UNTLing at SemEval-2020 Task 11: Detection of Propaganda Techniques in English News Articles

Maia Petee and **Alexis Palmer**
The University of North Texas
Denton, Texas
maiapetee@my.unt.edu, alexis.palmer@unt.edu

Abstract

Our system for the PropEval task explores the ability of syntactic and semantic features to detect and label propagandistic rhetorical techniques in English news articles. For Subtask 2, labeling identified propagandistic fragments with one of fourteen technique labels, our system attains a micro-averaged F1 of 0.40; in this paper, we take a detailed look at the fourteen labels and how well our model detects each of them. We also propose strategies to fill the gaps.

1 Introduction

This paper describes the UNT Linguistics contribution to the Propaganda Evaluation (PropEval) shared task at SemEval-2020 (Da San Martino et al., 2020). Our system uses non-neural classification methods to address the two linked subtasks: 1) detection of variable-length English-language news article spans that employ propagandistic rhetorical techniques; and 2) 14-way labeling of identified spans. The system focuses on semantically-oriented features, including word embeddings, sentiment/valence analysis, and named entity features.

The amount of text available in our vast global landscape, both content created by companies for public use and user-generated content, has grown far beyond the reach of human moderation. The prevalence of content with misleading rhetorical techniques, often promoting partisan political or social agendas, calls for automated detection and labeling of propaganda in textual media. Currently, systems like this are rare (for example, Proppy (Da San Martino et al., 2019)), and to our knowledge there are none in wide use. Widely-available, browser-based, automated, interactive detection of online propaganda could reduce the quantity of that propaganda and promote the formation of a better-informed public, acting as a bulwark against misleading content in real time. Such systems could influence rhetorical trends in the presentation of online news to skew away from the unnecessarily sensational and toward the factual.

Our system approaches the task of propaganda detection primarily from a semantic perspective, following the intuition that propagandistic text is different from non-propagandistic text in ways that are linguistically nuanced. The semantic information encoded in word and sentence embeddings, for example, has been shown to capture such nuances (Chen et al., 2013). We focus primarily on the Subtask 2 (Technique Classification) and made an official submission for this task only. Our Subtask 1 system achieved a macro-averaged F1 score of 0.349, but only after the end of the official submission period. Our system for Subtask 2 came in 27th out of 32 teams on the test set (F = 0.391) and 34th out of 42 teams on the development set (F = 0.409). Detailed results and error analysis appear below.

2 Data and System Overview

Data. The task data are English-language news articles scraped from the web. Within each article, gold standard propagandistic fragments are indicated via character offsets and then labeled with one of the 14 technique labels. In all, 18 different propagandistic techniques are represented in the data.[1] Because

[1]Descriptions and examples of the 18 techniques can be found here: https://propaganda.qcri.org/annotations/definitions.html.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1847–1852
Barcelona, Spain (Online), December 12, 2020.

Feature	Description
word	Text of token
prevfirstword	Text of previous token
nextfirstword	Text of next token
pos	Fine-grained PoS of token
prevwordpos	PoS of prevfirstword
nextwordpos	PoS of speech of nextfirstword
word.dependency	Syntactic dependency of token
prevword.dependency	Syntactic dependency of prevfirstword
nextword.dependency	Syntactic dependency of nextfirstword
bow_beginning	BoW preceding token, up to 5 tokens
bow_end	BoW following token, up to 5 tokens
semantic_left	Vector of bow_beginning
semantic_right	Vector of bow_end
v1 - v300	GloVe embedding for token

Feature	Description
text	Text of all words in fragment
semantic	Propaganda fragment vector
ents	Named entities (NEs) in fragment
enttypes	Types (PER, LOC, etc.) of NEs
intensity	# of words in VAD / fragment length
valence	All valence ratings in fragment
arousal	All arousal ratings in fragment
dominance	All dominance ratings in fragment

Table 1: Features used for Subtask 1 (left) and Subtask 2 (right).

some of the techniques occur with very low frequency in the task data, the task organizers collapsed the 18 techniques into 14 labels. One extremely low-frequency technique was eliminated, and two of the 14 labels merge similar low-frequency techniques into a single label class (1. Bandwagon; *Reductio ad Hitlerum*, and 2. Whataboutism; Straw Man; Red Herring). Training data comprise 371 articles with 5468 propagandistic spans. Development data contains 940 gold-labeled spans in a total of 75 articles.

