Detection of Propaganda Using Logistic Regression

Jinfen Li
College of Arts and Sciences,
Syracuse University
jli284@syr.edu

Zhihao Ye
College of Information
Science and Engineering,
Hunan University
zhihaoye.chn@qq.com

Lu Xiao
School of Information Studies,
Syracuse University
lxiao04@syr.edu

Abstract

Various propaganda techniques are used to manipulate peoples perspectives in order to foster a predetermined agenda such as by the use of logical fallacies or appealing to the emotions of the audience. In this paper, we develop a Logistic Regression-based tool that automatically classifies whether a sentence is propagandistic or not. We utilize features like TF-IDF, BERT vector, sentence length, readability grade level, emotion feature, LIWC feature and emphatic content feature to help us differentiate these two categories. The linguistic and semantic features combination results in 66.16% of F1 score, which outperforms the baseline hugely.

1 Introduction

Attributes of social media communication make it challenging for a user to interpret someones comment and to examine the truthfulness of the information. For example, a social media message can be anonymous, from real people, or automatically generated, making it difficult to identify its source. Because of this challenge to interpret and evaluate a social media message, social media users are found to be persuaded by views that have no factual basis (Guo et al., 2018). They are influenced by misinformation and disinformation.

Various definitions are given in the literature to explain what propaganda is (for a list of such definitions, please see: https://publish.illinois.edu/mirasotirovic/whatispropaganda). Focusing on the techniques in propaganda, we adopt Elluls definition that propaganda is "A set of methods employed by an organized group that wants to bring about the active or passive participation in its actions of a mass of individuals, psychologically unified through psychological manipulation and incorporated in an organization" (Ellul, 1966). People use propaganda techniques to purposely shape information and foster predetermined agenda (Miller, 1939; Weston, 2018). With the fast and wide spread of online news articles, it is much desired to have computing technologies that automatically detect propaganda in these texts.

This study presents our approach to a shared task that is aimed at detecting whether an given sentence from a news article is propagandistic. The shared tasks are part of 2019 Workshop on NLP4IF: censorship, disinformation, and propaganda , co-located with the EMNLP-IJCNLP conference. We focused on one of the task, which is referred to as SLC (Sentence- level Classification). In our approach, we came up with various features and classified the sentences using Logistic Regression.

2 Our Approach

Our model includes a list of linguistic features and semantic features extracted from BERT. After experiments on the BERT model and other machine learning models, we got the best performance using Logistic Regression.

2.1 Data

(Da San Martino et al., 2019a) provided with a corpus of about 500 news articles and splited the corpus into training, development and test, each containing 350, 61, 86 articles and 16,965, 2,235, 3,526 sentences. Each article has been retrieved with the newspaper3k library and sentence splitting has been performed automatically with NLTK sentence splitter (Da San Martino et al., 2019a).

2.2 Our Features

We identified a list of features and selected the top 98% using feature selection tool SelectKBest of Sklearn with score funtion of f_classif (https://scikit-learn.org/stable/modules/generated/sklearn.

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 119–124

feature_selection.SelectKBest.html). Our final features including TF-IDF, length, readability grade level, emotion, LIWC and emphatic features, and the semantic features extracted from BERT.

2.2.1 TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) (Jones, 2004) gives us the information of term frequency through the proportion of inverse document frequency. Words that have small term frequency in each document but have high possibility to appear in documents with similar topics will have higher TF-IDF, while words like function words though frequently appear in every document will have low TF-IDF because of lower inverse document frequency. We used feature selection tool of sklearn based on ANOVA to select top 100 features from over 40,000 words.

2.2.2 Sentence Length

We found that the propagandistic sentences are more likely to be longer than the non-propagandistic ones, so we came up some features to capture this information. We have categorical feature **Short or Long Document** and used 1 to denote that it is a long document. A sentence belongs to a short document if it has less than eight tokens; otherwise, it belongs to a long document. We also have discrete features including **Text Length**(the number of characters in a sentence), **Word Count** and **Word Count Per Sentence**.

