Journal of Theoretical and Applied Information Technology
15th October 2020. Vol.98. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION AND CONSTRUCTION OF ARABIC CORPUS: FIGURATIVE AND LITERAL

NOUH SABRI ELMITWALLY[1,2], AHMED ALSAYAT[1]

[1]Dept. of Computer Science, College of Computer and Information Sciences Jouf University KSA

[2]Dept. of Computer Science, Faculty of Computers and Artificial Intelligence - Cairo University, Egypt

Email: , nselmitwally@ju.edu.sa[1], nouh.sabri@fci-cu.edu.eg[2], asayat@ju.edu.sa[1]

ABSTRACT

Annotation of Arabic texts is an imperative task that is also costly and time-consuming. To overcome these obstacles to creating Arabic resources for analysis and training, we have built an integrated Arabic corpus. Constructing an Arabic corpus including various rhetorical and unusual sentences is a challenging task for the classification methods. Such a wide, reliable Arabic corpora doesn't exist, which has motivated us to create a corpus for further analysis. The main contributions of this paper are twofold: (1) We construct the Arabic Figurative Sentiment Analysis (AFSA) corpus, consisting of the annotated figurative sentiment texts for the Arabic Saudi Dialect and Modern Standard Arabic (MSA). The construction of this corpus is based on the Arabic Language Sentiment Analysis (ALSA) framework, which involves annotated literal and figurative texts. The collected data contains 2000 texts from the Holy Quran, Al-Hadeeth, and the Arabic Saudi Dialect Dataset, which is comprised of 1000 literal and 1000 figurative annotations. (2) The process developed classifies the collected texts into figurative categories resulting in F1-scores reaching 92%, 93%, and 94% using the Multinomial Naïve Bayes (MNB) classifier, logistic regression (LR), and the Bernoulli Naïve Bayes (BNB) approach, respectively.

Keywords: *Machine Learning; Arabic Sentiment Analysis; Figurative Language*

1. INTRODUCTION

Arabic is one of the most widely spoken and extensively studied languages in the world. Arabic language sciences are divided into linguistics, morphology, grammar, rhetoric, presentation science, and rhymes. Aesthetic expressions are relevant to Arabic language learners studying the essential famous books that define rhetoric and examine its branches and sections. The distinct rhetorical methods vary in the Arabic language based on what relates to the correct articulation of the letters. These techniques also consider the existence of linguistic and figurative expressions, such as metaphor, hyperbole, omission, substitution, presentation, delay, euphemism, praise, defamation, spelling, and logic [1].

Arabic Language Sentiment Analysis (ALSA), also referred to as opinion mining or emotion detection, is defined as accurately identifying the sentiment of a short or long text. Also, ALSA reviews texts as positive, negative, or neutral; however, such reports may also include other modifiers that provide additional sentimental meaning for texts, such as high positive or low negative. These analyses are used to reflect people's opinions about certain services or products. The approach of sentiment analysis is categorized into the lexicon-based approach, the machine learning–based approach, and the hybrid approach. In the lexicon-based approach, the polarity of a document is calculated from the polarity of its words by using dictionaries. Furthermore, this approach reveals a low recall in manually constructing a lexicon while maintaining precision. The independent machine learning–based approach counts on building classifiers such as the Support Vector Machine (SVM), Neural Network (NN), etc. On the other hand, the hybrid approach combines the lexicon-based approach and the machine learning–based approach [2].

Work on ALSA requires considering many natural language processing features, such as hyperbole, sarcasm, irony, metonymy, metaphor, and simile [3]. A general binary text classification for the literal and non-literal (figurative) was derived by specifically investigating text classification for the two Arabic figurative devices of hyperbole and simile [1]. Multiple figurative sentiment analysis studies exist on Italian [4], Chinese [5], and English [6], whereas to the best of our knowledge, this

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

current research is the first 1 dedicated research on figurative sentiment analysis using ALSA [3]. We annotated Arabic corpus texts into literal and figurative sections and implemented machine learning classifiers in texts collected from social media and other sources that we manually annotated to validate this first corpus.

