Advanced NLP - Propaganda detection and Its Classification

Candidate number-277307

Introduction

Problem Outline:

The concept of propaganda implies the expression of a particular viewpoint and undertaking an action by people or groups who are directed to the opinions and behavior change of yet another group for achieving their goals (Institute for Propaganda Analysis 1938). We are interested in propaganda from a journalistic point of view: how media framing, which involves the exclusion of some aspects to produce a favorable impression, can shape news by highlighting the positive or negative aspects (Jowett and O'Donnell 2012, p. 1). Propaganda moves to the mind and affects the rationality and approximates, no matter how bluntly or bluntly mysterious and ingenious it is. It creates the effect only when it goes unnoticed in the soul and mind. Accordingly propagandist can achieve high level of success influence because of that. Especially there, the case of disinformation and propaganda was probably seen in the recent circumstances, for instance, the Brexit and 2016 elections of the country of the US. 1. Along with web emergence, a freedom of expression in combination with a simple publication of contents online created a sufficient amount of contexts for information bargain. It was also exploited in social media which by now has made a million people obtain information in real-time. To an extent, this explains the capability to detect propaganda in written content that may hold the nuanced reading of language, the ability to perceive the subtleties that propaganda employs as the rhetorical strategy.

However, arguably the principal stumbling block dealing with context interpretation must be pointed out. Propaganda-like texts frequently provide contextual clues and framing to get across their serious meaning, giving relevance to AI algorithms to understand the situation in which these words have arise. The requirement of complicated NLP implementations which are effectively capable of context-sense and even able to read real meaning between the lines of the text is the outcome.

On top of that, one algorithm that does the work of propaganda detection has to be able to detect hidden clues and symbolic meanings in the message text. Propaganda scripts can be full of meanings mimics through the use of such linguistic tools as speaker arousing words, rhetorical devices, and narrative techniques intended to trigger emotional reactions or just mentality twist. Identifying these implicit signs of propaganda involve the computational processing of the text beyond a simple grammar level to a complex interpretation involving the natural language processing of terms and phrases to find underlying patterns indicative of the propagandistic intent.

Propaganda is also manipulating discussions, making sincere dialogues harder to identify next. To differentiate genuine dialogues from falsified ones, it is one of the biggest challenges. Contexts that exploit propaganda and all of a sudden evolve in the form of legitimate discourse usually struggle to differentiate between the authenticity and the deceptive message sent out. Intelligent algorithms have to understand nuances among persuasive rhetoric’s and authentic conversations by advanced semantic understanding and context recognition in order to correct the conclusion.

Dataset Overview:

For this particular task, the dataset is given with tab separated values (tsv) files, in each of which the annotated text samples are arranged and labeled according to the whether they contain propaganda or not. Such annotation plays a pivotal role in categorizing text into separate areas and that in turn facilitates examination of machine learning algorithms for propaganda detection.

Each sample in the dataset consists of two main components: "label" and pointing out the "tagged_in_context " .The column called "Label" is the key information when it comes to whether the propaganda is included in the specified text snippet or not. It explains the description for nine specific propaganda techniques, such as "flag-waving" and " loaded_language" to indicate the various ways the message is intended to deceive the public. Also, the dataset will have a label attributed to not propaganda, so as to denote instances where the content is completely free from any propaganda bias.

Major Significance of Propaganda Detection:

The exposure of propaganda results in various consequences, both in media, politics, and other regions. Today's information era is a breeding ground for misinformation and disinformation, therefore, the capability to automatically detect propaganda can be a vital defensive wall against being deceived and manipulated. Through using advanced algorithms to properly detect propagandistic elements, including texts’ contexts, these automated means become very effective as they protect information.

Moreover, successful propaganda appraisal enables people to critically analyse news increasing their authority over these manipulation instruments and making the public discussion healthier. Attentiveness to information and the powers used in propaganda cultivates critical thinking skills, which in effect increases the ability of people to differentiate facts from fiction and media literacy among the audience.

To the same extent, propaganda detection is not just about individual empowerment, but is also a crucial factor that contributes to the larger society as well. Supporting democratic institutions against deception and manipulation are key to preserving the rule of law, which includes the principles of transparency, accountability, and informed decision-making. Through these means, they uphold the balance of the information landscape that is considered to be crucial for the survival of democratic institutions.

