UAIC1860 at SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles

Vlad Ermurachi[1], Daniela Gîfu[1,2]
[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi
[2]Institute of Computer Science, Romanian Academy - Iasi Branch
{vlad.ermurachi, daniela.gifu}@info.uaic.ro

Abstract

The "Detection of Propaganda Techniques in News Articles" task at the SemEval 2020 competition focuses on detecting and classifying propaganda, pervasive in news article. In this paper, we present a system able to evaluate on fragment level, three traditional text representation techniques for these study goals, using: tf*idf, word and character n-grams. Firstly, we built a binary classifier able to provide corresponding propaganda labels, propaganda or non-propaganda. Secondly, we build a multiclass model to identify applied propaganda.

1 Introduction

Propaganda is a strong component of media ideas making it easier for reputation of people with high stature and to organizations (Thota *et al.*, 2018; Gifu *et al.*, 2014). Research on detecting propaganda has focused, especially, on news articles (Fitzmaurice, 2018; Gifu and Dima, 2014). Using a range of psychological and rhetorical techniques, propaganda intends to manipulate deliberately people's beliefs, attitudes or actions (Al-Hindawi and Kamil, 2017). Consequently, automatic detection and classification of propaganda in news articles is a challenging work (Martino et al., 2019).

The goal of this paper is to implement an automatic system, which imply two fragment level classifications for the presence of propaganda in news articles. First, a binary classification model, able to provide corresponding propaganda labels, propaganda or non-propaganda. Second, a multilabel multiclass model in order to identify applied propaganda.

The rest of the paper is structured as follows: section 2 describes other works related to propaganda identification, section 3 presents the dataset and methodology of this study, section 4 briefly relates an analysis and the results we have obtained, followed by section 5 with the conclusions.

2 Related Work

This topic has attracted significant attention in recent years, evidenced by increasing number of workshops of the same competition (e.g. Fake News Challenge Stance Detection Task 2018, SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection). Thus, work on this topic was never followed by high results, as this problem is highly subjective and text classification even for humans is very controversial and biased. Most of the authors used Bag of Words features, usually normalized with tf*idf (Saleh et al., 2019; Barro´n-Ceden˜o et al., 2019a) or character n-grams features for stylistic purposes (Stamatos, 2009).

For this research, we focused more on the character-level features, which are capturing various style markers (e.g. prefixes, suffixes, etc.) found in recent research (Barro´n-Ceden˜o et al., 2019b). Martino and his team mention 18 most important techniques used in news articles, considered with significant values for this task: loaded language, red herring, obfuscation, intentional vagueness, confusion, and

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1835–1840
Barcelona, Spain (Online), December 12, 2020.

straw man (Weston, 2018); name calling or labelling (Miller, 1939); repletion and black- and-white fallacy, dictatorship (Torok, 2015); exaggeration or minimization and appeal to authority (Jowett and

O'Donnell, 2012); appeal to fear/prejudice; flag-waving and bandwagon (Hobbs and Mcgee, 2008); causal oversimplification; whataboutism (Richer, 2017); reductio ad hitlerum (Tenibaum, 2009), etc.

Because for SemEval-2020 Task 11 there were not enough instances to represent valuable information, some of these techniques had been merged (Whataboutism with Straw men and Red herring; Bandwagon with reduction ad hitlerum) and other were eliminated (Obfuscation, Intentional Vagueness, Confusion).

3 Dataset and Methods

This section contains details about both datasets built as part of SemEval-2020 Task 11 "Detection of Propaganda Techniques in News Articles" and the study methodology, which was the basis for solving both sub-tasks.

3.1 Dataset

The dataset consists of news articles, retrieved with the newspaper3k library, in plain text format, split in two parts. The first part has two folders, train-articles and dev-articles, and the second part, a third folder for the test set. Each article has the following structure: a title followed by an empty row and the body content, starting with the third row, one sentence per line. For automated sentence splitting, NLTK was used. For binary classification issue, we trained four models, using 370 news articles manually annotated by six annotators. The indexes for the fragments containing propaganda were in separate .TSV file.