The Arabic figurative is widely used in different lifestyle domains. Figurative language in Arabic plays a basic but beneficial role in ecommerce websites, which demonstrate how significant using Arabic figurative language is in e-business texts [7]. As a result, a rich usage of Arabic figurative language has spread all over the Internet. Thus, our research paper intends to develop an Arabic figurative corpus for use as a language resource for teaching ALSA in figurative sentiment analysis. The corpus is composed of 2,000 Arabic texts collected from the existing datasets of the Holy Quran, Hadeeth, and the Arabic Saudi Dialect Dataset, SS2030 (https://www.kaggle.com/snalyami3/arabic-sentiment-analysis-dataset-ss2030-dataset). Then one must apply three classification algorithms to evaluate the created dataset. We used this work as a starting point for this vital field of research.

In this section, we describe the understanding and implementation of figurative language when applied to these texts. Using machine learning to analyze Arabic texts requires entering annotated data into a corpus-based model built using a manual rhetoric annotation that classifies texts as literal or figurative. Most previous research about Arabic is based on annotated texts separated into either a positive or a negative for the machine learning classification algorithms. The most widely used classification algorithms are the support vector machine (SVM) and the Naïve Bays (NB) classifiers [8]. However, much of this research considers the sentiment analysis of dialect languages without reviewing MSA. In performing our analysis, the Multinomial Naïve Bayes (MNB), logistic regression (LR), and the Bernoulli Naïve Bayes (BNB) machine learning classification models, as well as two vectorization methods, Term Frequency/Inverse Document Frequency (TF/IDF) and Vectorization and Bag of Words (BOW) [9].

The remainder of the paper is organized as follows: Section 2 reviews related work on figurative analysis by considering past studies performed with literal and rhetorical analyses, previous methods of classifications applied to different Arabic datasets, and recent challenges facing Arabic Sentiment Analysis (ASA). Section 3 presents the corpus annotation, and Section 4 provides an overview of our model for Arabic text representation and its components. Section 5 presents the results of this research, and Section 6 concludes the work and makes recommendations for future research.

2. RELATED WORK

This section is divided into two sub-sections: 2.1 describes the related studies to general Arabic texts analysis, and 2.2 demonstrates the related work based on our three selected classification algorithms, MNB, LR, and BNB.

2.1 Analysis and Classification of Arabic Text

Important previous research has focused on figurative language features such as sarcasm, feelings, emotions, and irony, as well as the polarity classification of social media texts and comments [10, 11]. Irony has been explicitly studied as a form of figurative language in Italian [12] and English [6], specifically for detecting the polarity of tweets. People convey messages, feelings, emotions, and opinions figuratively to more accurately express their intended meanings to others. For this research, using metaphors, irony, sarcasm, and figurative language presents significant challenges to machine learning algorithms and classifiers.

Ahmad et al. [13] provided a comprehensive review of available approaches for sentiment analysis, including lexicon-based and machine learningbased using the SVM classifier. Also, Iman et al. [14] constructed an automated Algerian corpus from Facebook that used the vectorized embedding technology of doc2vec and word2vec. Next, sentiment classification was extracted using a variety of machine learning classifiers, such as SVMs, NBs, and logistic regression (LR). The classification results of the f1-score showed that SVMs and NBs outperform other classifiers. On the other end of the spectrum, the worst result was obtained by multilayer perceptron neural networks (MLPs).

Malave et al. [15] constructed the SCUBA (Sarcasm Classification Using a Behavioral modeling Approach) to recognize sarcastic messages on Twitter, which raised interest in investigating figurative language. Similarly, Nguyen and Jung [16] built a system to analyze the figurative language of short texts collected from Social Networking Services (SNSs).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AL-Jumaili et al. [17] applied a hybrid method to a dataset of Arabic tweets collected for sentiment analysis. They divided the proposed hybrid method into three phases: the Arabic corpus, feature extraction phase, and classification. In the feature extraction phase, they used the stemming and POS for linguistic analysis, and TF-IDF was used for statistical analysis. In the classification section, they applied SVM, KNN, and ME to classify the tweets into positive, negative, objective, or neutral categories. Their results showed that the f-score of SVM outperformed the other classifiers.

A semi-supervised approach was introduced [18]. The technical implementation applied four classification algorithms, SVM, NB, Random Forest (RF), and K-Nearest Neighbor (KNN), to manually collected data from a Jordanian platform. To select the best classifier performance, the feature evaluation was tested by using three different feature selection methods of Correlation-based Feature Selection, Principle Components Analysis, and SVM Feature Evaluation. The results revealed that the SVM algorithm outperformed others, with 92.3% accuracy.