Subtask 1: Propagandistic Span Identification. Subtask 1 requires the identification of propagandistic spans from raw text. We treat this as a sequence labeling task, using a conditional random field (CRF, (Lafferty et al., 2001)) classifier and a beginning/inside-propaganda vs. not-propaganda labeling approach. In other words, the first token of a propagandistic span is labeled B-P, the subsequent tokens of that same span are labeled I-P, and all other tokens are labeled N.

CRFs naturally incorporate contextual factors when identifying spans, which in this case vary in length from one or two tokens to multi-sentence chunks. For example, we convert the following sentence (propagandistic span underlined) to the labeling shown below: *Manchin says Democrats acted like babies at the SOTU (video) Personal Liberty Poll Exercise your right to vote* becomes [B-P, I-P$\times$5, N$\times$10].[2] Results for this three-way labeling task were very bad (see discussion in Section 4).

In a second approach, we reduce the granularity to the sentence-fragment level, breaking the text into chunks by splitting at label boundaries and at sentence boundaries; recall that the input texts are usually multi-sentence documents. For labeling, we collapse the distinction to simply propaganda (P) or not-propaganda (N), such that the example above would be labeled as the sequence: [P$\times$6, N$\times$10].

Final features for Subtask 1 appear in the left-hand side of Table 1. Basic features include the word itself, its immediate textual context and fine-grained part of speech labels for both target and context words. We also include syntactic dependency information, a bag-of-words feature consisting of the token's close context, and the token's word embedding, using GloVe embeddings(Pennington et al., 2014). Finally, we include word embeddings for the tokens in the BoW features. For the chunk approach, we tried the first and last words of each chunk, plus a bag-of-words with all words in the chunk, plus the words in the two adjacent fragments, word embeddings for the chunk and its neighbors, and sentiment scores for the chunk and its neighbors.

Subtask 2: Technique Classification. For Subtask 2, we use a Logistic Regression classifier to label individual propagandistic fragments with one of fourteen technique labels. See Table 2 for a list of labels. We first build a simple baseline classifier for Subtask 2, with only two features: a bag-of-words feature that included all the tokens in a given fragment and a fragment embedding, averaging over all words in the fragment. This baseline achieves a micro-averaged F1 of 0.225. Our final system additionally incorporates named entity (NE) features and features from the VAD Lexicon (see Table 1, right-hand side). The NE features are a list of all NEs in the fragment and their types. The VAD Lexicon is a sentiment-analysis dictionary that uses the Valence-Arousal-Dominance model of affect (Warriner et al., 2013). We use the

[2]This sentence reflects the somewhat noisy data extraction process, as shown by the ad fragment in the second half.

Label	Precision	Recall	F1 Score (Dev.)
System 1: B/I/N			
N	0.992	1.000	0.996
B-P	0.167	0.014	0.025
I-P	0.280	0.014	0.028
System 2: Binary			
N	0.949	0.984	0.966
P	0.931	0.806	0.864

Label	F-Score (Test)	F-Score (Dev.)
Baseline	0.252	0.265
Our System: Weighted Microaverage	**0.391**	**0.409**
Appeal to Authority	0.043	0.083
Appeal to Fear/Prejudice	0.053	0.119
Bandwagon; *Reductio ad Hitlerum*	0.000	0.000
Black/White Fallacy	0.000	0.000
Causal Oversimplification	0.029	0.267
Doubt	0.326	0.354
Exaggeration, Minimization	0.118	0.123
Flag-Waving	0.405	0.517
Loaded Language	**0.626***	**0.594***
Name-Calling, Labeling	0.367	0.333
Repetition	0.078	0.118
Slogans	0.176	0.250
Thought-Terminating Clichés	0.000	0.000
Whataboutism; Straw Man; Red Herring	0.000	0.059

Table 2: Results for Subtask 1 (left, no official submission) and Subtask 2 (right, official submission).

