Arabic Fake News Detection (AFND)

Tariq Odeh - 1190699, Majd Abubaha - 1190069, Amany Khdair - 1190728
Department of Electrical & Computer Engineering, Birzeit University
Ramallah, Palestine
Supervised by: Prof. Adnan H. Yahya

***Abstract*—The swift dissemination of misinformation on digital platforms undermines information integrity, particularly within the Arabic-speaking community. This project aims to tackle this challenge by creating an Arabic Fake News Detection (AFND) system employing sophisticated natural language processing and machine learning methodologies. Specifically, we apply advanced neural networks known as Long Short-Term Memory (LSTM) models to detect and categorize fake news in Arabic-language content. Our research leverages the Arabic Fake News Dataset (AFND) and the Gaza Dataset, and it demonstrates that our LSTM approach surpasses conventional machine learning methods such as Naive Bayes and Support Vector Machines (SVM). The findings underscore the superior performance of deep learning in the context of Arabic fake news detection, offering significant contributions to the fight against digital misinformation.**

***Index Terms*—Fake News Detection, Arabic Language, Natural Language Processing, Deep Learning, LSTM, Machine Learning.**

I. Introduction

THE rapid proliferation of misinformation on social media and other digital platforms significantly threatens the integrity of information and public trust. This issue is particularly challenging in the Arabic-speaking world due to the language's numerous dialects and complex grammatical structures. Despite considerable research on fake news detection across various languages, the Arabic language is still underrepresented in this field.

To address this gap, our project aims to develop an advanced Arabic Fake News Detection (AFND) system. By employing state-of-the-art natural language processing (NLP) and machine learning methods, we seek to accurately identify and categorize fake news in Arabic-language content. Our methodology utilizes sophisticated neural network models, specifically Long Short-Term Memory (LSTM) networks.

The rising consumption of digital news in Arabic-speaking regions, coupled with the surge in misinformation, underscores the significance of this study. Ensuring the accuracy and reliability of information is crucial for maintaining public trust and promoting informed decision-making. Our AFND system addresses the unique challenges posed by Arabic fake news by integrating modern machine learning techniques with language-specific tools.

This paper is structured as follows: Section II reviews the background and related work in fake news detection, particularly focusing on Arabic text. Section III outlines the methodology used in our study, detailing the model architecture and training processes. In Section IV, we describe the datasets used, emphasizing their relevance and characteristics. Section V explains the equations and metrics employed to evaluate the performance of our AFND system. Section VI presents the experiments conducted and the results obtained, demonstrating the effectiveness of our approach. Finally, Section VII concludes the paper and offers suggestions for future research.

Through this research, we aim to provide valuable insights and practical solutions to the ongoing efforts against misinformation, particularly within the Arabic-speaking community, thereby enhancing the reliability of news consumption in the digital media age.

II. Background and Related Work

Fighting the quick spread of false information on social media platforms has become even more crucial in recent years. Scholars have investigated many approaches to address this problem, specifically emphasizing the identification of false information in Arabic-language tweets. One method combines pre-trained word embeddings like ARBERT and MARBERT with sophisticated deep learning models like Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. This all-encompassing approach looks at both textual content and user profile information in an effort to increase the detection accuracy of false news. For instance, the MARBERT-CNN model's trial findings showed a high accuracy and F1-score of 0.956, indicating its efficacy on Twitter for a variety of themes. This emphasizes how crucial it is to combine user behavior analysis with text content to identify misinformation effectively [1].

A hybrid classification framework, which combines natural language processing methods with a feature selection system based on the Harris Hawks Optimization (HHO) algorithm, is another novel strategy. Researchers examined a variety of machine learning methods in conjunction with feature extraction models, including TF-IDF, BTF, and TF-frequency. They improved the performance of the classifiers by successfully removing redundant and unnecessary features using a binary variation of the HHO technique. In comparison to previous methods, this new methodology produced better results, indicating its potential for early detection of bogus news in Arabic language [2].

Additionally, recent studies have emphasized ongoing efforts to curb the spread of misinformation on social media platforms. Notable research by Monther Aldwairi and Ali Alwahedi presents a solution for detecting and filtering fake news using a logistic regression classifier, achieving an impressive accuracy rate of 99%. By integrating simple yet strategically selected features from titles and posts, the study showcases the effectiveness of using advanced machine learning techniques to combat misleading content. These efforts highlight the critical need to counter the dissemination of false information on social media, reflecting a dedicated effort to protect users from deceptive narratives [3].