Methodology

A methodology for solving tasks of propaganda detection and categorization of propaganda techniques was created with high accuracy with the help of a multi-phase approach. The preparations described are expected to be exhaustive and include the existing machine learning techniques together with modern deep learning technologies.

For both tasks, I commenced by loading the training and validation datasets from CSV files, setting the stage for subsequent preprocessing steps aimed at cleansing the textual data. This preprocessing pipeline encompassed a series of operations, including the removal of special tokens, conversion of text to lowercase, tokenization, elimination of stopwords and punctuation, and stemming to reduce words to their root forms. This meticulous preprocessing ensured that the text data was suitably prepared for subsequent analysis and model training.

Task 1: As part of propaganda detection task approach in question involves adopting a different range of classification approaches.

1. Traditional Machine Learning with CountVectorizer and TF-IDF Features:

- Binary Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) models are used. Hyperparameters such as alpha for MNB and the regularization parameter (C) for SVM were set based on initial experimentation and grid search.

- To turn textual information into numerical data the vectorizer containing both CountVectorizer and TF-IDFVectorizer was chosen, since the second one returns term frequency-inverse document frequency of all mentioned terms in documents instead of per-term frequency as CountVectorizer does.

- The models are then trained on the vectorized data and evaluated on the training and validation sets to measure the model's ability to distinguish prop samples from real news.

2. Deep Learning with BERT:

- The BERT (Bidirectional Encoder Representations from Transformers) model is adjusted to be competent in a sequence classification task. Key hyper parameters included the learning rate, batch size, number of epochs, and dropout rate. These parameters were chosen based on prior research and experimentation, with adjustments made during model training to optimize performance.

- Applying Hugging Face's transformers library, a pre-trained BERT model is tweaked in relation to the subject of work. Tokenization of the document data is fast and accurate enough to enable the use of the custom data loaders for training BERT model via validation sets.

- Training epochs that undergo iterative trainings are run and training and validation loss measures are monitored along with accuracy scores that assist in evaluating the success of the model in propaganda detection.

Task 2: Classification of Propaganda Techniques

In addressing the classification of propaganda techniques, a similar methodology is employed, encompassing diverse classification methodologies: In addressing the classification of propaganda techniques, a similar methodology is employed, encompassing diverse classification methodologies:

1. Traditional Machine Learning with CountVectorizer and TF-IDF Features:

- MNB classifier draws start from the multinomial data structure and uses Bayes' theorem to determine propaganda pattern in the textual corpus. Hyper parameters such as alpha chosen through experimentation and cross-validation. The impact of different alpha values on model performance was assessed during training and validation.

- The technique used is based on the count vectorizer and also TF-IDF vectorizers. Also, text data is converted into the numerical features suitable for classification.

- The classifier is trained on the vectorized features and its performance on both the train and validation datasets are evaluated as one of the indicating ways for proper identification of propaganda techniques

2. Deep Learning with BERT:

- After proving success in using BERT model to approach the matter of propaganda techniques classifications, again this model will be applied to solve this problem and hyper parameters such as learning rate, batch size, and number of epochs selected based on prior research and experimentation. The impact of varying these hyper parameters on model convergence and performance was monitored closely.

- A BERT for sequence classification task is fine-tuned with the model architecture excellently using the transformer-based design.

- Tokenizing the text input, setting up the data loaders and using the BERT model as a part of the classifier pipeline are done in the initial stages of the machine learning process.

- In reality, the model performance is repeatedly reached with the profiling, focusing on its ability to discover the concealed subtleness of propaganda in the corpus of texts.

The proposed methods mean that we try to find a middle ground between iteration and practicality. They intend to both use established methods and the latest techniques for the aforementioned tasks. The way the things are done though, is not pretentiously ornate but rather is sophisticatedly executed to guarantee the means of the productive ends in solving the issues of propaganda identification and classification.

Hyper-parameter settings

1.Traditional Machine Learning with TF-IDF Features:

Multinomial Naive Bayes (MNB) Classifier:

Alpha (Smoothing Parameter): Alpha is a hyperparameter that indicates the magnitude of smoothing that is applied to the probabilities in an uninformed Bayes classifier. It handles the case when features are absent in the training data which is the case with zero probabilities.

