# USTHB at ArAIEval23 Shared Task: Disinformation Detection System based on Linguistic Feature Concatenation

**Mohamed Lichouri**
LCPTS-USTHB, Algiers, Algeria
mlichouri@usthb.dz

**Khaled Lounnas, Aicha Zitouni**
LCPTS-USTHB, Algiers, Algeria
CRSTDLA, Algiers, Algeria
{k.lounnas, a.zitouni}@crstdla.dz

**Houda Latrache**
CRSTDLA, Algiers, Algeria
h.latrache@crstdla.dz

**Rachida Djeradi**
LCPTS-USTHB, Algiers, Algeria
rdjeradi@usthb.dz

## Abstract

In this research paper, we undertake a comprehensive examination of several pivotal factors that impact the performance of Arabic Disinformation Detection in the ArAIEval'2023 shared task. Our exploration encompasses the influence of surface preprocessing, morphological preprocessing, the FastText vector model, and the weighted fusion of TF-IDF features. To carry out classification tasks, we employ the Linear Support Vector Classification (LSVC) model. In the evaluation phase, our system showcases significant results, achieving an $F_1$ micro score of 76.70% and 50.46% for binary and multiclass classification scenarios, respectively. These accomplishments closely correspond to the average $F_1$ micro scores achieved by other systems submitted for the second subtask, standing at 77.96% and 64.85% for binary and multiclass classification scenarios, respectively.

## 1 Introduction

In recent years, the detection of disinformation in digital content has become a critical challenge at the intersection of natural language processing and information security, spurred by the growing influence of online platforms (Shu et al., 2020). The Arabic-speaking digital landscape, in particular, has witnessed an alarming increase in susceptibility to the dissemination of false or misleading information, a phenomenon well-documented in recent research (Harrag and Djahli, 2022). The ramifications of disinformation extend beyond individual deception; they cover broader societal consequences, affecting public opinion, social cohesion, and even national security.

Recognizing the gravity of this issue, we actively participate in the inaugural shared task organized by ArAIEval'2023, which focuses on disinformation detection in Arabic text (Hasanain et al., 2023). Our engagement in this task reflects our commitment to addressing this pressing challenge. By harnessing advanced natural language processing techniques and machine learning models, we endeavor to contribute to the development of effective disinformation detection systems tailored to the nuances of the Arabic language. Through rigorous experimentation and evaluation, we aim to enhance our understanding of the complexities involved and offer practical solutions to safeguard the integrity of digital discourse and information dissemination in the Arabic-speaking world.

To combat the proliferation of disinformation in Arabic text, a growing number of research has been dedicated to developing robust and effective detection systems (Alam et al., 2022; Mubarak et al., 2023). Much like the endeavors undertaken in the field of Arabic dialect identification (Lichouri et al., 2021b), disinformation detection in Arabic requires a nuanced understanding of the language's intricacies (Nagoudi et al., 2020), as well as the ability to sift through vast amounts of textual data (Himdi et al., 2022) to identify instances of deceptive or misleading content.

In this paper, we embark on an extensive exploration of disinformation detection in Arabic, drawing inspiration from the methodologies and techniques employed in previous shared tasks (Lichouri et al., 2020). Leveraging these insights, we aim to build upon existing research and contribute to the ongoing efforts to enhance the accuracy and effectiveness of disinformation detection systems in Arabic text.

Our study encompasses a comprehensive analysis of various factors influencing the performance of Arabic disinformation detection, including surface and morphological preprocessing techniques (Lichouri et al., 2021a), feature engineering strategies (Fouad et al., 2022), and the implementation of



*Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, pages 647–651
December 7, 2023 

state-of-the-art machine learning models. Through rigorous experimentation and evaluation, we seek to provide valuable insights and practical solutions that can aid in the identification and mitigation of disinformation.

This paper is organized as follows: Section 2 offers insights into the dataset we have employed. Moving on to Section 3, we introduce our proposed system, which includes details about the cleaning and preprocessing steps discussed in Section 3.1. The process of feature engineering is elucidated in Section 3.2. Section 3.3 is dedicated to a comprehensive discussion of our findings. Finally, we wrap up the paper in Section 4 with a conclusive summary of our contributions and key findings.

