Propaganda Detection and Categorization Using Word2Vec with SVM, and BERT Sequence Classification

Anu Oluwatuyi

July 6, 2024

Abstract

This research focuses on the detection of propaganda within text, a crucial task for ensuring information integrity across various media platforms. Our study establishes a baseline using Support Vector Machines (SVM) combined with TF-IDF vectors, followed by more sophisticated models incorporating Word2Vec embeddings and BERT sequence classification. The SVM model with Word2Vec targets contextual word similarities while the BERT sequence classification model aims for comprehensive semantic analysis. We assessed the performance of each model based on precision, recall, and F1-score metrics. The findings demonstrate that the BERT sequence classification model surpasses the TF-IDF-SVM and Word2Vec-SVM combinations. Although the latter two models achieved comparable F1-scores of 0.68 and 0.69 respectively on the detection task, they were less effective in classifying propaganda techniques. BERT excelled in both tasks, achieving F1-scores of 0.94 in propaganda detection and 0.78 in propaganda categorization, offering enhanced detection capabilities for complex propaganda techniques. This suggests that BERT-based models are more effective in the nuanced identification of propaganda, promoting more reliable and precise text classification.

Keywords Propaganda Detection, Text Classification, SVM, Word2Vec, BERT

1 Introduction

In today's world, information spreads rapidly across various media platforms, but so does propaganda. Consider a relatively recent example of the social media campaign spreading false information about the newly developed COVID-19 vaccines at the time. Despite lacking scientific evidence, the campaign gained traction, leading to widespread fear and reluctance among the public to utilize the potentially life-saving vaccines (Skafle et al. 2022).

The impact can also be seen in religious and political spheres where propaganda is used to manipulate the view of the public. For instance, the current conflict in Gaza is often accompanied by propagandistic narratives aimed at shaping public opinion and garnering support for one side over the other (Amer 2017, Lopatin et al. 2017).

The ability to automatically detect and categorize different types of propaganda can aid in ensuring that the information reaching the public is credible. However, this is a complex task. Propaganda often uses subtle language tricks and emotional appeals that can be hard to spot without a deep understanding of how texts are written. The varied ways propaganda appears across different texts and settings add to this challenge, making it difficult for automated systems to consistently recognize and classify it.

This paper discusses the use of Support Vector Machines (SVM) combined with Word2Vec and BERT embedding techniques to detect the presence of propaganda in text and differentiate between various types of propaganda.

2 Methodology

This study employs a series of techniques to detect and categorize propaganda within textual data. Our methodology spans several stages, from data preprocessing and feature extraction to model training and evaluation across different tasks—firstly, detecting the presence of propaganda, and secondly, identifying specific types of propaganda.

The dataset comprises sentences manually labeled with eight distinct propaganda techniques (including **(i)** flag waving, **(ii)** appeal to fear/prejudice, **(iii)** causal simplification, **(iv)** doubt, **(v)** exaggeration/minimization, **(vi)** loaded language, **(vii)** name-calling/labeling, **(viii)** repetition) and a control group labeled as *non-propaganda*. The data is split into a training set for developing the models and a validation set for testing their effectiveness.

Embedding Techniques

Eembedding techniques convert text into numerical data that machine learning algorithms can interpret, as computers do not inherently understand human language

(Pennington et al. 2014). This paper explores various embedding methods including TF-IDF, Word2Vec, and BERT. TF-IDF transforms text into high-dimensional, sparse vectors, capturing term importance within documents. Word2Vec and BERT, on the other hand, generate dense, lower-dimensional vectors, encapsulating deeper semantic meanings and relationships. These embeddings significantly enhance the ability of Natural Language Proccessing (NLP) models to process and analyze textual data, forming the backbone of our approaches for detecting and categorizing propaganda.

2.1 TF-IDF with SVM Approach

TF-IDF (Term Frequency-Inverse Document Frequency) combined with Support Vector Machines (SVM) provides a robust baseline due to its simplicity and proven track record in text classification tasks. TF-IDF effectively highlights the most relevant terms in the texts, while SVM is capable of finding the optimal boundary between classes with its margin-based approach.