The limitation of the Arabic corpus inspired authors in [19] to build a reference for Arabic tweets from social media; specifically, Twitter. The created corpus was named the Arabic Tweets Speech Act Corpus (ArTSAC), which was a new, enhanced version of modern standard Arabic (MSA) for tweets called ArSAS. The developed Arabic Tweets Speech Act Corpus has much richer annotation features than does MSA. The ArTSAC corpus was classified by using SVM algorithm. An average precision result achieved 90%, as well as an F-score of 89.6%, which revealed that using the ArTSAC corpus reached higher scores than using the SArSAS corpus.

Al-Smadi et al. [20] used supervised machine learning to improve the Aspect-Based Sentiment Analysis (ABSA) for a referenced Arabic dataset containing hotel reviews. The proposed approach addressed the aspect level, opinion mining, and polarity identification of sentiments. The implementation of classification used SVM, NB, Bayes Network, Decision Tree (DT), and KNN classifiers. Based on related work exploring the same dataset, the supervised learning approach has proven superior. Evaluation of the results showed that SVM had the best performance in aspect level, opinion mining, and aspect category identification.

Al-Barhamtoshy et al. [21] presented an intelligent model to analyze and classify Arabic tweets in five different corpora. This model is split them into four main phases: pre-processing, feature selection, language model, and classification. The employment task of linguistic pre-processing and the similarity function were used to outperform the cluster of Arabic tweets. The output results of the six used classifiers method explains that the proposed model increased the accuracy of the sentiment classifications.

Eldefrawi et al. [22] employed the subfield of opinion mining called comparative opinion mining, which is used to compare a group of entities to each other. The study entailed using unsupervised sentiment analysis on comparative opinions in the Arabic language to select the most preferred entity. To analyze the Arabic comparative opinions, the proposed model put these three steps under consideration: comparative keywords type, the features existing in opinion, and the position of entities through comparative keywords. Comparative keywords are classified into five proposed categories to simplify analyzing each comparative sentence. The total average of the f-measure was 96.5%, which indicates the correct identified sentiment.

Madhfar et al. [23] performed an experiment applying six classification methods, including NB, SVM, RF, LR, DT, and Stochastic Gradient Descent (SGD), to evaluate their performance. These classifiers were applied to an Arabic dataset collected from newspaper articles. The results showed that the LR outperformed other classifiers by obtaining the highest weighted F1-score. In addition, this analysis implied that the performance for each classifier was affected by the sizes of the feature and of the corpus.

Another study by Jarrar et al. [24] proposed a method of diacritic-based matching for Arabic words. The experiment concerns three alternative algorithms: subsume knowledge-based algorithms, imply rule-based algorithm, and alike machine learning based algorithms. Since this work involved such a substantial dataset, the evaluation of the algorithms was tested for soundness, completion, and accuracy. The final results showed that the accuracy of subsume was 100% that of imply 99.32%, that of alike 99.53%.

Atoum and Nouman [25] built a corpus of 1000 Arabic Jordanian tweets, in which the annotation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classes were positive, negative, and neutral. The SVM, and NB classification algorithms were used to separate the tweets into these three classes. The result suggested that the performance of SVM was better than that of NB.

The authors of [26] built an Arabic sarcasm dataset out of existing Arabic sentiment analysis datasets. After calibrating the dataset by using the BilSTM deep learning model, an F1-score of 0.46 was achieved, indicating a nature up to the challenge of the task and paving the way for future research.

According to authors in [27] the Arabic corpora have not been studied extensively in SA. Their analysis indicates that the reason there have been so few studies on the Arabic corpora is that the Arabic language presents challenges to SA. This study also compared SA papers written before May 2017 and published in four different databases with the papers related to SA of the Arabic language. This resulted in only 48 papers related to the Arabic language out of 1458 total papers. This study thus indicated that there were still many gaps needing to be filled in Arabic-based SA.

Another experiment explored why ASA research has been so limited. The authors of this study [28] concluded their work by addressing the importance of creating and using lexicons rather than focusing on Arabic corpora. Another intensive recent review [29] found that ASA still requires more research on its preprocessing process, feature selection, and classification methods. The study also showed that building a standardized Arabic dataset may occur in future work, along with applying the most suitable classification method. In this research, our goal is to establish the first study for analyzing figurative sentiments in Arabic and then to apply classification algorithms to discover which performs best.

