ARSEN-20: A NEW BENCHMARK FOR ARABIC SENTIMENT DETECTION

Yang Fang
School of Computer Science and Technology
Huaibei Normal University
Huaibei, China
20211209024@chnu.edu.cn

Cheng Xu
School of Computer Science
University College Dublin
Dublin, Ireland
cheng.xu1@ucdconnect.ie

1 INTRODUCTION

Sentiment analysis (SA), also referred to as opinion mining, stands as one of the most successful tasks within the realm of Natural Language Processing (NLP) (Marreddy & Mamidi, 2023; Hussein, 2018). It entails the detection, extraction, and classification of opinions and emotions expressed in textual input by users(Ravi & Ravi, 2015). Despite Arabic experiencing exponential growth in online users (Al-Ayyoub et al., 2019), the complexity of the language, its divergence from English language characteristics, and various cultural and historical factors have led to a scarcity of studies on Arabic sentiment analysis (ASA)(El-Masri et al., 2017).

Arabic can be categorized into three main types: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) (Ghallab et al., 2020). Among these, the majority of online users tend to utilize a combination of Modern Standard Arabic (MSA) and Dialectal Arabic (DA), presenting considerable challenges in Arabic text processing(Abdulla et al., 2013).

This study reveals a paucity of Arabic datasets dedicated to sentiment analysis, many of which are outdated(Alayba et al., 2018) , and most suffer from relatively small sizes (less than 20,000)(Alharbi et al., 2020) . For instance, datasets such as ArTwitter (Abdulla et al., 2013), introduced in 2013, ASTD (Nabil et al., 2015), proposed in 2015, and the Arabic Health Services (AHS) dataset (Alayba, 2019), released in 2019, are notable examples. Remarkably, Arabic sentiment analysis benchmark datasets related to COVID-19 are notably scarce. This dearth not only hinders ASA research but also fails to reflect the latest advancements in the field. To address these gaps, this paper proposes, ArSen-20[1], a COVID-19 themed Arabic benchmark dataset curated through manual annotation by our trained and professional annotation team. By doing so, we aim to mitigate the aforementioned issues and furnish ASA research with an updated benchmark dataset featuring more precise data annotations and a larger volume of data, thereby contributing to the ASA mission within the NLP domain.

The remainder of this paper is structured as follows: Section 2 delineates the pertinent details of the proposed ArSen-20 dataset, while Section 3 presents a summary of the article and outlines avenues for future research.

2 DATASET DESCRIPTION

The ArSen-20 dataset originates from the AROT-COV23 dataset, as introduced by (Xu & Yan, 2023). The AROT-COV23 dataset encompasses approximately 500,000 original Arabic COVID-19-related tweets spanning from January 2020 to January 2023. From this extensive corpus, we randomly selected 20,000 tweets and meticulously categorized them into three distinct classes: positive, neutral, and negative sentiments.

After classifying the sentiment information of each tweets, we counted the number of each category and divided the dataset into training set, verification set and test set in a ratio of 8:1:1. The specific statistical information is shown in Table 1. Furthermore, for comprehensive insights into the label distribution and dataset characteristics, we furnish the meanings of labels and details of the annotation process in Appendix A, with further elaboration provided in Appendix B.

[1]This dataset is available at https://github.com/123fangyang/ArSen-20.

Table 1: The ArSen-20 dataset statistics.

Statistics	Num
Training set size	16000
Validation set size	2000
Testing set size	2000
Neutral	17262
Positive	878
Negative	1860

It is our aspiration that the ArSen-20 dataset will serve as a valuable resource for the research community, facilitating deeper investigations into the emotional landscape of individuals in the Arab region during the COVID-19 pandemic. Through this endeavor, we endeavor to elevate the prominence of Arabic within the realm of NLP research.

3 CONCLUSION

This paper introduces the ArSen-20 dataset, a curated collection comprising 20,000 original Twitter data entries spanning from January 2020 to January 2023. Through meticulous manual labeling, we have effectively addressed prevalent issues within existing Arabic datasets, including staleness, limited volume, and suboptimal quality. By doing so, we have not only provided an updated, larger, and more accurate benchmark dataset for the field of ASA but also laid the groundwork for advancing research in this domain.