Data Preprocessing

Data Preprocessing involved several key steps to prepare the text for analysis. First, text normalization was conducted by converting all texts to lowercase to ensure uniformity across the dataset. Next, tokenization was performed where sentences were broken down into individual words to facilitate processing. Lastly, stopword removal was employed to eliminate common words that contribute minimal informational value to the context of text classification, streamlining the dataset for more effective analysis.

Feature Extraction using TF-IDF

TF-IDF vectorization was applied to convert text data into a numeric form understandable by the machine learning algorithms. The TF-IDF value is calculated as follows:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

where: $\text{TF}(t, d)$ is the term frequency of term t in document d; $\text{IDF}(t, D)$ is the inverse document frequency of term t across the document set D; and $\text{IDF}(t, D) = \log \left(\frac{N}{|\{d \in D : t \in d\}|} \right)$ N is the total number of documents.

TF-IDF Training with SVM

An SVM classifier with a linear kernel was chosen for its effectiveness in high-dimensional spaces, typical of text data. Hyperparameters like the regularization parameter C were tuned using grid search to find the optimal balance between bias and variance.

Handling Imbalanced Data

For the second task of classifying types of propaganda, the training data was notably imbalanced across different classes. To address this, Synthetic Minority Oversampling Technique (SMOTE) was employed to artificially augment the minority classes in the training set, improving the classifier's ability to learn from underrepresented classes (Chawla et al. 2002).

2.2 Word2Vec Approach

Word2Vec is a compelling approach for both tasks—propaganda detection and type classification—due to its ability to capture semantic meanings of words from large text corpora. Unlike TF-IDF, which merely accounts for the frequency of words, Word2Vec learns contextually enriched vector representations that uncover semantic relationships between words (Mikolov et al. 2013).

Word2Vec Training and Sentence Vectorization

Word2Vec was implemented using the skip-gram model, which predicts the context given a target word, enhancing its ability to handle rare words effectively. The parameters set for the model training were: *Vector size* of 300, to capture a wide range of semantic information; *Window size* of 15, allowing the model to consider a broader context around each word; and *Min count* of 1, including all words in the training.

Each sentence in the dataset was transformed into a fixed-length vector by averaging the Word2Vec embeddings of all words in the sentence:

$$v_{\text{sentence}} = \frac{1}{N} \sum_{i=1}^{N} v_{w_i}$$

where v_{w_i} represents the vector for word i in the sentence, and N is the total number of words.

Enhanced Weighting Within Propaganda Spans

To enhance model sensitivity to propaganda-specific language, words located within identified propaganda spans received increased weighting during the vector averaging process. If a word was within a propaganda span, its vector was repeated in the averaging process, effectively increasing its influence on the resulting sentence vector. The weighting factor was empirically set to 3, based on preliminary tests that optimized detection accuracy.

Model Training with SVM

The sentence vectors served as input for training an SVM classifier. An RBF kernel was selected for its capability to handle non-linear data distributions, which is typical in text data:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

where γ is a parameter that defines the influence of a single training example. Hyperparameter tuning was conducted using grid search to find the optimal values for C (the regularization parameter) and γ (Cortes & Vapnik 1995).

Handling Imbalanced Data

For the second task of classifying types of propaganda, similarly to TF-IDF approach SMOTE was employed to artificially augment the minority classes in the training set.

2.3 BERT Sequence Classification Implementation

BERT (Bidirectional Encoder Representations from Transformers) represents a significant advancement in the domain of natural language processing, particularly for tasks involving deep semantic understanding of text (Devlin et al. 2019). The model's architecture, which pre-trains on a large corpus using both left and right context in all layers, is especially suitable for nuanced tasks such as propaganda detection and type classification. The bidirectional nature of BERT makes it adept at understanding context, which is essential for accurately classifying sentences that may contain subtle propaganda techniques.

