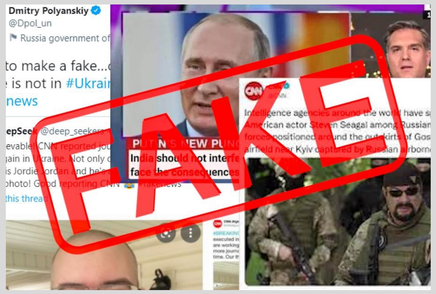
Advanced Techniques in Propaganda Detection:

Leveraging Language Models and AI for Social Media Analysis

IST.664.M003.FALL23.Natural Language Processing 16692.1241 Prof. Lu Xiao
Members: Shreya Zope, Janhavi Ghuge, Anvitha Shasidhar, Aatmaj Janardan

INTRODUCTION

- In the digital age, distinguishing truth from propaganda, a tool designed to shape opinions by presenting biased perspectives, is crucial. This challenge is magnified by the Internet's power to amplify messages.
- Our exploration focuses on using Natural Language Processing (NLP) to detect propaganda. We delve into techniques that discern factual information from manipulative content. This presentation highlights the blend of technology and critical thinking in identifying and understanding propaganda in modern media.

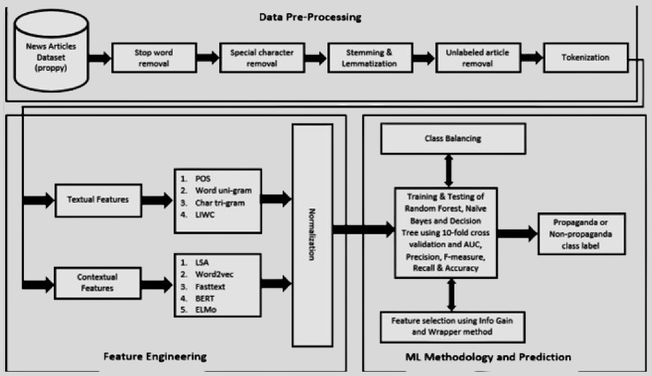


OBJECTIVE

- Our aim is to compare various natural language processing (NLP) algorithms for identifying propaganda in digital media. It assesses model methodology and performance measures spanning from traditional machine learning to advanced deep learning techniques such as GPT-4.
- The study emphasizes the merits and limits of each paradigm, emphasizing its efficacy in various settings. This study intends to provide insights into NLP's capabilities in tackling the issues of propaganda detection, hence helping in the creation of more accurate and unbiased information processing technologies in the digital era.

METHODOLOGY

The research papers use inventive methods to detect propaganda. One study, "Detecting Propaganda Techniques in Code-Switched Social Media Text," analyzes English and Roman Urdu code-switched text. It creates an annotated corpus and experiments with monolingual, multilingual, and cross-lingual models, revealing the challenges of code-switching.



MODEL COMPARISON

Summarized table which includes the model related details from the research papers and articles we studied

Model	Approach	Performance Metrics	Summary
BERT-based attention	Used BERT uncased for contextual understanding, proposed threshold-based classification	Precision: 60.1%, Recall: 66.5%, F1 Score: 63.2%	Reasonable performance, potential for a more sophisticated model
RoBERTa	Integrated with BiLSTM for sentence/span classification, achieved >97.60% recall, f1-score, and AUC	Recall, f1-score, AUC >97.60%	Exceptional performance, robust handling of complexity
BERT (various features)	Utilized various feature models and BERT, achieved recall, f1-score, and AUC ~90.10%	Recall, f1-score, AUC ~90.10%	Outperformed existing baselines, superior performance
Ensemble	Explored diverse algorithmic approaches, employed ensemble techniques	High-performance measures	Highlighted synergy of features and models
AraBERT	Leveraged for Arabic Language Understanding, primarily assessed by Micro-F1 score	Micro-F1 score	Tailored for Arabic Language Understanding tasks
XLM-RoBERTa	Achieved Micro-average F1-Score of 53%, excelled in Precision, Hamming score	Micro-average F1-Score, Precision, Hamming score	Superior performance in specific metrics
GPT-4 ('base' and 'chain')	Base: Precision: 52.86%, Recall: 64.52%, F1 Score: 58.11%. 'Chain': Precision: 56.86%, Recall: 57.82%, F1 Score: 57.34%	Moderate Precision and Recall	Reasonable but not superior performance compared to others

