https://jsras.rcc.edu.ly/
Vol.1 No.2 (2024), 30-37
Article history:
Received: 22 Aug. 2024
Accepted: 18 Sep. 2024
Published: 21 Sep. 2024

Journal of Sustainable Research in Applied Sciences

A review of the available Arabic dialects datasets for Sentiment Analysis

Abdullah Habberrih[1], Mustafa Ali Abuzaraida[2*]
[1]Misurata University, Faculty of Information Technology, Department of Computer Science, Libya

*** Corresponding author email address**: abuzaraida@it.misuratau.edu.ly

Abstract

In recent decades, the resources available for Arabic natural language processing have undergone a significant increase and development. This includes the exploration of Arabic Language Sentiment Analysis from Arabic utterances in both Modern Standard Arabic (MSA) and different Arabic dialects (DA). With the prevalence of internet usage among Arab people, communication in dialect languages has become common, and as such poses a challenge in analyzing sentiments due to the different dialects used across Arab countries. MSA is notable for its publicly available rich corpus of written resources such as news articles, books, and academic papers, whereas Arabic dialects lack such publicly available resources. Consequently, researchers have focused their investigations on DA rather than MSA since the majority of Arabic exchanges on social media are generated in local dialects. The objective of this study is to examine recent research endeavors that have made their datasets publicly available and to determine the frequently utilized resources and domains in the realm of sentiment analysis for Arabic dialects. The findings reveal that Twitter is the most commonly employed source for researchers to obtain their datasets, while politics, sports, and movies are the most frequently utilized domains for these datasets.

Keywords: Sentiment Analysis, Arabic Language, Arabic Dialects, Arabic Dialect Datasets, Public Arabic Datasets.

1. Introduction

The surge in social media usage as a platform for individuals to express their opinions on various products and topics [1] has created a demand for the analysis of textual data. To meet this demand, many researchers have employed Natural Language Processing (NLP) methods to investigate people's attitudes and opinions [2]. One particularly useful NLP method is Sentiment Analysis (SA). SA is a task that involves extracting the sentiment from a given text, which can be classified into three categories: positive, negative, and neutral. Some classification schemes also include strongly negative and strongly positive sentiments [1].

Sentiment analysis has become increasingly significant in various domains, particularly in the realms of business, marketing, and politics. This is due to its ability to provide a comprehensive understanding of individuals' preferences, dislikes, and attitudes towards products, concepts, and services, which is crucial in gauging public sentiment [26].

As noted by sources [3, 5], the Arabic language is presently ranked as the fifth most widely spoken language in the world, boasting over 350 million native speakers. Within the Arabic language, there exist three distinct variations, namely Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) [2]. CA, the language utilized in writing the Quran, is considered a classical form of Arabic. MSA, on the other hand, is commonly employed in political, journalistic, literary, and educational contexts. DA is an informal variant of Arabic that is used in daily communication, and it varies from country to country and even from city to city.

The majority of Arabic exchanges on social media are generated in local dialects, prompting many researchers to concentrate on dialectal Arabic DA rather than MSA. However, there is a dearth of tools and resources available for DA, as most existing resources focus solely on MSA. Consequently, numerous researchers have resorted to independently collecting data from platforms like Twitter to advance their research. This paper seeks to examine publicly available datasets for Arabic dialects and identify the most commonly used resources and domains [28].

On Twitter, a significant social media platform, a vast number of tweets rich in data are exchanged daily, including those in Arabic dialects. Notably by [28], in March 2014, active Arabic users posted over 17 million tweets per day. A vast quantity of tweets is produced every minute, with a significant portion of them being in Arabic.

2. Available Arabic dialect datasets

There are six prevalent types of Arabic Dialect, including Maghrebi (commonly spoken in northern Africa like Libya, Tunis, Morocco, and Algeria), Khaliji (spoken in the Arab Gulf area), Shami (Levantine) spoken in (Jordan, Lebanon, Palestine, and Syria), Egyptian (spoken in Egypt), Sudanese, and Iraqi, which are spoken in Sudan and Iraq [4], as shown in Figure 1.