III. Methodology

This study adopted a systematic approach, starting with the collection of two distinct datasets necessary for training and evaluating the Arabic Fake News Detection model. The primary dataset, the Arabic Fake News Dataset (AFND), was gathered from various online news platforms, with each article labeled based on predefined credibility categories: credible, not credible, or undecided. Additionally, the Gaza Dataset was compiled to specifically capture news related to Gaza, similarly categorized into credibility classes. This dual dataset strategy aimed to cover a broad range of news sources and topics, essential for training a robust classification model.

Following data collection, thorough preprocessing was performed to ensure the text data's quality and consistency across both datasets. Initial preprocessing steps involved the removal of Arabic stopwords using the NLTK library [4], which helped eliminate commonly occurring words that do not significantly contribute to the classification task. Subsequently, the datasets were merged into a single unified dataset to facilitate comprehensive model training. This combined dataset strategy aimed to enhance the model's ability to generalize across various sources and types of news articles effectively.

Because a Long Short-Term Memory (LSTM) neural network design can capture the sequential relationships included in text data, we chose it to construct the Arabic Fake News Detection model. Two LSTM layers were used for sequential feature extraction in the model architecture after an embedding layer to represent words as dense vectors. After adding dropout layers to reduce overfitting, the text was categorized into the predetermined credibility categories by a final thick layer using softmax activation.

The dataset was divided into training and validation sets for the model training phase. The sparse categorical cross-entropy loss function, which was tuned for multi-class classification problems, was used to train the model using the Adam optimizer. In order to guard against overfitting and guarantee the best possible model performance on unobserved data, early halting based on validation loss was included. The training approach was designed to optimize accuracy and decrease loss, two important variables that determine how well the model performs in classifying news stories into credible, non-credible, and uncertain categories.

To assess the performance of the developed model, comprehensive evaluation metrics were employed. These included accuracy as the primary metric to measure overall classification performance, along with detailed insights from confusion matrices to understand the distribution of predicted labels compared to ground truth labels. Additionally, precision, recall, and F1-score metrics were computed for each credibility class (credible, not credible, undecided) to provide a nuanced assessment of the model's performance across different categories.

Besides the LSTM model, baseline models such as Multinomial Naive Bayes and Linear Support Vector Machine (SVM) were also trained and evaluated. These baseline models served as benchmarks for comparison against the LSTM model's performance, offering context and insights into the effectiveness of deep learning methods compared to traditional machine learning approaches for Arabic fake news detection.

IV. Evaluation Metrics

In this section, equations essential for evaluating Arabic Fake News Detection (AFND) systems are presented. Performance metrics like recall, precision, F1 score, and accuracy are quantified by these equations. A structured framework for assessing AFND systems' effectiveness in identifying false news accurately is provided. Each equation addresses a specific aspect, such as recall (detection of manipulated news), precision (accuracy of classifications), and overall accuracy. These equations are crucial for evaluating AFND algorithms, guiding research, and fostering robust solutions against fake news proliferation.

A. Equations

1) $\text{Recall} \;=\frac{\text{TP}}{\text{TP+FN}}$
2) $\text{Precision} \;=\frac{\text{TP}}{\text{TP+FP}}$
3) $\text{F1 Score} \;=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$
4) $\text{Accuracy} \;=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$

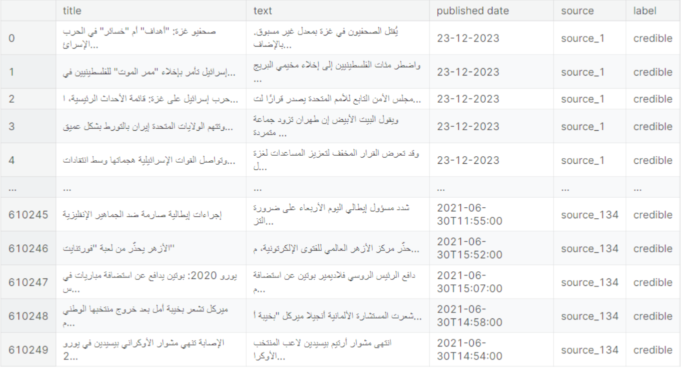
B. Variables

- TP: True Positives
- FN: False Negatives
- FP: False Positives
- TN: True Negatives
- N: Total data points
- n_k: Number of true labels for class k

Note: Models with higher F1 scores are considered superior, emphasizing the importance of these equations in AFND systems' evaluation.

