UNIVERSITY
OF SUSSEX

School of Engineering and Informatics

Advanced Natural Language Processing

Assessed Coursework

Candidate No:

276236

Module Convenor:

Julie Weeds

Module:

968G5: Advanced Natural Language Processing

INDEX

Assessed Coursework

Introduction:

In today's digital age, the spread of propaganda has become a significant challenge. Propaganda is a form of communication that aims to influence the opinions, beliefs, and behaviors of individuals or groups towards a specific cause or ideology. It often involves the use of clear techniques, such as emotional appeals, logical misconceptions, and manipulative language, to shape public perception and promote a particular agenda.

The rise of social media and online platforms has amplified the reach and impact of propaganda. It has become easier for malicious actors to share misleading or false information with a wide audience, potentially leading to social unrest, political polarization, and decay of trust in institutions. Therefore, detecting propaganda is crucial for maintaining a healthy and informed public discourse.

Propaganda detection is a critical task that involves identifying instances of propaganda in various forms of communication, such as news articles, social media posts, and advertisements [1]. By accurately detecting propaganda, we can raise awareness about its presence, counter its effects, and promote media literacy among individuals.

However, detecting propaganda poses several challenges. Firstly, propaganda can take diverse forms and employ a wide range of techniques, making it difficult to establish clear criteria for identification [2]. Secondly, propaganda often blends with legitimate forms of influence, such as journalism and public relations, blurring the lines between informative and manipulative content. Thirdly, the subjective nature of propaganda perception can lead to disagreements and inconsistencies in labeling and annotation [1]. As a result, developing robust and adaptable detection methods is an ongoing challenge that requires continuous research and innovation.

Overview of the provided dataset:

To tackle the task of propaganda detection, a dataset was provided for this coursework. The dataset consists of two files: **"propaganda_train.tsv"** and **"propaganda_val.tsv"**, representing the training and validation sets, respectively. Each file is in tab-separated value (TSV) format and contains the following columns:

- **"index"**: A unique identifier for each example.
- **"tagged_in_context"**: The text of the example with propaganda techniques tagged using XML-like tags.
- **"label"**: The label indicating whether the example contains which type of propaganda or not.

The training set is used to develop and train the propaganda detection models, while the validation set is used to evaluate their performance and compare different approaches. The dataset provides a diverse collection of examples sourced from various news articles and online sources, covering a range of topics and propaganda techniques.

Goals and objectives of the coursework:

1. Develop and implement multiple approaches for binary classification of propaganda vs. non-propaganda text.
2. Develop and implement multiple approaches for multiclass classification of specific propaganda techniques.
3. Evaluate the performance of each approach using appropriate metrics and compare their effectiveness.
4. Analyze the strengths and weaknesses of each approach and identify common errors and challenges.
5. Provide insights and recommendations for future improvements in propaganda detection methods

Methods:

Preprocessing

The text data undergoes several preprocessing steps to prepare it for the classification tasks:

1. **Tokenization:** The text is split into individual tokens using word-level tokenization for Naive Bayes and the BERT tokenizer for BERT.
2. **Lowercasing:** All tokens are converted to lowercase to reduce vocabulary size and treat variations of the same word as equivalent.
3. **Punctuation Removal:** Punctuation marks are removed from the tokens to focus on the semantic content of the text. However, the decision to remove punctuation should be carefully considered based on the specific characteristics of the dataset and the propaganda techniques being studied.
4. **Encoding (BERT):** For the BERT approach, the preprocessed text is encoded into input IDs and attention masks. Input IDs represent the numerical representation of tokens, while attention masks indicate which tokens should be attended to by the model.

Task 1: Propaganda vs. Not Propaganda Classification

Approach 1: Naive Bayes with Bag of Words

The first approach for binary classification of propaganda vs. non-propaganda text uses the Naive Bayes algorithm with the Bag of Words (BoW) representation. The preprocessed tokens are converted into a BoW representation, where each document is represented as a vector of token frequencies. A Multinomial Naive Bayes classifier is then trained on the BoW features, learning the conditional probabilities of each token given the class label.