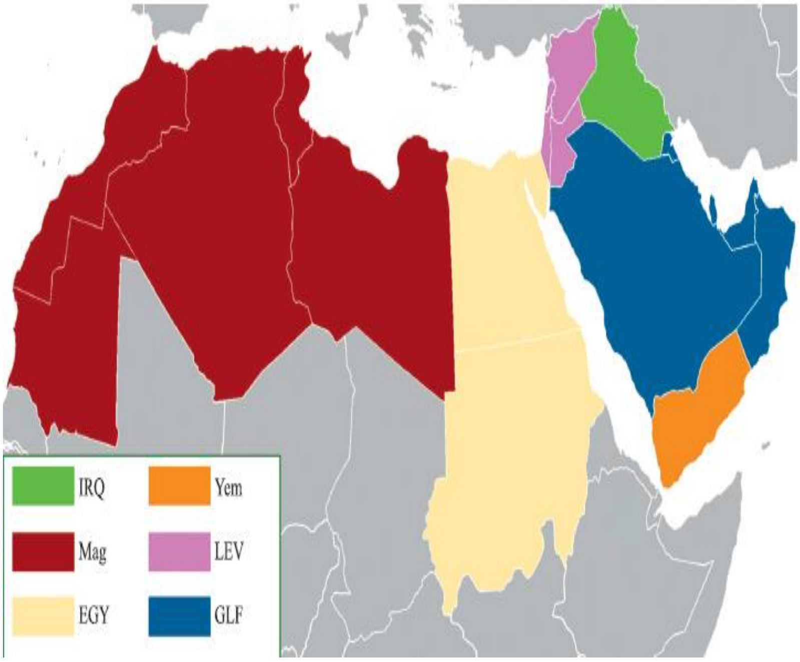

Fig. 1. Geographical Arabic Dialects Map [6].

As noted earlier, researchers from Arab countries tend to rely on social media platforms for collecting their datasets. According to [27], Twitter is the most commonly used platform for constructing datasets related to Arabic dialects. This observation is supported by the studies surveyed in this research, including [6, 7, 10, 12, 15, 16, 18, 19, 21, 22, 23, 24]. Additionally, Facebook and YouTube have also been utilized as resources for some researchers to collect their corpuses, as demonstrated in [8, 14, 17, 22, 25, 26]. Nevertheless, websites that feature reviews of restaurants, movies, hotels, and products may be preferred resources for researchers looking to accumulate their datasets, as seen in [9, 11, 13, 20, 25]. The availability of Arabic dialect datasets on the internet has had a significant impact on Arabic sentiment analysis research in recent years. See Figure 2 and 3.

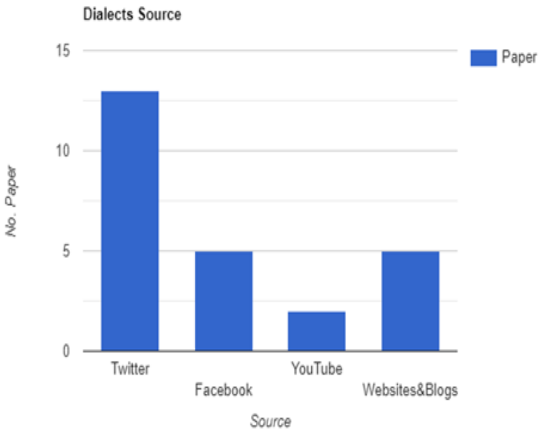

Fig. 2. Most Sources used in Arabic Dialects

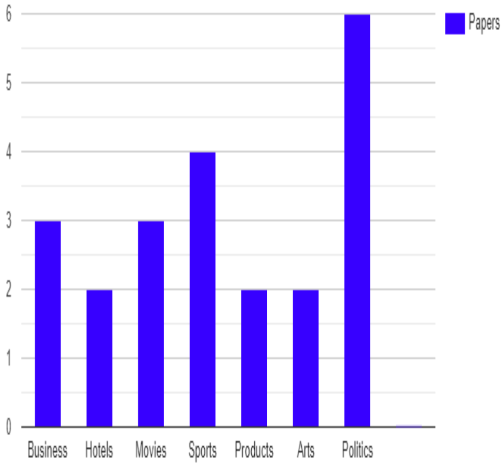

Fig. 3. Most Commonly Used Domains

Moreover, the Table1 below illustrates a selection of previous studies that have publicly shared their Arabic dialect datasets for the purpose of sentiment analysis.

The table includes: author's names, year of publication, size of datasets, dialect type, dataset source, and dataset domains. Note: (N/A: Not Available).