The significance of the ArSen-20 dataset transcends the confines of ASA, offering potential applicability in various other NLP tasks. For instance, its robustness and diversity make it suitable for tasks such as guiding marketing strategies, managing public opinion, and more. By providing this comprehensive resource, we endeavor to catalyze advancements not only in ASA but also in the broader spectrum of NLP research.

In conclusion, the ArSen-20 dataset represents a significant contribution to the field of Arabic sentiment analysis and NLP research at large. By leveraging this resource, researchers can unlock new insights, address pressing challenges, and propel the field towards new frontiers of discovery and impact.

REFERENCES

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pp. 1–6. IEEE, 2013.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N. Al-Kabi. A comprehensive survey of arabic sentiment analysis. *Information Processing Management*, 56(2): 320–342, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.07.006. URL https://www.sciencedirect.com/science/article/pii/S0306457316306689. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Abdulaziz Alayba. *Twitter Sentiment Analysis on Health Services in Arabic*. PhD thesis, Coventry University, 2019.

Abdulaziz M Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. Improving sentiment analysis in arabic using word representation. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pp. 13–18. IEEE, 2018.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. Asad: A twitter-based benchmark arabic sentiment analysis dataset. *arXiv preprint arXiv:2011.00578*, 2020.

Mazen El-Masri, Nabeela Altrabsheh, and Hanady Mansour. Successes and challenges of arabic sentiment analysis research: a literature review. *Social Network Analysis and Mining*, 7:1–22, 2017.

Abdullatif Ghallab, Abdulqader Mohsen, and Yousef Ali. Arabic sentiment analysis: A systematic literature review. *Applied Computational Intelligence and Soft Computing*, 2020:1–21, 2020.

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

Mounika Marreddy and Radhika Mamidi. Chapter 6 - learning sentiment analysis with word embeddings. In Dipankar Das, Anup Kumar Kolya, Abhishek Basu, and Soham Sarkar (eds.), *Computational Intelligence Applications for Text and Sentiment Data Analysis*, Hybrid Computational Intelligence for Pattern Analysis and Understanding, pp. 141–161. Academic Press, 2023. ISBN 978-0-323-90535-0. doi: https://doi.org/10.1016/B978-0-32-390535-0.00011-2. URL https://www.sciencedirect.com/science/article/pii/B9780323905350000112.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2515–2519, 2015.

Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2015.06.015. URL https://www.sciencedirect.com/science/article/pii/S0950705115002336.

Maite Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2: 325–347, 2016.

Cheng Xu and Nan Yan. AROT-COV23: A dataset of 500k original arabic tweets on COVID-19. In *4th Workshop on African Natural Language Processing*, 2023. URL https://openreview.net/forum?id=aUZhVQBl2W.

Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 129–136, 2003.

A Annotation Process

During the sentiment analysis (SA) task of tweet text, our annotation process categorizes tweets into three distinct classes: positive, neutral, and negative. Positive tweets typically exhibit emotional tendencies such as altruism, abundance, accolades, while neutral tweets may convey tones such as accentuation or alertness. Conversely, negative tweets often express sentiments like abolition, addiction, or alienation (Taboada, 2016).

Our researchers adhered to specific guidelines throughout the labeling process:

1. Classification of Tweet Types: Tweets categorized as "News" and "Business" are treated as factual statements, whereas those labeled as "Editorial" or "Letter to the Editor" are regarded as opinions(Yu & Hatzivassiloglou, 2003).
2. Objective Facts: Irrespective of whether they convey positive or negative news, factual statements are labeled as neutral.
3. Author's Perspective: Sentiment annotations should reflect the viewpoint of the tweet's author rather than that of the annotator.
4. Contextual Consideration: Label selection is contingent upon the context of the tweet's content.
5. Handling Mixed Content: Tweets containing a mixture of positive and negative words require careful consideration during labeling.
6. Emoticon Influence: Emoticons are considered as cues for determining tweet sentiment.

Table 2: Tweets field feature information.

Field	Type	Description
tweet id	string	The unique identifier of the requested Tweet.
label	string	Sentiment Classification of this tweet.
created_at	data	Creation time of the Tweet.
lang	string	Language of the Tweet,if detected by Twitter.
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.
name	string	The name of the user.
username	string	The Twitter screen name,handle,or alias.
user_created_at	data	The UTC datetime that the user account was created.
description	string	The text of this user's profile description(bio).