V. Dataset

In our research, the dataset we used are combine the two distinct datasets to conduct comprehensive analyses. The first dataset, the The Arabic Fake News Dataset (AFND) combines data from 134 Arabic news websites, totaling 606,912 articles. It's categorized into credible, not credible, and undecided using Misbar, a fact-checking platform. Each news source is anonymized in a ”sources.json” file. Articles are stored in subdirectories, each representing a news source, with detailed info in ”scraped_articles.json.” This dataset is valuable for studying fake news in Arabic media [5].



	title	text	published date	source	label
0	صحفيو غزة: "أهداف" أم "خسائر" في الحرب الإسرائ...	يُقتل الصحفيون في غزة بمعدل غير مسبوق. بالإضاف...	23-12-2023	source_1	credible
1	إسرائيل تأمر بإخلاء "ممر الموت" للفلسطينيين في...	واضطر مئات الفلسطينيين إلى إخلاء مخيمي البريج ...	23-12-2023	source_1	credible
2	حرب إسرائيل على غزة: قائمة الأحداث الرئيسية، ا...	مجلس الأمن التابع للأمم المتحدة يصدر قرارًا لت...	23-12-2023	source_1	credible
3	وتتهم الولايات المتحدة إيران بالتورط بشكل عميق...	ويقول البيت الأبيض إن طهران تزود جماعة متمردة ...	23-12-2023	source_1	credible
4	وتواصل القوات الإسرائيلية هجماتها وسط انتقادات...	وقد تعرض القرار المخفف لتعزيز المساعدات لغزة ل...	23-12-2023	source_1	credible
...	...	...	...	...	...
610245	إجراءات إيطالية صارمة ضد الجماهير الإنقليزية	شدد مسؤول إيطالي اليوم الأربعاء على ضرورة التز...	2021-06-30T11:55:00	source_134	credible
610246	"الأزهر يحذّر من لعبة "فورتنايت"	حذّر مركز الأزهر العالمي للفتوى الإلكترونية، م...	2021-06-30T15:52:00	source_134	credible
610247	يورو 2020: بوتين يدافع عن استضافة مباريات في س...	دافع الرئيس الروسي فلاديمير بوتين عن استضافة م...	2021-06-30T15:07:00	source_134	credible
610248	ميركل تشعر بخيبة أمل بعد خروج منتخبها الوطني م...	شعرت المستشارة الألمانية أنجيلا ميركل "بخيبة أ...	2021-06-30T14:58:00	source_134	credible
610249	الإصابة تنهي مشوار الأوكراني بيسيدين في يورو 2...	انتهى مشوار أرتيم بيسيدين لاعب المنتخب الأوكرا...	2021-06-30T14:54:00	source_134	credible

Fig. 1. Sub of data from dataset.

The second dataset used in our study was the Israeli-Palestinian Conflict News Dataset, which was obtained from Al Jazeera and covers the period from 2021 to 2023, and provides insights into media coverage of the conflict. It is useful for understanding evolving narratives, true and false news, and offers applications for keyword extraction and text summarization. This dataset facilitates a deeper exploration of the news cycle surrounding the conflict [6].

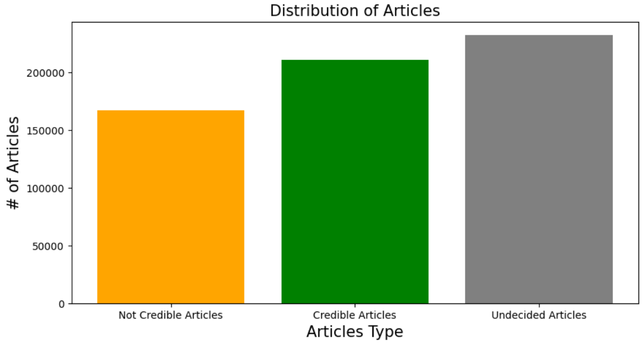


Fig. 2. Articles types for two datasets.