VAD Lexicon to calculate sentiment of the lexemes in each fragment across three dimensions rather than on a simple positive-negative scale. [3]

3 Models and Experimental Setup

For data preprocessing, we use spaCy's[4] NLP pipeline to get part-of-speech, dependency, named entity, and word embedding features. Specifically, we use GloVe embeddings (Pennington et al., 2014). For sentiment features, we access each token's valence, arousal, and dominance values from the VAD lexicon. Classifiers are trained and evaluated using Scikit-Learn[4] with default parameter settings, and all code has been released on Github.[5]

In the first phase of development, we use a 70/30 split of the 371 articles in the training set. In a later phase, we use all 371 articles for training, and we evaluate the system on the development set. Results in Table 2 are for the second configuration.

4 Results and Analysis

Subtask 1. For Subtask 1, we implemented two systems, one with B-P/I-P/N labeling, and the other with binary labeling. See results in Table 2, left-hand side. Here we mostly discuss the B-P/I-P/N system.

The B-P/I-P/N system struggles with the variability of the target spans' length and content. The label N is by far the most frequent in the training data, and the system is especially unlikely to correctly label B-P tokens. This underlabeling of B-P tokens consequently leads to underlabeling of I-P tokens, since without a correctly identified B-P, subsequent I-P tokens cannot be identified.

Feature analysis reveals that the most informative feature is the token's word embedding; 10 out of the 30 most informative features are individual vector elements. Another important set of features concerns the token's left-hand context: both the bag-of-words before the target token and the BoW's word embedding (bow_beginning and semantic_left) account for 13 out of 30 most informative features. These results are contrasted with their right-hand counterparts (bow_end and semantic_right), which when measured similarly account for only 3 out of 30 informative features.

It seems that the semantic content of tokens used in propagandistic rhetoric differs to some degree from tokens in non-propagandistic fragments. Additionally, it seems that the left-hand context is more important for signaling propaganda than the right-hand context. This could be due to the fact that it is easier for the system to detect a continuation of propagandistic speech (i.e., I-P tokens) than it is to detect the critical B-P tokens. This is possibly a result of the highly-skewed distribution.

[3] We also tried sentiment polarity and subjectivity features from spaCy, but these decreased overall performance and were thus not included in the final system.

[4] SpaCy: `https://spacy.io`; Scikit-Learn: `https://scikit-learn.org`

[5] `https://github.com/movetomars/PropEval`

	Baseline	BL+NE Feat.	BL+VAD Feat.	Final System
Overall	**0.225**	**0.391**	**0.344**	**0.409**
Appeal to Authority	0.138	0.087	0.000	0.083
Appeal to FearPrejudice	0.071	0.036	0.034	0.119
Bandwagon/Reductio ad Hitlerum	0.000	0.000	0.000	0.000
Black & White Fallacy	0.000	0.000	0.000	0.000
Causal Oversimplification	0.000	0.171	0.077	0.267
Doubt	0.250	0.214	0.295	0.354
Exaggeration/Minimization	0.121	0.027	0.056	0.123
Flag-Waving	0.180	0.506	0.476	0.517
Loaded Language	0.239	0.566	0.539	0.594
Name Calling/Labeling	0.145	0.318	0.286	0.333
Repetition	0.355	0.064	0.118	0.118
Slogans	0.182	0.182	0.255	0.250
Thought-Terminating Clichés	0.000	0.000	0.000	0.000
Whataboutism/Straw Man/Red Herring	0.000	0.056	0.065	0.059

Table 3: Feature ablation (on development set) of Subtask 2 Feature Groups

The binary system performs significantly better. We speculate on two possible contributing factors. First, this labeling removes a potentially false distinction between first tokens of propagandistic fragments and later tokens. Second, chunking the data in this manner slightly shifts the label distribution toward a more balanced state. This may also help.