Exploration: After several tests and trial and error, I am able to tune the level of alpha to get maximum bias-variance trade-off using my current model. I will use a grid search strategy to ensure that the optimal alpha is chosen, this is to minimize overfitting but at the same time to provide enough smoothing to deal with rare or unseen feature.

Impact: Smaller values of alpha lead to fewer smoothing, which, in turn, may cause the model to overfit and assign too high a probability to rare events. But quite the contrary, higher alpha values tend to create the bias and at the same time contribute to overconfidence in making predictions. It may as well lead to an improved generalization that saves the model from being overconfident. Alpha selection is also very important because, as it exerts a direct impact on bias-variance trade-off in the model.

Vectorization Parameters (max_df and min_df):In the TF-IDF vectorization process, max_df and min_df parameters are two parameters that determine the inclusion of terms in the vocabulary based on their document frequency.

max_df: This acts as an upper bound on the documents which the terms will participate in, and when the number of documents is smaller than this threshold, the terms will be included in the vocabulary. Documents that have a document frequency over more than max_df will be excluded since they could provide too much repetitive information with limited added value.

min_df: On the contrary, min_df sets the minimum document frequency threshold below which terms will be filtered out of the vocabulary. The words that have less than min_df appear often not to much to be helpful indicators of class membership.

Exploration: Data filtering process has been adjusted according to the observed experiment and my previous knowledge. I intention to eliminate words that are too frequent or too rare, so that I can cut off the noise and increase the discriminative power of the features extracted from the text data.

Impact: Efficient choice of max_df and min_df is inevitable on the way to the desired feature representation. He min_df which is too high might keep the stopwords or such other words that are highly frequent, although, if min_df is too low it will tend to include noisy or irrelevant words. The adjustment of the parameters in this way helps the model to identify the significant patterns in the text data while eliminating the insignificant information.

2. Deep Learning with BERT:(BERT Model Hyperparameters)

Learning Rate -Exploration: In view of the fact that we have done a thorough research about the learning formulas we have reached an equilibrium between convergence speed and stability. These trials were meant to try out the values of the learning rate from very small to big ones to check the model training features.

Optimization: During the experimentation process, we were very attentive to the training and validation loss curves to find the learning rate that leads to a stable convergence to the best solution. The methods we adapted included schedules and automatically changing learning rate algorithms like Adam for dynamic learning rate adjustment during the training process.

Impact: Very high learning rates may lead to fluctuations in behavior, like divergence or jumping over you is the optimal solution, and very low, on the other hand, may result in slow convergence or even being stuck in local minima.

Batch Size - Exploration: Our investigation of batch sizes was carried out by trying different values, ranging from small mini-batches to large batch sizes, to discover their effect on the training dynamics and the model performance.

Considerations: We did a close examination of the trade-off between the accuracy and effectiveness of computations, memory usage and the ability to generalize. Larger batch sizes can speed up training by taking advantage of parallelism, the improved hardware utilization, but in this way we can obtain suboptimal results because of the noisy updates. On the other hand, the smaller batches provide more stable updates and possibly better generalization but it may increase the training time and memory requirements.

Observations: By means of empirical observations, the training dynamics, such as the convergence speed, training stability, and generalization showing capability, were monitored to isolate the batch size which found the best balance for the peculiar nature of task and computational resources.

Evaluation

This evaluation will include reviewing the classifiers' performance for both Task 1 (propaganda detection) and Task 2 (classification of techniques applied in propaganda) in terms of their accuracy, specificity, and sensitivity issues. This assessment comprises of a methodology of assessment, metrics of evaluation, evaluation strategy and presentation of results.

Method of Evaluation:

In both cases, the assessment compromised calculating the classifier's exactness on both the training and future performance datasets. In Task 1, three different classifiers were employed: MNB with CountVectorizer, SVM with TfidfVectorizer for the pre-trained SVM model, and a fine-tuned BERT model. Similarly, in Task 2, two classifiers were used: A multinomial naive bayes (MNB) classifier used CountVectorizer and TfidfVectorizer as coupled with a BERT model that has been fine-tuned.

