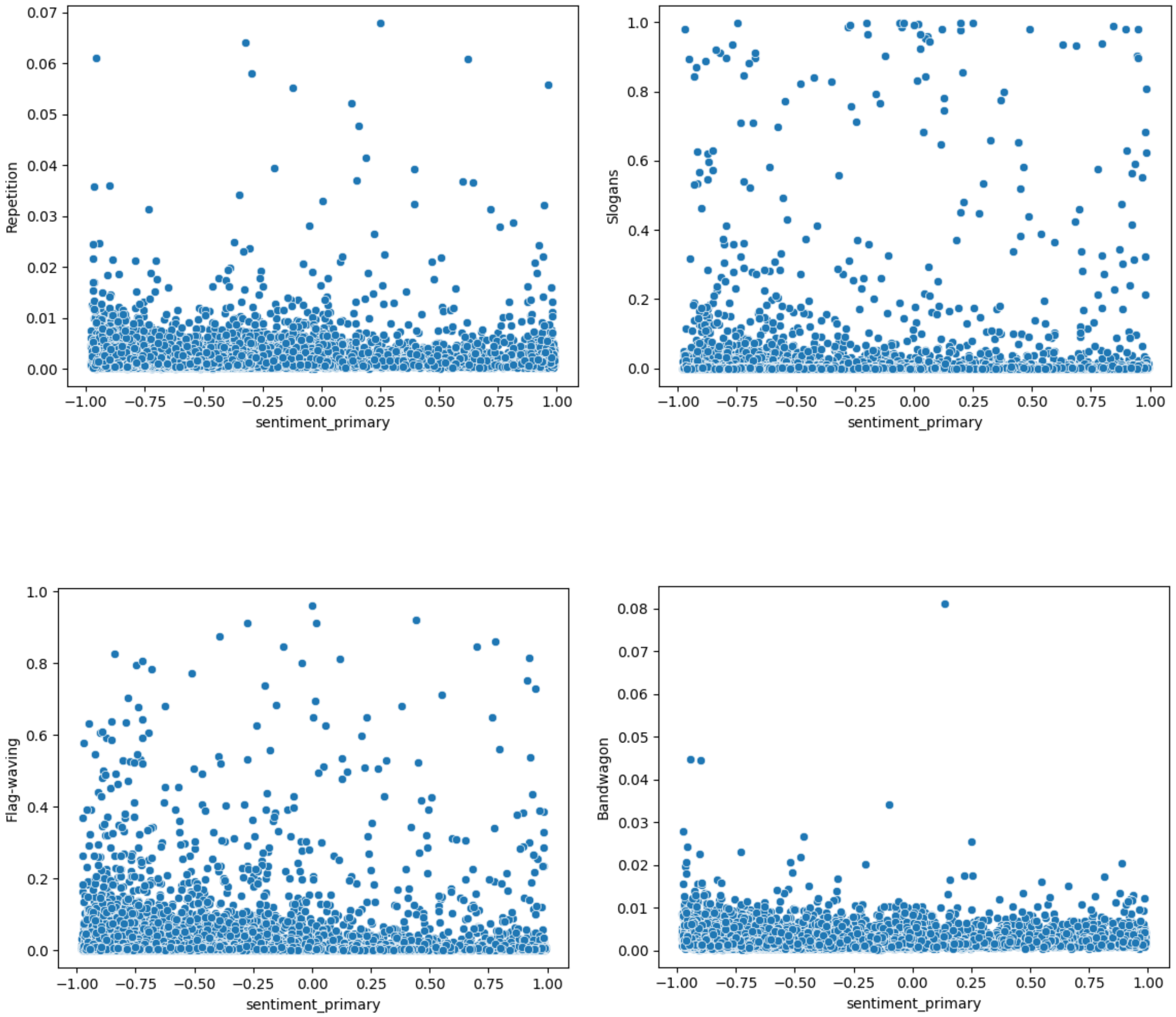




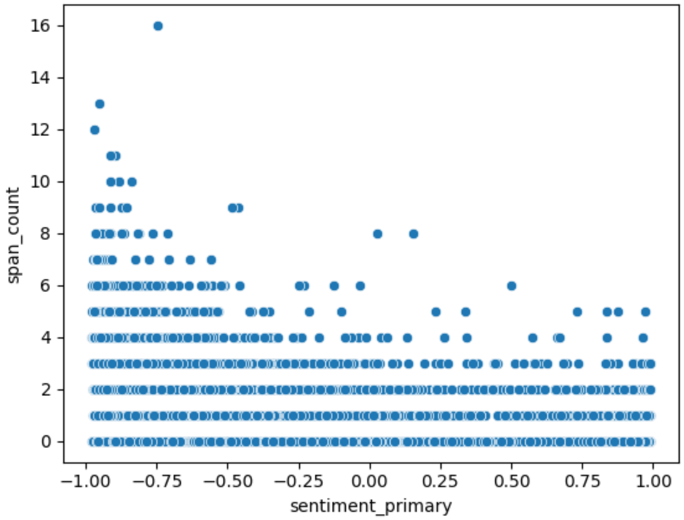
A general trend can be observed in the graphs above. The text entry that has a lower sentiment score has higher class confidence (more likely to have a persuasion technique). An interesting note is that the classes “Name Calling” and “Smears” had a wide range of class probability when the sentiment score was negative but when the score was positive the class probability was either really high or really low with little in between.