Subtask 2. For Subtask 2, we ranked 34th out of 42 teams on the development set and 27th out of 32 teams on the test set. Results across all fourteen technique labels are displayed in Table 2, right-hand side. One of our goals is to achieve robust performance across labels, to achieve non-zero results for as many propagandistic techniques as possible. Largely, we accomplished this: for eleven of fourteen labels — all except (Bandwagon; *Reductio ad Hitlerum*) and Thought-Terminating Clichés — we achieve some accuracy on the development set (10/14 on the test set: no (Whataboutism; Straw Man; Red Herring) fragments are correctly identified for this dataset). The system performs worst on the two combined labels, plus two other low-frequency labels.

Results of a feature ablation study (Table 3) show the greatest increase in overall performance from the named entity features (ents and enttypes): accuracy on two labels (Flag-Waving and Loaded Language) increases from .18 and .24 to .51 and .57, respectively. The increase for Flag-Waving fragments could be explained by the fact that these fragments, characterized by blind nationalism and by definition almost requiring the inclusion of named entities, almost always include a reference to the idealized entity ("the UK," "the new Sweden," "Zionism," and numerous uses of "America" and "American"). Conversely, the increase in detection of Loaded Language fragments might be caused by a relative lack of named entities in these fragments. Performance on Name Calling/Labeling also increases with the addition of named entity features, for perhaps obvious reasons (if a fragment employs the use of pejorative name-calling, it is more likely that the name/label will be an NE).

The VAD features — valence, arousal, and dominance — boost performance on categories related to the cultivation of a particular emotion in the audience, such as Appeal to Fear/Prejudice and Doubt. This makes sense, given that rhetoric designed to instill a particular emotion in others often models that emotion to "set the mood," as in fragments like "warning that an outbreak could occur at any time" and "evolve into something far more dangerous." The sentiments in these fragments have (expert) speculation as supporting evidence, but their vaguely- and anxiously-written premonitions mirror anxiety about worst-case outcomes and do not limit themselves to presenting factual information. Certain labels, such as Appeal to Authority and Repetition, achieve their highest scores with the baseline implementation.

Error Analysis. For Subtask 1, our system often incorrectly labels B-P tokens as N, and these errors snowball into a large number of predicted N labels for tokens whose true label is I-P. When the system correctly predicts a B-P label, it generally labels the next 10+ tokens as I-P, despite the fact that span length is highly variable. In fact, many propagandistic spans consist of just one or two tokens (e.g., in the Name-Calling/Labeling category: "futuristic," "despotic leader," etc. and the Loaded Language "very,

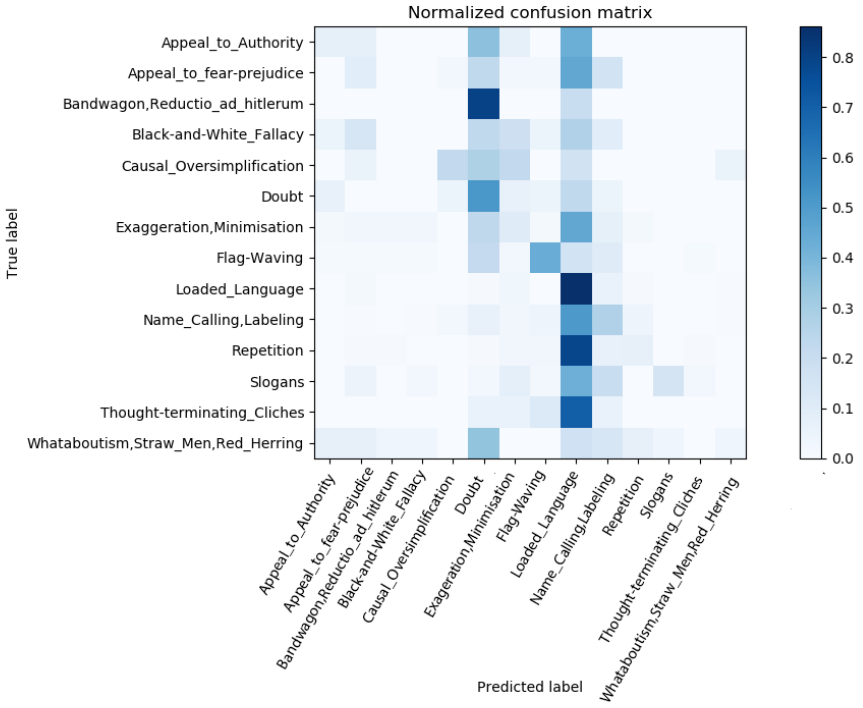

Figure 1: Normalized confusion matrix for Technique Classification (Subtask 2).

very"), meaning that the system's enthusiasm for the I-Plabel results in many misclassifications of what are actually non-propagandistic tokens.

