

Recent developments in information extraction approaches from Arabic tweets on social networking sites



Abdullah Ibrahim Abdullah Alzahrani ^{1,*}, Syed Zohaib Javaid Zaidi ²

¹Department of Computer Science, College of Science and Humanities, Al-Quwayyah, Shaqra University, Shaqraa, Saudi Arabia

²Institute of Chemical Engineering and Technology, University of the Punjab, Lahore, Pakistan

ARTICLE INFO

Article history:

Received 5 February 2022

Received in revised form

5 May 2022

Accepted 18 June 2022

Keywords:

Natural language processing

Naïve Bayes

K-NN

Support vector machines

ABSTRACT

Information extraction from Arabic tweets has attracted the attention of researchers due to the huge data accessibility for the swift expansion of social media platforms. With the increasing use of social web applications, information extraction from the various platforms has gained importance for understanding the trending post and events predictions based on those sentiments written by the users on certain news feeds. The Arabic Language is mostly used in Middle Eastern and African countries and most users tweet on social media using the Arabic language, therefore Arabic text classification and sentiment analysis aimed to predict information extraction from social media platforms. This research provides a more detailed critical review of the information extraction presented in the literature focused on using different tools, methods, and techniques like k-NN, support vector machines, Naïve Bayes, and other machine learning tools for the data extraction and processing.

© 2022 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

This century creates different channels of communication out of which social media is the most widely known platform for information sharing, discussion of views, content description, and marketing of products and goods. An important media of communication, the shared posts and information may be offensive, unsuitable, rude, or foul language (Asur and Huberman, 2010). Detection of rude and unsuitable content on social networks can be helpful for a number of important applications. For instance, the community using social media might be of views on separating out indecent or unacceptable, or hateful content from their media profile or separating out such content for their underage family members. Moreover, filtering disrespectful or indecent messages or conversations on social media platforms may denote the discussion of vehement/argumentative subjects/content or the presence of disrespectful speech that may be resulted in or contribute to crimes and anarchy in society and on social media


platforms (Abo et al., 2019; Yue et al., 2019). Social web networking sites for instance Twitter allow profiles to separate out depending on a user's given a list of specific keywords. Likewise, famous websites, for instance, Google and Yahoo, and other important event coverage webpages, like YouTube, got frames for restricted modes that separate out immodest and smut contents (Ahmed et al., 2015; Alhumoud et al., 2015). One way to filter out such desirable content is to maintain a list of obscene words to filter content against immodest content. However, the manual construction and maintenance of such lists are arduous. A specialized approach is required to understand the complexity of detection of disrespectful messages which requires architecture for the categorization of tweets and posts on social media (Abdullah and Hadzikadic, 2017). There are different important approaches for obtaining useful information from different posts. One of the techniques is sentiment analysis, which has been found to be effective for acquiring insights from a large number of posts and information shared by targeted users (Balahur, 2013; Yu et al., 2013).

The classification of these posts is based on different opinions like constructive, destructive, or impartial categories. Now depending on the Arabic language classification, two important categories are divided in literature i.e. machine learning and semantic methods (Al-Ayyoub et al., 2018; Joulin et al., 2016). One reasonable approach for obtaining information from tweets provided by multiple users

* Corresponding Author.

Email Address: a.alzahrani@su.edu.sa (A. I. A. Alzahrani)

<https://doi.org/10.21833/ijaas.2022.09.018>

 Corresponding author's ORCID profile:

<https://orcid.org/0000-0002-4718-7568>

2313-626X/© 2022 The Authors. Published by IASE.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

is the sentiment analysis method. The classification of people's opinions and sentiments can be divided into optimistic, pessimistic, or impartial categories. Particularly there are two important approaches for the classification of Arabic text. One is machine learning and the other is semantic methods. In these methods, machine learning models are trained by using data sets obtained from different platforms e.g. Facebook, Instagram, and Twitter (Al-Radaideh, 2020). Further, the model is employed to forecast information from selected tweets. In semantic methods sentiments, and dialect of the Arabic language are transformed. Further dialects are provided with levels of optimism and pessimism. These levels of dialects specified their classification based on optimism, pessimism, and impartiality. Any sentence indicated in these dialects is dependent on the sum of levels of all words used in the sentences or tweets (Al-Twairish et al., 2014; Badaro et al., 2019). During the previous 5 years, these two techniques have widely attracted the attention of research scientists. However, very limited attention is given to an important language of the Middle East and Africa i.e. Arabic language. According to an important report, there are about 350 million peers who spoke Arabic as their mother and official language (Boukil et al., 2018; Kaseb and Ahmed, 2016). Sentiment analysis is reported as an important tool to classify sentiments in text and further categorized it into different categories based on pessimism, optimism, and impartiality. However, it is observed that these categories are not valuable in decision-making (Baier et al., 2019; Elhassan and Ahmed, 2015; Zhang et al., 2011) and provide enigmatic results.