An exemplary tweet annotation is provided in Table 4. For tweets with mixed content, we employ a quantitative approach, tallying the number of positive and negative words present. The predominant sentiment is then determined based on this count, with tweets containing more positive words being labeled as positive, and vice versa.

B Dataset Details

By utilizing the Python 'random' module with a seed value of 42, we conducted random sampling on the AROT-COV23 dataset, resulting in the selection of 20,000 data points to constitute our dataset, named ArSen-20. The rationale behind this random sampling approach is to extract a subset from the entire dataset that effectively captures its overarching characteristics. This methodology not only reduces the dataset size to a manageable extent but also maintains its representativeness, thereby facilitating expedited analysis and processing speeds.

Our ArSen-20 dataset not only includes the textual content of original tweets from Twitter users but also integrates essential metadata related to both users and tweets. This comprehensive approach enriches the contextual understanding of Arabic sentiment analysis (ASA). Table 2 provides detailed information about the tweet field features. Through this table, users can easily discern the significance conveyed by each data field in our dataset, facilitating its utilization.

In the context of conducting ASA research using the ArSen-20 dataset, not all data fields may be utilized. Therefore, to streamline subsequent research endeavors, we have curated a selection of the most pertinent data fields that are likely to be employed. These selected fields are represented using functional symbols, as illustrated in Table 3. For instance, the field information can be categorized into three components: tweet text information f_1, context information$\{f_2, f_3\}$, and numerical context information$\{f_4, \ldots, f_{12}\}$. Researchers can then integrate these three components as inputs into their models, facilitating seamless experimentation and analysis.

The distribution of sentiment across the dataset is illustrated in Figure 1. As depicted, the majority of tweets exhibit a neutral sentiment (86.3%), followed by negative sentiment (9.3%), with positive sentiment being the least prevalent (4.4%).

Table 3: An example of ArSen-20 dataset fields that can be used in ASA.

#	Field	Value
-	tweet id	1268339838186594048
-	label	1
f_1	tweet	نمط الانتاج الزراعي الحالي والفقر ينتجان الشرو...
f_2	description	Neurologist.Tweets are notes ...
f_3	created_at	2020-06-04 00:32:36+00:00
f_4	like_count	1
f_5	quote_count	0
f_6	reply_count	0
f_7	retweet_count	0
f_8	followers_count	228
f_9	following_count	897
f_{10}	tweet_count	9029
f_{11}	listed_count	0
f_{12}	user_verified	False

Table 4: Labels used in annotation and examples of each.

Labels	Example in Arabic	English Translation
Positive	بيان بحمد من الله وفضله والشكر له ...	With the praise of God, His grace, thanks...
Neutral	وفاة رئيس الاتحاد البوليفي لكرة القدم ...	The President of the Bolivian Football...
Negative	يوجد العديد من الطالبات المصابةمخا...	There are many female students who are...

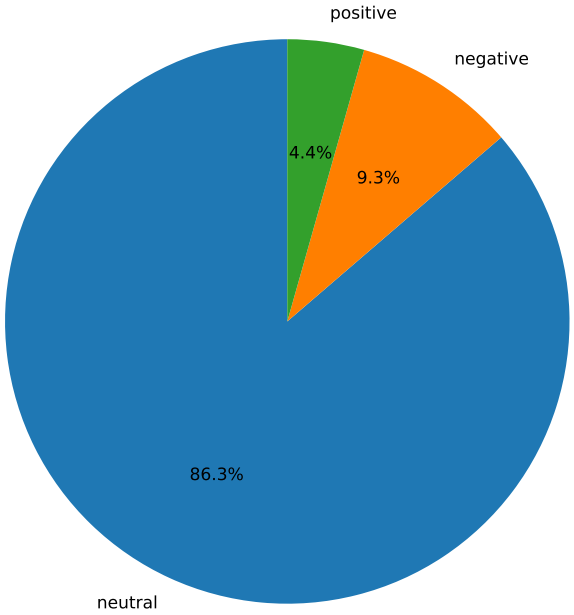

Figure 1: The proportion of different sentiment in ArSen-20 dataset.