Evaluation Strategy:

When it comes to traditional machine learning classifiers (MNB and SVM), the accuracy score function from Scikit-learn is the tool used directly to calculate the training and testing accuracies. Whereas in case of BERT model, training and validation losses, as well as accuracies, are being monitored at each epochs during training. Through this iterative evaluation approach the checking of the model convergence as well as the generalization performance is performed across the epochs.

Results Presentation:

The results are clearly presented in tabular format for easy interpretation. For Task 1, a table is provided summarizing the training and testing accuracies of the MNB and SVM classifiers:

Classifier	**Vectorizer**	**Training Accuracy**	**Testing Accuracy**
MultinomialNB	CountVectorizer	0.891	0.683
SVM	TfidfVectorizer	0.742	0.566

Additionally, precision, recall, and F1-score for each propaganda technique class obtained from the BERT model are provided in the following table:

Propaganda Technique	**Precision**	**Recall**	**F1-Score**	**Support**
Appeal to Fear/Prejudice	0.00	0.00	0.00	43
Causal Oversimplification	0.00	0.00	0.00	31
Doubt	0.00	0.00	0.00	38
Exaggeration/Minimisation	0.00	0.00	0.00	28

Propaganda Technique	Precision	Recall	F1-Score	Support
Flag Waving	0.00	0.00	0.00	39
Loaded Language	0.00	0.00	0.00	37
Name Calling/Labeling	0.00	0.00	0.00	31
Not Propaganda	0.52	1.00	0.68	301
Repetition	0.00	0.00	0.00	32

Task 1 Results:

The Multinomial Naive Bayes (MNB) classifier was markedly better (corresponding testing accuracy of 68). Accuracy is at 49% when compared to the actual classifier SVM (56%). Lower LDA scores (1. 88%) when compared different vectorization techniques (CountVectorizer and TfidfVectorizer, respectively). This indicates that the MNB classifier could have been a better option that was more effective at detecting the trends in the data for this task. Nonetheless, it is necessary to state that although the work of both traditional machine learning classifiers is subpar to what would be required for any real-world scenario; it highlights the complexity of the task or the requirement of more competent models.

The BERT classifier, a neural network classifier being more complicated and able to obtain contextual information, was observed to vary outcomes across different propaganda technique classes. In spite of the fact that some courses managed to reach a decent performance level with respect to precision, recall and F1-score, several others were incapable of being precisely classified. Such finding indicates that BERT could probably benefit from finetuning or additional data as it better recognizes words from classes that are rare in the dataset. Furthermore, model's accuracy might depend on vectorization technique (TfidfVectorizer), which might leads, to the conclusion that other preprocessing methods should be tuned up in order to get the best classifier.

Similarly, for Task 2, separate tables display the training and testing accuracies of the MNB classifiers with both CountVectorizer and TfidfVectorizer:

Classifier	Vectorizer	Training Accuracy	Testing Accuracy
MultinomialNB	CountVectorizer	0.866	0.398
MultinomialNB	TfidfVectorizer	0.912	0.409

Additionally, the training and validation losses and accuracies for each epoch of the BERT model are presented in the following table:

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	1.972	0.133	1.851	0.269
2	1.734	0.131	1.747	0.333
3	1.485	0.126	1.657	0.398
4	1.198	0.124	1.633	0.437
5	0.917	0.126	1.635	0.448
6	0.611	0.150	1.715	0.455
7	0.449	0.130	1.759	0.480
8	0.300	0.138	1.922	0.470
9	0.228	0.116	1.974	0.477
10	0.169	0.117	2.141	0.466

Task 2 Results:

For the task 2, the MNB classifiers have complied with those average testing accuracies which are 39. 8% and 40. 9%) on the other hand, the application of the CountVectorizer and TfidfVectorizer it has as a result (0. 9%). This implies that the classifiers had some difficulties to successfully classify propaganda techniques based on provided variables and consequently the intelligent usage of features with more valuable information or complex models could result in better performance.

Nevertheless, although the BERT classifier outperformed other models with regard to training loss, which decreased over the course of epochs, and validation accuracy, which increased, the latter improved steadily throughout the course of training. This implies that it was the machine that understood the algorithm and was therefore learning to better identify the propaganda techniques. Consequently, the achieved final precision (~45%) is also showing crash spots to be improved which can be done perhaps by fine-tuning or modification of network architecture.