Table 1. Summary of the publicly available Arabic dialects datasets

Author's names, Year, Ref.	**Dataset size**	**Dialect Type**	**Source**	**Domain**
Al-Jawad et al, 2022, [6]	1170 tweets	Iraqi	Twitter	Politics & Influencers
Baly et al, 2017, [7]	14,400 tweets	Multi-dialects	Twitter	1. Business 2. Multimedia 3. Personal 4. Politics 5. Product Reviews 6. Religion 7. Sports 8. Check-ins
ElSahar and El-Beltagy, 2015, [9]	33,000 reviews	MSA and DA	Different Reviewing Websites	1. Movies 2. Hotels 3. Restaurants 4. Products
Rushdi-Saleh et al, 2011, [11]	500 reviews	MSA and DA	Different web pages & blogs	Movies
Nabil et al, 2015, [12]	10,000 tweets	Egyptian	Twitter	Hashtag: EgyptTrends
Fourati et al, 2020, [14]	Over 9,000 comments	Tunisian	YouTube	1. Sports 2. Politics 3. Comedy 4. TV shows, series, and arts 5. Tunisian music videos
Almuqren and Cristea, 2021, [15]	20,000 tweets	Saudi	Twitter	Different Saudi Telecom Companies
Omar et al, 2022, [16]	16,730 tweets	Libyan	Twitter	Different Libya telecom companies
Al-Twairesh et al, 2017, [21]	17,573 tweets	Saudi	Twitter	N/A
Jarrar et al, 2022, [22]	48,198 documents	Iraqi, Libyan, Sudanese, & Yemeni	Twitter, Facebook, and YouTube	N/A
Mekki et al, 2022, [8]	About 26000 comments	Tunisian	Facebook	COVID-19
Mdhaffaret al, 2017, [17]	17,000 comments	Tunisian	Facebook	Different Tunisian radios and TV channels pages

Bayazed et al, 2020, [23]	8923 tweets	Five Saudi Arabia sub-dialects	Twitter	N/A
Abdulla et al, 2013, [10]	2,000 tweets	MSA & Jordanian	Twitter	Politics & Arts
Aly and Atiya, 2013, [13]	Over 63,000 reviews	MSA & DA	GoodReads	Book reviews
Salameh et al, 2015, [24]	2,000 tweets	Syria	Twitter	N/A
Al-Moslmi et al., 2017, [25]	8861 reviews	Several DA	Jeeran & qaym websites, Twitter, Facebook, & Google Play	Multiple domains like: Hotels, Cafes, Shopping & Health Care
MIHI et al, 2020, [19]	35,000 tweets	Moroccan	Twitter	1. Sports 2. Arts 3. Politics 4. Education 5. other social issues
Garouani and Kharroubi, 2022, [18]	18,000 tweets	MSA & Moroccan	Twitter	N/A
Abdelli et al, 2019, [26]	49864 comments	Algerian	Facebook	N/A
Rahab et al, 2019, [20]	N/A	Algerian	three Algerian Arabic newspaper websites	1. News 2. Political 3. Religion 4. Society 5. Sports

3. Conclusion and Future Work

Sentiment analysis is a vital application of NLP that can provide valuable insights by analyzing textual data. Nevertheless, developing NLP models and tools for Arabic dialects remains challenging due to the limited written resources available for many of these dialects. Through this study, we have examined recent research endeavors for Arabic dialects that have made their datasets publicly available. The results reveal that most researchers collect their datasets from social media platforms such as Twitter, Facebook, and YouTube, with dataset sizes ranging from 1000 to 65000. Saudi Arabia and Tunisia have the most publicly available datasets, while Libyan, Syrian, and Yemeni dialects have fewer datasets. As future work, the authors aim to collect a large corpus of Libyan dialect and make it available to the research community.

References

[1] A. Abugharsa, “Sentiment Analysis in Poems in Misurata Sub-dialect: A Sentiment Detection in an Arabic Sub-dialect”, IJCT, vol. 21, pp. 103–114, Sep. 2021.

[2] S. Alyami, A. Alhothali, and A. Jamal, “Systematic literature review of arabic aspect-based sentiment analysis,” J. King Saud Univ. Inf. Sci., vol. 34, no. 9, pp. 6524–6551, 2022.

[3] Habberrih, A., Abuzaraida, M.A. (2024). Sentiment Analysis of Arabic Dialects: A Review Study. In: Zakaria, N.H., Mansor, N.S., Husni, H., Mohammed, F. (eds) Computing and Informatics. ICOCI 2023. Communications in Computer and Information Science, vol 2001. Springer, Singapore. https://doi.org/10.1007/978-981-99-9589-9_11.

[4] A. Elnagar, S. Yagi, A. B. Nassif, I. Shahin, and S. A. Salloum, “Sentiment analysis in dialectal Arabic: a systematic review,” Adv. Mach. Learn. Technol. Appl. Proc. AMLTA 2021, pp. 407–417, 2021.

[5] Habberrih, A., & Ali Abuzaraida, M. (2024). Sentiment Analysis of Libyan Dialect Using Machine Learning with Stemming and Stop-words Removal. 5th International Conference On Communication Engineering And Computer Science (Cic-Cocos’24), 259–264. https://doi.org/10.24086/cocos2024/paper.1171

[6] H. A.-J. MM, H. Alharbi, A. F. Almukhtar, and A. A. Alnawas, “Constructing Twitter Corpus Of Iraqi Arabic Dialect (CIAD) For Sentiment Analysis,” Научно-технический вестник информационных технологий, механики и оптики, vol. 22, no. 2, pp. 308–316, 2022.