(Less observable relation)



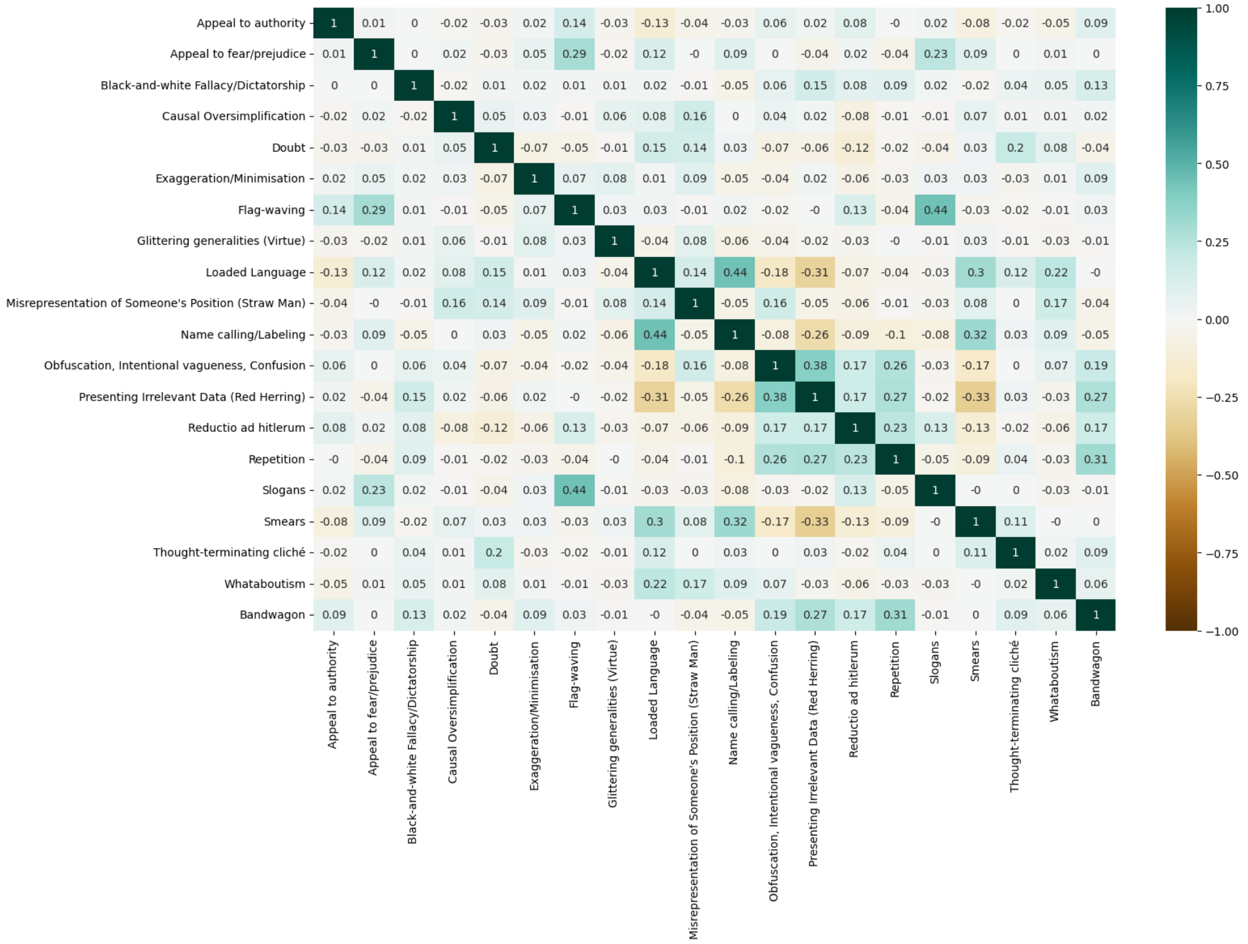

Repetition
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
sentiment_primary
Slogans
1.0
0.8
0.6
0.4
0.2
0.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
sentiment_primary
Flag-waving
1.0
0.8
0.6
0.4
0.2
0.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
sentiment_primary
Bandwagon
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
sentiment_primary