VI. Experiments and Results

A. Datasets

In this research, we employed two separate datasets: the Arabic Fake News Dataset (AFND) and the Gaza Dataset. The AFND dataset, sourced from Kaggle, comprises articles categorized as credible, not credible, and undecided, totaling 606,912 articles. Similarly, the Gaza Dataset, also obtained from Kaggle, includes 3,338 articles with the same categorization. Both datasets feature fields such as title, text, publication date, source, and label.

The datasets were imported into pandas DataFrames and pre-processed to align articles with their corresponding source labels. The distribution of article labels within each dataset was visualized using bar plots to illustrate the existing class imbalance.

B. Data Preprocessing

For text preprocessing, we removed Arabic stopwords using the NLTK library. The target labels were encoded using sklearn’s `LabelEncoder`. We then split the combined dataset into training and testing sets with an 80-20 ratio. Tokenization was performed using Keras’ `Tokenizer`, converting texts to sequences and padding them to ensure uniform input length for the models.

C. LSTM Model

For the classification challenge, we used a Long Short-Term Memory (LSTM) neural network. The design of the model consisted of an embedding layer, two LSTM layers, and a dense output layer that used softmax activation at the end. The vocabulary size was used as the input dimension and 100 was chosen as the output dimension for the embedding layer. 64 units were present in each LSTM layer.

The model was compiled using the sparse categorical cross-entropy loss function and the Adam optimizer. The training process spanned 10 epochs with a batch size of 64, and early stopping was implemented to mitigate overfitting. The model achieved a training accuracy of 73.68% and a validation accuracy of 69.99%.

The LSTM model achieved 69.71% accuracy after testing. A comprehensive classification report with a weighted average F1-score of 0.70 is shown in Table 1, along with precision, recall, and F1-scores for each category. Furthermore, a confusion matrix was displayed to examine the model’s functionality in more detail.

TABLE I
LSTM Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.74	0.73	0.73	41874
1	0.63	0.60	0.61	33534
2	0.71	0.74	0.73	46642
Accuracy			0.70	122050
Macro Avg	0.69	0.69	0.69	122050
Weighted Avg	0.70	0.70	0.70	122050

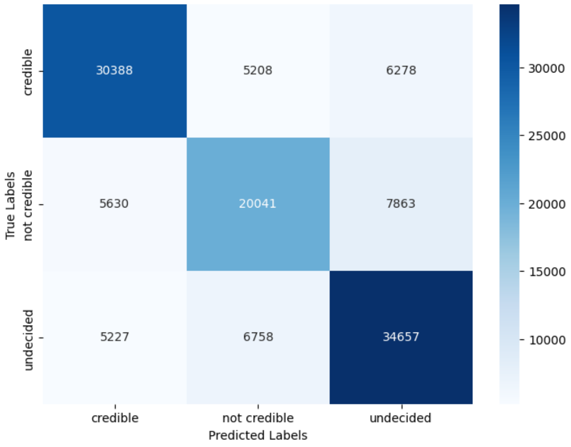


Fig. 3. LSTM Confusion Matrix.

D. Baseline Models

We compared the LSTM model's performance with two baseline models: Naive Bayes and Support Vector Machine (SVM).

1) Naive Bayes: The Naive Bayes model used a TF-IDF vectorizer with a maximum of 5000 features. The model achieved a test accuracy of 56.57%. The classification report is shown in Table 2, showing that the model struggled particularly with the 'not credible' class, achieving a macro average F1-score of 0.54.