2.2 Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Logistic Regression Classification Algorithms in Arabic Text

In light of the limited number of published studies in the field of Arabic natural language processing, we collected the most recent publications that related our three selected classification algorithms to only Arabic datasets. These collected related studies may include one, two, or all three of these algorithms either as main classifier(s) used in the published studies or to validate the published study.

Qadi et al. [30] constructed a large MSA dataset collected from Arabic news articles along with their tags. This dataset contained four different categories: business, sports, technology, and the Middle East. The dataset was scrubbed of numbers, punctuation, Latin characters, and stop words. To classify the data into its proper categories, a number of classification algorithms were applied, as well as the majority voting classifier. The best F1-score was achieved by S, which proved superior to the other classifiers, including MNB and LR. In addition, this study compared the results of the constructed dataset with those from another reported dataset, Akhbarona, which contained 7 different categories.

Zahir et al. [31] developed an Arabic dataset from a group of YouTube comments. They generated a classification model by applying the MNB classification algorithm to the genders of the comments' authors. The results of the classification model were promising, displaying high accuracy, average precision, and a high F1-score.

Qwaider et al. [32] extended a reported Arabic dataset called Shami by annotating one of its subsets. This reconstructed Levantine corpus was named Shami-Sent, which was capable of processing SA. For annotating the Shami-Senti corpus, both lexiconbased annotation and human annotation were used, classifying sentences as positive, negative, or neutral. The study extended to evaluate the performance of the model by comparing the results of some classifier algorithms with the results obtained from different datasets. Using the Shami-Senti feature engineering, the accuracy of MNB reached 75.2%.

Authors in [33] constructed an Arabic corpus from four different well-known platforms to detect hate speech. The evaluation of the generated dataset was tested by applying twelve machine learning algorithms, including MNB, BNB, and LR. Moreover, two of the deep learning algorithms were used, in which the Recurrent Neural Network outperformed all of the other classifiers, yielding 98.7% accuracy.

Sadik and Aberkane [34] constructed an Arabic corpus from the reviews on one of the popular online newspapers in Algeria, Echorouk. Next, they applied six classification algorithms to sort the opinions into groups: positive, negative, or neutral. The data were tested from these 3-classes as well as 2-classes, employing features of unigram or bigram

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

analysis and sometimes both. The results of MNB classification were superior to those using other classifiers. In addition, using the combination of both unigram and bigram features across two or three classes worked better than using each one separately. The accuracy of MNB of 2-classes reached 85.57%, and it reached 65.64% for 3-classes.

The authors of [35] collected tweets from Twitter using a crawler to construct a large Arabic corpus. Their research approach concentrated on increasing user interest, which had fallen in all of the six categories measured: sports, religion, technology, health, economics, and literature. The MNB classification algorithm was implemented to classify the interests of each user. The validation of this study was compared to that of other studies, and here it achieved a high level of accuracy.

Lichouri et al. [36] sought to build a global Algerian corpus, which focused on Algerian dialect regions according to word-level and sentence-level approaches. Three classification algorithms were applied at both levels and then combined to produce better results using both majority and minority voting procedures. In addition to the constructed Algerian dialect dataset, the experiments were also carried out using a set of Arabic dialects called PADIC. The study showed that Algerian dialects performed well at the word level using minority voting; however, better results occurred for PADIC at the sentence level. The average accuracy of Arabic dialects was 92%, while that of Algerian dialects was 76%.

Al-Tamimi et al. [37] collected Arabic comments from YouTube to build another dataset. They collected comments from videos popular in different Arab countries. After preprocessing this dataset, several algorithms were applied to classify the comments into positive, negative, or neutral. The experiments were then initiated, testing the dataset from different perspectives, including balanced and unbalanced for two classes and three classes. Beyond this, the study tested the dataset based on dataset types, including raw datasets, related datasets, normalized datasets, and related normalized datasets. The highest F1 results showed that SVM with Radial Bases Function outperformed KNN and BNB, achieving an average accuracy of 88% when using unbalanced 2- classes with normalized datasets.

Gamal et al. [38] extracted their Arabic Dialects Opinion Mining dataset from Twitter to classify tweets into positive or negative opinions. They manipulated the extracted data in a preprocessing phase by removing all redundant tweets to yield more streamlined results. Six classification algorithms, including BNB, NB, MNB, SGD, LR, and SVM, were applied, resulting in a high validation score of data classification. The highest accuracy occurred with SVM, reaching 93.5%.