Hyperparameter settings:

- Smoothing parameter alpha: Tuned using grid search with values [0.1, 0.5, 1, 10]. The best value is selected based on the F1 score metric.

Justification: Naive Bayes is a simple yet effective algorithm for text classification tasks. Despite its assumption of feature independence, it often performs well on high-dimensional sparse data like text. The BoW representation captures token frequency information, which can be indicative of propaganda presence. The Multinomial variant is suitable for text data, as it models discrete feature counts.

Approach 2: BERT

The second approach utilizes the BERT model for binary classification. BERT is a pre-trained language model that learns contextual word representations. The preprocessed and encoded text data is used as input to the BERT model, and a sequence classification head is added on top for fine-tuning.

Hyperparameter settings:

- Maximum sequence length: 128
- Batch size: 32
- Learning rate: 2e-5
- Number of epochs: 5

Justification: BERT has demonstrated remarkable performance on various text classification tasks due to its ability to capture rich contextual information and learn meaningful representations. Fine-tuning BERT allows the model to learn task-specific patterns and details related to propaganda detection. The self-attention mechanism in BERT enables it to capture long-range dependencies and contextual information, which is crucial for identifying subtle propaganda techniques.

Task 2: Propaganda Technique Classification

Approach 1: Naive Bayes with Bag of Words

For the multiclass classification of specific propaganda techniques, the first approach follows a similar methodology as in Task 1. The preprocessed text data is converted into BoW features, and a Multinomial Naive Bayes classifier is trained to predict the propaganda technique for each example.

Hyperparameter settings:

- Smoothing parameter alpha: Tuned using grid search with values [0.1, 0.5, 1, 10]. The best value is selected based on the weighted F1 score metric, which takes into account class imbalance.

Justification: Naive Bayes with BoW serves as a simple baseline approach for multiclass propaganda technique classification. It assumes that the presence of certain words or phrases is indicative of specific propaganda techniques. However, the Naive Bayes assumption of feature independence may be limiting for capturing the complex dependencies and relationships between words that are characteristic of propaganda techniques.

Approach 2: BERT

The second approach for multiclass propaganda technique classification employs the BERT model, following a similar methodology as in Task 1. The snippet text data is tokenized and encoded using the BERT tokenizer, and a pre-trained BERT model with a sequence classification head is fine-tuned on the encoded data. The number of output classes in the classification head is set to the number of unique propaganda techniques in the dataset.

Hyperparameter settings:

- Maximum sequence length: 128
- Batch size: 32
- Learning rate: 2e-5
- Number of epochs: 5

Justification: By optimizing BERT for the classification of multiclass propaganda techniques, the model can be trained to recognize the hidden trends and traits particular to each approach. BERT's ability to capture contextual information and long-range dependencies is particularly valuable for distinguishing between different propaganda techniques, which often rely on subtle linguistic cues and rhetorical devices.

However, it is important to consider the limitations of BERT, such as its computational complexity and the potential for overfitting, especially when dealing with a limited amount of labeled data for specific propaganda techniques. Regularization techniques and careful hyperparameter tuning can help mitigate these issues.

Result and Analysis:

Evaluation metrics and methodology

To evaluate the performance of the different approaches for both Task 1 and Task 2, several evaluation metrics are employed:

1. **Accuracy**: Accuracy is a straightforward metric that tells you the percentage of correct predictions made by the model. It's calculated by dividing the number of correct predictions by the total number of predictions. However, accuracy alone may not give you the full picture, especially when you have uneven class distribution.
2. **Precision**: Precision focuses on the quality of positive predictions. It answers the question, "Out of all the instances the model predicted as positive, how many were actually positive?" A high precision means that when the model predicts something as positive, it's likely to be correct.

