Advancing Arabic Sentiment Analysis: ArSen Benchmark and the Improved Fuzzy Deep Hybrid Network

Yang Fang[1] **Cheng Xu**[2*] **Shuhao Guan**[2] **Nan Yan**[3] **Yuke Mei**[4]
[1] Huaibei Normal University [2] University College Dublin
[3] Georgia Institute of Technology [4] Wuhu Institute of Technology
cheng.xu1@ucdconnect.ie

Abstract

Sentiment analysis is crucial in Natural Language Processing as it enables the extraction of opinions and emotions from text. However, Arabic sentiment analysis is often overlooked. Current benchmarks for Arabic sentiment analysis tend to be outdated or lack comprehensive annotations, which limits the development of more accurate and reliable models for the Arabic language. To address these challenges, we introduce ArSen, a meticulously annotated Arabic dataset centered on COVID-19, along with IFDHN, a novel model that employs fuzzy logic for more precise sentiment classification[1]. ArSen offers a robust and contemporary benchmark, and IFDHN achieves state-of-the-art performance in Arabic sentiment analysis, with 78.12% accuracy, an F1-Macro score of 55.83%, and an F1-Micro score of 78.12% on the test set. Notably, by using only 0.23% of the computational resources of large language models, IFDHN achieved performance comparable to LLaMA-3-8B, showcasing significant improvements over existing methods.

1 Introduction

Sentiment analysis (SA), also known as opinion mining, is a critical task in Natural Language Processing (NLP) that involves detecting, extracting, and classifying opinions and emotions expressed in text (Marreddy and Mamidi, 2023; Hussein, 2018). In recent years, the advent of social media platforms like Twitter (X for now) has provided a rich data source for SA. Building on this, sophisticated models such as RoBERTa-LSTM and KEAHT have emerged, further promoting the development of the SA field (Tan et al., 2022; Tabinda Kokab et al., 2022; Tiwari and Nagpal, 2022).

Despite these advancements in sentiment analysis, the complexity of the Arabic language, combined with its significant differences from English, has led to a scarcity of studies and resources in Arabic sentiment analysis (ASA) (El-Masri et al., 2017; Yan and Xu, 2024). The widely used ASA benchmarks, such as Gold Standard (Refaee and Rieser, 2014) and SemEval (Rosenthal et al., 2017), are often outdated and small in scale (less than 10,000). To address this gap, we leveraged a large volume of Arabic tweets generated during the COVID-19 pandemic. During this pandemic, Arabic-speaking users widely shared their emotions and experiences. This large-scale public sharing made it possible to construct a comprehensive and diverse dataset. Therefore, we introduce **Ar**abic **Sen**timent (ArSen), a COVID-19-themed Arabic benchmark created through meticulous manual annotation by trained professionals. The ArSen benchmark aims to address the previously mentioned challenges and provide ASA research with a modern, comprehensive resource featuring accurate data annotations, thus advancing the field of ASA within NLP.

Additionally, we propose a new model called the **I**mproved **F**uzzy **D**eep **H**ybrid **N**etwork (IFDHN), designed specifically to enhance sentiment classification through the integration of fuzzy logic. Fuzzy logic has been effectively applied in sentiment analysis to handle the ambiguity and nuances of language (Zadeh, 1996; Vashishtha et al., 2023). Our IFDHN model demonstrates state-of-the-art (SOTA) performance in ASA tasks, validating the effectiveness of incorporating fuzzy logic to improve classification accuracy.

Our **contributions** are twofold: (1) We proposed ArSen, a robust and contemporary benchmark for ASA tasks, addressing the lack of up-to-date and high-quality benchmarks in this domain; (2) we introduced IFDHN, a novel model that integrates fuzzy logic to better handle ambiguous sentiments, improving overall classification performance.

The paper is organized as follows: Section 2 introduces the ArSen benchmark, detailing its con-

* Corresponding author.
[1]Resources are available at: https://github.com/123fangyang/ArSen.

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 507–516

struction and significance. Section 3 discusses the architecture and features of IFDHN model. Section 4 presents comprehensive evaluations of the IFDHN model against leading SOTA models using the ArSen dataset. Finally, Section 5 summarizes our findings and proposes directions for future research in ASA.