Abuelenin et al. [39] built an Arabic corpus by collecting tweets and used 400 terms from the Arabic Slang Lexicon to classify these tweets. To minimize the limitation techniques of machine learning (ML) and semantic orientation (SO) and to enhance their proposed framework, they proposed a hybridize schema, consisting of preprocessing, feature extraction, lexical classifiers, machine learning classifiers, and evaluation. With different applied types of feature extraction techniques, they concluded that count-vectorizer had better results than TF-IDF and hashing-vectorizer. The overall accuracy resulting from applying Linear-SVM reached 92.98%.

Al Omari et al. [40] applied the LR approach with TF-IDF to group the Arabic collected comments as either negative, positive, or neutral. The dataset consisted of customer comments on various public services, which were collected from Google reviews and Zomato. The study show that the LR classifier produced promising results in SA when combined with positive prediction.

Table 1 summarizes this section by describing the size of each study dataset along with its type and class. The best classifier reflects the classification algorithm, which gives better results than other classifiers. However, this column sometimes is not BNB, MNB, nor LR. Therefore, this study was chosen because it considered one of our selected classification algorithms for comparison validity.

Journal of Theoretical and Applied Information Technology
15th October 2020. Vol.98. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

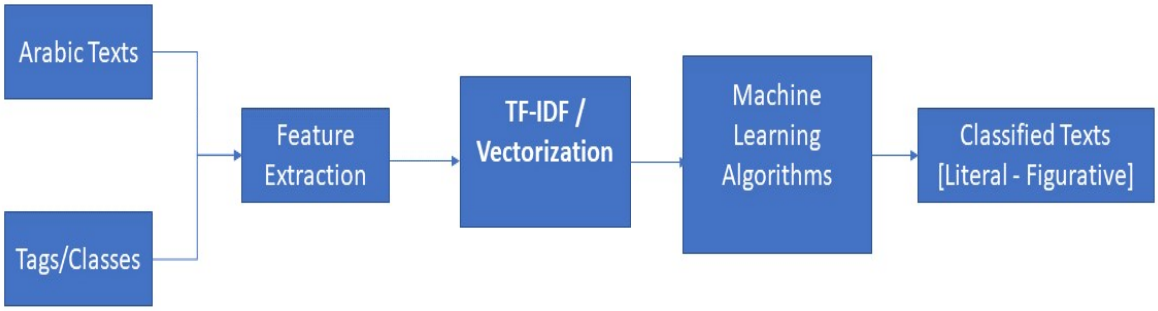

Figure 1 The Architecture Framework Of The Arabic Figurative Sentiment Analysis (AFSA)

3. CORPUS ANNOTATION

Here we present the corpus structure for the ALSA with a scheme based on an annotated figurative corpus we constructed manually. Figure 1 presents this proposed figurative approach, consisting of four steps: First, collect texts and build the corpus. Second, perform feature extraction. Third, apply Term Frequency/Inverse Document Frequency (TF-IDF) and Vectorization, and fourth, implement machine learning algorithms to classify the texts as literal or figurative. These steps are detailed further in subsequent sections.

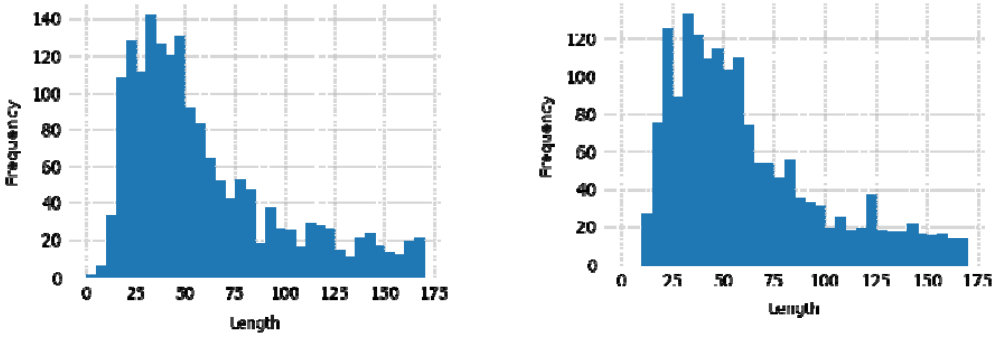

Figure 2 The Corpus (A) Without And (B) With Arabic

3.1 Text Cleaning and Prepossessing

Several techniques exist for cleaning texts, including:

- remove mentions of hashtags (e.g., #التعليم_والصحة) and links,
- remove unwanted characters, e.g., ignore non-alphanumeric characters,
- remove unnecessary whitespace and punctuations (e.g., '.','!', and '?'),
- remove copied texts in the corpus to keep a single occurrence of each word,
- remove excessively long words, such as "جمييييل جداااا" that is transformed into "جميل جدا",
- remove consecutive Tatweel ('ـ') within Arabic characters, and
- remove emojis and other special characters

Figure 2 illustrates the effects of removing Arabic stop-words using the Natural Language Tool Kit (NLTK) library, from nltk.corpus.stopword s.words("arabic"). The total corpus words decreased by 3%.