3. **Recall**: Recall, on the other hand, measures the model's ability to find all the positive instances. It addresses the question, "Out of all the actual positive instances, how many did the model correctly identify?" A high recall indicates that the model is good at detecting positive instances and minimizing false negatives.
4. **F1 score**: The F1 score combines precision and recall into a single metric. It's the harmonic mean of precision and recall, which means it gives equal weight to both metrics. The F1 score is particularly useful when you have imbalanced classes because it balances the importance of false positives and false negatives.

For Task 1 (binary classification), these metrics are computed directly based on the binary predictions and ground truth labels.

For Task 2 (multiclass classification), the metrics are computed using two averaging approaches:

- **Micro-averaging:** Micro-averaged metrics are calculated by treating each example equally, regardless of its class. The true positives, false positives, and false negatives are summed up across all classes, and then the metrics are computed using these aggregated values. Micro-averaging gives equal weight to each example and is useful when the class distribution is imbalanced.
- **Macro-averaging:** Macro-averaged metrics are calculated by computing the metrics independently for each class and then taking the unweighted mean across classes. Macro-averaging gives equal weight to each class, regardless of its size, and is useful when the performance on minority classes is of particular interest.

The evaluation methodology involves training each approach on the training set and evaluating its performance on the validation set. The best-performing model for each approach is selected based on the relevant metric (F1 score for Task 1, weighted F1 score for Task 2).

Results and Comparison

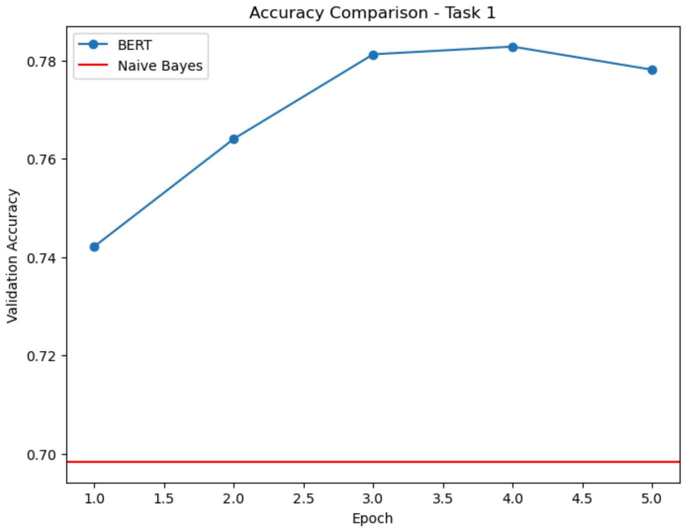

Fig 1: Accuracy Comparison - Task 1

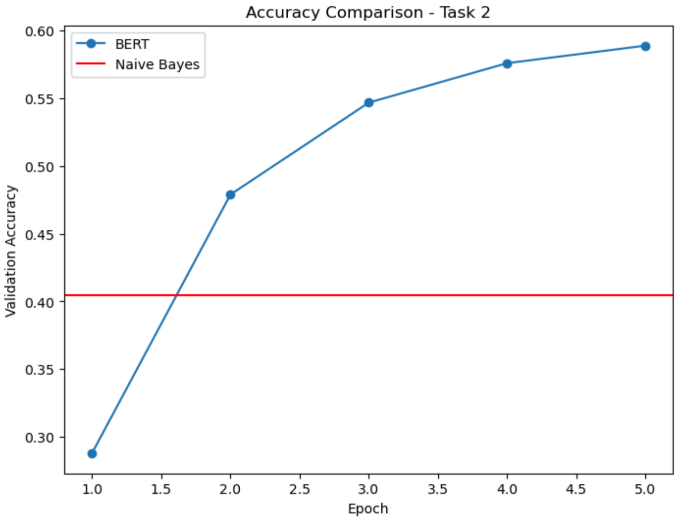

Fig 2: Accuracy Comparison - Task 2

Task 1: Propaganda vs. Not Propaganda Classification

Approach 1 (Naive Bayes) results:

- Validation Accuracy: 0.6984
- Validation F1 Score: 0.7293

Approach 2 (BERT) results:

- Best Validation Accuracy: 0.7828
- Best F1 Score: 0.7747

The results for Task 1 demonstrate a significant performance difference between the Naive Bayes approach and the BERT approach, with BERT consistently outperforming Naive Bayes in terms of both accuracy and F1 score. This suggests that BERT's ability to capture contextual information and learn meaningful representations of words is highly beneficial for propaganda detection. The fine-tuning process allows BERT to adapt its pre-trained knowledge specifically for the task of binary propaganda classification.