## 2 Description of the Dataset

A disinformation dataset constitutes a crucial resource for studying and comprehending the multifaceted landscape of misinformation, misleading content, and fabricated information within various digital platforms. Such datasets encompass a diverse array of textual, visual, and multimedia content intentionally designed to deceive, mislead, or manipulate audiences. These datasets serve as invaluable assets for researchers, data scientists, and machine learning practitioners engaged in the development of advanced algorithms and models aimed at detecting, analyzing, and combating disinformation. By analyzing patterns, linguistic cues, and contextual elements within disinformation datasets, researchers gain insights into the tactics, strategies, and evolving nature of disinformation campaigns, thereby contributing to the enhancement of society's ability to discern and mitigate the harmful impacts of deceptive content in an increasingly interconnected information landscape.

Additional information regarding this dataset can be found in Table 1, where we took part for the first time this year in both editions of the Disinformation Detection Definition shared task. This task involves classifying binary and fine-grained disinformation categories based solely on the text of a tweet. Please note that these statistics pertain to the dataset after we removed punctuation and emojis. Imbalanced datasets can have a pronounced effect on system performance, causing the development of biased models that prioritize the dominant class (e.g., "no-disinformation" in binary classification and "HS" in multi-class classification). This can result in decreased predictive accuracy for the underrepresented classes, such as "disinformation" in binary classification, "Rumor", and "Spam" in the multi-class scenario, and compromised decision-making in applications like fraud detection or medical diagnosis. Addressing class imbalance through techniques like oversampling, undersampling, or using appropriate evaluation metrics is crucial for more equitable and accurate model outcomes.

## 3 Proposed system

### 3.1 Data Cleaning and Preprocessing

In the challenging domain of disinformation detection within Arabic text, it becomes imperative to adeptly capture essential information while efficiently removing undesirable elements. This task is known for its complexity and nuance, demanding a detailed approach. To address this challenge, we have implemented a two-phase preprocessing strategy:

**Phase 1: Surface Preprocessing** - In this initial phase, we execute a range of foundational procedures:

- *Arabic Letter Normalization*: Ensuring consistency in Arabic script characters (Sallam et al., 2016).
- *Punctuation and Emoji Removal*: Eliminating punctuation marks and emoticons (Shiha and Ayvaz, 2017).
- *Stop Words Removal*: Handling common words that do not contribute substantially to meaning.
- *Diacritics Removal*: Removing diacritical marks for text clarity (Jbara et al., 2009).
- *Exclusion of Non-Arabic Content*: Ensuring that only Arabic text remains (Omar et al., 2021).

These collective measures ensure text clarity, uniformity, and the removal of any distractions.

**Phase 2: Morphological Preprocessing** - In this phase, our focus shifts to the intricacies of language. Here, we employ the following techniques:

- *Lemmatization*: Simplifying word forms to their base or dictionary form (El Kah and Zeroual, 2021).
- *Stemming*: Reducing words to their root forms, aiding in the identification of core word meanings and structures (Atwan et al., 2021).

Table 1: ArAIEval (Task2A/2B) dataset statistics where : Task2A for Binary classification whereas Task2B for Multiclass classification problem.

| | Train | Dev | Test |
|---|---|---|---|
| # sentences | 14147/2656 | 2115/397 | 3729/876 |
| # words | 324727/68073 | 48917/10062 | 100646/27312 |
| Max # word per sentence | 65/67 | 65/59 | 62 /62 |
| Min # word per sentence | 0 / 1 | 0 / 1 | 1 / 1 |
| Max # char per sentence | 280/290 | 280 /285 | 311 /311 |
| Min # char per sentence | 0 / 3 | 0 / 3 | 2 / 2 |

Table 2: The various combinations and parameter used in our work

| Settings | Range |
|---|---|
| ngram_range | (m,n) with m=1 to 3 and n=1 to 10 |
| tfidf_weights | 0.5 - 1 |
| tfidf max_features | 1000 -25000 |
| SVM | C=100, gamma=1-10 |
| fasttext_supervised | epoch=100, loss='ova' |
| fasttext_unsupervised | epoch=100, ws=6 model='skipgram' dim=1000 |

Throughout both phases, we intricately harmonize and fine-tune various techniques to arrive at the optimal configuration for our preprocessing pipeline.