2.2.3 Readability Grade Level

We used The Flesch Grade Level readability formula, which is also commonly referred to as the Flesch-Kincaid Grade Level to calculate the readability grade of each text (Kincaid et al., 1975). The Flesch-Kincaid Grade Level outputs a U.S. school grade level, which indicates the average student in that grade level can read the text. For example, a score of 9.4 indicates that students in the ninth grade are able to read the document. The formula is as follow.

$$FKRA = (0.39 * ASL) + (11.8 * ASW) - 15.59$$

where, **FKRA** = Flesch-Kincaid Reading Age, **ASL** = Average Sentence Length (i.e., the number of words divided by the number of sentences), **ASW** = Average number of Syllable per Word (i.e., the number of syllables divided by the number of words). The average grade level is eighth and twelfth for non-propagandistic and propagandistic sentences, respectively.

2.2.4 Emotion Feature

Studies about the relationship between emotion and propaganda techniques are conducted. For example, (Kadir et al., 2016) found out that propaganda techniques in YouTube conjure peoples emotion that could affect unity. We took advantage of these studies by adding emotion features for SLC task.

- **NRC VAD Lexicon** (Mohammad, 2018); **NRC Emotion Lexicon** (Mohammad and Turney, 2013); **NRC Affect Intensity Lexicon** (Mohammad and Bravo-Marquez, 2017). We calculated the total score of the words listed in these lexicons respectively, and normalized the score between zero and one for each sentence.

- **MPQA** (Wilson et al., 2005), **Bing Liu** (Hu and Liu, 2004), and **AFINN** (Nielsen, 2011). We calculated the percentage of words with positive and negative emotions respectively in these lexicons for each sentence.

- **Insult** Noted that insult words are likely to be used in Name Calling and Labeling techniques, we refer to a lexicon that contains insult words from the http://www.insult.wiki/wiki/Insult_List. We calculated the count of insult words appearing in a sentence and normalized it by the token counts.

- **LIWC Emotion Lexicon**
  Affect the LIWC dictionary includes the overall affect including positive emotions, negative emotions, anxiety, anger and sadness; **Negative Emotions** it also includes negative emotion words correspond with human ratings of the writing excerpts (Alpers et al., 2005); **Anger** and some anger words without considering the context like 'hate, kill, annoyed'. We combined these three emotion information provided by LIWC emotion lexicon with the others provided by the lexicons mentioned above as the final emotion features.

2.2.5 LIWC Feature

- **Dictionary Words**: Percentage of all words captured by the dictionary, which refers to the collection of words that define one particular of the 80 categories (Tausczik and Pennebaker, 2010).

- **Article** The use of article can tell us some information about gender and the personality. (Newman et al., 2008) found that males had higher use of large words and articles than women. (Pennebaker and King, 1999) showed that articles were less frequent in the writing of people who scored high on extraversion.

- **Conjugations** Depth of thinking is reflected in complexity, and people use conjunctions to join multiple complex thoughts together to deepen their thoughts (Graesser et al., 2004).

- **Quote** The use of quote distracts us from the main body of the text to the content in the quotes. For example, ironic content (e.g. "A researcher with the organisation, Matthew Collins, said it was 'delighted' with the decision."), slogans (e.g. "Time for US to do the same.") and and loaded language (e.g. "Muslin Invaders") are put in the double quotes.

2.2.6 Emphatic Content in Double Quote

Researchers have identified many standard techniques (Koob, 2015; Zollmann, 2019) used in propaganda, such as slogans, name calling and loaded language, which often include the emphatic content in the title format(every word begins with capital letter) or every letter of the word is capitalized in the double quote. Therefore, our model includes a feature that reflects this aspect.

- **Slogans.** A slogan is a brief and striking phrase that may include labeling and stereotyping (Da San Martino et al., 2019b). Slogans tend to act as emotional appeals (Dan, 2015). Ex.: President Donald Trump Proposes "Simple Immigration Plan": Illegals Have To Go!

- **Name Calling.** Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises (Miller, 1939). Ex.: Democrats Friend Louis Farrakhan In Iran: "Death to America!" America Is The "Great Satan" Neither Manafort nor these "Russians" are in the visitor logs.

- **Loaded Language** Using words/phrases with strong emotional implications (positive or negative) to influence an audience (Weston, 2018). Ex.: Dem Candidate Ilhan Omar Defending Tweet On "The Evil Doings Of Israel" by Frank Camp, Daily Wire, October 28, 2018:

To translate the emphatic content in double quote into feature, we used a feature called "isEmphatic". If we found the stressed content in double quote in the format of title or upper letter in a sentence, we would use 1 to denote the sentence has emphatic content in it.

