CUET_sstm at ArAIEval Shared Task: Unimodal (Text) Propagandistic Technique Detection Using Transformer-Based Model

Momtazul Arefin Labib, Samia Rahman, Hasan Murad, Udoy Das
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u1904111, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd,
u1804109@student.cuet.ac.bd

Abstract

In recent days, propaganda has started to influence public opinion increasingly as social media usage continues to grow. Our research has been part of the first challenge, Unimodal (Text) Propagandistic Technique Detection of ArAIEval shared task at the ArabicNLP 2024 conference, co-located with ACL 2024, identifying specific Arabic text spans using twenty-three propaganda techniques. We have augmented underrepresented techniques in the provided dataset using synonym replacement and have evaluated various machine learning (RF, SVM, MNB), deep learning (BiLSTM), and transformer-based models (bert-base-arabic, Marefa-NER, AraBERT) with transfer learning. Our comparative study has shown that the transformer model "bert-base-arabic" has outperformed other models. Evaluating the test set, it has achieved the micro-F1 score of 0.2995 which is the highest. This result has secured our team "CUET_sstm" first place among all participants in task 1 of the ArAIEval.

1 Introduction

Propaganda is the use of manipulative ideas or news to influence the behavior or point of view of humans in order to serve a specific agenda. It has become widespread in media, newspaper articles, social media posts, and broadcasts. Individuals have rarely been informed without bias. Propaganda is a powerful political tool that attracts large groups of people. Social media has often been used to spread propaganda and misinformation, to divert attention from more serious issues. Detecting propaganda has been crucial to prevent false news from circulating.

Propaganda detection in Arabic text spans is quite challenging as the language has been rich in both morphology and syntax also it has a vast number of dialects (as mentioned in Elnagar et al., 2021). There has been a noticeable gap in the resources, annotated datasets, and NLP tools available for Arabic compared to languages such as English. All of this complicates NLP tasks in Arabic.

Primarily, this paper has intended to detect propaganda techniques in Arabic on social media (tweets) and in paragraphs. The Arabic-NLP 2024 conference, along with ACL 2024, has organized ArAIEval introducing a dataset that contains text with 23 types of propaganda techniques, to detect propaganda applied in text span (Hasanain et al., 2024b).

To achieve our goal, we have augmented underrepresented propaganda techniques in the dataset using synonym replacement from the nlpaug library and have evaluated a variety of models, including three machine learning techniques (RF, SVM, and MNB), one deep learning approach (BiLSTM), and three transformer-based models (bert-base-arabic, Marefa-NER and AraBERT). During our comparative analysis, each model has been trained and assessed on the provided dataset. The most effective model has been"bert-base-arabic", attaining the micro-F1 score of 0.2995 which is the highest.

The core contributions of our research work are augmenting underrepresented techniques in the dataset using synonym replacement with nlpaug and finetuning "bert-base-arabic" transformer model to detect 23 propaganda technique classes by specifying the span in the text. Detailed implementation information is available in the linked GitHub repository below- `https://github.com/Aref111n/CUET_SSTM-AraiEval`.

2 Related Work

Propaganda detection can be splitted into three categories based on previous research works: traditional ML techniques, advanced DL models, and transformer-based architectures.

ML-based techniques have been well-known for

Proceedings of The Second Arabic Natural Language Processing Conference, pages 507–511
August 16, 2024 ©2024 Association for Computational Linguistics