Task 2: Propaganda Technique Classification

Approach 1 (Naive Bayes) results:

- Validation Accuracy: 0.4045
- Validation Weighted F1 Score: 0.3971

Approach 2 (BERT) results:

- Best Validation Accuracy: 0.5890
- Best Weighted F1 Score: 0.5844

The results for Task 2 reveal a more challenging scenario for both the Naive Bayes approach and the BERT approach compared to Task 1. The multiclass classification of specific propaganda techniques proves to be a more complex task, as evidenced by the lower performance metrics. While BERT demonstrates superior performance compared to Naive Bayes, there is still significant room for improvement in this challenging task.

Discussion of results:

The superior performance of BERT in both tasks can be attributed to several factors. Firstly, BERT's self-attention mechanism enables it to capture long-range dependencies and contextual information, which is crucial for identifying subtle propaganda techniques that may span across multiple sentences or paragraphs. Secondly, BERT's deep architecture and large-scale pre-training on diverse text corpora allow it to learn rich and transferable language representations that can be effectively fine-tuned for specific tasks.

However, it is important to note that the BERT approach comes with higher computational requirements and longer training times compared to the Naive Bayes approach. The choice between the two approaches may depend on the available resources and the trade-off between performance and efficiency.

The lower performance of both approaches in Task 2 compared to Task 1 highlights the complexity of multiclass propaganda technique classification. The diverse linguistic patterns and nuances associated with each technique may not be adequately captured by the current approaches. To address these challenges, several strategies can be explored, such as increasing the size and diversity of the training dataset, incorporating domain-specific knowledge and linguistic features, and exploring more advanced architectures and techniques.

Overall, the evaluation results demonstrate the effectiveness of leveraging pre-trained language models like BERT for propaganda detection tasks.

Error Analysis:

Task 1 : Propaganda vs Not-propoganda Classification

Upon analyzing the misclassifications made by the Naive Bayes and BERT approaches in Task 1, several common errors were identified:

- **False positives:** Some non-propaganda texts were incorrectly classified as propaganda. These errors often occurred when the text contained clear or opinionated language that shared similarities with propaganda techniques. For example, texts expressing strong political views or emotionally charged statements were sometimes mislabeled as propaganda.
- **False negatives:** Propaganda texts were occasionally misclassified as non-propaganda. These errors typically involved subtle or less common propaganda techniques that were not adequately captured by the models.

Task 2: Propaganda Technique Classification

In Task 2, the multiclass classification of specific propaganda techniques, the following common errors were observed:

- **Misclassification of similar techniques:** Certain propaganda techniques that share similar linguistic patterns or strategies were often confused with each other. For example, "loaded_language" and "name calling,labeling" techniques, which both involve the use of emotionally charged or biased terms, were sometimes misclassified.
- **Confusion between rare techniques:** Propaganda techniques with limited training examples were more prone to misclassification. The models struggled to accurately identify and distinguish these techniques due to the scarcity of representative samples.
- **Misinterpretation of context:** In some cases, the models failed to fully grasp the context and details of the text, leading to misclassification of propaganda techniques. For instance, sarcasm or irony used in the text could be misinterpreted as a genuine propaganda technique.

Comparison of errors across approaches

Both approaches struggled with similar types of errors, such as distinguishing between closely related propaganda techniques or handling rare techniques with limited training data.

However, BERT generally made fewer errors compared to Naive Bayes. BERT's ability to capture contextual information and learn more sophisticated representations of text allowed it to better handle ambiguous or complex cases.