2 ArSen Benchmark

To address the aforementioned shortcomings in ASA, we introduce the ArSen benchmark. Firstly, our motivation for creating the ArSen benchmark is discussed in Section 2.1, where we outline the rationale for selecting COVID-19-themed tweets to develop the benchmark. We then move on to describe the benchmark construction process in Section 2.2, providing a thorough explanation of the data preprocessing and annotation steps involved. This section aims to provide a clear understanding of how ArSen was developed and the rigorous methodologies employed to ensure its quality.

2.1 Motivation

The COVID-19 pandemic disrupted daily life for everyone and became a trending topic on Twitter from 2020 to 2023 (Ali, 2021). For now, the COVID-19 crisis has largely subsided, the tweets from this period provide a comprehensive and complete picture of the real emotional states of Arabic-speaking users during the pandemic, such as fear, anxiety, hope, and solidarity (Lwin et al., 2020). Additionally, the pandemic led to discussions on a variety of topics, including health, economy, politics, and social interactions (Chandrasekaran et al., 2020), which enhances the dataset's comprehensiveness and enables the development of models that can handle a wide range of topics (Xu et al., 2022). This rich emotional context and topic diversity offer valuable insights for ASA in a contemporary and relevant setting. Therefore, we focus on using tweet data from the COVID-19 period to develop the ArSen benchmark for ASA. In our previous research, we introduced a similar benchmark, ArSen-20 (Fang and Xu, 2024), which included 20,000 tweets. However, ArSen-20 had limitations, such as a less rigorous annotation process, and no experiments were conducted using the dataset. To address these issues, we have implemented stricter annotation standards and performed extensive experiments to enhance the reliability and usefulness of our new benchmark.

Field	Type	Description
like_count	int	The number of likes on this tweet.
quote_count	int	The number of times this tweet has been quoted.
reply_count	int	The number of replies to this tweet.
retweet_count	int	The number of retweets to this tweet.
tweet	string	The actual UTF-8 text of the tweet.
user_verified	boolean	Indicates if this user is a verified Twitter User.
followers count	int	The number of followers of the author.
following count	int	The number of following of the author.
tweet_count	int	Total number of tweets by the author.
listed_count	int	The number of public lists that this user is a member of.
description	string	The text of this user's profile description (bio).
created_at	date	Creation time of the tweet.
label	string	Sentiment Classification of this tweet.

Table 1: Tweets field feature information.

2.2 Data Preprocessing and Annotation

Xu and Yan (2023) provided a suitable opportunity for our work with their proposed AROT-COV23[2] dataset, which collected approximately 500,000 original COVID-19-related tweets and contextual information, spanning from January 2020 to January 2023. These data can be accessed and used for research purposes, our ArSen dataset follows the same policy. To maintain representativeness while reducing dataset size for efficient analysis, we randomly selected ~10k tweets from AROT-COV23. Furthermore, to protect the privacy of Twitter users, we remove redundant features that could expose personal information during preprocessing, thereby streamlining the dataset. The detailed tweets field feature information is shown in Table 1.

Following this preprocessing phase, we annotated around 10,000 tweets into three classes: positive, neutral, and negative. Each tweet was annotated by three annotators, who are advanced Arabic speakers. They received thorough training in advance, following the same labeling guidelines. The annotation guidelines categorized tweets as follows:

Positive: Tweets expressing happiness, gratitude, affirmation, encouragement, and solidarity.

Neutral: Tweets conveying factual information, such as news updates, advertisements, suggestions, advice, and questions.

Negative: Tweets conveying sadness, condemnation, sarcasm, warnings, protests, regret, refutation, and obituaries.

Notably, in our annotation process, emojis helped as cues to label the tweets more quickly. For instance, a positive tweet often includes a 'smile emoji' or a 'red heart emoji' to express the author's

[2] https://github.com/chengxuphd/AROT-COV23

Labels	Example in Arabic	English Translations
Positive	.الحمد لله والشكر لله	Praise be to God and thanks be to God.
Neutral	.ارتفاع حصيلة وفيات فيروس كورونا إلى ستة	France: Coronavirus death toll rises to six.
Negative	.يوم حزين آخر في إيطاليا	Another sad day in Italy.