An important exploration in this regard is an analysis of tweets which is unclear to define an emotion of the person whether it reflects positivity or negativity. Tweets can have both optimistic and pessimistic emotions based on trending posts. Another issue of using sentiment analysis is to detect trenchancy in tweets, which has been identified as a critical research problem in language analysis (Castillo et al., 2011; Elhassan and Ahmed, 2015; Habash and Sadat, 2006). The level of tweets is also important in the classification of trending posts. A common tweet can be posted by many individuals on a single post but with different degrees of impact. Hence, the identification of tweets is dependent on the intensity of the words used in the tweets that can help to analyze an individual's intention about the trending post (Aggarwal and Zhai, 2012). The level of tweets is also important in the classification of trending posts. A common tweet can be posted by many individuals on a single post but with different degrees of impact. Hence, the identification of tweets is dependent on the intensity of the words used in the tweets that can help to analyze an individual's intention about the trending post (Janasik et al., 2009).

In this research work, we will critically review the studies pertaining to Arabic language detection tools to identify detailed analysis and hate speech in text

processing. Our objective is to analyze tools and techniques which has been reported in different literature research work. We will identify different machine learning algorithms for Arabic tweet processing which involves classifiers such as supervised learning, semi-supervised learning, hybrid methods dialects-based techniques, and unsupervised learning which identify the best methods for the information extraction from social media platforms.

2. Recent trends in information mining

Data obtained from social media like Twitter require a specialized approach for the identification sequence in data. The main reason is morphology volume and the nature of the pattern in the text. The main focus of research is the identification of the sequence of linguistics behavior in social speech. The sequence may depend upon the classification of comments as pessimistic, optimistic, or impartial sequences (Dalal and Zaveri, 2011). The common path for the assessment of sequence is initiated with data collection from social media by using programming interface hour, and web application. Further, propagation and transformation to compile a data set by encoding data mining and language processing methods. Finally, the data sets are terminated by a specialized algorithm to further interpreted the performance of the algorithm (Almuqren et al., 2017).

In order to understand the scheme for pattern identification analysis of linguistic sequence involving language detection and argument behavior and NLP techniques required to extract from literature.

3. Analysis of linguistic sequence

Various language processing tools are utilized to identify examine and derive linguistic sequences which are reported in up-to-date literature. Different linguistic patterns use machine learning and or classifier approach for language detection (Husain, 2020). They use a classifier like support vector machine Naïve Bayes and passive regression (Akaichi et al., 2013) as represented in Table 1. Also, there are other approaches that have utilized argument derivation and facts detection. Few approaches have been conducted to users' identification and attributions while using social networks. Table 1 shows the recent linguistic sequence of updated literature review related to text detection on social media.

4. Linguistic classification for the Arabic language

Arabic text classification has been studied by several researchers using different methods like regression, and Naïve Bayes. Support vector machine while identifying the most informative features of

social media text. For example, the research work utilized Naïve Bayes, a corpus-based strategy for the extraction of features from social media tweets written in the Arabic language (Alomari et al., 2017).

The classification of tweets was done by using discriminative assessment in three categories further data sets were trained using the Naïve Bayes model.

Table 1: Literature review (analysis of linguistic sequence) data sets obtained from social media, pre-processing methods applied, techniques used, and percentage accuracy of techniques

Research Article	Social Media Platform	Dialect	Data	Methods	Techniques	Accuracy
(Atoum and Nouman, 2019)	Twitter	Jordanian Arabic	2500 tweets	normalization, tokenization, encoding, annotation	Support vector machine and Naïve Bayes	82 %
(Al-Horaibi and Khan, 2016)	Twitter	Syrian Arabic	2000 tweets	Feature extraction, sentiment analysis, words representation	Naïve Bayes decision tree	65%
(Jardaneh et al., 2019)	Twitter	Jordanian Arabic	2300	Sentiment analysis features detection	Logistic regression, Random Forest, AdaBoost, Decision Tree	76%
(Alsanad, 2018)	Twitter	Saudi Arabic	2000 tweets	Stemming, tokenization, n-grams	Naïve Bayes	67 %
(Duwairi et al., 2015)	Twitter	Jordanian Arabic	2500 + tweets	Sentiment Analysis, Crowd Sourcing	Naïve Bayes, k-NN, SVM	69 %
(Salamah and Elkhilfi, 2014)	Twitter	Kuwaiti Arabic	340,000 tweets	Tokenization, segmentation	SVM, Naïve Bayes	76 %
(Ismail et al., 2018)	Twitter	Sudanese Arabic	4712 tweets	Labeling, Annotation	SVM, NB, k-NN	92 %
(Al-Osaimi and Badruddin, 2014)	Twitter	Saudi Arabic	1700 tweets	Sentiment Analysis, Text Mining, Data Structuring	NB, Naïve Bayes, Decision Tree	63 % max

The data was obtained from public social media accounts using sentiments analysis to the experimented results were validated by employing performance assessment tools (Alotaibi et al., 2019). The categories of Arabic tweets were distinguished into pessimistic and optimistic categories which consist of about more than 1500 Arabic tweets. There was a certain classifier that was based on machine learning tools. The optimal results were obtained by the VEKA machine learning classifier which showed maximum efficiency by improving the overall accuracy by 0.3% (El-Halees, 2008). Additionally, few studies discovered machine learning approaches for the purpose of sentiment

investigation of Arabic text more than 2500 tweets returned in the Arabic language were compiled by employing crowdsourcing and labeled into data sets. Further, the data was trained by using Naïve Bayes, a support vector machine, and K nearest neighbor classifier. Further, the data was validated and processed for the removal of unwanted words the process is depicted in Fig. 1. In this study the best classifier proved the accuracy of nearly 70% under the Naïve Bayes classifier. A bag of words was used with k-NN nearest which showed better performance than support vector machine classifiers (Duwairi and El-Orfali, 2014).