The errors made by BERT tended to be more minute and involved cases where the context was highly ambiguous or the propaganda technique was borderline. In contrast, Naive Baye's errors were more frequent and often stemmed from its simplistic assumptions based on word frequencies.

Conclusion:

In this coursework, we explored the task of propaganda detection using two approaches: Naive Bayes with bag-of-words representation and BERT fine-tuning. We tackled two sub-tasks: binary classification of propaganda vs. non-propaganda text and multiclass classification of specific propaganda techniques.

For Task 1, the binary classification task, both approaches demonstrated reasonable performance, with BERT outperforming Naive Bayes. BERT's ability to capture contextual information and learn meaningful representations of text proved beneficial for distinguishing propaganda from non-propaganda content.

In Task 2, the multiclass classification of propaganda techniques, the performance of both approaches was lower compared to Task 1. The increased complexity of identifying specific propaganda techniques posed challenges for both Naive Bayes and BERT. However, BERT still exhibited superior performance, showcasing its capability to handle more subtle and diverse linguistic patterns.

The error analysis revealed common challenges, such as distinguishing between similar propaganda techniques, handling rare techniques with limited training data, and capturing the ambiguity and subjectivity inherent in propaganda detection. The limitations of the training dataset and the complexity of propaganda language were identified as potential factors contributing to the errors.

Limitations and challenges:

- **Limited size and diversity of the training dataset:** The provided dataset, while valuable, may not fully represent the wide range of propaganda techniques and their variations.
- **Ambiguity and subjectivity in propaganda detection:** Propaganda detection often involves subjective judgments and interpretations. The boundary between propaganda and persuasive language can be blurry, leading to potential disagreements among annotators.

- **Complexity of propaganda language:** Propaganda techniques employ sophisticated linguistic devices and rhetorical strategies, making them challenging to detect automatically.
- **Evolving nature of propaganda:** Propaganda tactics and techniques continuously evolve, adapting to new platforms, technologies, and socio-political contexts. Keeping pace with these changes and developing models that can generalize to new forms of propaganda is an ongoing challenge.
- **Computational resources:** Training and fine-tuning large language models like BERT requires significant computational resources and time.

Future work and improvements:

- **Expanding the dataset:** Collecting and annotating a larger and more diverse dataset, covering a wider range of propaganda techniques and sources, can help improve the robustness and generalization of the models.
- **Incorporating linguistic and domain knowledge:** Integrating linguistic features, rhetorical devices, and domain-specific knowledge into the models could enhance their ability to capture the nuances of propaganda language.
- **Exploring advanced architectures and techniques:** Investigating and adapting state-of-the-art architectures and techniques from related fields, such as deep learning, transfer learning, or multi-task learning, could potentially improve the performance of propaganda detection models.
- **Developing explainable models:** Designing models that provide interpretable explanations for their predictions could aid in understanding the specific linguistic patterns and features that contribute to the classification of propaganda techniques. Explainable models can facilitate error analysis, model refinement, and user trust in the system.
- **Cross-domain and multilingual propaganda detection:** Extending the scope of propaganda detection to multiple languages and domains can broaden its applicability and impact. Developing models that can handle propaganda in different linguistic and cultural contexts is essential for addressing the global nature of propaganda.
- **Continuous monitoring and adaptation:** Establishing mechanisms for continuous monitoring and adaptation of propaganda detection models is crucial to keep pace with the evolving nature of propaganda tactics.

By addressing these limitations and pursuing the suggested future directions, the field of propaganda detection can continue to advance, developing more accurate, reliable, and explainable models. The ultimate goal is to empower individuals and society with tools and knowledge to critically evaluate information, resist manipulation, and promote a more informed and rational public discourse.

Bibliography:

[1] Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R. and Nakov, P., 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

[2] Martino, G.D.S., Yu, S., Barrón-Cedeño, A., Petrov, R. and Nakov, P., 2019. Fine-grained analysis of propaganda in news articles. *arXiv preprint arXiv:1910.02517*.