2.2.7 BERT Features

In order to further extract the semantic information of text, we apply sentence vectors generated by the state-of-the-art models, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). Specifically, we use pretrained BERT model to predict text category, but we do not directly adopt BERT results as our final results because of the better performance of Logistic Regression. We use the vector obtained by BERT's hidden layer which can represent the semantic feature. The experimental result shows that BERT features can improve hugely on F1 score on the development dataset.

3 Experiment

3.1 Data Cleaning

For the input of BERT model, we removed the punctuation, and changed all the uppercase letters to lowercase. Also, we changed all clitics to full words (e.g. "isn't" becomes "is not"). For the linguistic features extraction part, we did not apply the same method as above, because uppercase letter and quotes are important features for this task.

3.2 Model

We used two models, one is the pretrained BERT model and the other is Logistic Regression. The architecture of our model is shown in Figure 1.

3.3 Model Setup

We used the pretrained uncased BERT-Base model and fine-tuned it using the following hyper-

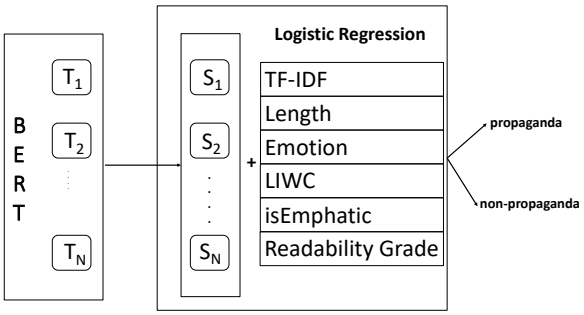

Figure 1: The architecture of our model

parameters: batch size of 16, sequence length of 70, weight decay of 0.01, and early stopping on validation F1 with patience of 7. For optimization, we used Adam with a learning rate of $2e-5$. We tuned our models on the train dataset and we report results on the development dataset. For the Logistic Regression, we used the solver of LBFGS, penalty of l2, C of 1.0 and we used "balanced" mode to automatically adjust weights inversely proportional to class frequencies in the input data.

4 Results

Table 1 shows the ablation study results for the SLC task. We used the Logistic Regression with sentence length(the number of characters) feature to be the baseline. To test the importance of each individual feature in the classification, we applied them to Logistic Regression one at a time, including readability grade level, sentence length, LIWC, TF-IDF, emotion and BERT. Among these features, readability and sentence length increased 3.13% and 5.34% of F1 score, while LIWC, TF-IDF and emotion features increased 7.28%, 12.76% and 12.92% of F1 score respectively. These results suggest that the length and the complexity of a sentence is effective to differentiate propagandistic sentences from the non-propagandistic ones, but not as effective as LIWC, TF-IDF and emotion do. The implication is that while propaganda techniques are likely to appear in a complex and longer sentences, there are also long non-propagandistic sentences containing complex words. In addition, some propaganda techniques like slogan are not necessarily expressed in long sentences. The difference of language use, reflected by the words, punctuations (LIWC), term frequency inverse document frequency (TF-IDF) and the emotional expression (emotion) shapes a more fit boundary between propagandistic and non-propagandistic sentences.

We further explored the efficiency of semantic features extracted from BERT. The BERT feature improves the most among all the features in Logistic Regression by 18.05% of F1 score. This indicates that the higher granularity representation of a sentence better capture the presence of propaganda techniques. We conducted experiment using the pretrained and fine-tuned BERT and obtained huge improvements on the SLC task. As shown in Table 1, BERT performed better than LR_bert but worse than LR†‡, which indicates that the transfer learning when considering single semantic variable is not as effective as the combination with other linguistic features. Furthermore, we explored the effect of the isEmphatic feature introduced in Section 2.2.6. The isEmphatic feature is extremely sparse. We compared the performances of two classifiers that had the same feature set except the presence of isFmphatic, i.e., LR† and LR†‡ . The isEmpahtic feature improved the performance as evidenced by the slightly increase from 65.08% to 66.16%.