Table 2: Labels used in annotation and examples of each.

happiness or well-wishes to others. In addition, the tweet's sentiment must reflect the author's emotion when they posted the tweet, rather than the annotators' opinion.

In the annotation process, we employ a voting mechanism. If two out of the three annotators agree on a label, we accept that label (Rosenthal et al., 2017; Alharbi et al., 2021). Otherwise, this tweet will be deleted. Furthermore, Table 2 provides examples of tweets from each sentiment category as part of the annotation process.

We present the detailed statistics for the ArSen dataset in Table 3, offering insights into the data size and label classifications, which indicate that neutral sentiments dominate the dataset. This is primarily because most tweets aimed to inform the public about the latest developments in the pandemic by sharing neutral news updates, while only a smaller portion expressed the authors' genuine emotional responses (positive or negative).

Statistics	Num	Proportion
Data size		
Training set	8153	80%
Validation set	1020	10%
Testing set	1020	10%
Avg. tweet length (tokens)	146	-
Labels		
Neutral	7069	69.4%
Positive	1564	15.3%
Negative	1557	15.3%

Table 3: The ArSen dataset statistics.

3 Proposed Model

Researchers have long recognized the unique advantages of fuzzy logic in capturing the ambiguities and uncertainties of real-world data (Das et al., 2020). Zadeh (1996) introduced the concept of computing words using fuzzy logic. In this approach, sentiment polarity is determined by calculating fuzzy membership values ranging from 0.0 to 1.0. Each word in the text is assigned a score within this range, reflecting the realistic scenario where sentiment is not always binary but often ambiguous and uncertain (Vashishtha et al., 2023). In recent years, fuzzy logic has also drawn significant attention in the field of SA (Huyen Trang Phan and Nguyen, 2023; Golondrino et al., 2023; Sun et al., 2024; Alzaid and Fkih, 2023). Moreover, the ArSen dataset contains contextual information, so we would like to construct a multi-channel fuzzy model to test the ArSen dataset. A recent study in the field of fake news detection provides an opportunity for this work. Xu and Kechadi (2023, 2024) introduced the FDHN model, which uses fuzzy logic and multiple input types: news text, textual context, and numerical context. The text inputs are processed by TextCNNs, while numerical context is handled by CNN and Bi-LSTM layers, then processed by a Fuzzy Layer. The model's outputs are concatenated and integrated in the final layer, achieving SOTA performance metrics on the LIAR dataset (Wang, 2017), which includes multi-class labels such as pants-fire, false, barely-true, half-true, mostly-true, and true. This use of fuzzy multi-class labels shows a strong similarity to our ArSen benchmark. Beyond this, they both require contextual information. Therefore, we believe that the strengths of the FDHN model allow us to adequately analyze the ArSen benchmark. In order to transfer FDHN to the ASA task, we tailored and improved the architecture of FDHN to propose the IFDHN model, aiming to better utilize its fuzzy logic and context-dependent properties. Through our experiments, we found that introducing textual context information, specifically the *created_at* and *description* features in the ArSen dataset, was redundant and decreased the model's performance, leading us to remove these features. Furthermore, we designed a separate TextCNN to process tweet text and then fine-tuned the CNN-BiLSTM module for numerical context.

Although Large Language Models (LLMs) like GPT-3.5/4 (OpenAI, 2024) and LLaMA-3 (Dubey et al., 2024) have achieved impressive results in many NLP tasks like question answering and text generation, they fall short in interpretability and computational efficiency for fuzzy classification tasks (Chang et al., 2024; Bang et al., 2023; BehnamGhader et al., 2024). For example, GPT-4 achieved only 28.1% accuracy on the LIAR dataset, whereas FDHN achieved 46.5% (Pelrine et al., 2023), and FDHN requires only about 3 seconds to train an epoch on a single A100 GPU, while LLMs generally require more than 4 A100 GPUs to be fine-tuned for hours to train for downstream

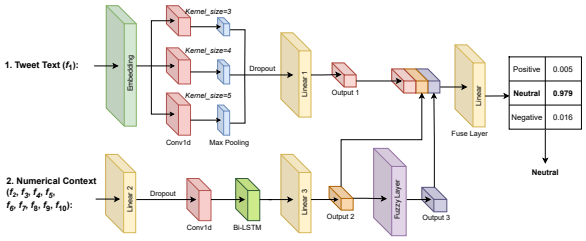

Figure 1: The IFDHN model structure.