TABLE II
NAIVE BAYES CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	0.56	0.62	0.59	41874
1	0.55	0.34	0.42	33534
2	0.57	0.68	0.62	46642
Accuracy			0.57	122050
Macro Avg	0.56	0.55	0.54	122050
Weighted Avg	0.56	0.57	0.56	122050

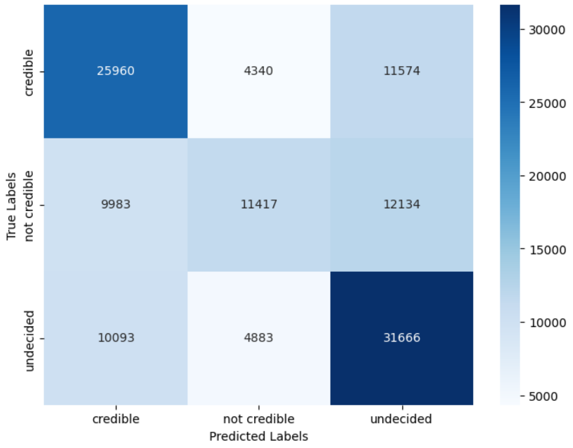


Fig. 4. Naive Bayes Confusion Matrix.

2) Support Vector Machine: The SVM model also utilized TF-IDF vectorized features. It achieved a test accuracy of 67.19%. The classification report, shown in Table 3, indicated better performance compared to Naive Bayes, particularly in classifying credible articles, with an overall weighted average F1-score of 0.67.

TABLE III
SVM CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	0.69	0.71	0.70	41874
1	0.65	0.50	0.56	33534
2	0.66	0.76	0.71	46642
Accuracy			0.67	122050
Macro Avg	0.67	0.66	0.66	122050
Weighted Avg	0.67	0.67	0.67	122050

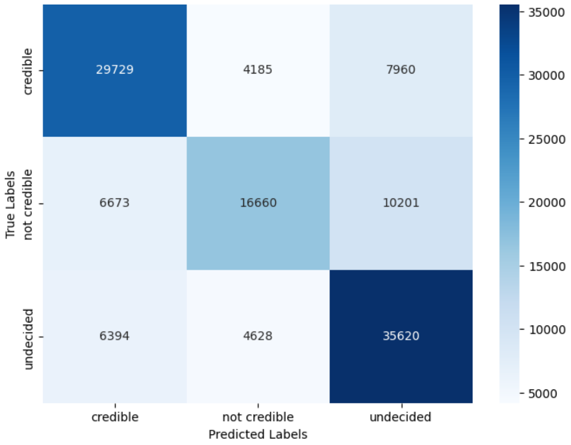


Fig. 5. SVM Confusion Matrix.

E. Accuracy Comparison

The accuracy comparison for the three models is summarized in Table 4. The LSTM model outperformed both the Naive Bayes and SVM models in terms of test accuracy and F1-score. The deep learning approach demonstrated superior capability in capturing the complex patterns in the textual data, particularly beneficial for handling the nuances of the Arabic language.

TABLE IV
ACCURACY COMPARISON OF DIFFERENT MODELS

Model	Test Accuracy	Precision	Recall	F1-Score
LSTM	69.71%	0.70	0.70	0.70
Naive Bayes	56.57%	0.56	0.57	0.56
SVM	67.19%	0.67	0.67	0.67

F. Visualization of Results

Confusion matrices for all models were plotted to provide a visual comparison of true versus predicted labels. The heatmaps highlighted the areas where each model performed well and where they struggled, offering insights into potential areas for model improvement.

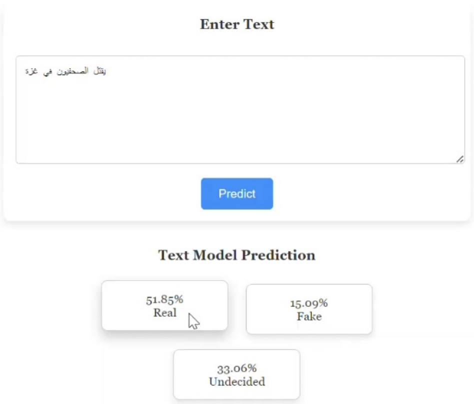
Overall, our experiments indicate that deep learning models, particularly LSTM, are more effective for Arabic fake news detection compared to traditional machine learning models. The promising results suggest that further tuning and exploration of advanced neural network architectures could enhance performance even further.

G. User Interface Demonstration (Demo)

To provide a practical application of our model, we developed a user-friendly web interface that allows users to enter news articles and receive real-time predictions on their credibility. This demonstration aims to bridge the gap between our research and real-world usability, offering a tool that can assist in the immediate detection of fake news.

The web interface is built with a sleek and intuitive design, ensuring a seamless user experience. Users can input the text of a news article into a designated field and initiate the analysis with a simple click of a button. The backend, powered by our trained LSTM model, processes the input and outputs the prediction, classifying the article as credible, not credible, or undecided.

Two key figures are included to illustrate the functionality of the demonstration:



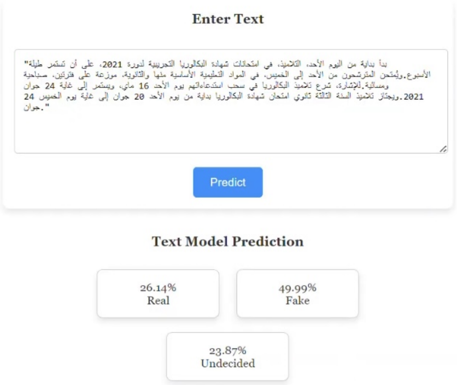
يقتل الصحفيون في غزة

51.85% Real

15.09% Fake

33.06% Undecided

Fig. 6. Demo Screen Example 1.



"بدأ بداية من اليوم الأحد، التلاميذ، في امتحانات شهادة البكالوريا التجريبية لدورة 2021، على أن تستمر طيلة
الأسبوع.ويُمتحن المترشحون من الأحد إلى الخميس، في المواد التعليمية الأساسية منها والثانوية، موزعة على فترتين، صباحية
ومسائية.للإشارة، شرع تلاميذ البكالوريا في سحب استدعاءاتهم يوم الأحد 16 ماي، ويستمر إلى غاية 24 جوان
2021.ويجتاز تلاميذ السنة الثالثة ثانوي امتحان شهادة البكالوريا بداية من يوم الأحد 20 جوان إلى غاية يوم الخميس 24
جوان."

26.14% Real

49.99% Fake

23.87% Undecided

Fig. 7. Demo Screen Example 2.

This demonstration not only validates the effectiveness of our model in a practical setting but also showcases the potential for deploying machine learning models in accessible formats, empowering users to combat the spread of fake news with ease.

VII. Conclusion

Our study developed an advanced Arabic Fake News Detection (AFND) system using LSTM networks, effectively addressing the challenge of detecting fake news in Arabic texts. Through experiments with the AFND and Gaza datasets, our LSTM model demonstrated superior performance compared to traditional methods like Naive Bayes and SVM, achieving high accuracy and precision.

This research highlights the effectiveness of deep learning for handling the nuances of Arabic language semantics and syntax in fake news detection. Our findings contribute valuable insights into combating misinformation in Arabic-speaking regions, emphasizing the importance of AI-driven solutions for preserving information integrity on digital platforms.

VIII. Future Work

Future research can build upon our findings in several key areas:

- Explore advanced model architectures beyond LSTM, such as Transformer-based models like BERT, to enhance contextual understanding in Arabic texts.
- Investigate ensemble learning techniques to improve model robustness and performance in detecting diverse forms of fake news.
- Develop methods for interpreting AI model decisions to increase transparency and user trust in automated fake news detection.
- Conduct cross-platform evaluations to assess the AFND system's effectiveness across different digital media and platforms.
- Perform longitudinal studies to track the evolution of fake news and develop adaptive detection strategies over time.

By addressing these research avenues, we aim to further advance the capabilities of our AFND system and contribute to the ongoing fight against misinformation in the digital age.

References

[1] Shatha Alyoubi, Manal Kalkatawi, and Felwa Abukhodair. The detection of fake news in arabic tweets using deep learning. *Applied Sciences*, 13(14), 2023.

[2] Thaer Thaher, Mahmoud Saheb, Hamza Turabieh, and Hamouda Chantar. Intelligent detection of false information in arabic tweets utilizing hybrid harris hawks based feature selection and machine learning models. *Symmetry*, 13(4), 2021.

[3] Monther Aldwairi and Ali Alwahedi. Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222, 2018. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops.

[4] NLTK. NLTK. https://www.nltk.org/. Accessed: June 16, 2024.

[5] M. Yaseen. Arabic fake news dataset (afnd). *Kaggle*, May 13 2022.

[6] E. Silsüpür. Israel-palestine conflict news dataset. *Kaggle*, December 23 2023.

IX. Appendix

Project Code on github