BERT Implementation

In this paper, BERT was employed using its sequence classification capabilities. We also utilized the *cased* version of the BERT model because case sensitivity is critical in understanding certain propaganda phrases which may include proper nouns or acronyms that are case-sensitive.

To enhance the model's ability to focus on specific segments within texts that might contain propaganda, special tokens (`<BOS>` for beginning of span and `<EOS>` for end of span) were introduced. These tokens were added to the BERT tokenizer as additional special tokens, allowing the model to recognize and give special attention to the text enclosed by these markers during training and prediction phases.

BERT Input Processing

The BERT model processes input text sequences converted into tokens $X = [x_1, x_2, ..., x_N]$, where N is the sequence length. Each token is embedded and then passed through the BERT layers, producing contextualized token embeddings $E = [e_1, e_2, ..., e_N]$. The embeddings for the special tokens associated with propaganda spans help to modify the contextual embeddings, focusing the model's attention on critical segments:

$$E' = \text{BERT}(X)$$

Where E' is the output embedding matrix from BERT.

Fine-tuning the BERT Model

For sequence classification, the embedding of the `[CLS]` token, typically the first token in a processed sequence, was used as the aggregate sequence representation for classification tasks. This embedding was fed into a simple classifier layer:

$$p = \text{softmax}(W \cdot e_{[\text{CLS}]} + b)$$

Where W (weights) and b (biases) are trainable parameters of the classifier, and p represents the probability distribution over the target classes.

Hyperparameter Tuning

Fine-tuning involved adjusting several hyperparameters, including the learning rate, batch size, and number of epochs, to optimize the performance on the propaganda detection tasks. Trough iterative trainings, the best settings were empirically determined to maximize the accuracy while avoiding overfitting. Using 5 epochs with batch size 8 gave slightly better accuracy but took a longer time to run than 3 epoch with the same batch size.

3 Results

The evaluation of the three approaches—TF-IDF with SVM, Word2Vec with SVM, and BERT sequence classification—was conducted for two distinct tasks: (1) detecting the presence of propaganda and (2) identifying specific types of propaganda. In tables 1 and 2 below, the performances of these models are detailed for each task. The effectiveness of the models on both tasks were evaluated using precision, recall, and F1-score as the primary metrics. Although accuracy was utilized during the hyperparameter tuning process to select the optimal parameters for the models, it was not emphasized in the final analysis due to its limitations when dealing with imbalanced datasets.

Evaluation Metrics Employed

- **Accuracy**: Defined as the ratio of correctly predicted observations to the total observations, accuracy measures the overall correctness of a model.

It is especially suitable for balanced datasets. The formula for accuracy is expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) represent the counts of each prediction outcome.

- **Precision:** Precision measures the accuracy of positive predictions, quantifying the number of true positives over the sum of true and false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision is vital in situations where the cost of a false positive is high, such as in the identification of specific types of propaganda where misclassification can lead to inappropriate responses.

- **Recall (Sensitivity):** Recall measures the ability of a model to find all the relevant cases within a dataset, calculated as the number of true positives divided by the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

High recall is crucial in ensuring that all potential instances of propaganda are detected, minimizing the risk of false negatives, which in the context of propaganda could mean failing to identify harmful content.

- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balance between the two by taking into account both false positives and false negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The macro F1 score is the arithmetic mean of the F1 scores of all classes:

$$\text{Macro F1} = \frac{\sum_{i=1}^{N} F1_i}{N}$$

Where $F1_i$ is the F1 score for class i and N is the number of classes.

Justification for Evaluation Methods: Accuracy was initially considered and used during the hyperparameter tuning phase to select optimum parameters, as it provides a quick and straightforward assessment of overall performance. However, in the final analysis, accuracy was not heavily emphasized. This decision is supported by the nature of the task: propaganda detection often deals with imbalanced classes (propaganda vs. non-propaganda), where accuracy might not fully capture the effectiveness of a model in dealing with minority classes. Models with high accuracy might still perform poorly on the less frequent but critical propaganda class.