Model	Validation			Test		
	Accuracy	F1-Macro	F1-Micro	Accuracy	F1-Macro	F1-Micro
RoBERTa	0.6889	0.2719	0.6889	0.6850	0.2710	0.6850
AraT5-Tweet-Base	0.7134	**0.6604**	0.7134	0.7723	**0.6837**	0.7723
FNet	0.7233	0.5081	0.7233	0.7429	0.4960	0.7429
LLaMA-3-8B	0.7428	0.6236	**0.7428**	0.7595	0.6240	0.7595
FDHN	0.7350	0.4888	0.7306	0.7753	0.5575	0.7753
IFDHN	**0.7478**	0.5113	0.7368	**0.7812**	0.5583	**0.7812**

Table 5: Comparison of various state-of-the-art models on ArSen dataset. The highest scores are highlighted in bold, while the second-highest scores are highlighted with an underline.

tasks. Taking these considerations into account, we decided to use FDHN as the baseline model in this work. However, we also included the results of LLaMA-3-8B for comparison.

As illustrated in Figure 1, the IFDHN model comprises two primary channels: *Tweet Text* and *Numerical Context*. The tweet text is fed into a distinct TextCNN, while the numerical context is processed by a combination of CNN and Bi-LSTM layers before being passed through a Fuzzy Layer. The model produces three outputs: output 1 is derived from the Tweet Text channel, output 2 is derived from the Numerical Context channel, and output 3 is the Fuzzy Layer-processed version of output 2. These three output representations are then concatenated and integrated in the final layer. In particular, an example data point used in our IFDHN model is shown in Table 4, with f_1 representing the tweet text and $\{f_2, \ldots, f_{10}\}$ representing the numerical context. More detailed component analysis is provided in Appendix A.

#	Field	Value
-	label	positive
f_1	tweet	شكرا لقيادتنا الحكيمة.
f_2	like_count	6
f_3	quote_count	0
f_4	reply_count	0
f_5	retweet_count	6
f_6	followers_count	10977
f_7	following_count	356
f_8	tweet_count	9029
f_9	listed_count	108232
f_{10}	user_verified	False

Table 4: An example data point from ArSen dataset used in IFDHN model.

4 Experimental Results

In this section, we present a comprehensive analysis of our experiments, which are divided into two main parts: a performance evaluation on the ArSen dataset using the IFDHN model and other SOTA models (Section 4.1). In addition, an ablation study was performed to investigate the impact of different features on the performance of the IFDHN model (Section 4.2). Performance evaluation metrics are detailed in Appendix B. Detailed information about our experimental setup, including the development environment and hyperparameter configurations, can be found in Appendix C.

4.1 Performance Comparison

We evaluated the performance of the IFDHN model with several SOTA models on the ArSen dataset. Table 5 presents a comparison of the accuracy and F1 scores for the validation and testing sets.

The RoBERTa model (Liu et al., 2019) is an optimized BERT (Devlin et al., 2019) variant trained with more data and longer sequences. Despite its robust architecture, RoBERTa yielded the lowest performance in our experiments, with particularly low F1-Macro scores of 0.2719 on the validation set and 0.2710 on the test set.

Nagoudi et al. (2022) evaluated both Dialectal Arabic and Modern Standard Arabic, introducing the AraT5-Tweet-Base model. This model achieved the highest F1-Macro scores in both validation and testing sets among the evaluated models, with scores of 0.6604 and 0.6837, respectively. AraT5-Tweet-Base's ability to handle both common language forms in tweets allows it to better capture the diverse sentiment labels present in the dataset. This flexibility in processing both language forms likely contributed to its superior performance in F1-Macro compared to our IFDHN model.

The FNet model(Lee-Thorp et al., 2022) replaces the self-attention mechanism in Transformer encoders with unparameterized Fourier Transforms. In our ArSen dataset, the FNet model delivered average performance across various metrics.

The LLaMA-3 model (Dubey et al., 2024) is a decoder-only LLM with a 128K token vocabulary, optimized for efficient language encoding and pre-

trained on over 15 trillion tokens. This structure makes the LLaMA-3 model less suitable for our sentiment classification task. It features grouped query attention, offering strong performance across diverse NLP tasks. In our experiments, this model achieved the highest validation set F1-Micro score of 0.7428 without any fine-tuning. This result may be due to the relatively small scale of our benchmark.