However, the BERT classifier's result was also showing a peak where the number of epochs was reached to a certain point, showing that additional training beyond this could be ineffective. Having this fact in mind means that we need to be very attentive on model convergence and stop training when it is necessary to ensure adequate generalization.

Analysis

Error Analysis: Evaluating the errors committed by both the approaches recognition is important towards learning about the effectiveness of each method and what needs to be improved for better results.

Task 1: Conventionally ML Mode with TF-IDF Characteristics

Common Misclassifications: MNB being the Naive Bayes Multinomial which has been trained on the TF-IDF features, the classifier may have faced the problem that it was unable to distinguish between the propaganda techniques like "exaggeration" and "loaded language" since there was not a significant difference between them.

Error Modes: One of the root causes of such misclassifications is the fact that most machine learning algorithms that are traditional in nature show difficulty in natural language processing, especially with context. The method of the model focusing on representing the certain words through the bag model might be the reason for the heavy occupation by the model of the machine in noticing the tricky details on different propaganda manipulations.

Deep Learning with BERT

Common Misclassifications: The study had some limitations, despite the advanced capabilities of BERT the technique misclassification was found for example in the propagandican strategies like "" and "".

Error Modes: One cause for misclassification in the study was the intricacy of fitting grammatical patterns in the project. BERT, which is the strength in understanding contexts, could have trouble in determining the propaganda approach that is expressed in a way too subtle or drawn out, therefore it might wrongly classify.

Task 2: The Regular Traditional Machine Learning using TF-IDF Feature

Common Misclassifications: The TF-IDF-based MNB classifier did, as well as the previous one, have problems to determine whether there is a difference between languages like "doubt" and "loaded language".

Error Modes: The imperfections in the capabilities of classical machine learning algorithms that did not provide a way of sensing the semantic shades represented the main reason for misclassifications, where propaganda strategies overlapped in the content of texts. Though the model was based on TF-IDF representations, it could have failed to give an idea about of propagandistic variations in the topics.

Deep Learning with BERT

Common Misclassifications: The BERT model sometimes misjudged situations of "flag flipping" as well as "name-calling/labeling," which implies the challenges in precisely underscoring the context of different words.

Error Modes: Many consequential errors were related to the training model reliance on the token-level representations which may not be enough to capture the meaning of propaganda tactics in longer pieces of text. Although it is able to portray contextual information, BERT could still be affected by some propaganda techniques, which are quite semantic-based and thus require deep and detailed understanding. As a result, it could still misclassify things.

Limitations and Improvement Options:

The analysis reveals several limitations and opportunities for enhancing the accuracy and robustness of propaganda detection algorithms: The analysis reveals several limitations and opportunities for enhancing the accuracy and robustness of propaganda detection algorithms:

Semantic Understanding: Both method of classic machine learning and advancement of deep learning were not able to have semantic discrimination of propaganda technique catchphrase. As for the more advanced models with better context understanding, such as contextual embeddings or attention mechanisms, they can deal with these restrictions. Thanks to introducing up the sophisticated processing methods into models, they start understanding the key semantics of propaganda techniques. That is why the classification becomes more accurate.

Data Imbalance: Class imbalances were serious difficulties that arose particularly frequently in the course of model training, when training techniques ran the risk of overfitting specific rare propaganda methods. Handling data asymmetry with techniques such as oversampling and creating loss functions beyond the simple ones can result in better models. Discernment of all propaganda's classes can be facilitated by enabling a more balanced representation of the techniques in the training data. Therefore, models can learn to be more specific in the classification.

Model Interpretability: Although they are very efficient tools, deep learning models can also exhibit a lack of interpretability at times, thus making us unable to grasp the reasons behind some predictions. Introducing post-hoc interpretability techniques, such as attention visualization or saliency mapping, into the training can definitely improve the interpretability and credibility of the model. These techniques help the users to better understand the reasoning behind the model, which in turn makes better predictions, this way enabling to be more informed and confident to use such model in real-world applications.