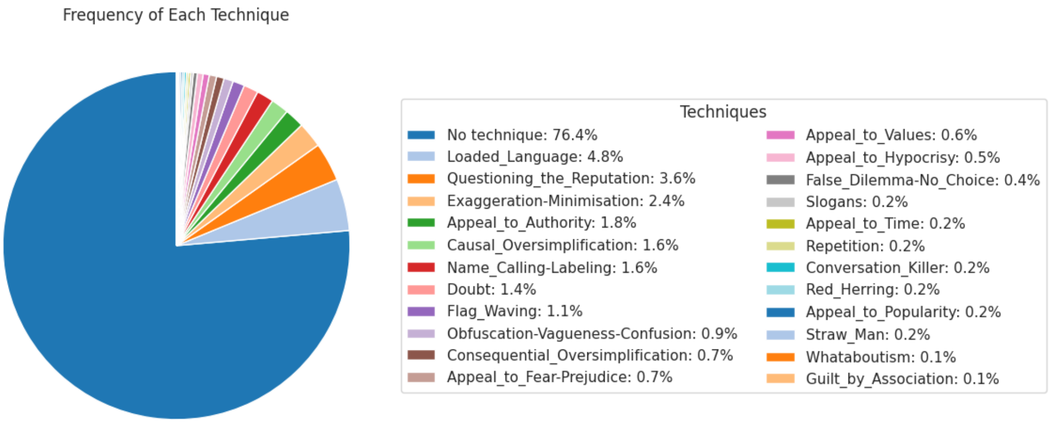

Figure 1: Frequency of propaganda techniques in the dataset.

detecting reliability in information. According to Rashkin et al., 2017, the truthfulness of information is based on common linguistic patterns. Naive Bayes Classifiers (Rashkin et al., 2017, Mihalcea and Strapparava, 2009), Support Vector Machine (Mihalcea and Strapparava, 2009) and Random Forest Classifier (Franklin et al., 2020) have been mostly used to understand those patterns.

DL-based algorithms have been deployed when data has become more available and affordable due to the growth of social networks and increased propaganda on social platforms. BiLSTM has been most commonly used in numerous works (Arsenos and Siolas, 2020).

Large language models like BERT, RoBERTa, and GPT have been used in detecting propaganda techniques (Hasanain et al., 2024b, Hasanain et al., 2023). span-level annotation has been proposed by Da San Martino et al., 2019, allowing specific portions of the text to coincide with any propaganda strategy. The top-performing models have employed BERT-based representations for a shared task[1] in 2019. XLM-RoBERTa-based model Marefa-NER (Hussein et al., 2022) and monolingual AraBERT (Attieh and Hassan, 2022) have proved their promising possibility in the field of propaganda detection. Several shared tasks (Hasanain et al., 2024b, Alam et al., 2022) have fueled the progress in the detection of reliable information in the Arabic Language.

3 Data

We have used the Arabic propaganda detection dataset (introduced in Hasanain et al., 2024a) from ArAIEval Shared Task 1 at ArabicNLP 2024 (Hasanain et al., 2024b), segmented into training (6997 samples), development (921 samples), and test sets (1046 samples). There are 995 tweets and 6002 paragraphs in the training set, 249 tweets and 672 paragraphs in the dev set, and 260 tweets and 786 paragraphs in the test set. The dataset includes 23 propaganda technique classes with notable distribution imbalances, as illustrated in Figure 1. About 74.1% of the text in the dataset contains no technique. To overcome this, we have augmented underrepresented techniques using the synonym replacement method of the nlpaug library (as mentioned in Coulombe, 2018). BeautifulSoup and Regex have also been used for additional preprocessing.

4 System

We have participated in only Task 1, which is an unimodal and multilabel sequence tagging task. The input is an Arabic multigenre text, particularly a news paragraph or tweet, and the required output is to detect propaganda techniques and their exact spans in the text.

4.1 Data Augmentation

To accurately identify all the propaganda techniques, that are imbalanced in the dataset, we have augmented the lowest-frequent 10 technique classes using the nlpaug library. This tool uses synonym replacement to create text variations, preserving their meanings but increasing training data. We have selected "wordnet" as the source and carefully adjusted the indices to maintain accurate labeling. As a result, our training dataset

[1] http://www.netcopia.net/nlp4if/2019/

<table>
<tr><th>Sentance</th><td colspan="5">رئيس كارثه اللهم عجل بزواله
(A disastrous leader. Oh God, hasten his demise)</td></tr>
<tr><th>Labels</th><td colspan="5">1. text: رئيس كارثه (disastrous leader), technique: Questioning the Reputation
2. text: كارثه (disastrous), technique: Loaded Language</td></tr>
<tr><th>Tokens splitted</th><td><b>1</b>
رئيس</td><td><b>2</b>
كارثه</td><td><b>3</b>
اللهم</td><td><b>4</b>
عجل</td><td><b>5</b>
بزواله</td></tr>
<tr><th>Label-1</th><td>B-Questioning the Reputation</td><td>I-Questioning the Reputation</td><td>O</td><td>O</td><td>O</td></tr>
<tr><th>Label-2</th><td>O</td><td>B-Loaded Language</td><td>O</td><td>O</td><td>O</td></tr>
<tr><th>Final tags</th><td>B-Questioning the Reputation</td><td>B-Loaded Language</td><td>O</td><td>O</td><td>O</td></tr>
</table>

Table 1: Example of a Sentance with it's labeling and the tagging process with BIO-encoding.

increased from 6997 entries to 9812.