The FDHN model (Xu and Kechadi, 2023, 2024), significantly contributed to the development of our IFDHN model. The FDHN model outperforms in all metrics while using fewer computational resources, which further motivated us to refine the model for our ASA task.

The IFDHN model outperformed all other models in accuracy and achieved the highest F1-Micro score on the test set. More importantly, we achieved comparable performance using just 0.23% of LLaMA-3's computational resources. As shown in Table 8, the IFDHN model has the lowest time cost, taking only 0.44 seconds. This outstanding result might be due to our multi-channel structure, which combines more information than just the tweet text, making it well-suited for the ArSen benchmark.

4.2 Ablation Experiment

To evaluate the impact of different features on the overall performance, we conducted a series of ablation experiments on the ArSen dataset. Table 6 summarizes the results.

Our ablation study included three sets of experiments: (1) evaluating each feature individually, (2) assessing the impact of excluding each feature one at a time, and (3) analyzing the model's performance with all features combined. This study provided critical insights into the role of various features in sentiment analysis for Arabic text. These experiments led to the following findings:

1. The tweet feature emerged as the most critical for accurate sentiment detection. It achieved the highest performance scores when used alone and caused the most significant performance drop when excluded. This underscores the importance of the tweet as the primary source of sentiment information.

2. The interaction metric was identified as the second most crucial feature. Although its standalone performance was similar to that of the meta-data, it yielded the highest scores when the meta-data feature was excluded, highlighting its value in sentiment detection.

3. The meta-data feature contributed significantly to the model's performance. Its inclusion improved the model's ability to generalize and provided context that complemented the tweet's content.

Feature	Validation			Test			Mean
	Accuracy	F1-Macro	F1-Micro	Accuracy	F1-Macro	F1-Micro	
Interacting metric	0.6850	0.2755	0.6862	0.7164	0.2830	0.7164	0.5604
Meta-data	0.6869	0.2715	0.6869	0.7164	0.2783	0.7164	0.5594
Tweet	**0.7390**	**0.4976**	**0.7319**	**0.7772**	**0.5655**	**0.7772**	**0.6814**
All without							
Tweet	0.6869	0.2715	0.6869	0.7164	0.2783	0.7164	0.5594
Interacting metric	0.7272	**0.4749**	0.7244	0.7713	0.5294	0.7713	0.6664
Meta-data	**0.7380**	0.4680	**0.7260**	**0.7753**	**0.5464**	**0.7753**	**0.6715**
All	**0.7478**	**0.5113**	**0.7368**	**0.7812**	**0.5583**	**0.7812**	**0.6861**

Table 6: Ablation Experiment Results on the ArSen dataset. The interaction metric includes numerical features of *like_count*, *quote_count*, *reply_count*, and *retweet_count*. The meta-data feature comprises *followers_count*, *following_count*, *tweet_count*, *listed_count*, and *user_verified*. In the first experiment, we individually tested our packed features. Next, we excluded one feature at a time. Finally, all features were included to observe their combined performance.

The ablation study highlights the importance of combining multiple features to improve the robustness and accuracy of Arabic sentiment analysis models. While tweet content is key, interaction metrics and metadata provide valuable context that enhances sentiment detection.

5 Conclusion

In this paper, we introduced a novel Arabic sentiment analysis benchmark focused on the COVID-19 pandemic and presented the IFDHN model, tailored specifically for sentiment analysis within this context. Our model demonstrated substantial performance improvements over other SOTA models. Compared to the large language model LLaMA-3-8B, our model achieved a 0.5% and 2.17% increase in accuracy on the validation and test sets, respectively, and a 2.17% increase in F1-Micro on the test set. More notably, the IFDHN model reduced processing time by approximately 422 times compared to LLaMA-3-8B, achieving a remarkable processing speed of just 0.44 seconds. This comprehensive evaluation highlights the IFDHN model's capability to effectively capture nuanced sentiments, making it a valuable tool for understanding public sentiment.