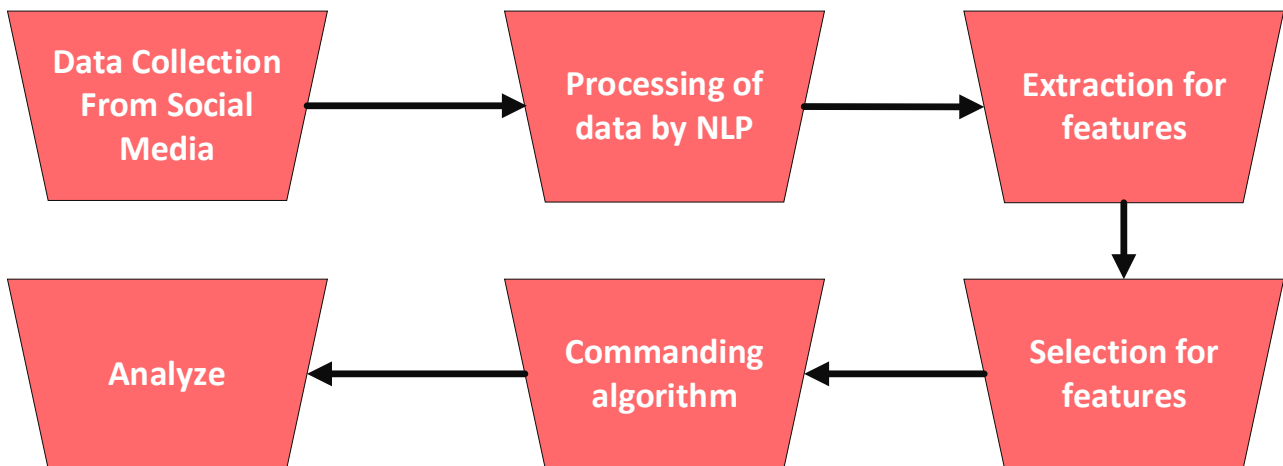


Fig. 1: Critical model of linguistic classification

5. Natural language processing (NLP) classification of techniques for Arabic data

Abdullah and Hadzikadic (2017) created data sets obtained from social media sites from Arabic tweets. They used sentiment analysis and annotated the Arabic text for data pre-processing. Further NLP

classifications were used for the processing of Arabic data which clarify different aspects of Arabic dialects in detail. Some studies also discussed the mining of Arabic sentiments using NLP. Those studies discussed systematic mining systems which involve important approaches advanced methods and implement sentiment analysis (Abdullah and

Hadzikadic, 2017). For example, Badaro et al. (2019) discussed the certain type of symbols code words hyperlinks, and punctuation marks for the purpose of annotation of special tags for specific dialects in social media sites which also mentioned hashtags, emojis, and hyperlinks. Similarly, Arabic grammar and a set of different dialects which summarized with employed for the annotation which linked with the traits of parts of speech on Arabic varieties like

standard Arabic, Romans script Arabic or classical Arabic (Alshargi et al., 2019). NLP task for the Arabic language has been used for various Arabic corpora like Tunisian Arabic (Guellil et al., 2021), Quranic Arabic (Dukes and Habash, 2010), and Arabic corpus (Traboulsi, 2009).

Fig. 2 shows common natural language processing tasks for data from social networking media.

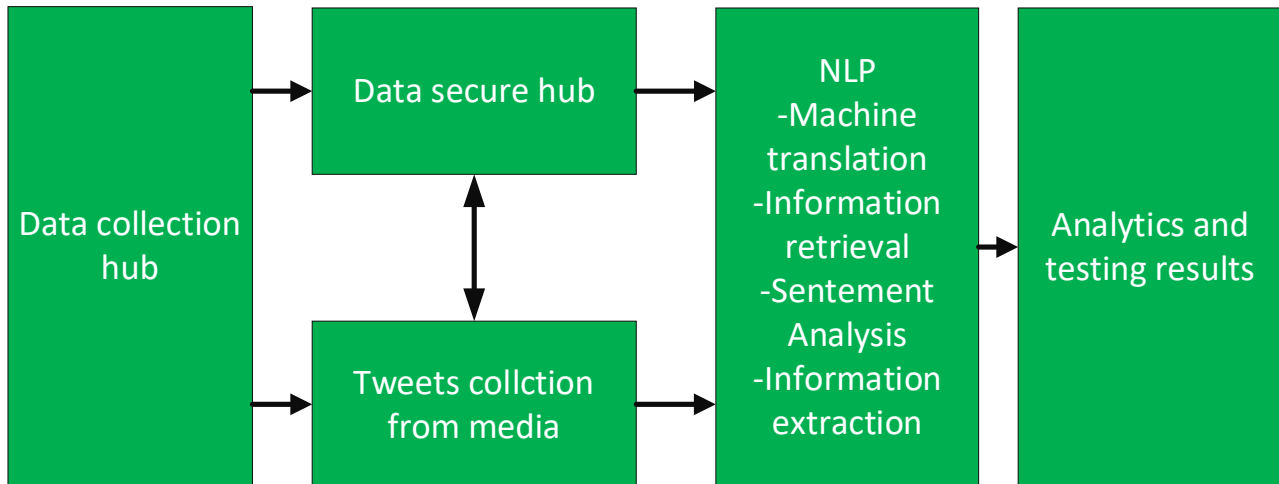


Fig. 2: Common natural language processing tasks for data from social networking media

The NLP for these categories provides users with the option for searching words transliteration stem and exact phrase seeking as shown in Fig. 2. Further mapping of words was done by sentiment classification and named entity transformation which further proposed annotated tags sets to classify specific words of users wish. This will help researchers in pre-processing the task in ambiguous circumstances which provide assistance with the separation of important information from different idioms, oral speech, dialects, and glowing of multiple words from fake accounts than in those from public stability accounts. Additionally, experiments were performed for different data sets, and the prediction of fake information from different data sets was identified. Particularly the tweets with similar hashtags, emojis, and suffixes at the beginning and the conclusion of a tweet are difficult for the accuracy and precision of training sets (Castillo et al., 2011; Habash and Sadat, 2006).

6. Linguistic dialect and Corpora

In the field of linguistic corpora various linguistic dialects and Arabic corpora or reported like Quranic Arabic corpus (Dukes and Habash, 2010), Tunisian Arabic corpus (McNeil, 2018), and Arabic corpus (Traboulsi, 2009). All these corpora have different dialects and linguistic patterns, for instance, the Quranic corpus is an annotated structure that involves each pattern of the Holy Quran with its syntax, tree bank, semantic morphology, and syntax. The source code for the data collection and testing of corpus data consist of the number of sentences manually annotated and can be further preprocessed

for keywords, punctuation, and Unicode character which provides an ability to look for structures and grammatical patterns in Quranic Arabic (Dukes and Habash, 2010).

Similarly, Tunisian corpus is a freeware that is available in 17 categories having about 0.9 million words. The 17 important categories involve Internet platforms, blogs, cell phone conversations, jokes, and many different forums. The searching option is stretched based on three categories like regex, stem words, and exact pattern (McNeil, 2018). Similarly, Arabic corpus is an annotated corpus having data sets with frequent words, grammatical patterns, and searching option for root words as shown in Fig. 3. Moreover, Arabic corpora are expected to be available in near future and can be used for information extraction task from different social media platforms (Alsaedi and Burnap, 2015).

7. Argument detection

Argument detection is required extensively for the NLP task to make develop restructuring and correction of data to be evaluated in the learning models (Abend et al., 2009). The examination of collected data is done by retrieval of data insides from social text tweeted in the Arabic language (Comunello and Anzera, 2012). In the process of argument detection, the collection of documents forming a corpus involving sentences is collected from social media of users' choice followed by annotation and representation of entities representing claims or segments in favor and against the argument (Al-Laith and Shahbaz, 2021). Further, tokenization, splitting of information, tagging of root

words, and generation of vectors were done for the structuring of data (Abdul-Mageed et al., 2014). Further, unnecessary words like weblinks, special characters, stock words, and wide spaces are removed then the tags were associated with

argumentative sentences, cue words indicated words, and adjectives (Itani, 2018). Further machine learning classifiers are used to process the data (Abudalfa and Ahmed, 2017).

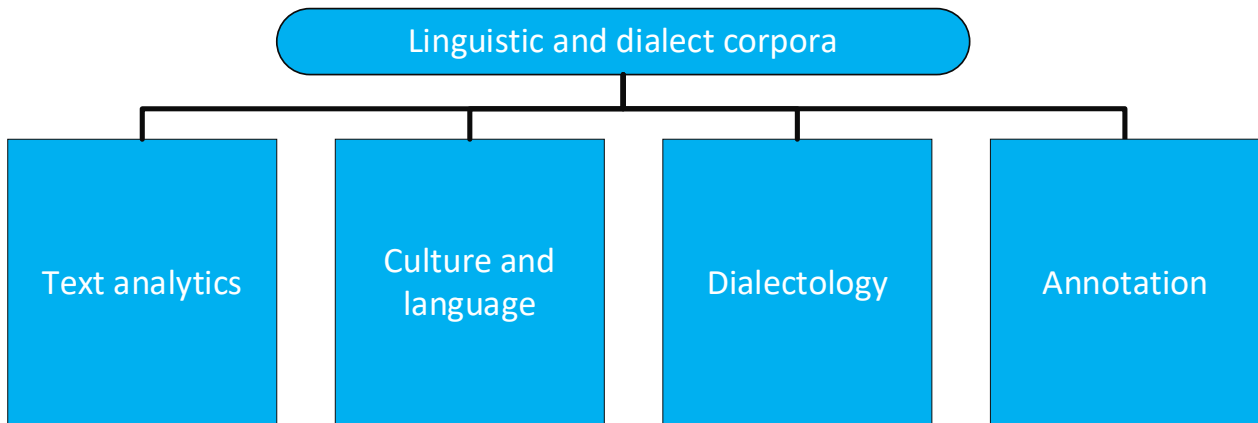


Fig. 3: Elements of corpora of linguistic and information characterization task

8. Offensive language detection

Offensive language has been termed as hostile thoughts, insults, obscenity, and curses that emerged from offensive behavior which hurt the emotions of people living in a society (Li, 2020). The emergence of new communication platforms like instant messaging e.g. Instagram, Facebook, YouTube, and Twitter arise a problem related to the negative impacts of offensive language during virtual communication (KhosraviNik and Esposito, 2018). Alakrot presented informative aspects of modeling for the direction of hate speech during online communication in the Arabic language at the first place the data was collected at a large scale from virtual social media comments in the Arabic language having offensive and flaming words and comments on the vlogs (Alakrot, 2019). The data sets were pre-processed using tokenization, normalization, and filtering. Further, the data sets were trained by using SVM classifiers and stimulated by N-gram and word-level features. They reported a precision of 90% for the detection of offensive language on Arabic media platforms. Several researchers worked on the detection of abusive comments on social media. They used different classification approaches like decision trees, log odds, and ratio accounts for the detection of hateful and offensive information on Arabic social media (Guellil et al., 2020).

9. Discussion

Recent developments on information extraction for the social web data sets on Arabic text mining has been worked on in previous studies (Abdulla et al., 2013). Alsaleem (2011) used SVM and NB for the detection of text categorization from Arabic newspapers. They classify information into different categories which involve culture, politics, economics, information technology, sports, and current affairs.

They reported that the accuracy obtained from the support vector machine algorithm was found to be more improved than the NB algorithm. The maximum precision of 96% was obtained by using the SVM classifier (Alsaleem, 2011). There are few other studies that were particularly focused on dialect-based methods for the text classification of the Arabic language. One of the studies was focused on it information retrieval from Twitter. They used predicate calculus for the classification of tweets (Al-Smadi et al., 2019). They reported an accuracy of 86% but the overall simulation time was very long. Similarly, Mataoui et al. (2016) studied the data sets from Algerian Arabic corpus. They reported an efficiency of 78% using three different lexicons. Abdulla et al. (2014) employed a dictionary-based technique while taking data from Twitter, Yahoo, and Maktoob. The reported accuracy was 70.5% for Twitter data. Data from Yahoo presented lower accuracy of 63% (Abdulla et al., 2014). Some of the studies reported lower accuracies due to the use of corpus-based features on various datasets (Badaro et al., 2015; Mohammad et al., 2016) as represented in Table 2. We may estimate from the above information that corpus-linked methods are not so good in terms of accuracy, however, dictionary-based approaches perform with improved accuracy in comparison to lexicon-based approaches.

10. Conclusion

Advancement of communication technologies in recent times increased caused information processing from Arabic dialect and extremely complex issues. This information from social media tweets has valuable information for decision-makers, especially in economic politics, media observers, and government agencies. This work reviewed based studies related to the Arabic text information classification using different machine learning techniques. Further, the accuracy and contribution of

various methods in Arabic classification were identified. SVM and Naïve Bayes are found to be reliable approaches for information extraction from social media text. Data mining techniques like NLP also quite interesting for the issues relating to text

mining from political sentiments. This paper will assist the obviously to research community for the employment of correct techniques for Arabic language information extraction.

Table 2: Literature review (information extraction) machine learning and other tools, result accuracy of techniques used, contribution, and miscellaneous open issues

Author	Contribution	Open issues	Tools	Results	Algorithms
(Duwairi et al., 2015)	Validation of Arabic data sets at multiple events tagging of multiple dialects	Multiple classifiers were used for small data sets accuracy is inefficient, mapping of controversial events	Sentiment analysis dialects	Accuracy with 75 % efficiency	Support vector machine, Naive Bayes, k-NN
(Abdul-Mageed et al., 2014)	Argument extraction by using determinant features, specific corpus were used for social text contribution	Normalization of unsampled data only one algorithm was used with limited features	Corpus for text mining lexicon based classifier	69 %	SVM
(Harrag et al., 2009)	Data sets with unique features were used bilingual dialects provide informative data sets	The training time was much higher for segregated data sets	Hadith corpus lexicon	93 %	Decision Tree
(Saad and Ashour, 2010)	Different methods were used for pre-processing of data gathered from specialized category	Data sets were limited with single classification of text	Sentiment analysis, annotation with distributed representation	94 %	Decision tree
(Shoukry and Rafea, 2012)	Unigrams and bigrams were employed for negation and switching phrases	Accuracy was very limited dual opinion arises during testing of tweets which gives arise ambiguous sentiment	Sentiment Analysis, switching phrases, 10 fold validation	0.721	Naïve Bayes, SVM
(Al-Osaimi and Badruddin, 2014)	Twitter	Saudi Arabic	1700 tweets	Sentiment Analysis, Text Mining, Data Structuring	NB, Naïve Bayes, Decision Tree

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abdulla NA, Ahmed NA, Shehab MA, Al-Ayyoub M, Al-Kabi MN, and Al-rifai S (2014). Towards improving the lexicon-based approach for Arabic sentiment analysis. *International Journal of Information Technology and Web Engineering*, 9(3): 55-71. <https://doi.org/10.4018/ijitwe.2014070104>

Abdulla NA, Ahmed NA, Shehab MA, and Al-Ayyoub M (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. In the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, IEEE, Amman, Jordan: 1-6. <https://doi.org/10.1109/AEECT.2013.6716448>

Abdullah M and Hadzikadic M (2017). Sentiment analysis on Arabic tweets: Challenges to dissecting the language. In the International Conference on Social Computing and Social Media, Springer, Vancouver, Canada: 191-202. https://doi.org/10.1007/978-3-319-58562-8_15

Abdul-Mageed M, Diab M, and Kübler S (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech and Language*, 28(1): 20-37. <https://doi.org/10.1016/j.csl.2013.03.001>

Abend O, Reichart R, and Rappoport A (2009). Unsupervised argument identification for semantic role labeling. In the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, 1: 28-36. <https://doi.org/10.3115/1687878.1687884>

Abo MEM, Raj RG, and Qazi A (2019). A review on Arabic sentiment analysis: State-of-the-art, taxonomy and open research challenges. *IEEE Access*, 7: 162008-162024. <https://doi.org/10.1109/ACCESS.2019.2951530>

Abudalfa S and Ahmed M (2017). Survey on target dependent sentiment analysis of micro-blogs in social media. In the 9th IEEE-GCC Conference and Exhibition (GCCCE), IEEE, Manama, Bahrain. <https://doi.org/10.1109/IEEEGCC.2017.8448158>

Aggarwal CC and Zhai C (2012). A survey of text classification algorithms. In: Aggarwal C and Zhai C (Eds.), *Mining text data: 163-222*. Springer, Boston, USA. https://doi.org/10.1007/978-1-4614-3223-4_6

Ahmed NA, Shehab MA, Al-Ayyoub M, and Hmeidi I (2015). Scalable multi-label Arabic text classification. In the 6th International Conference on Information and Communication Systems (ICICS), IEEE, Amman, Jordan: 212-217. <https://doi.org/10.1109/IACS.2015.7103229>

Akaichi J, Dhouioui Z, and Pérez MJLH (2013). Text mining Facebook status updates for sentiment classification. In the 17th International Conference on System Theory, Control and Computing, IEEE, Sinaia, Romania: 640-645. <https://doi.org/10.1109/ICSTCC.2013.6689032>

Alakrot A (2019). Detection of anti-social behaviour in online communication in Arabic. Ph.D. Dissertation, University of Limerick, Limerick, Ireland.

Al-Ayyoub M, Nuseir A, Alsmearat K, Jararweh Y, and Gupta B (2018). Deep learning for Arabic NLP: A survey. *Journal of Computational Science*, 26: 522-531. <https://doi.org/10.1016/j.jocs.2017.11.011>

Al-Horaibi L and Khan MB (2016). Sentiment analysis of Arabic tweets using text mining techniques. In the 1st International Workshop on Pattern Recognition, International Society for Optics and Photonics, Tokyo, Japan, 10011: 288-292. <https://doi.org/10.1117/12.2242187>

- Alhumoud SO, Altuwaijri MI, Albuhairei TM, and Alohaideb WM (2015). Survey on Arabic sentiment analysis in Twitter. *International Science Index*, 9(1): 364-368.
- Al-Laith A and Shahbaz M (2021). Tracking sentiment towards news entities from Arabic news on social media. *Future Generation Computer Systems*, 118: 467-484. <https://doi.org/10.1016/j.future.2021.01.015>
- Almuqren L, Alzammam A, Alotaibi S, Cristea A, and Alhumoud S (2017). A review on corpus annotation for Arabic sentiment analysis. In the *International Conference on Social Computing and Social Media*, Springer, Vancouver, Canada: 215-225. https://doi.org/10.1007/978-3-319-58562-8_17
- Alomari KM, ElSherif HM, and Shaalan K (2017). Arabic tweets sentimental analysis using machine learning. In the *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, Arras, France: 602-610. https://doi.org/10.1007/978-3-319-60042-0_66
- Al-Osaimi S and Badruddin KM (2014). Role of emotion icons in sentiment classification of Arabic tweets. In the *6th International Conference on Management of Emergent Digital Ecosystems*, Association for Computing Machinery, Buraidah, Al Qassim, Saudi Arabia: 167-171. <https://doi.org/10.1145/2668260.2668281>
- Alotaibi S, Mehmood R, and Katib I (2019). Sentiment analysis of Arabic tweets in smart cities: A review of Saudi dialect. In the *Fourth International Conference on Fog and Mobile Edge Computing*, IEEE, Rome, Italy: 330-335. <https://doi.org/10.1109/FMEC.2019.8795331>
- Al-Radaideh Q (2020). Applications of mining Arabic text: A review. In: Sadollah A and Sinha T (Eds.), *Recent trends in computational intelligence*: 91-109. BoD-Books on Demand, Norderstedt, Germany. <https://doi.org/10.5772/intechopen.91275>
PMCID:PMC7447403
- Alsaedi N and Burnap P (2015). Arabic event detection in social media. In the *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Cairo, Egypt: 384-401. https://doi.org/10.1007/978-3-319-18111-0_29
- Alsalem S (2011). Automated Arabic text categorization using SVM and NB. *The International Arab Journal of e-Technology*, 2(2): 124-128.
- Alsanad A (2018). Arabic topic detection using discriminative multi nominal Naïve Bayes and frequency transforms. In the *International Conference on Signal Processing and Machine Learning*, Association for Computing Machinery, Shanghai, China: 17-21. <https://doi.org/10.1145/3297067.3297095>
- Alshargi F, Dibas S, Alkhereyf S, Faraj R, Abdulkareem B, Yagi S, and Rambow O (2019). Morphologically annotated corpora for seven Arabic dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In the *4th Arabic Natural Language Processing Workshop*, Association for Computational Linguistics, Florence, Italy: 137-147. <https://doi.org/10.18653/v1/W19-4615>
- Al-Smadi M, Talafha B, Al-Ayyoub M, and Jararweh Y (2019). Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics*, 10(8): 2163-2175. <https://doi.org/10.1007/s13042-018-0799-4>
- Al-Twairesh N, Al-Khalifa H, and Al-Salman A (2014). Subjectivity and sentiment analysis of Arabic: Trends and challenges. In the *IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, Doha, Qatar: 148-155. <https://doi.org/10.1109/AICCSA.2014.7073192>
- Asur S and Huberman BA (2010). Predicting the future with social media. In the *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, Toronto, Canada, 1: 492-499. <https://doi.org/10.1109/WI-IAT.2010.63>
- Atoum JO and Nouman M (2019). Sentiment analysis of Arabic Jordanian dialect tweets. *International Journal of Advanced Computer Science and Applications*, 10(2): 256-262. <https://doi.org/10.14569/IJACSA.2019.0100234>
- Badaro G, Baly R, Akel R, Fayad L, Khairallah J, Hajj H, and El-Hajj W (2015). A light lexicon-based mobile application for sentiment mining of Arabic tweets. In the *2nd Workshop on Arabic Natural Language Processing*, Association for Computational Linguistics, Beijing, China: 18-25. <https://doi.org/10.18653/v1/W15-3203>
- Badaro G, Baly R, Hajj H, El-Hajj W, Shaban KB, Habash N, and Hamdi A (2019). A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18(3): 1-52. <https://doi.org/10.1145/3295662>
- Baier L, Jöhren F, and Seebacher S (2019). Challenges in the deployment and operation of machine learning in practice. In the *27th European Conference on Information Systems*, Stockholm-Uppsala, Sweden: 1-15.
- Balahur A (2013). Sentiment analysis in social media texts. In the *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Atlanta, Georgia: 120-128.
- Boukili S, Biniz M, El Adnani F, Cherrat L, and El Moutaouakkil AE (2018). Arabic text classification using deep learning technics. *International Journal of Grid and Distributed Computing*, 11(9): 103-114. <https://doi.org/10.14257/ijgcd.2018.11.9.09>
- Castillo C, Mendoza M, and Poblete B (2011). Information credibility on Twitter. In the *20th International Conference on World Wide Web*, Hyderabad, India: 675-684. <https://doi.org/10.1145/1963405.1963500>
- Comunello F and Anzera G (2012). Will the revolution be tweeted? A conceptual framework for understanding the social media and the Arab Spring. *Islam and Christian-Muslim Relations*, 23(4): 453-470. <https://doi.org/10.1080/09596410.2012.712435>
- Dalal MK and Zaveri MA (2011). Automatic text classification: A technical review. *International Journal of Computer Applications*, 28(2): 37-40. <https://doi.org/10.5120/3358-4633>
- Dukes K and Habash N (2010). Morphological annotation of Quranic Arabic. In the *7th International Conference on Language Resources and Evaluation*, European Language Resources Association, Valletta, Malta: 2530-2536.
- Duwairi R and El-Orfali M (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science*, 40(4): 501-513. <https://doi.org/10.1177/0165551514534143>
- Duwairi RM, Ahmed NA, and Al-Rifai SY (2015). Detecting sentiment embedded in Arabic social media-A lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*, 29(1): 107-117. <https://doi.org/10.3233/IFS-151574>
- El-Halees AM (2008). A comparative study on Arabic text classification. *Egyptian Computer Science Journal*, 30(2): 1-11.
- Elhassan R and Ahmed M (2015). Arabic text classification on full word. *International Journal of Computer Science and Software Engineering*, 4(5): 114-120.
- Guellil I, Adeel A, Azouaou F, Chennoufi S, Maafi H, and Hamitouche T (2020). Detecting hate speech against politicians in Arabic community on social media. *International Journal of Web Information Systems*, 16(3): 295-313. <https://doi.org/10.1108/IJWIS-08-2019-0036>
- Guellil I, Saâdane H, Azouaou F, Gueni B, and Nouvel D (2021). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5): 497-507. <https://doi.org/10.1016/j.jksuci.2019.02.006>

- Habash N and Sadat F (2006). Arabic preprocessing schemes for statistical machine translation. In the Human Language Technology Conference of the North American Chapter of the ACL, Association for Computational Linguistics, New York, USA: 49-52.
- Harrag F, El-Qawasmeh E, and Pichappan P (2009). Improving Arabic text categorization using decision trees. In the 1st International Conference on Networked Digital Technologies, IEEE, Ostrava, Czech Republic: 110-115. <https://doi.org/10.1109/NDT.2009.5272214>
- Husain F (2020). Arabic offensive language detection using machine learning and ensemble machine learning approaches. ArXiv Preprint ArXiv:2005.08946. <https://doi.org/10.48550/arXiv.2005.08946>
- Ismail R, Omer M, Tabir M, Mahadi N, and Amin I (2018). Sentiment analysis for Arabic dialect using supervised learning. In the International Conference on Computer, Control, Electrical, and Electronics Engineering, IEEE, Khartoum, Sudan: 1-6. <https://doi.org/10.1109/ICCCEE.2018.8515862>
PMCID:PMC5811579
- Itani M (2018). Sentiment analysis and resources for informal Arabic text on social media. Ph.D. Dissertation, Sheffield Hallam University, Sheffield, UK. <https://doi.org/10.1016/j.procs.2017.10.101>
- Janasik N, Honkela T, and Bruun H (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3): 436-460. <https://doi.org/10.1177/1094428108317202>
- Jardaneh G, Abdelhaq H, Buzz M, and Johnson D (2019). Classifying Arabic tweets based on credibility using content and user features. In the IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, IEEE, Amman, Jordan: 596-601. <https://doi.org/10.1109/JEIT.2019.8717386>
- Joulin A, Grave E, Bojanowski P, and Mikolov T (2016). Bag of tricks for efficient text classification. ArXiv Preprint ArXiv:1607.01759. <https://doi.org/10.48550/arXiv.1607.01759>
- Kaseb GS and Ahmed MF (2016). Arabic sentiment analysis approaches: An analytical survey. *International Journal of Scientific and Engineering Research*, 7(10): 712-723.
- KhosraviNik M and Esposito E (2018). Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics*, 14(1): 45-68. <https://doi.org/10.1515/lpp-2018-0003>
- Li W (2020). The language of bullying: Social issues on Chinese websites. *Aggression and Violent Behavior*, 53: 101453. <https://doi.org/10.1016/j.avb.2020.101453>
- Mataoui MH, Zelmato O, and Boumechache M (2016). A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. *Research in Computing Science*, 110(1): 55-70. <https://doi.org/10.13053/rcs-110-1-5>
- McNeil K (2018). *Tunisian Arabic corpus: Creating a written corpus of an 'unwritten' language*. Edinburgh University Press, Edinburgh, UK. <https://doi.org/10.1515/9780748677382-004>
- Mohammad SM, Salameh M, and Kiritchenko S (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55(1): 95-130. <https://doi.org/10.1613/jair.4787>
- Saad MK and Ashour WM (2010). Arabic text classification using decision trees. In the 12th international workshop on computer science and information technologies CSIT, Moscow, Russia, 2: 75-79.
- Salamah JB and Elkhlifi A (2014). Microblogging opinion mining approach for Kuwaiti dialect. In *The International Conference on Computing Technology and Information Management, Society of Digital Information and Wireless Communication*, Dubai, UAE: 388-396.
- Shoukry A and Rafea A (2012). Sentence-level Arabic sentiment analysis. In the *International Conference on Collaboration Technologies and Systems*, IEEE, Denver, USA: 546-550. <https://doi.org/10.1109/CTS.2012.6261103>
- Traboulsi H (2009). Arabic named entity extraction: A local grammar-based approach. In the *International Multiconference on Computer Science and Information Technology*, IEEE, Mragowo, Poland: 139-143. <https://doi.org/10.1109/IMCSIT.2009.5352809>
- Yu Y, Duan W, and Cao Q (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4): 919-926. <https://doi.org/10.1016/j.dss.2012.12.028>
- Yue L, Chen W, Li X, Zuo W, and Yin M (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2): 617-663. <https://doi.org/10.1007/s10115-018-1236-4>
- Zhang J, Zhan ZH, Lin Y, Chen N, Gong YJ, Zhong JH, and Shi YH (2011). Evolutionary computation meets machine learning: A survey. *IEEE Computational Intelligence Magazine*, 6(4): 68-75. <https://doi.org/10.1109/MCI.2011.942584>