4.2 Data Preprocessing

We have excluded emojis, special characters, HTML tags, URLs, numbers, and punctuations, retaining only Arabic letters and spaces for clarity and relevance. Arabic text normalization has removed Tashkeel (diacritical marks) and unified character variations: different forms of "alif", "taa" marbuta, and "non-standard kaf" to their normalized version. Multiple spaces have been replaced with a single space. Table 2 shows examples of how our preprocessing impacts Arabic texts.

As sequence tagging can be considered a variation of Named Entity Recognition (NER), we have adopted the "BIO" tagging technique from NER. Before applying BIO tagging, we have to tokenize texts into smaller segments to label with annotated spans. This led to misalignments between token spans and labeled propaganda technique spans. We have explored 3 tokenization techniques: tokenizing with marefa-ner, tokenizing with bert-base-arabic, and word-by-word splitting. The best-achieved consistency is by splitting word by word, though some misalignments remained. We have handled this by labeling all words that share the region of a propaganda span. Another problem with this BIO-tagging is spans can overlap, as shown in Table 1. There have been two propaganda technique labels for the given sentence. After applying word-by-word splitting on this sentence, 5 tokens have been found. Label-1 contains the first and the second split tokens. So these two tokens are tagged with the technique given in Label-1. On the other hand, Label-2 contains only the second split token. So the Second split token has been tagged again with the technique given by Label-2. This creates an overlapping issue in the given dataset.

4.3 Initial Experimentation

We have tried some ML and DL strategies to better understand the problem and to get the baseline. We have applied Random Forest, SVM, and Multinomial Naive Bayes Classifier, using DictVectorizer for feature extraction. Our deep learning architecture contains an embedding layer using Word2Vec, a BiLSTM layer, a Dense layer with ReLU activation, and a CRF layer.

4.4 Overview of the Adopted Model

We have implemented 3 transformer-based models. Marefa-NER is a large NER model targeted to extract 9 different entities. AraBERT, on the other hand, is a model trained on 60M Arabic tweets and introduced by Antoun et al., 2020. "bert-base-arabic" Safaya et al., 2020 is a BERT based Arabic pre-trained model which is available on HuggingFace. This uncased model has been pre-trained on a corpus consisting of 8.2 billion Arabic words covering both Modern Standard Arabic and various dialects.

After BIO-encoding the labels, we have 47 target classes, making it a multi-class token classification challenge. We have used the models experimenting with various hyperparameters to optimize training. A pre-trained tokenizer has been used and tokenized samples with a maximum length of 128 and ensured truncation. For training, we have used the Trainer API.

5 Results and Analysis

In this section, we present performance comparisons among the various ML, DL, and transformer-

Before Preprocessing	Preprocessing Actions	After Preprocessing
تحديثات ٢٤/٧ على موقعنا	Remove numbers and punctuation	تحديثات على موقعنا
جاء من إيران	Standardize “alif” forms	جاء من ايران
هذه مهمة خاصة	Replace “taa” marbuta	هذه مهمه خاصه

Table 2: Examples of Preprocessing Actions on Arabic Text.

based approaches we have examined during our study.

5.1 Parameter Setting

In our best-performing transformer model, “bert-base-arabic”, the parameters have been set as follows: learning rate as 0.000001 and weight decay as 0.001, 15 training epochs with conditions to save the best model with the lowest evaluation loss, training and evaluation batch size both set to 8 and the optimizer to Adam. We have also evaluated the performance of AraBERT and Marefa-NER with varying parameters. The best performance with AraBERT was achieved with a learning rate of 0.0001 and a weight decay of 0.01, training and evaluation batch size was same as “bert-base-arabic” and have ran for 10 epochs. In Marefa-NER, learning rate and weight decay was same as AraBERT, batch size for both the train and evaluation have been set to 4 and have ran for 4 epoch.

5.2 Evaluation Metrics

The metric used to evaluate this task is a modified micro-average F1 score that considers the matching between the gold-standard labels and the predictions. Additionally, we have focused on measuring other metrics such as precision scores (P) and recall (R) scores.