Limitations

While our IFDHN model shows substantial promise in Arabic sentiment analysis, several limitations must be noted. Our experiments were limited to a COVID-19-focused dataset, which may affect generalizability across other domains within Arabic sentiment analysis. The model's robustness in diverse contexts remains unexplored as it was not tested on established benchmarks like LABR (Aly and Atiya, 2013) and ASTD (Nabil et al., 2015). Additionally, despite the potential of LLMs like GPT-4 for nuanced language understanding (Guan and Greene, 2024a,b; Guan et al., 2024), their high resource demands, challenges in fuzzy classification, data contamination issues (Xu et al., 2024), and susceptibility to illusions (Schaeffer et al., 2023) precluded their inclusion in our study. Our benchmark, primarily sourced from Twitter, may not fully represent broader Arabic language use, potentially introducing platform-specific biases. Ethical considerations also arise in the use of this dataset, particularly regarding the potential for misuse in surveillance, censorship, or other harmful activities, underscoring the importance of adhering to strict ethical guidelines. Furthermore, the model does not address the complexities of Arabic dialects, which vary significantly in vocabulary and syntax. Future work should include comprehensive evaluations across diverse datasets, explore the integration of LLMs, and account for dialectal variations to enhance the accuracy and generalizability of Arabic sentiment analysis.

Acknowledgments

This research was supported by Science Foundation Ireland and Anhui Province university natural science research key project (Grant no.2023AH050333).

References

Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2018. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of Computational Science*, 27:386–393.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2021. Asad: A twitter-based benchmark arabic sentiment analysis dataset. *Preprint*, arXiv:2011.00578.

Manal Mostafa Ali. 2021. Arabic sentiment analysis about online learning to mitigate covid-19. *Journal of Intelligent Systems*, 30(1):524–540.

Mohamed Aly and Amir Atiya. 2013. LABR: A large scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, Sofia, Bulgaria. Association for Computational Linguistics.

Maryam Alzaid and Fethi Fkih. 2023. Sentiment analysis of students' feedback on e-learning using a hybrid fuzzy model. *Applied Sciences*, 13(23).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Ranganathan Chandrasekaran, Vikalp Mehta, Tejali Valkunde, and Evangelos Moustakas. 2020. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. *J Med Internet Res*, 22(10):e22624.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Rangan Das, Sagnik Sen, and Ujjwal Maulik. 2020. A survey on fuzzy deep neural networks. *ACM Comput. Surv.*, 53(3).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mazen El-Masri, Nabeela Berardinelli, and Hanady Ahmed. 2017. Successes and challenges of arabic sentiment analysis research: a literature review. *Social Network Analysis and Mining*, 7:22.

Yang Fang and Cheng Xu. 2024. Arsen-20: A new benchmark for arabic sentiment detection. In *5th Workshop on African Natural Language Processing*.

Gabriel Elías Chanchí Golondrino, Manuel Alejandro Ospina Alarcón, and Luz Marina Sierra Martínez. 2023. Determination of the satisfaction attribute in usability tests using sentiment analysis and fuzzy logic. *Int. J. Comput. Commun. Control*, 18.

Shuhao Guan and Derek Greene. 2024a. Advancing post-OCR correction: A comparative study of synthetic data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6036–6047, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shuhao Guan and Derek Greene. 2024b. Synthetically augmented self-supervised fine-tuning for diverse text ocr correction. In *27th European Conference on Artificial Intelligence*, ECAI 2024, Santiago de Compostela.

Shuhao Guan, Cheng Xu, Moule Lin, and Derek Greene. 2024. Effective synthetic data and test-time adaptation for OCR correction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Maha Heikal, Marwan Torki, and Nagwa El-Makky. 2018. Sentiment analysis of arabic tweets using deep learning. *Procedia Computer Science*, 142:114–122. Arabic Computational Linguistics.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338.

Dinh Tai Pham Huyen Trang Phan and Ngoc Thanh Nguyen. 2023. Fedn2: Fuzzy-enhanced deep neural networks for improvement of sentence-level sentiment analysis. *Cybernetics and Systems*, 0(0):1–17.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2022. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

May Oo Lwin, Jiahui Lu, Anita Sheldenkar, Peter Johannes Schulz, Wonsun Shin, Raj Gupta, and Yinping Yang. 2020. Global sentiments surrounding the covid-19 pandemic on twitter: Analysis of twitter trends. *JMIR Public Health Surveill*, 6(2):e19447.