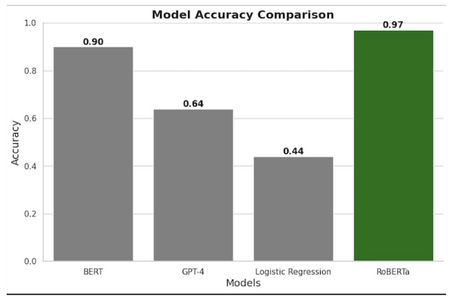
LIMITATIONS

Scarcity of Annotations & Models: Insufficient high-quality annotations and fine-tuned models in the domain
Language Scope: Limited to English articles, reducing applicability to other languages.
Computational Constraints: Possible limitations due to computational resources for larger-scale tasks.
Imbalanced Datasets: Fewer instances of propaganda articles versus non-propaganda, addressed using oversampling.

SOLUTION

Methodical Annotations: Consistent and reliable annotation approaches.
Balanced Datasets: Employing oversampling for imbalanced data.
Advanced Feature Selection: Using wrapper-based methods for feature refinement.
Rigorous Model Evaluation: Careful model selection and metric assessment.
Diverse Dataset Curation: Creating diverse datasets for better model generalization.

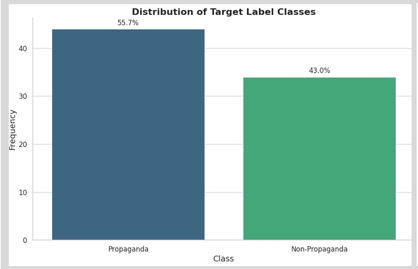
RESULTS



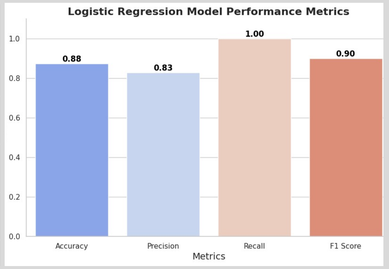
The following bar graph showcases the comparative accuracies of different models utilized in our analysis: BERT, GPT, Logistic Regression, and RoBERTa. Each model has undergone rigorous evaluation to determine its effectiveness in handling the given task. The bar graph provides a visual representation of their respective accuracies, aiding in the comparison and selection of the most suitable model for our specific task.

COMPUTATIONAL RESULTS

- **Data collection** - Data was collected from various sources like Twitter, news articles, etc.
- **Data annotation** - We manually annotated the data for each text
- **Data validation** - Crosschecked each other's annotated values of data.
- **Data pre-processing** - Data cleaning was done using packages in python.
- **Data exploration** - Explored the data and visualized them to get a better understanding.
- **Data splitting** - Data was split into training and test for evaluation.
- **Model selection and building** - Based on the research, we selected the Roberta model and we built it.
- **Model evaluation** - Evaluated the accuracy, precision, and F-1 score for the selected model.



Distribution of Target Variable: Propoganda and Non-propoganda



Logistic regression model showcased a accuracy of 90% while RoBERTa gave 100% accuracy.

CONCLUSION

RoBERTa Outperformed: The conclusion favors RoBERTa as it showed superior performance in detecting propaganda in code-switched texts. Its direct modeling of multilingualism led to better results.
Performance Measures: The primary evaluation measure, micro-average F1-score, highlighted RoBERTa's effectiveness, addressing the dataset's imbalance.
Future Scope: The study emphasizes expanding annotated data, particularly in low-resource languages, to enhance model capabilities. Future work involves improving models with a more balanced dataset and exploring subtler propaganda detection, especially in image-based and social media contexts.
Comparative Analysis: While GPT-4 showcased potential comparable to the state-of-the-art methods, RoBERTa-BiLSTM-CRF demonstrated the highest F1-scores, particularly in detecting propaganda fragments in bilingual texts. The ensemble approach and RoBERTa fine-tuning significantly enhanced detection accuracy.

ACKNOWLEDGMENT

Our heartfelt thanks to Professor Lu Xiao for their invaluable guidance throughout this project. Her deep knowledge in Natural Language Processing and insightful perspectives have not only enriched our understanding but have also been instrumental in shaping our approach towards detecting propaganda.
We're grateful to iSchool for their support.
Additionally, our appreciation extends to all contributors for their assistance and input.