In capturing the performance of the models accross all classes, the macro F1 score was employed instead of the weighted F1 score. The macro F1 score treats all classes equally, averaging the F1 scores independently calculated for each class, which is particularly useful in datasets with class imbalances or when minority classes are as significant as majority ones. In contrast, the weighted F1 score might mask poor performance on less frequent classes by emphasizing the majority classes. The macro averages where similarly employed for precision and recall.

Given these considerations, precision, recall, and F1-score were deemed more suitable for a nuanced evaluation of model performance, aligning with best practices in fields where class imbalance is prevalent (Chawla et al. 2002, Sokolova & Lapalme 2009)

Evaluation and Analysis

On the propaganda Detection task (Task 1), the results show that BERT sequence classification significantly outperformed the other methods, with accuracy and F1-scores reaching as high as 0.94. This indicates the superior capability of BERT in understanding and processing the contextual nuances in text data for effective classification.

Task 1: Propaganda Detection

This task involved determining whether a given sentence contains propaganda. The performance metrics for each model are illustrated in Table 1 below:

Table 1: Propaganda Detection Metrics

Model	**Precision**	**Recall**	**F1-score**
TF-IDF (baseline)	0.68	0.68	0.68
Word2Vec	0.70	0.69	0.69
BERT	0.94	0.94	0.94

Task 2: Identifying Specific Types of Propaganda

In the more challenging task of classifying specific types of propaganda, BERT also excelled, acheiving F1-scores of up to 0.78. In contrast, traditional machine learning models (TF-IDF with SVM and Word2Vec with SVM) showed reduced effectiveness accross all metrics, which suggests difficulty in differentiating between various propaganda types. The performance metrics for each model on this task are illustrated in Table 2 below:

Table 2: Propaganda Classification Metrics

Model	Precision	Recall	F1-score
TF-IDF (baseline)	0.31	0.28	0.29
Word2Vec	0.25	0.29	0.26
BERT	0.79	0.78	0.78

4 Discussion

The results of this study highlight the advanced performance of the BERT sequence classification model, in the detection and classification of propaganda. BERT's performance far exceeds that of traditional models like TF-IDF with SVM and Word2Vec with SVM in both tasks. This significant advantage demonstrates the utility of deep learning and transformer models in complex language processing tasks that require nuanced understanding of context and semantic relationships.

Advantages of BERT over Traditional Models

BERT's architecture, which integrates deep bidirectional processing, allows it to capture context from both sides of a token within a sequence. This capability is crucial for accurately interpreting the meaning of text in tasks such as propaganda detection, where context heavily influences interpretation. Traditional models, by contrast, typically focus on surface-level text statistics and are less adept at understanding deeper semantic nuances.

Limitations of TF-IDF and Word2Vec:

TF-IDF is adept at basic text classification but falls short when it comes to understanding context. It processes words in isolation, failing to account for the context that influences meaning, which is particularly limiting when words have different meanings based on their usage. Word2Vec, while embedding words in a semantic vector space and providing deeper insights, encounters issues with new words not included in its training set and with polysemous words, reflecting problems with fixed vocabulary and contextual ambiguity.

Challenges with BERT:

BERT, recognized for its ability to deeply analyze contextual information, demands extensive computational resources, limiting its deployment in resource-constrained environments. It is also prone to overfitting to the peculiarities of its training data, potentially amplifying existing biases. Additionally, BERT's complex model can misinterpret texts that contain nuanced or coded language, as its sensitivity may lead it to misread subtleties designed to mislead.

Common Error Patterns Across Models:

All methods consistently face difficulties in accurately classifying subtle propaganda techniques such as "Loaded Language" and "Name Calling." These challenges stem from the models' struggles with nuances in word usage, which are critical for distinguishing between different types of propaganda. Furthermore, sarcasm and irony pose significant challenges across all methods, as these linguistic styles rely heavily on contextual and implied meanings that are often beyond the literal interpretation capabilities of the models.