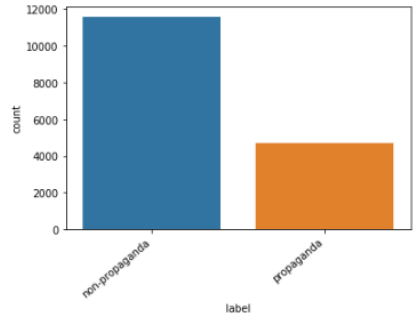

Figure 1. Binary classification with training dataset.

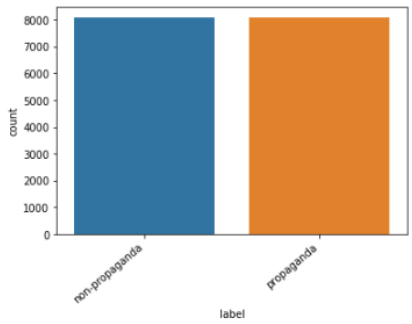

Figure 2. Balanced dataset binary classification.

Retrieving the information from each article based on the given indexes, we identified 5468 sentences containing at least one of the propaganda techniques and 10577 sentences that do not contain propagandistic content (see Figure 1). We noticed that there is a high imbalance between distributions of classes in the dataset, which may lead to poor results when training the model. In order to solve this high data disproportion, we oversampled the minority class (see Figure 2). For multinomial classification problem, we have used 6129 propagandistic fragments distributed as in Table 1.

Label	**Instances**
Appeal to Authority	144
Appeal to fear-prejudice	294
Bandwagon, Reductio ad hitlerum	72
Black-and-White Fallacy	107
Causal Oversimplification	209

Doubt	493
Exaggeration, Minimisation	466
Flag-Waving	229
Loaded Language	2123
Name Calling, Labeling	1058
Repetition	621
Slogans	129
Thought-terminating Cliches	76
Whataboutism, Straw Men, Red Herring	108
Total	6129

Table 1: Class distribution for propaganda technique classification task.

3.2 Methodology

For the sub-task Span Identification (SI), the first objective was to retrieve the fragments from the articles and classify them into two categories: those containing propaganda labeled with 'propaganda' and all the other fragments labeled with 'non-propaganda'. For the sub-task Technique Classification (TC), the first objective was to retrieve the fragments from the articles and classify them into multiple categories, considering those 14 propaganda techniques (Martino *et al*., 2019). Once our data frame was created, we pursued to the dataset pre-processing. In order to create a reliable dataset, we automatically striped the redundant information, like stop words and special characters using NLTK library. In addition, we created a custom transformer for removing initial and end white spaces and converting text into lower case.

Once our dataset was cleaned, we took our next step to feature engineering. Features are generally designed by examining the training set with an eye to linguistic intuitions and the linguistic literature on the domain (Jurafsy and Martin, 2019). Given the consistent use of linguistic attributes for training machine learning models and results from previous papers for propaganda detection, we considered bag of words and tf*idf scores appropriate for this task.

Using bag-of-words model the text was converted into a matrix of occurrence of words within the given fragments. It focuses on whether given words appear or not in the text, and generates a document term matrix. Applying scikit- learn's CountVectorizer function and defining character n-grams in range (1, 6) we got the numerical representation of the texts.

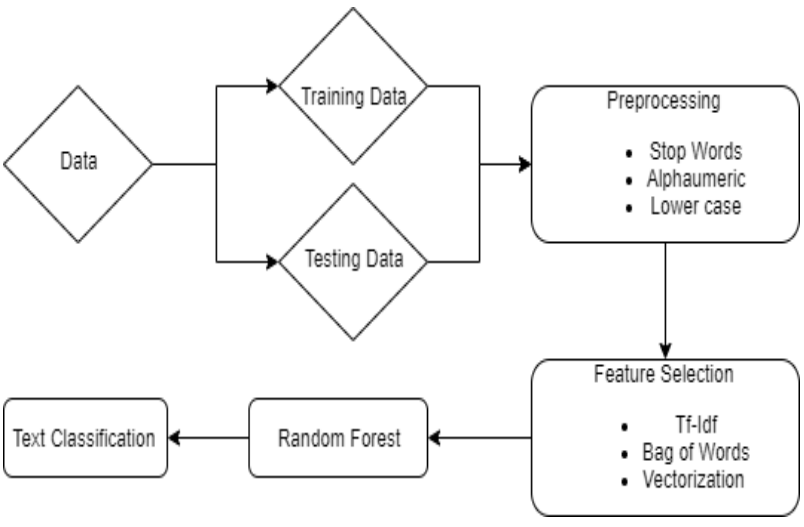

Figure 3: The system architecture.