Fine-tuning Strategies: There is no good fine-tuning BERT models without high-quality and sufficient dataset for training. The robustness of the models hand tuned could be increased either through augmenting of the training data or through the combination of semi supervised or transfer learning. Through large scale and diverse database exploitation, much of the performance of fine-tuned models can improve by using pre-trained models and representing propaganda principles in a manner that makes them and their approaches more generalizable and applicable in various contexts.

Conclusion

Summary of Findings: An assessment and analysis of the different classification categorizations showed a number of precious details about the monitoring of propaganda as well as techniques of propaganda. The way traditional machine learning methods, e. g. , Multinomial Naive Bayes classifier and Support Vector Machine, have shown a certain level of effectiveness in detecting propaganda; however, they have always posed a great challenge when it comes to capturing semantic subtleties. However, fine-tuned BERT models proved more accurate; these generated complications when it came to model interpretability and the application of the learned approach to uncommon propaganda methods.

Implications and Insights: For each way of thinking, strength and weakness are inseparable, which has real world consequences on how the technology is utilized. The very basic of the machine learning techniques are understood to be simple and interpretable but there is very low that they have on catching the complex language patterns effectively. Unlike fully connected networks, deep learning algorithms tend to be more accurate and able to perceive intricate meaning, but these systems are inconspicuous and their performance efficiency depends a lot on the amount of compute resources. Such a settled standpoint calls for a strategy that combines the sophisticated process and strong of each deep learning and traditional techniques enough to compensate for the voids of the two types.

Future Directions: The future propagating the detection work is that the coming research should be engaged in several specific directions, which as follows. Firstly, there is a requirement to examine hybrid methods that blend convincingly the benefits of conventional machine learning, neural networks and deep learning for higher productivity, and at the same time maintaining the model transparent. Besides that, a part of experiment work has to do with making instruments which can deal with some kind of data imitation and instances of methods that are stored rarely to a better extent. Furthermore, it is crucial to make modifications to the models so that they will fit well in different domains and languages. This should ensure that there is no barrier to their relevance across many areas. At last, the research should be applied to interpretability and transparency model because these elements are highly relevant when it comes to acceptance and trust in the end-use via real-world implementation of propaganda detection algorithms. Through such tackling of these challenges and by way of exploring new options, the following field of propaganda detection can consequently be sure to make great strides in the act of combating misinformation and educating people on digital literacy.

Further Work:

Research Directions: The identification and eradication of propaganda in the future can be improved by pursuing and examining the following topics in more detail:

Fine-tuning Model Architectures: The aim is to study how to improve the architecture of the deep learning models very much like BERT to sharpen ability to detect minute linguistic cues and enhance the models' output.

Exploring Novel Feature Representations: It is essential to try different types of feature representations such as graph-based or contextual embeddings that generalize convey rich information about underlying semantic intention and finally are useful in the discriminative power of classifiers.

Integrating Domain-Specific Knowledge: The model development process may include integration of domain-related information such as linguistic theory or psychological concepts into classification algorithm to obtain a better comprehension of propaganda language and enhanced accuracy of its categorization.

Improvement Opportunities: There are a number of areas which we could focus on when the techniques are being reviewed:

Optimizing Hyperparameters: Running various hyperparameter tuning experiments with a view of identifying "optimal settings" for the models that are based on "traditional machine learning" and "deep learning", where learning rates, batch sizes, and regularization techniques included.

Increasing Training Data Volume: Growing the amount and variation in the composition of the training set data so that the models are not over specialized and become more flexible and robust by enriching the annotations samples or by employing data augmentation methods.

Incorporating Ensemble Techniques: Researching ensemble learning techniques, like averaging models or stacking, to get the decision strategies from the different classifiers and enhance the overall classification performance as a result of the capability to utilize the diversified model aspects.

Application Extensions: Apart from the limits discussed in this paper, propaganda detection tools can be applicable in different fields,

For example:

Detecting Propaganda in Multimedia Content: Including machine learning models to deal with multimedia content, say images and videos, and making a binary choice about whether something is visual propaganda or misinformation, across different media formats.

Identifying Emerging Propaganda Tactics: Creating contemporary detection models which can recognize new and developing human behaviors in propaganda and real-time monitoring of online discussions, social media trends and news reports.