3.2 Examples of the Annotation Scheme

We compared the annotation of the annotators A1 and A2 on all 2000 texts, calculating that the two annotators achieved agreement for 1600 texts and disagreement in the remaining 400. To calculate Cohen's kappa coefficient, κ, for measuring interannotator agreements (IAA), we compared the annotation of A1 and A2 on all 1000 text. The agreement produced interesting results with the IAA between A1 and A2 concerning the choice between literal and figurative having $\kappa = 0.50$, i.e., a moderate agreement (see Table 3). In this study,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Cohen's kappa was calculated for A1 and A2 using the following equation:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

4. TEXT REPRESENTATION

Since computers remain unable to read text with the same level of comprehension as human beings, two fundamental word representation approaches allow for the transforming of words into a format that can be used in machine learning models (such as BOW and TF-IDF). Figure 3 illustrates a word-cloud containing the most common words literal and figurative classifications.

Table 2 Examples Of Corpus Texts Representing A Variety Of Figurative Categories

Figurative Categories	Count of Texts	Examples
Sarcasm	96	محل القهوة الجديد سيفتتح قريبا فى الشارع العام، هذه قصة جديدة! [mahalu alqahwat aljadid sayaftatih qaribaan fa alshsharie aleami, hadhih qisat jadyd!] [The new coffee shop will open soon on the main street, this is a new story!]
Literal	1000	حقوق المراه التي تضمنها لها وزاره العدل [huquq almarah alty tudaminuha laha wizaruh aleadl] [Women's rights guaranteed by the Ministry of Justice]
Hyperbole	637	الرحمن الرحيم [alruhmun alrahim] [Most Merciful]
Simile	267	مثلهم كمثل الذى استوقد نارا [mathaluhum kamathal aldhaa astawqad narana] [They are like a man who burned fire]

4.1 Bag-of-Words Featurization/Vectorization

The corpus text represented in a bag-of-words contains only the frequency of each word (i.e., word count) to measure how many times each word appears in the corpus regardless of word positions. The word representation of the text includes similar vectors for the same tokens or words, and the bag-ofwords approach provides a good representation of the text. The scikit-learn module of sklearn.feat ure extraction.text.CountVectorizer transforms the corpus texts into features vectors (i.e., [0 1 0 1 0 1 1 0 1 0]) as follows:

$$BoW = w_1 \cup w_2 \cup \ldots \cup w_n = \bigcup_{i=1}^{n} w_i \quad (2)$$

Table 3 Inter-Annotator Agreement (IAA)

Metrics	IAA
kappa	0.5081967213114753
Fleiss	0.5081967213114753
alpha	0.5128205128205129
scotts	0.48717948717948717

4.2. TF-IDF Featurization/Vectorization

These texts are represented with fixed length numeric word vectors comprised of zeros and ones.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The term frequency-inverse document frequency (TF-IDF) value represents the essential words in a corpus and is proportional to the frequency of the words. The TF-IDF also provides the normalized BOW word count as shown below:

$$tfidf_{i,a} = tf_{i,a} * \log\frac{N}{df_i} \quad (3)$$

where $tf_{i,a}$ represents the number of times a token *i* appears in document *d* divided by the total tokens in *d*, *N* represents the total documents found in the corpus, and df_i represents the number of texts containing the token or word *i*.

Table 4 . Performance Results Of The Literal And Figurative Analysis.

Model	Accuracy (%)	Recall (%)	F1-score
MultinomialNB	93%	92%	92%
LogisticRegression	93.25%	93.2%	93%
BernoulliNB	95%	94%	94%

5. EXPERIMENTS AND RESULTS

We built an annotated AFSA corpus and developed and applied machine learning algorithms to the annotated sentiment figurative language by focusing on the corpus containing 2000 texts. The corpus consisted of 1000 literal and 1000 figurative texts and further divided into 1000 positive and 1000 negative messages for training and testing.