Mounika Marreddy and Radhika Mamidi. 2023. Chapter 6 - learning sentiment analysis with word embeddings. In Dipankar Das, Anup Kumar Kolya, Abhishek Basu, and Soham Sarkar, editors, *Computational Intelligence Applications for Text and Sentiment Data Analysis*, Hybrid Computational Intelligence for Pattern Analysis and Understanding, pages 141–161. Academic Press.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Alexander Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2268–2273, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Bingli Sun, Xiao Song, Wenxin Li, Lu Liu, Guanghong Gong, and Yan Zhao. 2024. A user review data-driven supplier ranking model using aspect-based sentiment analysis and fuzzy theory. *Engineering Applications of Artificial Intelligence*, 127:107224.

Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. 2022. Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14:100157.

Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. 2022. Roberta-lstm: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525.

Dimple Tiwari and Bharti Nagpal. 2022. Keaht: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis. *New Gen. Comput.*, 40(4):1165–1202.

Srishti Vashishtha, Vedika Gupta, and Mamta Mittal. 2023. Sentiment analysis using fuzzy logic: A comprehensive literature review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

Cheng Xu and M-Tahar Kechadi. 2023. Fuzzy deep hybrid network for fake news detection. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT '23, page 118–125, New York, NY, USA. Association for Computing Machinery.

Cheng Xu and M-Tahar Kechadi. 2024. An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12:88006–88021.

Cheng Xu, Jing Wang, Tianlong Zheng, Yue Cao, and Fan Ye. 2022. Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Archives of Medical Science*, 18(5):1208–1220.

Cheng Xu and Nan Yan. 2023. AROT-COV23: A dataset of 500k original arabic tweets on COVID-19. In *4th Workshop on African Natural Language Processing*.

Nan Yan and Cheng Xu. 2024. Decolonizing african NLP: A survey on power dynamics and data colonialism in tech development. In *5th Workshop on African Natural Language Processing*.

L.A. Zadeh. 1996. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2):103–111.

A Component Analysis

In this section, we present a comprehensive component analysis of our proposed model for the ASA task. The performance of the IFDHN models is evaluated using various metrics, including accuracy, F1-macro, and F1-micro, on both validation and testing sets. Importantly, all training phases of the models are finished within 10 epochs.

Table 7 provides a summary of the performance of our models in both sets. This table includes three fundamental components: TextCNN (TC), CNNBiLSTM (CB), and Fuzzy (FZ).

In the first row of this table, we use only the TextCNN module to process our dataset. This module proved to be the most significant part of the IFDHN model, achieving high scores across all evaluation metrics, with the highest F1-Macro score on the testing set. Additionally, in the third row, when the TextCNN module is excluded, the F1-Macro score is the lowest. Moreover, when comparing row two with row four, adding the Fuzzy layer leads to improved performance across all metrics. If the Fuzzy layer is not employed, alternative methods such as a self-attention mechanism or a probabilistic approach like Bayesian Neural Networks may also be effective in handling uncertainty and enhancing model performance.

B Evaluation Metrics

For our experiments, we utilize Accuracy and F1-score to evaluate the performance of models. Specifically, due to the class imbalance of our dataset, we report F1-Macro and F1-Micro to capture the model's performance across all classes. These evaluation metrics are widely used in many research studies (Heikal et al., 2018; Al-Smadi et al., 2018).

Accuracy is the simplest and most intuitive performance metric. It is defined as the ratio of correctly predicted instances to the total number of instances in the dataset. The formula for Accuracy is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

F1-Score combines precision and recall into a single metric by taking their harmonic mean, providing a balance between the two. The formula for F1-Score is:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model	Validation			Test			Mean
	Accuracy	F1-Macro	F1-Micro	Accuracy	F1-Macro	F1-Micro	
TC	0.7429	0.5022	0.7342	0.7743	**0.5621**	0.7743	0.6817
TC + CB	0.7341	0.4946	0.7318	0.7782	0.5553	0.7782	0.6787
CB + FZ	0.6869	0.2715	0.6869	0.7164	0.2783	0.7164	0.5594
TC + CB + FZ	**0.7478**	**0.5113**	**0.7368**	**0.7812**	0.5583	**0.7812**	**0.6861**

Table 7: Performance comparison of different sub-models of the IFDHN on validation and testing sets.

where Precision is defined as $\frac{\text{TP}}{\text{TP+FP}}$ and Recall is defined as $\frac{\text{TP}}{\text{TP+FN}}$.