[7] R. Baly et al., “Comparative evaluation of sentiment analysis methods across Arabic dialects,” Procedia Comput. Sci., vol. 117, pp. 266–273, 2017.

[8] A. Mekki, I. Zribi, M. Ellouze, and L. H. Belguith, “A Tunisian benchmark social media data set for COVID-19 sentiment analysis and sarcasm detection,” 2022.

[9] H. ElSahar and S. R. El-Beltagy, “Building large arabic multi-domain resources for sentiment analysis,” in Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16, 2015, pp. 23–34.

[10] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, “Arabic sentiment analysis: Lexicon-based and corpus-based,” in 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), 2013, pp. 1–6.

[11] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, “OCA: Opinion corpus for Arabic,” J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 10, pp. 2045–2054, 2011.

[12] M. Nabil, M. Aly, and A. Atiya, “Astd: Arabic sentiment tweets dataset,” in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2515–2519.

[13] M. Aly and A. Atiya, "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), 2013, pp. 494–498.

[14] C. Fourati, A. Messaoudi, and H. Haddad, "TUNIZI: a Tunisian Arabizi sentiment analysis Dataset," arXiv Prepr. arXiv2004.14303, 2020.

[15] L. Almuqren and A. Cristea, "AraCust: a Saudi Telecom Tweets corpus for sentiment analysis," PeerJ Comput. Sci., vol. 7, p. e510, 2021.

[16] A. Omar, M. Essgaer, and K. M. S. Ahmed, "Using Machine Learning Model To Predict Libyan Telecom Company Customer Satisfaction," in 2022 International Conference on Engineering & MIS (ICEMIS), 2022, pp. 1–6.

[17] S. Mdhaffar, F. Bougares, Y. Esteve, and L. Hadrich-Belguith, "Sentiment analysis of tunisian dialects: Linguistic ressources and experiments," in Third Arabic Natural Language Processing Workshop (WANLP), 2017, pp. 55–61.

[18] M. Garouani and J. Kharroubi, "MAC: an open and free Moroccan Arabic Corpus for sentiment analysis," in Innovations in Smart Cities Applications Volume 5: The Proceedings of the 6th International Conference on Smart City Applications, 2022, pp. 849–858.

[19] S. Mihi, B. Ait, I. El, S. Arezki, and N. Laachfoubi, "MSTD: Moroccan sentiment twitter dataset," Int. J. Adv. Comput. Sci. Appl, vol. 11, no. 10, pp. 363–372, 2020.

[20] H. Rahab, A. Zitouni, and M. Djoudi, "SANA: Sentiment analysis on newspapers comments in Algeria," J. King Saud Univ. Inf. Sci., vol. 33, no. 7, pp. 899–907, 2021.

[21] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets," Procedia Comput. Sci., vol. 117, pp. 63–72, 2017.

[22] M. Jarrar, F. A. Zaraket, T. Hammouda, D. M. Alavi, and M. Waahlisch, "Lisan: Yemenu, Irqi, Libyan, and Sudanese Arabic Dialect Copora with Morphological Annotations," arXiv Prepr. arXiv2212.06468, 2022.

[23] A. Bayazed, O. Torabah, R. AlSulami, D. Alahmadi, A. Babour, and K. Saeedi, "SDCT: multi-dialects corpus classification for Saudi Tweets," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, 2020.

[24] M. Salameh, S. M. Mohammad, and S. Kiritchenko, "Sentiment after translation: A case-study on arabic social media posts.," in HLT-NAACL, 2015, pp. 767–777.

[25] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis," J. Inf. Sci., vol. 44, no. 3, pp. 345–362, 2018.

[26] A. Abdelli, F. Guerrouf, O. Tibermacine, and B. Abdelli, “Sentiment analysis of Arabic Algerian dialect using a supervised method,” in 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), 2019, pp. 1–6.

[27] A. A. Al Shamsi and S. Abdallah, “A Systematic Review for Sentiment Analysis of Arabic Dialect Texts Researches,” in Proceedings of International Conference on Emerging Technologies and Intelligent Systems: ICETIS 2021 Volume 2, 2022, pp. 291–309.

[28] Abdullah Habberrih and Mustafa Ali Abuzaraida. "Sentiment Analysis of Libyan Middle Region Using Machine Learning with TF-IDF and N-grams." International Conference for Information and Communication Technologies. Cham: Springer Nature Switzerland, 2023.