5.3 Comparative Analysis

We have found that among all our strategies, the “bert-base-arabic” has achieved the highest micro-F1 of 0.2994 whereas “Marefa-NER” achieved the highest F1 score of 0.28. Table 3 presents the class-wise F1 scores. “bert-base-arabic” performed better than all models securing 1^{st} rank in the leaderboard.

Analysing the F1 scores of the transformed-based models shown in Table 3, It has been found that the F1 scores are very close. So a question has been arised whether the differences are actually significant. So we have splitted our test dataset into 4 split ensuring random shuffling and named them as “prop_test-1”, “prop_test-2”, “prop_test-3” and “prop_test-4”. We have ran a statistical difference test with the 4 split of test dataset with the transformer-based models. The result obtained has been shown in Table 4. It shows that, “bert-base-ner” performing better then the other models for almost all the splits of test data. Hence, we can assure the significance of the minor differences between our best performing model with the other models.

Classifier		Micro Average		
		P	**R**	**F1**
ML	MNB (baseline)	0.153	0.088	0.112
	RF	0.161	0.133	0.146
	SVM	0.214	0.128	0.16
DL	BiLSTM	0.266	0.175	0.211
TF	bert-base-ner	**0.314**	**0.286**	**0.299**
	Marefa-NER	0.312	0.256	0.281
	AraBERT	0.301	0.26	0.279

Table 3: Performance of different systems on test dataset

5.4 Discussion

As mentioned before, the annotation contains Propaganda technique spans that overlap each other. However, our model does not consider overlapping and provides output as nonoverlapping spans. This is a major reason of error in most of the cases. Also despite we have augmented the under-represented techniques, the dataset remains quite imbalanced. While augmenting, some dominant techniques such as “Loaded Language” have also increased. Our future works will be to overcome the overlapping span issue with a proper encoding strategy and to create a more balancing dataset.

6 Conclusion

Though Arabic language processing is a very challenging task, we have been able to identify 23 dis-

Model Name	prop_test-1	prop_test-2	prop_test-3	prop_test-4
Marefa-NER	0.264	0.295	0.258	0.272
AraBERT	0.288	0.306	0.237	0.291
bert-base-ner	0.301	0.296	0.277	0.303

Table 4: Statistical difference test with the transformer-based models

tinct propaganda techniques with their spans. We have contributed to augmenting underrepresented techniques and testing with a variety of ML, DL and transformer-based models. Among these, "bert-based-ner" outperformed others achieving a micro-F1 score of 0.2995. However, our model has struggled with overlapping spans and has been biased toward the majority class.

Ethics Statement

During the analysis, preprocess, and implementation of our systems, we have been committed to keeping the highest ethical standard. Our goal is to share and contribute positively toward the development of a propaganda technique detection system in the Arabic language.

References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the The Seventh WANLP*, Abu Dhabi, United Arab Emirates (Hybrid). ACL.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Anastasios Arsenos and Georgios Siolas. 2020. Ntuaails at semeval-2020 task 11: Propaganda detection and classification with bilstms and elmo. In *Proceedings of the Fourteenth Workshop on SemEval*, pages 1495–1501.

Joseph Attieh and Fadi Hassan. 2022. Pythoneers at wanlp 2022 shared task: Monolingual arabert for arabic propaganda detection and span extraction. In *Proceedings of the The WANLP*.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of EMNLP-IJCNLP 2019*. ACL.

Ashraf Elnagar, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. 2021. Systematic literature review of dialectal arabic: identification and detection. *IEEE Access*.

Paul Franklin, Donald Cooper, Jan Danel, and Tiger Hu. 2020. Russian facebook propaganda detection with classification models.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2023. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024a. Can gpt-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation*, LREC-COLING 2024, Torino, Italy.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024b. Araieval shared task: Propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok. Association for Computational Linguistics.

Ahmed Samir Hussein, Abu Bakr Soliman Mohammad, Mohamed Ibrahim, Laila Hesham Afify, and Samhaa R El-Beltagy. 2022. Ngu cnlp atwanlp 2022 shared task: Propaganda detection in arabic. In *Proceedings of the The Seventh WANLP*.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the EMNLP 2017*.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on SemEval*.