Model	Precision	Recall	F1
LR_base	38.80	49.42	43.47
LR_read	41.15	53.45	46.50
LR_length	42.49	57.38	48.82
LR_liwc	42.11	63.87	50.75
LR_tfidf	45.76	72.94	56.23
LR_emotion	49.58	65.36	56.39
LR_bert	55.50	69.01	61.52
BERT	67.00	63.19	65.04
LR†	57.10	75.64	65.08
LR†‡	58.00	77.00	**66.16**

Table 1: Sentence-level (SLC) results. † represents the inclusion of features other than isEmphatic into the model. ‡ represents the inclusion of isEmphatic features into the model

5 Related Work

There are a number of researchers applying machine learning to automatically identify Propagandistic news articles. (Barrón-Cedeño et al., 2019) presented PROPPY, the first publicly available real-world, real-time propaganda detection system for online news and they show that character n-grams and other style features outperform existing alternatives to identify propaganda based on word n-grams. (Ahmed et al., 2017) proposed

a fake news(propagandistic news articles) detection model that use n-gram analysis and machine learning techniques. (Orlov and Litvak, 2018) presented an unsupervised approach using behavioral and text analysis of users and messages to identify groups of users who abuse the Twitter microblogging service to disseminate propaganda and misinformation.

Most relevant to our study, (Da San Martino et al., 2019b) proposed a BERT based technique to identify propaganda problems in the news articles. Specifically, the researchers trained a Multi-Granularity BERT model that includes multiple levels of semantic representations on two tasks. One task FLC identifies which of 18 propaganda techniques is/are present in the given fragment of the text. The other, namely, SLC is about classifying whether the given sentence is propagandistic. Different from their approach, we focused on the SLC task, and used the fine-tune BERT vectors combining various linguistic features, and fitted into a Logistic Regression model. Also, we only used the vectors extracted from the hidden layers of BERT to be part of our features. With a similar but smaller dataset, the researchers' model achieved 60.98% of F1 score, while ours is 66.16%. In this competition, our team ranked 9th out of 29 teams on the development set, with the F1 score of the top team being 2.7% higher than ours.

6 Conclusion and Future Work

In this paper, we focused on the sentence-level propaganda detection task and developed an automatic system based on some effective features. We got features including TF-IDF, length, emotion, readability level, LIWC, emphatic and BERT. Our ablation study shows that the length and complexity of sentence help to improve the performance slightly, comparing to the use of language reflected in specific term, frequency and emotional expression, which captures more propagandistic information. The semantic information extracted from BERT is crucial in detecting propaganda techniques, which improves the F1 score the most. The combination of these features and the BERT feature achieved the best performance with the Logistic Regression model. The F1 score is 66.16%.

Compared to (Da San Martino et al., 2019b), our approach focus more on the linguistic features combined with semantic features extracted from BERT, and use machine learning model , while they use the deep learning model with a high granularity task to improve performance on low granularity task.In terms of the performance, our F1 score is 66.16% whereas theirs is 60.98%. On the other hand, we noted that the two studies used different versions of the propaganda datasets, which may contribute to the observed difference in the performances.

In the future, we plan to embed the features we designed in the BERT model or studied more features from the propaganda techniques to improve the performance.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138. Springer.

Georg W Alpers, Andrew J Winzelberg, Catherine Classen, Heidi Roberts, Parvati Dev, Cheryl Koopman, and C Barr Taylor. 2005. Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior*, 21(2):361–376.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Giovanni Da San Martino, Alberto Barron-Cedeno, and Preslav Nakov. 2019a. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom (NLP4IF): Censorship, Disinformation, and Propaganda*, NLP4IFEMNLP '19, Hong Kong, China.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, Hong Kong, China.

Lavinia Dan. 2015. Techniques for the translation of advertising slogans. In *Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue*, volume 3, pages 13–21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacques Ellul. 1966. *Propaganda*. Knopf New York.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Lei Guo, Jacob A Rohde, and H Denis Wu. 2018. Who is responsible for twitters echo chamber problem? evidence from 2016 us election networks. *Information, Communication & Society*, pages 1–18.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

S Abd Kadir, A Mohd Lokman, and Toshio Tsuchiya. 2016. Emotion and techniques of propaganda in youtube videos. *Indian journal of science and technology*, 9:S1.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Jeff Koob. 2015. *Ad Nauseam: How Advertising and Public Relations Changed Everything*. iUniverse.

Clyde Raymond Miller. 1939. *How to detect and analyze propaganda*. Town Hall, Incorporated.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

FÅ Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblog,[in:] m. rowe et al. In *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages 718 in CEUR Workshop Proceedings*.

Michael Orlov and Marina Litvak. 2018. Using behavior and text analysis to detect propagandists and misinformers on twitter. In *Annual International Symposium on Information Management and Big Data*, pages 67–74. Springer.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Florian Zollmann. 2019. Bringing propaganda back into news media studies. *Critical Sociology*, 45(3):329–345.