Implications for Future Research

The findings encourage several directions for future research:

1. **Model Enhancement** Optimizing BERT by adjusting its training procedures, exploring alternative pre-trained models such as RoBERTa, or incorporating advanced linguistic features could enhance its effectiveness.
2. **Hybrid Models** There is potential in developing hybrid models that combine the computational efficiency of traditional models with the semantic depth of BERT, possibly through ensemble techniques or integrated architectures.
3. **Broader Context Incorporation** Enriching models with broader contextual information, like text source and historical usage of propaganda techniques, could refine their predictive accuracy.
4. The study's reliance on labeled data might introduce annotation biases, and the performance of models, especially sophisticated ones like BERT, is dependent on the quality and volume of the training data. Expanding the dataset and utilizing unsupervised or semi-supervised learning methods, along with stricter annotation guidelines, could help mitigate these issues.

Overall, this research highlights the potential of sophisticated NLP methods for critical societal applications such as media content analysis, setting a foundation for more informed and effective strategies against misinformation and propaganda in various media outlets.

5 Conclusion

This study has demonstrated the great capabilities of advanced natural language processing techniques, particularly the BERT sequence classification model, in the tasks of detecting and classifying propaganda within text. Our investigation revealed that while traditional models like TF-IDF with SVM and Word2Vec with SVM serve well as baselines, they are significantly outperformed by BERT in terms of accuracy, precision, recall, and F1-score across both tasks.

The deep learning approach, embodied by BERT's architecture, provides a significant advantage in capturing the nuanced semantic relationships and contextual cues necessary for effective propaganda detection. This suggests that the future of text analysis, particularly in critical applications such as misinformation identification and media content analysis, will be driven by further advancements in transformer-based models.

However the depoyment of robust technologies such as BERT is not without its challenges. The study also highlighted several challenges, such as the high computational demands of BERT and the dependency on large, well-annotated datasets, which could limit its applicability in resource-constrained settings. Future research should focus on addressing these challenges by optimizing model training processes, exploring lighter model architectures, and enhancing dataset quality and size through advanced data augmentation techniques.

In conclusion, our results support a move towards adopting deeper, contextually aware computational models in NLP to better address the sophisticated techniques used in propaganda.

References

Amer, M. (2017), 'Critical discourse analysis of war reporting in the international press: The case of the Gaza war of 2008–2009', *Palgrave Commun* **3**(1), 13.
URL: *https://www.nature.com/articles/s41599-017-0015-2*

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'SMOTE: Synthetic Minority Over-sampling Technique', *jair* **16**, 321–357.
URL: *http://arxiv.org/abs/1106.1813*

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Mach Learn* **20**(3), 273–297.
URL: *https://doi.org/10.1007/BF00994018*

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *in* J. Burstein, C. Doran & T. Solorio, eds, 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
URL: *https://aclanthology.org/N19-1423*

Lopatin, E., Samuel-Azran, T. & Galily, Y. (2017), 'A clash-of-civilizations prism in German media? Documenting a shift from political to religious framing of the Israeli–Palestinian conflict', *Communication and the Public* **2**(1), 19–34.
URL: *https://doi.org/10.1177/2057047316689795*

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient Estimation of Word Representations in Vector Space'.
URL: *http://arxiv.org/abs/1301.3781*

Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global Vectors for Word Representation, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543.
URL: *http://aclweb.org/anthology/D14-1162*

Skafle, I., Nordahl-Hansen, A., Quintana, D. S., Wynn, R. & Gabarron, E. (2022), 'Misinformation About COVID-19 Vaccines on Social Media: Rapid Review', *Journal of Medical Internet Research* **24**(8), e37367.
URL: *https://www.jmir.org/2022/8/e37367*

Sokolova, M. & Lapalme, G. (2009), 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management* **45**(4), 427–437.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S030645730900*