Span count vs sentiment score



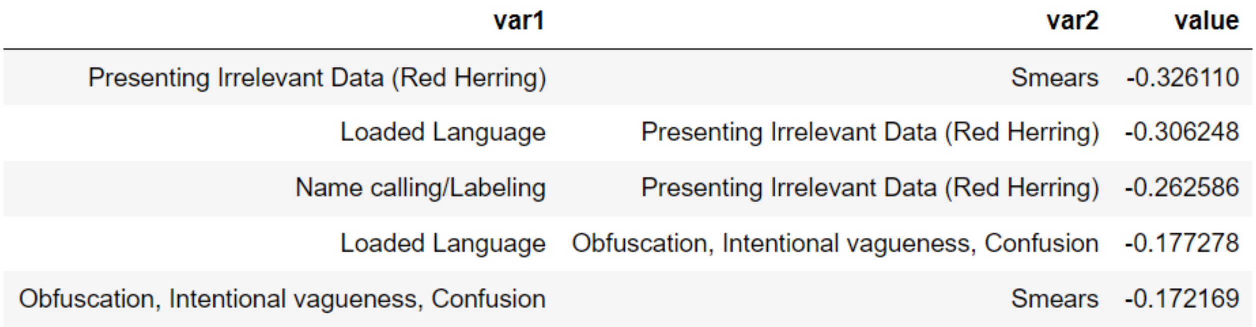
The text with a high span count has a lower sentiment score. It could be that the labeled class of persuasion techniques leans more toward negative sentiment, for example, “Doubt” and “Loaded Language”.