In addition, we included the statistical measure TF*IDF (*Term Frequency-Inverse Document Frequency*) in order to evaluate how relevant a token is to a document in a collection of news articles. This was a simple way of normalizing the Bag of Words by looking at each word's frequency in comparison to the document frequency. The reason of using tf*idf instead of the raw number of occurrences of a token in a given text is to scale down the influence of tokens that appear very often in the provided corpus and thus are generally less informative than features that occur in a smaller fragment of the training corpus.

Sub-Task 1: Span Identification (SI). We analyzed the training dataset to identify the fragments for propaganda, labeled with three distinguished tags: id (the identifier of the article), begin offset (the character where the covered span begins, being included), end offset (the character where the covered span ends, being not included). Based on these labels, we crafted a set of rules to identify propaganda and sentences of news articles were randomly generated. The application code was written in the Python programming language and the results are presented in Table 2.

Sub-Task 2: Technique Classification (TC). We analyzed the dataset in order to classify on the fragment-level into one of the 14 classes. The labelled file contained four columns: the article id, the propaganda technique, the begin offset, which is the character where the covered span begins (included) and end offset which is the character where the covered span ends (not included). We used this restriction and made the second submission, with the results presented in Table 3.

4 Results

Below, the results for each individual task using the development and test sets are presented. We report Precision (P), Recall (R) and F1-score (F1), for each baseline on all classes. The official submission for the SI task was 0.33 and 0.43 for the TC-task.

Sub-Task 1: SI

Dev Set			
Model	**F1**	**Precision**	**Recall**
Naive Bayes	0.16840	0.10923	0.36748
Logistic Regression	0.29666	**0.22792**	0.42477
Random Forrest Classifier	**0.30272**	0.20693	**0.56368**
Support Vector Machine	0.26479	0.18283	0.47994
Test Set			
Random Forrest Classifier	0.33210	0.24490	0.51574

Table 2. Span identification results.

In Table 2 we see that the Random Forrest model has the best performance on the development set for the detection of propaganda in news with a F1 of 0.30 and a Recall of 0.56, while the highest precision was achieved using Logistic Regression – 0.22. However, the final submission on the test-set was done using Random Forrest algorithm.

Analyzing the particular features, we observed that word and character n-grams perform almost identically, with character n-gram features performing slightly better in recall score for propaganda sentences and precision in non-propaganda phrases, while word n-grams achieving higher results in all the other cases. These two features correlate well with each other as well, reporting high results for both classes.

Sub-Task 2: TC

Label	Dev Set F1 Score	Test Set F1 Score
Appeal to Authority	0.10526	0.17391
Appeal to fear-prejudice	0.16129	0.22120
Bandwagon,Reductio ad hitlerum	0.33333	0.09756
Black-and-White Fallacy	0.00000	0.02899
Causal Oversimplification	0.13793	0.07595
Doubt	0.42029	0.36406
Exaggeration,Minimisation	0.23656	0.21008
Flag-Waving	0.60000	0.38776
Loaded Language	0.57879	0.62328
Name Calling,Labeling	0.40367	0.42966
Repetition	0.14118	0.11159
Slogans	0.04545	0.11429
Thought-terminating Cliches	0.06897	0.05556
Whataboutism,Straw Men,Red Herring	0.00000	0.04255
Micro-averaged F1 measure	**0.43744**	**0.41173**

Table 3. Technique classification results

Table 3 reflects the results for second sub-task, which yield an improvement in performance, especially for the classes with many instances. The overall F1 on dev-set is 0.43 and 0.41 on testing dataset. It seems like the system did not encounter any issues predicting Loaded Language or Name Calling. However, it found problematic to classify under-sampled techniques like Black-and-White Fallacy or Whataboutism, Straw Men, Red Herring.