The Subtask 2 classifier overwhelmingly predicts labels from a small subset (see Figure 1 for details), often incorrectly labeling fragments as Loaded Language or Doubt. When examining the gold-labeled Loaded Language fragments, we see high diversity in these segments, both semantically and in terms of length. There are many one- and two-word fragments, and many others that run for multiple sentences. Loaded Language seems to be a sort of super-category, overlapping with many other categories. Many fragments that are labeled with other techniques, such as Exaggeration/Minimization (e.g. "their dream of independence is all but dead"), also use high-affect language that elicits an emotional response in the reader, making it easy to miss pragmatic subtleties and mislabel fragments *containing* loaded language as belonging to the Loaded Language category. A similar effect is perhaps in play in the misclassification of many fragments as Doubt: almost all of the (Bandwagon; *Reductio ad Hitlerum*) fragments were labeled as Doubt. This could be due to similar sentiment readings across the two labels: both labels are likely to use low-valence (negative) and high-arousal (intense) lexemes. They are also likely to use many named entities of similar categories, given that the goal of both techniques is to cast aspersions on an individual or population.

Future Iterations. One significant shortcoming of our Subtask 1 implementation is its inability to address the skewed distribution of labels in the training data. At the token level, there is a marked imbalance between the number of propagandistic tokens and the number of non-propagandistic tokens. Roughly 50,000 tokens are labeled as B-P and I-P in the training articles, while almost 12 million tokens are labeled N. Specifically, only .004% of tokens in the training data are propagandistic. The ratio is likely to be similarly skewed in development and testing data. To address this problem in the future, we will try downsampling the N class.

Some of the label combinations seem to obfuscate critical characteristics of the data: for example, the combination of the previously-separate labels of Bandwagon and *Reductio ad Hiterlum*. These two techniques are not similar in presentation; the Reductio ad Hitlerum fragments are semantically distinct

from the Bandwagon fragments in that they tend to contain entities known for having been terrorists or terrorized, such as Adolf Hitler, the Jews, the Palestinians, ISIS, Henry VIII, and the Soviet Union. Bandwagon fragments do not share this characteristic. Since the combination of labels seems to have been undertaken for balancing purposes instead of true similarity, our system would in the future separate these combined labels.

Finally, the features our system utilizes are not focused around distinguishing between the pragmatic subtleties inherent to each task. A future round of feature engineering will revolve around finding the differences between frequently misclassified labels and their predicted counterparts, and modeling these to achieve greater coverage among all tasks.

5 Conclusion

Semantic information such as word embeddings, as we have demonstrated, can be used to some small effect to detect and label propagandistic rhetoric in news articles. However, relying upon semantic and sentiment information alone is far from a complete approach to this complex problem. In the future, we will downsample the skewed Subtask 1 data to maximize our classifier's performance. For Subtask 2, we will create a more granular labeling taxonomy and will seek to create semantic- and discourse-level features that distinguish the more specific technique labels (e.g., Thought-Terminating Clichés) from the more general (e.g., Loaded Language). Taking these next steps will ideally help to advance the automatic detection of propaganda in news articles.

Acknowledgments

We would like to thank Taraka Rama for helpful conversations and technical support throughout this project. Thanks also to Brian Cooper for useful suggestions. Finally, computational resources were provided by the University of North Texas High-Performance Computing Services, a division of the Research IT Services, University Information Technology, with additional support from UNT Office of Research and Economic Development.

References

Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, EMNLP-IJCNLP 2019, Hong Kong, China, November.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain, December.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.