Between-class correlation



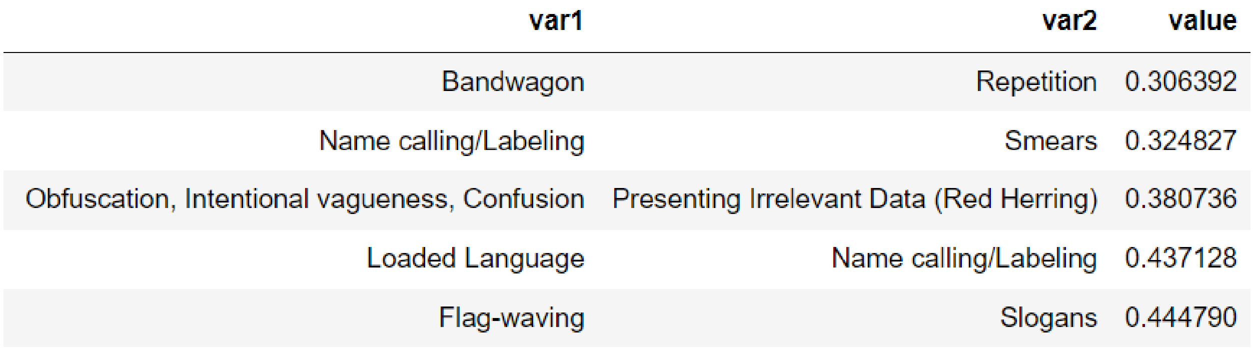
	Appeal to authority	Appeal to fear/prejudice	Black-and-white Fallacy/Dictatorship	Causal Oversimplification	Doubt	Exaggeration/Minimisation	Flag-waving	Glittering generalities (Virtue)	Loaded Language	Misrepresentation of Someone's Position (Straw Man)	Name calling/Labeling	Obfuscation, Intentional vagueness, Confusion	Presenting Irrelevant Data (Red Herring)	Reductio ad hitlerum	Repetition	Slogans	Smears	Thought-terminating cliché	Whataboutism	Bandwagon
Appeal to authority	1	0.01	0	-0.02	-0.03	0.02	0.14	-0.03	-0.13	-0.04	-0.03	0.06	0.02	0.08	-0	0.02	-0.08	-0.02	-0.05	0.09
Appeal to fear/prejudice	0.01	1	0	0.02	-0.03	0.05	0.29	-0.02	0.12	-0	0.09	0	-0.04	0.02	-0.04	0.23	0.09	0	0.01	0
Black-and-white Fallacy/Dictatorship	0	0	1	-0.02	0.01	0.02	0.01	0.01	0.02	-0.01	-0.05	0.06	0.15	0.08	0.09	0.02	-0.02	0.04	0.05	0.13
Causal Oversimplification	-0.02	0.02	-0.02	1	0.05	0.03	-0.01	0.06	0.08	0.16	0	0.04	0.02	-0.08	-0.01	-0.01	0.07	0.01	0.01	0.02
Doubt	-0.03	-0.03	0.01	0.05	1	-0.07	-0.05	-0.01	0.15	0.14	0.03	-0.07	-0.06	-0.12	-0.02	-0.04	0.03	0.2	0.08	-0.04
Exaggeration/Minimisation	0.02	0.05	0.02	0.03	-0.07	1	0.07	0.08	0.01	0.09	-0.05	-0.04	0.02	-0.06	-0.03	0.03	0.03	-0.03	0.01	0.09
Flag-waving	0.14	0.29	0.01	-0.01	-0.05	0.07	1	0.03	0.03	-0.01	0.02	-0.02	-0	0.13	-0.04	0.44	-0.03	-0.02	-0.01	0.03
Glittering generalities (Virtue)	-0.03	-0.02	0.01	0.06	-0.01	0.08	0.03	1	-0.04	0.08	-0.06	-0.04	-0.02	-0.03	-0	-0.01	0.03	-0.01	-0.03	-0.01
Loaded Language	-0.13	0.12	0.02	0.08	0.15	0.01	0.03	-0.04	1	0.14	0.44	-0.18	-0.31	-0.07	-0.04	-0.03	0.3	0.12	0.22	-0
Misrepresentation of Someone's Position (Straw Man)	-0.04	-0	-0.01	0.16	0.14	0.09	-0.01	0.08	0.14	1	-0.05	0.16	-0.05	-0.06	-0.01	-0.03	0.08	0	0.17	-0.04
Name calling/Labeling	-0.03	0.09	-0.05	0	0.03	-0.05	0.02	-0.06	0.44	-0.05	1	-0.08	-0.26	-0.09	-0.1	-0.08	0.32	0.03	0.09	-0.05
Obfuscation, Intentional vagueness, Confusion	0.06	0	0.06	0.04	-0.07	-0.04	-0.02	-0.04	-0.18	0.16	-0.08	1	0.38	0.17	0.26	-0.03	-0.17	0	0.07	0.19
Presenting Irrelevant Data (Red Herring)	0.02	-0.04	0.15	0.02	-0.06	0.02	-0	-0.02	-0.31	-0.05	-0.26	0.38	1	0.17	0.27	-0.02	-0.33	0.03	-0.03	0.27
Reductio ad hitlerum	0.08	0.02	0.08	-0.08	-0.12	-0.06	0.13	-0.03	-0.07	-0.06	-0.09	0.17	0.17	1	0.23	0.13	-0.13	-0.02	-0.06	0.17
Repetition	-0	-0.04	0.09	-0.01	-0.02	-0.03	-0.04	-0	-0.04	-0.01	-0.1	0.26	0.27	0.23	1	-0.05	-0.09	0.04	-0.03	0.31
Slogans	0.02	0.23	0.02	-0.01	-0.04	0.03	0.44	-0.01	-0.03	-0.03	-0.08	-0.03	-0.02	0.13	-0.05	1	-0	0	-0.03	-0.01
Smears	-0.08	0.09	-0.02	0.07	0.03	0.03	-0.03	0.03	0.3	0.08	0.32	-0.17	-0.33	-0.13	-0.09	-0	1	0.11	-0	0
Thought-terminating cliché	-0.02	0	0.04	0.01	0.2	-0.03	-0.02	-0.01	0.12	0	0.03	0	0.03	-0.02	0.04	0	0.11	1	0.02	0.09
Whataboutism	-0.05	0.01	0.05	0.01	0.08	0.01	-0.01	-0.03	0.22	0.17	0.09	0.07	-0.03	-0.06	-0.03	-0.03	-0	0.02	1	0.06
Bandwagon	0.09	0	0.13	0.02	-0.04	0.09	0.03	-0.01	-0	-0.04	-0.05	0.19	0.27	0.17	0.31	-0.01	0	0.09	0.06	1

Top 5 negative class correlation



var1	var2	value
Presenting Irrelevant Data (Red Herring)	Smears	-0.326110
Loaded Language	Presenting Irrelevant Data (Red Herring)	-0.306248
Name calling/Labeling	Presenting Irrelevant Data (Red Herring)	-0.262586
Loaded Language	Obfuscation, Intentional vagueness, Confusion	-0.177278
Obfuscation, Intentional vagueness, Confusion	Smears	-0.172169

Top 5 positive class correlation



var1	var2	value
Bandwagon	Repetition	0.306392
Name calling/Labeling	Smears	0.324827
Obfuscation, Intentional vagueness, Confusion	Presenting Irrelevant Data (Red Herring)	0.380736
Loaded Language	Name calling/Labeling	0.437128
Flag-waving	Slogans	0.444790

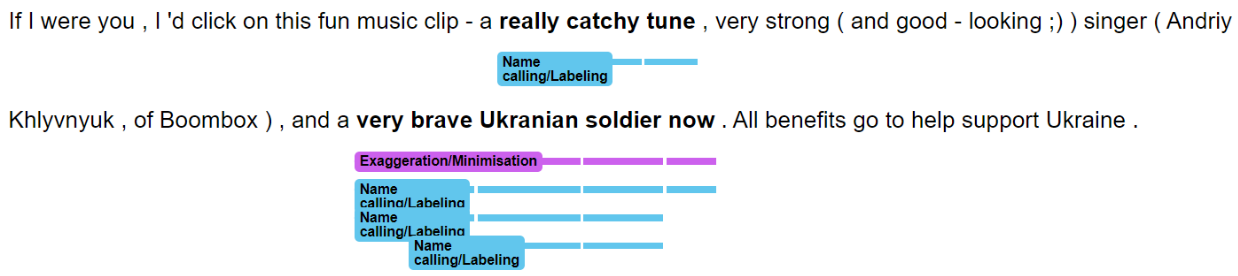
Considering that the “Loaded Language” is usually an insult. Such sentences would need a target for the insult. The entity presented would lead to a high class probability of “Name Calling” which could explain the high correlation between the two.

“Flag-waving” which is “playing on strong national feeling” often has a short message that aims to trigger an emotional response and is very similar to “Slogan”, for example, "let's make America great again" would be considered both “Flag-waving” and “Slogan.” The similarity between the two classes could lead to a high correlation value.

Span visualization

The following is the visualization on the Twitter dataset.

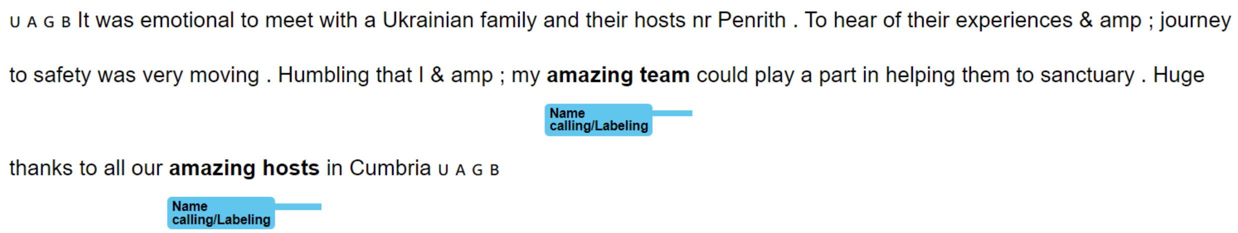
Positive sentiment text



If I were you , I 'd click on this fun music clip - a **really catchy tune** , very strong (and good - looking ;)) singer (Andriy

Khlyvnyuk , of Boombox) , and a **very brave Ukranian soldier now** . All benefits go to help support Ukraine .

https://t.co/InoFPryISQ

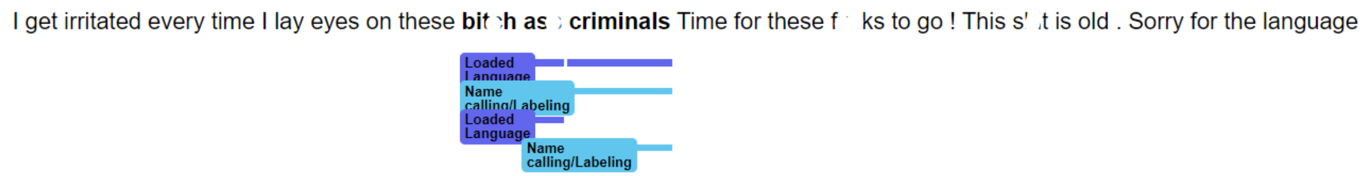


U A G B It was emotional to meet with a Ukrainian family and their hosts nr Penrith . To hear of their experiences & amp ; journey

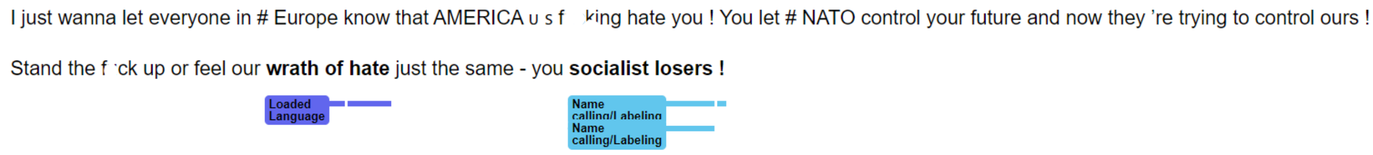
to safety was very moving . Humbling that I & amp ; my **amazing team** could play a part in helping them to sanctuary . Huge

thanks to all our **amazing hosts** in Cumbria U A G B

Negative sentiment text



I get irritated every time I lay eyes on these **bit ch as criminals** Time for these f ks to go ! This s' t is old . Sorry for the language



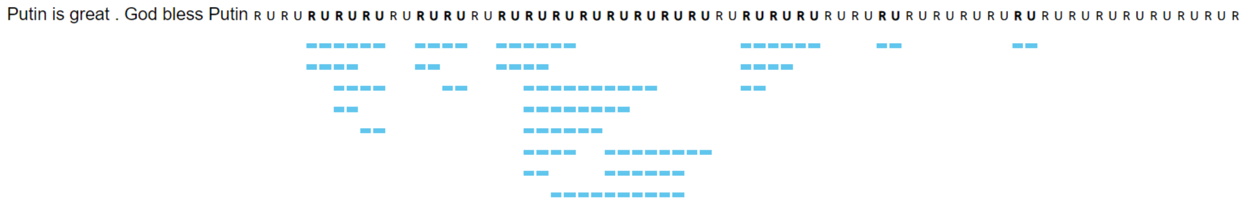
I just wanna let everyone in # Europe know that AMERICA U S f king hate you ! You let # NATO control your future and now they 're trying to control ours !

Stand the f ck up or feel our **wrath of hate** just the same - you **socialist losers !**

From the example, we can see the text span and the corresponding detected persuasion technique. We can see that the same persuasion can go in both positive directions (“amazing team”) as well as negative directions “socialist losers.” We can also see the effect of low recall in 2nd negative example, where the word “hate” in 2nd line is part of the detected span but the same word “hate” in 1st line does not get detected.

<u>Miss classified</u>

The following example denotes the detected span which could be considered misclassified. (The RU is the Russian flag emoji)



Putin is great . God bless Putin R U R U **R U R U R U** R U **R U R U** R U **R U R U R U R U R U R U R U R U** R U **R U R U R U** R U R U **R U** R U R U R U R U **R U** R U R U R U R U R U R U R U R

Conclusion

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

Both model was trained and evaluated on the SemEval dataset but we also ran the model on a much larger Twitter dataset to see the feasibility of using the model on social media corpus.

Both models suffer from low recall, possibly due to the small size of the training data. However, even with low recall performance, the model can still provide useful labeling on the Twitter text shown in the example above.

Overall, it is possible to use a machine learning model to detect the persuasion technique in social media texts. However, improvements should still be made to the model to achieve higher performance and more reliable detection. Some future work is discussed in the next section.

Future work

There are a lot of areas in this project that can be improved, for instance,

1. Improve the performance of the current model by extending the labeled dataset. This could be achieved by
    a. Hand-labeling each entry in the Twitter dataset (expensive)
    b. Adjust/verify/re-label the output of the current model on the Twitter data (less expensive)
2. Replacing the classification model for task 1 (to deep learning model, for example)
3. Considering that a big part of a persuasion message has an entity or organization the message is aiming at, those messages would be classified as using “Name calling/Labeling” techniques. Thus, we can run a NER (name-entity recognition) model to extract the name of the entity those messages are targeting.
4. More work related to the web and UI would be needed to overlay the detection result from the model into an actual display.

Limitation

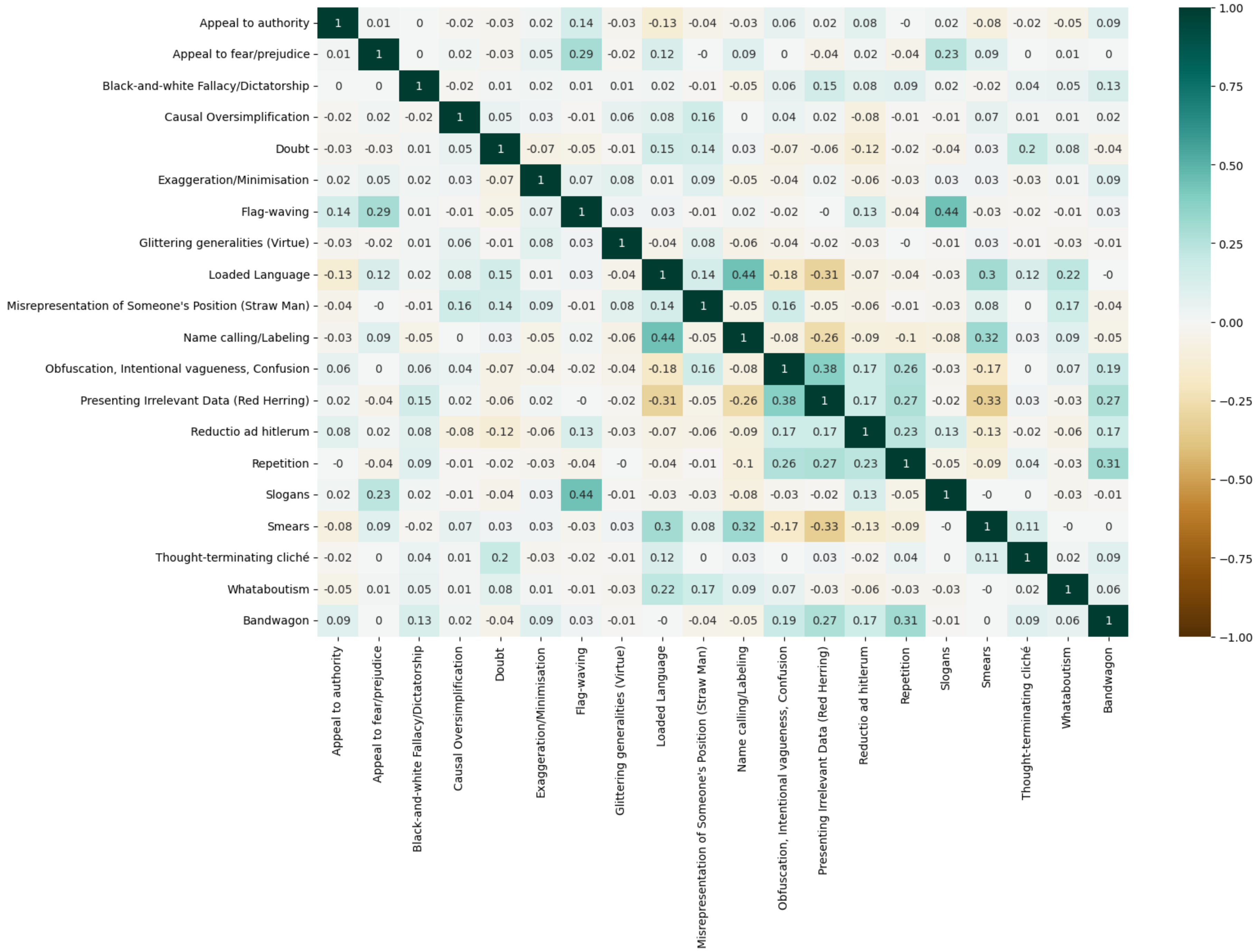
1. The techniques being used in this project cannot process the pictures or GIFs which are a large of social media messages—for example, memes, posters, pamphlets, etc.
2. The techniques cannot detect the missing information. It cannot detect that the message lacks a proper reference nor can it judge that the reference is sufficient. For example, a fake research claim.

Reference

Dimitar Dimitrov, Bin Ali, B., Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, & Da San Martino, G. (2021). SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the International Workshop on Semantic Evaluation*.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Petty, R. E., & Cacioppo, J. T. (1977). Forewarning, cognitive responding, and resistance to persuasion. *Journal of Personality and social Psychology*, *35*(9), 645.


Appeal to authority 1 0.01 0 -0.02 -0.03 0.02 0.14 -0.03 -0.13 -0.04 -0.03 0.06 0.02 0.08 -0 0.02 -0.08 -0.02 -0.05 0.09
Appeal to fear/prejudice 0.01 1 0 0.02 -0.03 0.05 0.29 -0.02 0.12 -0 0.09 0 -0.04 0.02 -0.04 0.23 0.09 0 0.01 0
Black-and-white Fallacy/Dictatorship 0 0 1 -0.02 0.01 0.02 0.01 0.01 0.02 -0.01 -0.05 0.06 0.15 0.08 0.09 0.02 -0.02 0.04 0.05 0.13
Causal Oversimplification -0.02 0.02 -0.02 1 0.05 0.03 -0.01 0.06 0.08 0.16 0 0.04 0.02 -0.08 -0.01 -0.01 0.07 0.01 0.01 0.02
Doubt -0.03 -0.03 0.01 0.05 1 -0.07 -0.05 -0.01 0.15 0.14 0.03 -0.07 -0.06 -0.12 -0.02 -0.04 0.03 0.2 0.08 -0.04
Exaggeration/Minimisation 0.02 0.05 0.02 0.03 -0.07 1 0.07 0.08 0.01 0.09 -0.05 -0.04 0.02 -0.06 -0.03 0.03 0.03 -0.03 0.01 0.09
Flag-waving 0.14 0.29 0.01 -0.01 -0.05 0.07 1 0.03 0.03 -0.01 0.02 -0.02 -0 0.13 -0.04 0.44 -0.03 -0.02 -0.01 0.03
Glittering generalities (Virtue) -0.03 -0.02 0.01 0.06 -0.01 0.08 0.03 1 -0.04 0.08 -0.06 -0.04 -0.02 -0.03 -0 -0.01 0.03 -0.01 -0.03 -0.01
Loaded Language -0.13 0.12 0.02 0.08 0.15 0.01 0.03 -0.04 1 0.14 0.44 -0.18 -0.31 -0.07 -0.04 -0.03 0.3 0.12 0.22 -0
Misrepresentation of Someone's Position (Straw Man) -0.04 -0 -0.01 0.16 0.14 0.09 -0.01 0.08 0.14 1 -0.05 0.16 -0.05 -0.06 -0.01 -0.03 0.08 0 0.17 -0.04
Name calling/Labeling -0.03 0.09 -0.05 0 0.03 -0.05 0.02 -0.06 0.44 -0.05 1 -0.08 -0.26 -0.09 -0.1 -0.08 0.32 0.03 0.09 -0.05
Obfuscation, Intentional vagueness, Confusion 0.06 0 0.06 0.04 -0.07 -0.04 -0.02 -0.04 -0.18 0.16 -0.08 1 0.38 0.17 0.26 -0.03 -0.17 0 0.07 0.19
Presenting Irrelevant Data (Red Herring) 0.02 -0.04 0.15 0.02 -0.06 0.02 -0 -0.02 -0.31 -0.05 -0.26 0.38 1 0.17 0.27 -0.02 -0.33 0.03 -0.03 0.27
Reductio ad hitlerum 0.08 0.02 0.08 -0.08 -0.12 -0.06 0.13 -0.03 -0.07 -0.06 -0.09 0.17 0.17 1 0.23 0.13 -0.13 -0.02 -0.06 0.17
Repetition -0 -0.04 0.09 -0.01 -0.02 -0.03 -0.04 -0 -0.04 -0.01 -0.1 0.26 0.27 0.23 1 -0.05 -0.09 0.04 -0.03 0.31
Slogans 0.02 0.23 0.02 -0.01 -0.04 0.03 0.44 -0.01 -0.03 -0.03 -0.08 -0.03 -0.02 0.13 -0.05 1 -0 0 -0.03 -0.01
Smears -0.08 0.09 -0.02 0.07 0.03 0.03 -0.03 0.03 0.3 0.08 0.32 -0.17 -0.33 -0.13 -0.09 -0 1 0.11 -0 0
Thought-terminating cliché -0.02 0 0.04 0.01 0.2 -0.03 -0.02 -0.01 0.12 0 0.03 0 0.03 -0.02 0.04 0 0.11 1 0.02 0.09
Whataboutism -0.05 0.01 0.05 0.01 0.08 0.01 -0.01 -0.03 0.22 0.17 0.09 0.07 -0.03 -0.06 -0.03 -0.03 -0 0.02 1 0.06
Bandwagon 0.09 0 0.13 0.02 -0.04 0.09 0.03 -0.01 -0 -0.04 -0.05 0.19 0.27 0.17 0.31 -0.01 0 0.09 0.06 1
Appeal to authority
Appeal to fear/prejudice
Black-and-white Fallacy/Dictatorship
Causal Oversimplification
Doubt
Exaggeration/Minimisation
Flag-waving
Glittering generalities (Virtue)
Loaded Language
Misrepresentation of Someone's Position (Straw Man)
Name calling/Labeling
Obfuscation, Intentional vagueness, Confusion
Presenting Irrelevant Data (Red Herring)
Reductio ad hitlerum
Repetition
Slogans
Smears
Thought-terminating cliché
Whataboutism
Bandwagon
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00