### 3.2 Feature engineering

Our system operates through a well-defined structure consisting of four distinct phases, offering the flexibility to be applied individually or collectively. The initial two phases, Surface Preprocessing and Morphological Preprocessing, have been expounded upon in the previous section. The subsequent phases are detailed as follows:

**Phase 3: Feature Extraction** - In this stage, we employ a dual-model approach. Firstly, the FastText model undergoes comprehensive training in two modes: supervised and unsupervised, drawing from the training dataset. Then, we use this model to extract features from both the development and test datasets. Secondly, we leverage the TF-IDFVectorizer, an adept tool offering three distinct analyzers (Word, Char, and Char_wb), each encompassing variable n-gram ranges. As a default configuration, we combine these three TF-IDF features, affording them equal weights, all set to 1.

**Phase 4: Weighted Fusion** - In this phase, we combined the three TF-IDF features, supported by a weight vector featuring three distinctive values (w1, w2, w3) that correspond to the Word, Char, and Char_wb TF-IDF features, respectively.

Having presented these four distinctive phases, we executed four designed experiments that were inspired by our prior works (Lichouri et al., 2018; Abbas et al., 2019; Lichouri and Abbas, 2020a), where each embody distinct configurations:

**Experiment 1 (Lichouri et al., 2021a; Lichouri and Abbas, 2020b):** In this first experiment, we initiated with the first phase, by considering all the possible permutations of surface processing techniques. Following this, we considered the third phase, marked by the employment of a union of TF-IDF features. During the feature extraction process, we explored a range of n-gram values, spanning from $n = 1$ to 10. Finally, we finished by the training of the SVC classifier.

**Experiment 2 (Lichouri et al., 2020):** In this specific scenario, we worked with the second phase, by exploring various combinations of morphological processing techniques. Similar to Experiment 1, we progressed to the third phase, where we concat the TF-IDF features, all while varying the n-gram parameters. We then finished this experiment by training of the SVC classifier.

**Experiment 3:** For this unique experiment, we focused on the third phase, where we used FastText model for feature extraction, followed by the rigorous training of the SVC classifier.

**Experiment 4 (Lichouri et al., 2021b):** In this distinctive scenario, we executed the fourth phase, by applying a weighted union of TF-IDF features for feature extraction. Then, we concluded with

| Task | Binary | | | | Multiple | | | |
|---|---|---|---|---|---|---|---|---|
| **Desc** | MP | SP | F-Vec | WF | MP | SP | F-Vec | WF |
| Run 1 | 81,08 | **81,23** | 48,45 | 81,13 | 56,92 | **57,43** | 27.57 | 56,93 |
| Run 2 | 81,08 | 81,18 | 48.27 | 78,91 | 56,92 | 56,68 | 27.68 | 56,92 |
| Run 3 | 81,08 | 81,09 | 46.54 | 75,74 | 56,92 | 56,93 | 22.44 | 56,68 |

Table 3: The F1-micro percentages obtained using the proposed system Where: SP (Surface Preprocessing), MP (Morphological Preprocessing), F-Vec (Vectorisation), and WF (Weighted Fusion)

the training of the SVC classifier.

Following many iterations of these four experiments on both the training and development datasets, we recorded the best results attained for each experiment, along with the precise configurations that yielded these outcomes, as presented in Table 2.

### 3.3 Results and Discussion

In this study, we conducted a series of experiments aimed at detecting Arabic disinformation. These experiments were centered around the utilization of various descriptors, encompassing Surface Preprocessing (SP), Morphological Preprocessing (MP), the vectorisation model (F-Vec), and Weighted Fusion of TF-IDF (WF).

To explore the effectiveness of these descriptors, we employed a range of combinations and settings. This involved modifying n-gram values and TF-IDF weights to investigate the impact of word sequence length on results and term weighting in the text, respectively. Table 2 provides a comprehensive summary of the different combinations and parameters used in our study, while Table 3 presents the results obtained using these combinations.

Our experiments yielded valuable insights into the efficacy of various techniques for disinformation detection, specifically in binary and multiclass classification tasks. Notably, for the binary subtask, Surface Preprocessing demonstrated the highest performance, achieving an impressive F1-score of 81.23%. It was closely followed by the Weighted Union of TF-IDF features, with an F1-score of 81.13%, while Morphological Preprocessing exhibited slightly lower performance, resulting in an F1-score of 81.08%. Intriguingly, the FastText model underperformed in this context, attaining the lowest F1-score at 48.45%.