F1-Macro is an extension of the F1-Score for multi-class problems. It is calculated by first computing the F1-Score for each class independently and then averaging these scores. The formula for F1-Macro is:

$$\text{F1-Macro} = \frac{1}{N}\sum_{i=1}^{N}\text{F1-Score}_i$$

where N is the total number of classes, and F1-Score_i represents the F1-Score of the ith class. F1-Macro treats each class equally, which is beneficial when assess the model's performance across all classes without being biased by class size.

F1-Micro, on the other hand, aggregates the contributions of all classes to compute the precision and recall before calculating the F1-Score. Unlike F1-Macro, F1-Micro gives more weight to the classes with more instances. The formula for F1-Micro is:

$$\text{F1-Micro} = \frac{2 \times \text{TP}_{\text{sum}}}{2 \times \text{TP}_{\text{sum}} + \text{FP}_{\text{sum}} + \text{FN}_{\text{sum}}}$$

In this formula, TP_{sum}, FP_{sum}, and FN_{sum} are the sums of true positives, false positives, and false negatives across all classes, respectively.

By utilizing these metrics, particularly F1-Macro and F1-Micro, we gain a comprehensive understanding of our model performance, especially in the context of the class imbalance present in the ArSen dataset.

C Experimental Setup

The model was implemented using PyTorch[3], and the experiment was conducted on a NVIDIA RTX 4090 GPU. Building on this setup, we provide details in this section on the specific configuration of the model utilized in our experiments.

Firstly, in our IFDHN model, each module's output sequence length is configured to 6, with a dropout rate of 0.5 and an embedding dimension set to 128, utilizing zero-padding where necessary to maintain consistency.

The TextCNN module, responsible for processing tweet text, includes an embedding layer followed by three parallel CNN layers, which use kernel sizes of 3, 4, and 5, all with a depth of 128. Each CNN layer's output is subjected to MaxPooling to capture the most significant features. These pooled feature maps are then concatenated and fed into a linear layer with dropout to prevent overfitting.

The CNNBiLSTM module, which handles numerical context, starts with a linear layer incorporating dropout. This is followed by a CNN layer with 32 output channels and a kernel size of 1. The processed output is then fed into a three-layer BiLSTM network with dropout to capture temporal dependencies. Finally, a linear layer is applied to generate the module's output.

Secondly, to evaluate the performance of our IFDHN model, we compared it with some SOTA models, including RoBERTa, AraT5-Tweet-Base, and FNet, using HuggingFace implementations for sequence classification. Specifically, we employed the pre-training weights `roberta-base`[4], `AraT5-tweet-base` [5], and `fnet-base`[6], respectively. Additionally, for the FDHN model, we incorporated the *description* and *created_at* features as inputs to the text context module. All models were trained for 10 epochs, with other parameters set to their default values, and the time spent to train one epoch for all models is presented in Table 8. All the code results represent the best outcomes from a single execution, with a random seed set to 42.

Finally, we also included LLaMA-3 for testing to explore the performance of LLMs on

[3] https://pytorch.org/

[4] https://huggingface.co/FacebookAI/roberta-base

[5] https://huggingface.co/UBC-NLP/AraT5-base

[6] https://huggingface.co/google/fnet-base

this task. During the experiments with LLaMA-3, we also used its HuggingFace implementation (`Meta-Llama-3-8B-Instruct`[7]) and utilized LLM2Vec (BehnamGhader et al., 2024) for sentence embedding.

Model	Avg. Epoch Time	Val Best Epoch
LLaMA-3-8B	185.82s	-
RoBERTa	50.13s	5/10
AraT5-Tweet-Base	29.66s	3/10
FNet	23.70s	3/10
FDHN	0.47s	4/10
IFDHN	0.44s	4/10

Table 8: The average time spent by all models to train one epoch. The third column indicates the best epoch on the validation set, where the minimum loss value was achieved. There is no best Epoch number on the validation set since LLaMA-3 uses only the inference mode.

[7] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct