Tech Science Press

Sentiment Analysis for Arabic Social Media News Polarity

Adnan A. Hnaif[1,*], Emran Kanan[2] and Tarek Kanan[1]

[1]Al-Zaytoonah University of Jordan, Faculty of Science and Information Technology, Amman, 11733, Jordan
[2]Amman Arab University, Faculty of Computing and Informatics, Amman, 11953, Jordan
*Corresponding Author: Adnan A. Hnaif. Email: adnan_hnaif@zuj.edu.jo

Abstract: In recent years, the use of social media has rapidly increased and developed significant influence on its users. In the study of the behavior, reactions, approval, and interactions of social media users, detecting the polarity (positive, negative, neutral) of news posts is of considerable importance. This proposed research aims to collect data from Arabic social media pages, with the posts comprising the main unit in the dataset, and to build a *corpus* of manually-processed data for training and testing. Applying Natural Language Processing to the data is crucial for the computer to understand and easily manipulate the data. Therefore, Stop-Word removal, Stemming, and Normalization are applied. Several classifiers, such as Support Vector Machine, Naïve Bayes, K-Nearest Neighbor, Random Frost, and Decision Tree are used to train the dataset, and their accuracy is determined by data testing. These two steps are carried out using the open-source WEKA tool. As a result, each post is categorized into three different classes: positive, negative, and neutral. This research concludes that among the classifiers, SVM reaches the highest level of accuracy with a percentage of 83% for the F1-measure.

1 Introduction

This section is categorized into five subsections: Arabic language, Arabic Sentiment Analysis, Classification Techniques, Arabic Natural Language Processing (NLP), and Evaluation.

1.1 Arabic Language

The most common tool that allows people to communicate, talk, discuss, write, and express feelings and ideas is language. The world has thousands of languages, and undoubtedly, each country has its own official language. One of these is Arabic, a Semitic language, in the same family as Hebrew and Aramaic. Approximately 260 million people use Arabic as their first language, and more understand it as a second language. Arabic has its own alphabet, which is written from right to left, similar to Hebrew. Given its wide usage throughout the world, Arabic is one of the six official languages of the United Nations, along with English, Spanish, French, Russian, and Chinese. Many countries have Arabic as an official

language, although it is not spoken the same way. This language has many dialects, or varieties, such as Modern Standard, Egyptian, Gulf, Moroccan, and Levantine.

Several of these dialects widely differ, which presents difficulties for the speakers to understand each other. The Arabic language contains 28 letters, each with varying forms and spelling depending on its position in the word. For instance, the letter (ض), pronounced as (Dhad), is written as (ضـ) at the beginning of the word, as (ـضـ) between two letters, and as (ـض) at the end of the word [1].

The Arabic language has a classical form, that is, Modern Standard Arabic (MSA). In the Arab world, MSA is the language of the Holy Quran, books, news, official publications, and journals. However, in daily life, a slang form of the language, known as "street language", is used when communicating with each other. For example, Jordanian people have their dialect and specific verbal communication. In the Arabian Gulf, citizens use the Khaliji dialect. People in Lebanon and Syria speak the Levantine dialect, also called (Shamii), whereas those in Libya, Tunisia, Algeria, and Morocco speak the Maghrebi dialect. Similarly, Sudanese people have a unique dialect [1].

1.2 Arabic Sentiment Analysis

The goal of sentiment analysis is to determine the attitudes of a group of people using one or more than one platform. In recent years, social media posts have rapidly increased. Specifically, Arabic social media news posts have developed considerable influence on social media users. On this basis, the behavior, reaction, acceptance, and interaction of social media users are observed and analyzed. Furthermore, the analysis to make our proposed system helps users to become more relaxed and allow relevant organizations to be more familiar with each other. Later on, the behaviors, reactions, approval, and interaction are mentioned as the user opinion in this research [1].

NLP has many types, one of which is sentiment analysis that is used for studying human language through logical computational probabilities. The main goal of using sentiment analysis is to classify the documents and describe their polarity, whether positive, negative, or neutral.

One of the greatest challenges of sentiment analysis is data gathering because of the huge data set required to obtain the results [2]. For this reason, many application programming interfaces (API) have emerged. The central key of these APIs is to easily collect datasets from different sources. APIs are defined as a group of applications containing a communication protocol that plays the main role in collecting data to build sentiment analysis or any other software that requires the creation of a huge amount of data [3].

Sentiment analysis has many benefits, especially in business. Mostly, sentiment analysis helps in the fields of business intelligence application and recommender systems [4]. One of these benefits is monitoring the social media pages of trademark brands and companies through analysis of the reactions, comments, feedback, and contributions of social media users.

The goal of sentiment analysis is to extract or predict the polarity of people's behavior, reactions, approval, and interactions.

Considerable interest has been recently paid on sentiment analysis to help predict the polarity of Internet content in many areas. The majority of systems is built for English and European languages rather than Arabic. Therefore, in this research, sentiment analysis is applied on social media posts and the news are classified into positive, negative, and neutral by building several classifiers, such as Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (K-NN), and Decision Tree (DT). Accordingly, the classification is carried out using the WEKA Tool – an open-source machine learning software.

1.3 Classification Techniques

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each item in the data. In machine

learning, the three types of classification techniques are supervised, unsupervised, and semi-supervised. The first type is used in this research.

1.4 Arabic Natural Language

Computer science combines several fields and has numerous subfields, such as Natural Language and Artificial Intelligence. These two significant subfields are related to the interactions that occur between computer and human (natural) languages, particularly in programming computers to enable the processing and analysis of large amounts of natural language data. Common challenges in the NLP include speech recognition, natural language understanding, and natural language generation.

Algorithms are used to enable computers to identify and extract the natural human languages. These algorithms convert natural language that is considered unstructured into a specific form, which the computer can understand.

1.5 Evaluation

Considerable research has focused on Sentiment Analysis and NLP. However, the most suitable and most accurate technique is yet to be identified. Therefore, evaluation measures are created and reach over nine types. The present research discusses the three most accredited and well-known measures, namely, Recall, Precision, and F-Score. Each measure has a formula and is developed through Information Retrieval. These formulas provide and present the accuracy percentage when they are solved and calculated. Thus, the accuracy percentage is the key determinant of the suitability of the given algorithm.

2 Literature Review

English is considered the dominant language of science in the world, and most studies and research are written as such. Recently, sentiment analysis has been applied on other languages, such as Arabic.

Sentiment analysis between English and Arabic show considerable difference. Fig. 1 shows that more research has been done on sentiment analysis of English compared with Arabic [5].
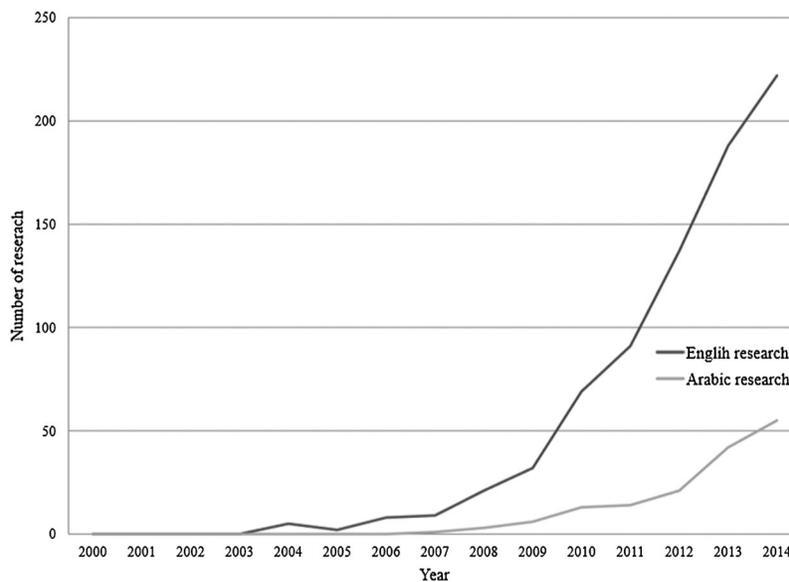

Figure 1: Disparity of research between Arabic and English [6]

Arabic still does not have a sufficient number of corpora. Fig. 2 shows the 10 most dominant languages on the Internet.
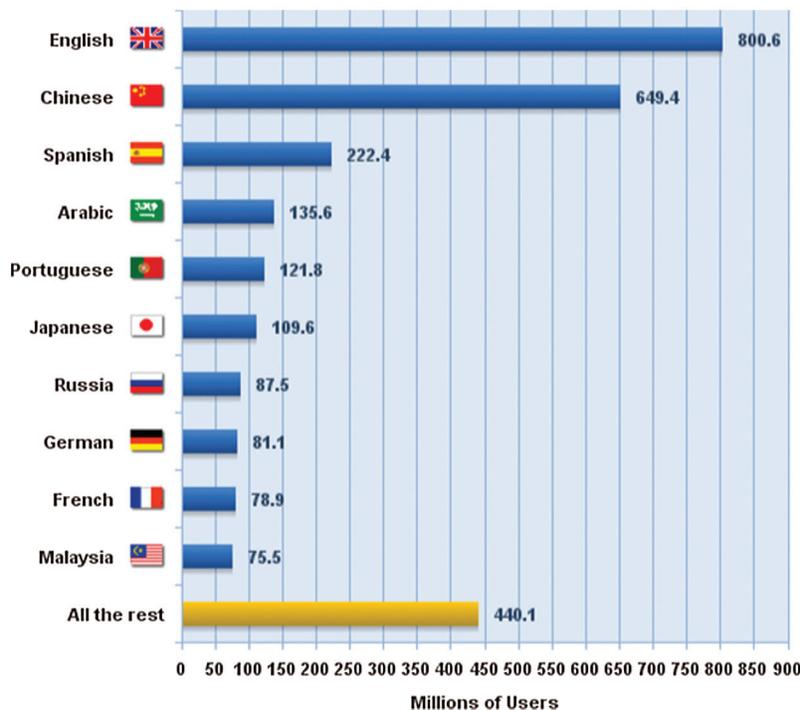

Figure 2: Top 10 languages by the number of users on the Internet [6]

A machine learning software called RapidMiner [6] is developed to support processing statement that is available in the Arabic language. According to RapidMiner, SVM and NB classification techniques are applied. SVM classifiers have better performance than NB when merged with stemming and TF-IDF scheme using Bigrams.

From Twitter, 10,006 tweets from Egyptian pages have been collected and then divided into four parts, as follows: Positive tweets (799); Negative tweets (1,684); Neutral tweets (832); and finally, Objective tweets (6,691). SVM and NB are used as classification techniques. NLP is applied and a deep learning model is used for opinion mining, Recursive Neural Tensor Network (RNTN), which is trained by a non-Twitter *corpus* and achieves better performance. Thus, the importance of lemmatization for handling the complexity and lexical sparsity of Arabic is confirmed [7].

API is used to collect data from Twitter [8] given that the real-time collection of each tweet is required. Subsequently, the pre-processing of data is applied to remove emails and hyperlinks. Collected from January 20 to February 21, 2014, the data set contains (7,503) tweets for training. Tweets with English words written using the Arabic alphabet posed a challenge for this study. A total of (1,365) data points are collected for testing. Then, sensitive and sentiments analysis (SSA) is built for Arabic Twitter feeds, which was the first in the Arabic language.

Pre-processing of Arabic language using supervised machine learning has been applied to assess polarity, specifically for the Saudi dialect. Redundant tweets and duplicates are initially removed to exclude all unnecessary data and redundant letters (e.g., duplicate letters). A total of 4,000 tweets are collected. Five human annotators built the *corpus*. The polarity of each tweet is labelled, and a Bag of

Words (BOW) is established. Therefore, the classification techniques used in this research are SVM and NB. The use of BOW is proven to enhance the accuracy of the analysis [9].

The *corpus* is built using 2000 Arabic statements, divided into 1000 MSA from Twitter, www.booking. com and www.ejabat.com. For the data collection, the API is used for Twitter from June 2012. In addition, 10k tweets and 10k comments and reviews in Arabic are collected. SVM is applied as a classification technique and the data are divided into 80% for training, 10% for developing, and 10% for testing. Furthermore, the results are compared before and after using lexicon expansion, which shows a positive effect on the classification [10].

A dataset that contains 4000 tweets is collected and shows that the accuracy increased by adding a BOW that consists of names of popular people in the area to all applied techniques; results show that the SVM achieves 98% accuracy and is therefore considered to be the best classifier [11].

Identification of irony in Arabic statements is attempted. A total of 2000 tweets are collected and labelled as positive and 4,783 as negative. Two classification techniques are used, namely, SVM and NB under WEKA tool. The collected data contain tweets with mentions of various famous people, such as H. Clinton, M. Morsi, D. Trump, or A. Alsissi, using an API for Twitter. Following that, several features such as surface, sentiment, shifter, and contextual were applied. The application is intended to help the classifiers in the detection and increase their accuracy, which is successfully achieved. Each technique shows a 72.36% accuracy [12].

Two classification techniques, SVM and NB, are used. The data are collected from Twitter using API. After data collection, normalization is applied to reduce the data size. If a short text is available, then the unigram is used. Unigram can help the machine-learning algorithm to detect data patterns, and thus, is more effective [13].

Another research used five steps (data collection, pre-processing, classifications, clustering, and summarization) to collect data using a Twitter API and concluded that further features need to be introduced to enhance the detection [14]. NB is used as the classification technique, trained and tested using the WEKA toolkit.

The procedure is also carried out in two phases [15]. The first one is pre-processing the data and deleting unnecessary entities, such as mentions and hyperlinks. The second phase is constructing a feature set from the data to use in classification with SVM, NB, and KNN. F-Measure is used to provide a score for each word. The data is collected from Twitter using API between April 26 and June 1, 2014. The classifiers are trained and tested using the WEKA tool. Results show that the achievement by (Part Of Speech (PoS) is not significant. Notwithstanding, using Twitter has become widespread and now includes various Arabic dialects.

3 Procedures and Methodology

Sentiment analysis is classified into three central fields, namely, lexicon, tools, and lexicon tools. Lexicon is defined as the words, phrases, meanings, and patterns that can be used to express subjectivity. Tools contain different types of classifiers that use text classification algorithms. NLP tools include Stemmer and Tagger, among others [16]. The fundamental part of tools is the corpora, which includes the annotated data with its polarity. The classification algorithm uses these corpora to analyze new content.

In this study, BOW that can be considered as a dictionary is built for sentiment analysis, specifying the words, phrases, and patterns used in the language. Thus, anything related to sentiment analysis must have a *corpus* [7].

Unlike Arabic, English is relatively consistent, whether slang or standard. Arabic has numerous dialects and is not as comprehensively studied. Unfortunately, Arabic Sentiment *Corpus* is limited. Therefore, relevant data that can help in this study on sentiment analysis is insufficient. Most available research

regarding the Arabic *Corpus* only involves one topic – movie reviews. Furthermore, these studies include inadequate data and mostly require purchase. As a result, specialists had to develop a new Arabic Sentiment *Corpus* to complete this study. Notably, the data of the produced *corpus* are collected from Arabic social media news pages, such as Facebook and Twitter, which might be local or global.

Contrary to possible initial assumptions, building a *corpus* for Arabic is challenging and complicated. Overall, building a *corpus* is carried out in several steps, starting with the data gathering and ending with pre-processing.

3.1 Data Preparation

The dataset in this *corpus* is gathered from 15 different news pages. The dataset currently contains 6,138 posts and tweets from Facebook and Twitter of two different domains, local and international. Worthy of mention is that the comments and retweets are collected from the same social media.

The dataset is annotated and categorized into three polarity categories, namely, positive, negative, and neutral. Hence, this dataset is gathered to build the *corpus*.

3.2 Data Annotation

Three student groups from the computer science division at Amman Arab University are chosen to classify the items in the dataset into positive, negative, or neutral. After the annotation procedure is completed, the dataset is sorted into folders and text files. The *corpus* is openly and freely accessible for research and analysts.

3.3 Cleaning Dataset

The raw dataset is not 100% pure and requires additions, such as affixes. The Arabic dataset collected from online sources may contain special characters, non-Arabic words, non-Arabic letters, numbers, symbols, or elongated words. As necessary, all non-Arabic letters and HTML links are removed from the dataset.

The second process focused on special characters. Every so often, people use these special characters to draw emoticons, such as sad ":(" or smiley faces ":)", which are considered a short sentiment. However, people occasionally use them for no reason or by mistake [17].

Any special character used to represent any other emoticon should not be removed from the text because of its powerful meaning that affects the polarity of the text. Nevertheless, special characters that do not have any importance or meaning are removed.

In this step, a good practice is to build a list that contains emoticons, represented by special characters, to be used as a guide [18]. Moreover, at times numbers have the power to express a feeling or sentiment. Accordingly, these numbers are included as parts of the text [19–24].

Word elongation refers to the addition of extra letters in a word for emphasis. An example is "I looooove Jordan", which implies that the user is emphasizing their feelings by repeating the letter "O" in the word "love". Word elongation is also used in Arabic. Consequently, that iteration affects the processing steps; therefore, all redundant letters in the word should be removed. See Tab. 1.

4 Results and Analysis

This study aims to extract the polarity from Arabic posts on social media, such as Facebook and Twitter, by applying classification algorithms. The accuracy of this dataset classification is measured by Precision, Recall, and F1-Measure. Five classification algorithms are used, namely, SVM, NB, K-NN, DT, and RF.

Table 1: Word elongation

	No elongation	Elongation
Sentence in Arabic	أهلا وسهلا	اهلا و سسسسسهلا
Transcription	Ahlnwasahln	Ahlaaaaaaan
English Translation	Welcome	Weeeelcome

The classifiers are trained on the collected dataset. Three groups of people divided the dataset into three classifications, namely, positive, negative, and neutral.
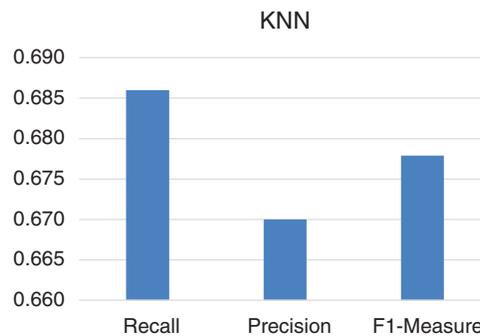
The classification algorithm is trained using a Cross-Validation tool. Notably, the dataset is divided into 80% for training and 20% for testing.

4.1 K-NN

Tab. 2 shows the accuracy measures of K-NN and Fig. 3 presents the results of K-NN classifier by Recall, Precision, and F1-Measure. The classifier categorizes the posts as Positive, Negative, and Neutral. The dataset was pre-processed by applying Arabic NLP tools, such as Normalization, Stop Word Removal, and Stemming.

Table 2: F1- measure for K-NN

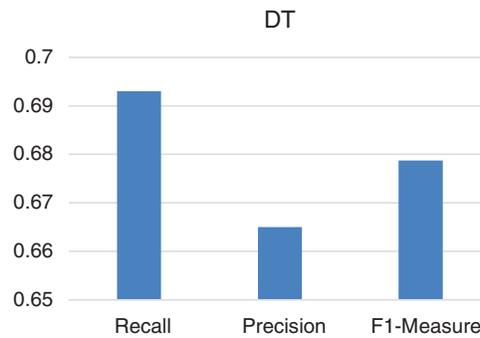| Scale	Value
Recall	0.686
Precision	0.670
F1-Measure	0.678

Figure 3: K-NN classifier results

4.2 DT

Tab. 3 shows the decision tree results, which presents the accuracy measures of DT, and Fig. 4 shows the DT classifier accuracy results, measured by Recall, Precision, and F1-Measure. The classifier categorizes the posts as Positive, Negative, and Neutral.

Table 3: Accuracy measures of DT

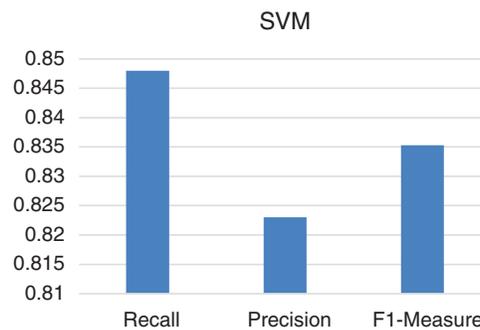| Scale	Value
Recall	0.693
Precision	0.665
F1-Measure	0.679

Figure 4: DT classifier results

4.3 SVM

Tab. 4 presents the accuracy measured of SVM and Fig. 5 shows the accuracy of the SVM classifier, measured by Recall, Precision, and F1-Measure. The classifier categorizes the dataset posts as Positive, Negative, and Neutral.

Table 4: Accuracy measures of SVM

Scale	Value
Recall	0.848
Precision	0.823
F1-Measure	0.835

Figure 5: SVM classifiers results

4.4 RF

Tab. 5 displays the accuracy measures of RF and Fig. 6 shows the accuracy of the RF classifier, measured by Recall, Precision, and F1-Measure. The classifier categorizes the posts as Positive, Negative, and Neutral.

Table 5: Accuracy measures of RF

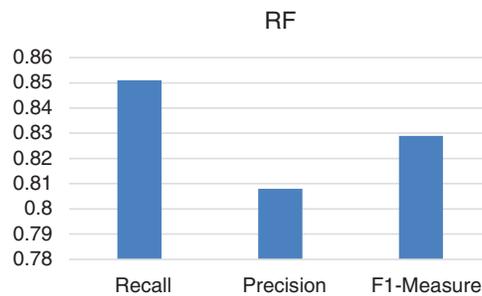| Scale	Value
Recall	0.851
Precision	0.808
F1-Measure	0.829

Figure 6: RF classifier results

4.5 NB

Tab. 6 shows the accuracy measures of NB and Fig. 7 displays the NB classifier accuracy, measured by Recall, Precision, and F1-Measure. The classifier categorizes the dataset posts into three classes as Positive, Negative, and Neutral.

Table 6: Accuracy measures of NB

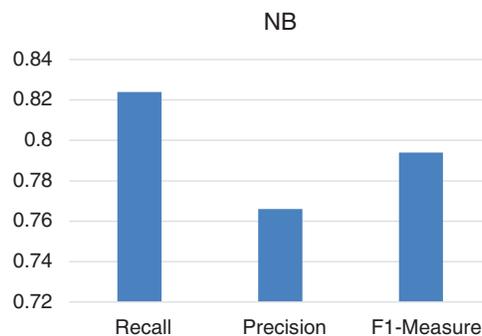| Scale	Value
Recall	0.824
Precision	0.766
F1-Measure	0.794

Figure 7: NB classifier results

5 Discussion

Tab. 7 shows the Recall, Precision, and F1-Measures of the five classifiers: KNN, DT, SVM, RF, and NB. Fig. 8 shows the Recall results of the five used classifiers. By comparison, the RF classifier shows the highest Recall accuracy while K-NN has the lowest Recall accuracy. These results are compatible with previous research.

Table 7: Recall, precision, and F1-measures for classifiers

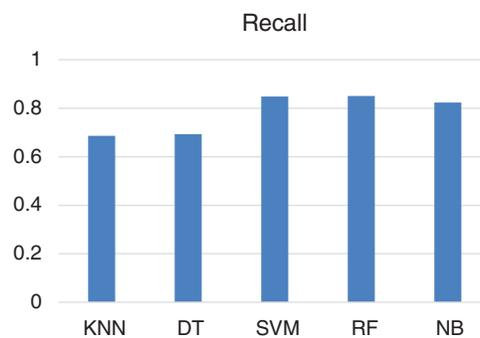| Type of Classifier	Recall	Precision	F1-Measure
KNN	0.686	0.67	0.678
DT	0.693	0.665	0.679
SVM	0.848	0.823	0.835
RF	0.851	0.808	0.829
NB	0.824	0.766	0.794

Figure 8: Recall results

Fig. 9 illustrates the precision results of the five classifiers. By comparison, SVM shows the highest Precision accuracy while K-NN has the lowest precision accuracy. These results are compatible with previous research.
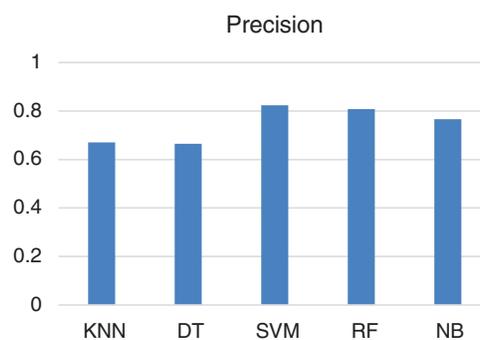

Figure 9: Precision results

Fig. 10 shows the F1-Measure results of the five classifiers. By comparison, SVM shows the highest F1-Measure accuracy while K-NN has the lowest F1-Measure accuracy. These results are compatible with previous research.
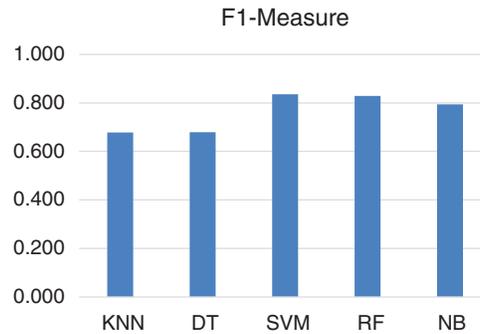

Figure 10: F1-measure results

Fig. 11 demonstrates the average results of all five classifiers. The first column represents Recall, the second is Precision, and the third is the F1-Measure. Compared with the other classifiers, SVM is superior in all measures with its level of accuracy reaching 83%.

Figure 11: Average results of all five classifiers

6 Conclusion

This study proposes sentiment analysis and extraction of the polarity from news pages by applying classification. The dataset is collected from news social media websites, such as Facebook and Twitter. A social medium is an open environment and allows people to express their opinions, which spread through different behaviors and beliefs. A dataset is collected and categorized into three classes, namely, Positive, Negative, and Neutral.

We propose the creation of news pages on social media to extract the polarity from the dataset. The dataset is compiled from Arabic social media, such as Facebook and Twitter, where approximately 6,138 post and tweets are collected. Three Arabic NLP tools are implemented, namely, Normalization, Stop-word removal, and Stemming. Five classification algorithms, SVM, NB, RF, K-NN, and DT are applied to extract the polarity from the datasets. The performance of the algorithms are calculated using Recall, Precision, and F1-measure standards.

Several experiments are carried out on all datasets including the classification algorithm, as follows: with the application of all NLP tools; with none of the NLP tools; and separately to the Facebook posts and then to the Twitter tweets using NLP tools.

When applying the classification algorithms to all datasets with all NLP tools, the results show that the SVM algorithm gives the highest accuracy for the F1-measure, followed by NB, DT, K-NN, and RF. When applying the classification algorithms to all datasets without Stemming, the results show that the RF algorithm provides the highest accuracy for the F1-measure, followed by SVM, NB, DT and K-NN. When applying the classification algorithms to all datasets without Stop-Word Removal, the results show that the RF provides the highest accuracy for the F1-measure, followed by SVM, NB, K-NN, and DT.

When applying the classification algorithm to the Facebook and Twitter datasets with all NLP tools, the results show that the SVM algorithm provides the highest accuracy for the F1-Measure, followed by RF, NB, DT, and K-NN.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Korayem, D. Crandall and M. Abdul-Mageed, "Subjectivity and sentiment analysis of Arabic: A survey," in *Proc. of the International Conference on Advanced Machine Learning Technologies and Applications*, Berlin, Heidelberg: Springer, pp. 128–139, 2012.

[2] S.-O. Proksch, W. Lowe, J. Wäckerle and S. N. Soroka, "Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches," *Legislative Studies Quarterly*, vol. 44, no. 1, pp. 97–131, 2019.

[3] N. Alsrehin, A. F. Klaib and A. Magableh, "Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study," *IEEE Access*, vol. 7, pp. 49830–49857, 2019.

[4] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton *et al.,* "Deriving marketing intelligence from online discussion," in *Proc. of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago Illinois, USA, pp. 419–428, 2005.

[5] M. Elhadad, K. F. Li and F. Gebali, "Sentiment analysis of Arabic and English tweets," in *Workshops of the International Conference on Advanced Information Networking and Applications (WAINA)*, Matsue, Japan, Cham: Springer, pp. 334–348, 2019.

[6] L. M. Abualigah, A. T. Khader, M. A. Al-Betar and O. A. Alomari, "Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering," *Expert Systems with Applications*, vol. 84, no. 3, pp. 24–36, 2017.

[7] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj and K. B. Shaban, "ArSentD-LEV: A multi-topic *corpus* for target-based sentiment analysis in arabic levantine tweets," in *Proc. of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, Miyazaki, Japan, 2018.

[8] E. Refaee and V. Rieser, "Can we read emotions from a smiley face? emoticon-based distant supervision for subjectivity and sentiment analysis of Arabic twitter feeds," in *Proc. 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, LREC*, Edinburgh, United Kingdom, pp. 1–5, 2014.

[9] G. Alwakid, T. Osman and T. Hughes-Roberts, "Challenges in sentiment analysis for Arabic social networks," *Procedia Computer Science*, vol. 117, no. 1, pp. 89–100, 2017.

[10] M. Ibrahim, M. Abu Al Magd, F. A. Annabi, S. Assaad-Khalil, E. M. Ba-Essa *et al.,* "Recommendations for management of diabetes during Ramadan: Update 2015," *BMJ Open Diabetes Research and Care*, vol. 3, no. 1, pp. e000108, 2015.

[11] A. A. Saifan, E. Alsukhni, H. Alawneh and A. Al-Sbaih, "Test case reduction using data mining technique," *International Journal of Software Innovation (IJSI)*, vol. 4, no. 4, pp. 56–70, 2016.

[12] P. Christoffersen, R. Goyenko, K. Jacobs and M. Karoui, "Illiquidity premia in the equity options market," *Review of Financial Studies*, vol. 31, no. 3, pp. 811–851, 2017.

[13] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal *et al.,* "Ensembl 2013," *Nucleic Acids Research*, vol. 41, no. D1, pp. D48–D55, 2013.

[14] T. Hayat, T. Muhammad, T. A.Alsaedi and M. S. Alhuthali, "Magnetohydrodynamic three-dimensional flow of viscoelastic nanofluid in the presence of nonlinear thermal radiation," *Journal of Magnetism and Magnetic Materials*, vol. 385, pp. 222–229, 2015.

[15] J. Athinarayanan, V. S. Periasamy, M. Alhazmi, K. A. Alatiah and A. A. Alshatwi, "Synthesis of biogenic silica nanoparticles from rice husks for biomedical applications," *Ceramics International*, vol. 41, no. 1, pp. 275–281, 2015.

[16] H. Saif, Y. He, M. Fernandez and H. Alani, "Contextual semantics for sentiment analysis of twitter," *Information Processing & Management*, vol. 52, no. 1, pp. 5–19, 2016.

[17] X. He and Z. Zhang, "GPK: An efficient special symbol input method for keyboards," in *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems 2019*, Glasgow, Scotland UK, pp. 1–6, 2019.

[18] A. Hamid, M. S. Mohsin and M. N. Khalid, "Effectiveness of Urdu reading braille characters with the help of tactile and visual clues," *Journal of Research in Psychology*, vol. 1, no. 1, pp. 16–20, 2019.

[19] T. Kanan and E. A. Fox, "Automated Arabic text classification with P-Stemmer, machine learning, and a tailored news article taxonomy," *Journal of the Association for Information Science and Technology*, vol. 67, no. 11, pp. 2667–2683, 2016.

[20] T. Kanan, X. Zhang, M. Magdy and E. A. Fox, "Big data text summarization for events: A problem-based learning course," in *Proc. of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, pp. 87–90, 2015.

[21] T. Kanan, S. Ayoub, E. Saif, G. Kanaan, P. Chandrasekarar *et al.,* "Extracting named entities using named entity recognizer and generating topics using Latent Dirichlet allocation algorithm for Arabic news articles," in *Proc. of the International Computer Sciences and Informatics Conference (ICSIC)*, Virginia Polytechnic Institute & State University, USA, pp. 1–22, 2015.

[22] T. Kanan, O. Sadaqa, A. Aldajeh, H. Alshwabka, S. AlZu'bi *et al.,* "A review of natural language processing and machine learning tools used to analyze arabic social media," in *Proc. of the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan, IEEE, pp. 622–628, 2019.

[23] T. Kanan, O. Sadaqa, A. Almhirat and E. Kanan, "Arabic light stemming: A comparative study between P-Stemmer, Khoja Stemmer, and Light10 Stemmer," in *Proc. of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, IEEE, pp. 511–515, 2019.

[24] T. Kanan, A. T. Obaidat and M. Al-Lahham, "SmartCert blockchain imperative for educational certificates," in *Proc. of the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan, IEEE, pp. 629–633, 2019.