However, a fascinating observation emerged when we transitioned to the multiclass classification subtask. Surprisingly, the same observation held true, but the obtained results dropped significantly, by approximately 20%, compared to the binary case. We hypothesize that this decline in performance could be attributed to the imbalanced nature of the dataset, which has a more pronounced impact in the multiclass scenario.

## 4 Conclusion

In conclusion, our comprehensive analysis of key factors in Arabic Disinformation Detection has shed light on critical aspects that significantly influence performance. Through a meticulous exploration of surface preprocessing, morphological preprocessing, the FastText vector model, and the weighted fusion of TF-IDF features, we have gained valuable insights into their impact on classification tasks.

Our system's noteworthy achievement of an $F_1$ micro score of 76.70% and 50.46% for binary and multiclass classification setups, respectively, closely aligns with the performance of other systems submitted for the second subtask. This not only reaffirms the significance of surface preprocessing and weighted TF-IDF feature fusion but also positions them as robust techniques in the domain of Arabic Disinformation Detection.

## References

Mourad Abbas, Mohamed Lichouri, and Abed Alhakim Freihat. 2019. St madar 2019 shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 269–273.

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 Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jaffar Atwan, Mohammad Wedyan, Qusay Bsoul, Ahmad Hamadeen, Ryan Alturki, and Mohammed

Ikram. 2021. The effect of using light stemming for arabic text classification. *International Journal of Advanced Computer Science and Applications*, 12(5).

Anoual El Kah and Imad Zeroual. 2021. The effects of pre-processing techniques on arabic text classification. *Int. J*, 10(1):1–12.

Khaled M Fouad, Sahar F Sabbeh, and Walaa Medhat. 2022. Arabic fake news detection using deep learning. *Computers, Materials & Continua*, 71(2).

Fouzi Harrag and Mohamed Khalil Djahli. 2022. Arabic fake news detection: A fact checking based deep learning approach. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–34.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hanen Himdi, George Weir, Fatmah Assiri, and Hassanin Al-Barhamtoshy. 2022. Arabic fake news detection based on textual analysis. *Arabian Journal for Science and Engineering*, 47(8):10453–10469.

Khitam Mahmoud Abdalla Jbara, Azzam T Sleit, and Bassam H Hammo. 2009. *Knowledge discovery in Al-Hadith using text classification algorithm*. University of Jordan.

Mohamed Lichouri and Mourad Abbas. 2020a. Simple vs oversampling-based classification methods for fine grained arabic dialect identification in twitter. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 250–256.

Mohamed Lichouri and Mourad Abbas. 2020b. Speechtrans@ smm4h'20: Impact of preprocessing and n-grams on automatic classification of tweets that mention medications. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 118–120.

Mohamed Lichouri, Mourad Abbas, and Besma Benaziz. 2020. Profiling fake news spreaders on twitter based on tfidf features and morphological process.

Mohamed Lichouri, Mourad Abbas, Besma Benaziz, Aicha Zitouni, and Khaled Lounnas. 2021a. Preprocessing solutions for detection of sarcasm and sentiment for Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 376–380, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mohamed Lichouri, Mourad Abbas, Abed Alhakim Freihat, and Dhiya El Hak Megtouf. 2018. Word-level vs sentence-level language identification: Application to algerian and arabic dialects. *Procedia Computer Science*, 142:246–253.

Mohamed Lichouri, Mourad Abbas, Khaled Lounnas, Besma Benaziz, and Aicha Zitouni. 2021b. Arabic dialect identification based on a weighted concatenation of TF-IDF features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 282–286, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and identifying the reasons for deleted tweets before they are posted. *Frontiers in Artificial Intelligence*, 6.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, Tariq Alhindi, and Hasan Cavusoglu. 2020. Machine generation and detection of arabic manipulated and fake news. *arXiv preprint arXiv:2011.03092*.

Ahmed Omar, Tarek M Mahmoud, Tarek Abd-El-Hafeez, and Ahmed Mahfouz. 2021. Multi-label arabic text classification in online social networks. *Information Systems*, 100:101785.

Rouhia M Sallam, Hamdy M Mousa, and Mahmoud Hussein. 2016. Improving arabic text categorization using normalization and stemming techniques. *Int. J. Comput. Appl*, 135(2):38–43.

Mohammed Shiha and Serkan Ayvaz. 2017. The effects of emoji in sentiment analysis. *Int. J. Comput. Electr. Eng.(IJCEE.)*, 9(1):360–369.

Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1385.