Taking a closer look at the misclassified examples can facilitate the development of machine learning models, pointing out instances that proved to be difficult in classification and can be analyzed for future improvements. Based on a short analysis of sentences, we assume that some of the model's errors are due to low number of examples in these classes or poor annotation, as it might be challenging to find specific patterns in highly biased sentences and even for a human it could be difficult to classify them correctly

5 Conclusion

This paper presents a system participating at SemEval Task 11. Since we performed an exhaustive investigation of propaganda detection at the fragment level, our experimental results showed that linguistic features like character and word n-grams are remarkably efficient for both tasks. The overall results are satisfactory and exceeds the baseline; however, there is still room for improvement, in predicting the techniques. Larger and well annotated dataset would provide more opportunities for exploring the issue of propaganda detection in news articles, in addition building a dataset sufficient in size and diversity will allow experiments with deep learning methods,

References

Al-Hindawi, F. H. and Kamil, S. I (2017). *The Pragmatic Nature of Manipulation*. In The Pragmatics of Manipulation in British and American Political Debates.

Barrón-Cedeño, A., Jaradat, I., Martino, G.D.S., Nakov, P. (2019a). *Proppy: Organizing News Coverage on the Basis of Their Propagandistic Content*". In Information Processing and Management, DOI:10.1016/j.ipm.2019.03.005.

Barrón-Cedeño, A., Martino, G.D.S., Jaradat, I., and Nakov, P. (2019b). *Proppy: A System to Unmask Propaganda in Online News*. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI'19), Honolulu, HI, USA.

Fitzmaurice, K.. (2018). *Propaganda*. In Brock Education Journal. 27. 10.26522/brocked.v27i2.580.

Gîfu, D., Teodorescu, M., Ionescu, D. (2014). *Pragmatical Rules for Success*. In International Letters of Social and Humanistic Sciences, vol. 26, 18-28.

Gîfu, D., Dima, I.C (2014). *An Operational Approach of Communicational Propaganda*. In International Letters of Social and Humanistic Sciences, vol. 23, 29-38.

Giovanni Da San Martino, Alberto Barr´on-Cede˜no, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov.2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. InProceedings of the 14thInternational Workshop on Semantic Evaluation, SemEval 2020, Barcelona, Spain, September.

Hobbs, R. and Mcgee, S. (2008). *Teaching about propaganda: An examination of the historical roots of media literacy*. In Journal of Media Literacy Education, 6(62):56–67.

Hunter, J. (2015). Brainwashing in a large group awareness training? The classical conditioning hypothesis of brainwashing. Master's thesis, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Jowett, G. S. and O'Donnell, V. (2012). *What is propaganda, and how does it differ from persuasion?* In Propaganda & Persuasion, chapter 1, Sage Publishing, 1-48.

Martino, G.D.S., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P. (2019). *Fine-Grained Analysis of Propaganda in News Articles*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China.

Miller, C. R. (1939). *The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall*. The Center for learning.

Richter, M. L. (2017). *The Kremlin's platform for 'useful idiots' in the West: An overview of RT's editorial strategy and evidence of impact*. Technical report, Kremlin Watch.

Saleh, A., Baly, R., Barrón-Cedeño, A., Martino, G.D.S., Mohtarami, M., Nakov, P., Glass, J. (2019). *Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection*. In Proceedings of the 13th International Workshop on Semantic Evaluation.

Stamatatos, E. (2009). *A Survey of Modern Authorship Attribution Methods*. In Journal of the American Society for Information Science and Technology 60:3, 538–556.

Thota, A., Tilak, P., Ahluwalia, S., Lohia, N. (2018). *Fake News Detection: A Deep Learning Approach*. In SMU Data Science Review, 1:3, art. 10, http://scholar.smu.edu/datasciencereview/vol1/iss3/10

Teninbaum, G. H. (2009). *Reductio ad Hitlerum: Trumping the judicial Nazi card*. In Michigan State Law Review, 541.

Torok, R. (2015). *Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming*. In Proceedings of the Australian Security and Intelligence Conference, 58–65.

Weston, A. (2018). *A rulebook for arguments*. Hackett Publishing.