Test data was extracted from the AFSA corpus with 1000 figurative sentences (e.g., hyperbole, sarcasm, and simile), along with the same number of non-figurative (e.g., literal) sentences to train our sentiment classification [5]. This approach to split the training and testing data ensured both an unbiased separation of the corpus and reliability of test results. We calculated the accuracy, recall, and F1- value metrics following the standard equations that follow:

$$F - value = \frac{2 * Accuracy * Recall}{Accuracy + Recall} * 100\% \quad (4)$$

$$Recall = \frac{The\ number\ of\ corrected\ texts\ of\ a\ specific\ class}{The\ number\ of\ texts\ of\ this\ class\ in\ testing\ data} * 100\% \quad (5)$$

$$Accuracy = \frac{The\ number\ of\ corrected\ texts\ in\ a\ specific\ class}{The\ number\ of\ texts\ in\ the\ class} * 100\% \quad (6)$$

Table 4 summarizes the performance of the AFSA sentiment analysis results for the classification algorithms Multinomial Naïve Bayes, logistic regression, and the Bernoulli Naïve Bayes in terms of the accuracy and F1-scores for the vectorization techniques of BOW and TF-IDF. Figure 4 presents the confusion matrix result for the logistic regression model. Based on these simulations and analyses, we observed that the vectorization using BOW provides better results than that using TF-IDF, suggesting the necessity of compiling the two complementary vectorization techniques of BOW and TF-IDF to improve results.

When comparing the accuracy, recall, and F1-scores for these machine learning models, the Bernoulli Naïve Bayes algorithm performed better than the Multinomial Naïve Bayes. However, each approach still suggested promising results for Arabic figurative sentiment analysis, but we cannot judge the training and testing data based solely on this variation.

Figure 3 (A) Literal Word-Cloud And (B) Figurative Word-Cloud.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

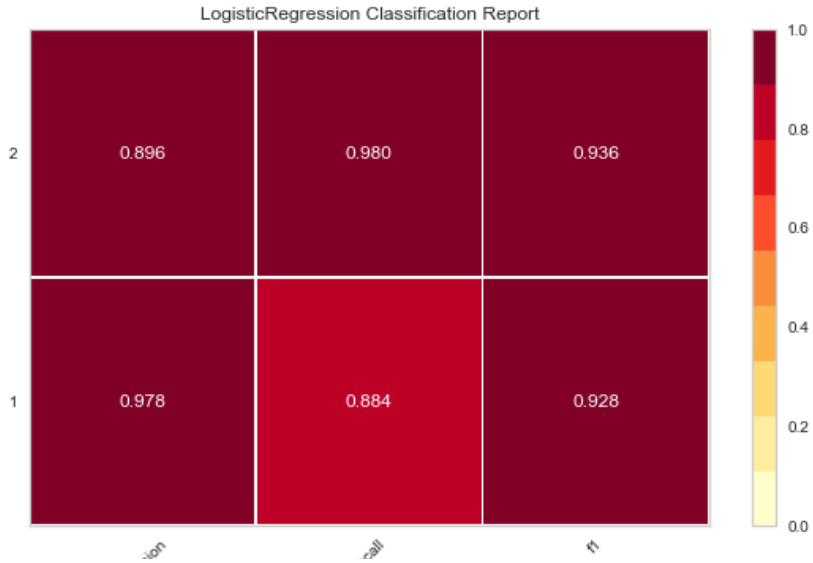

Figure 4 The Confusion Matrix Results For Logistic Regression.

6. CONCLUSIONS AND PERSPECTIVES

We presented the first attempt at modelling and analyzing the annotation of an Arabic Figurative Sentiment corpus based on manually collected text from the Holy Quran, Hadeeth, and publicly available tweets on Kaggle. This research concentrated on two different classes in Arabic literal and figurative language, which collectively contained 2000 texts (1000 literal and 1000 figurative). The experimental results for the figurative class achieved F1-scores of 92% for Multinomial Naïve Bayes, 93% for logistic regression, and 94% for Bernoulli Naïve Bayes. This result indicated that the classification algorithms may yield promising results when applied to figurative annotation.

Moreover, this work presented machine learning models that helped to analyze Arabic literal and figurative language classifications. They also provide a valuable resource for considering the sentiment meaning of figurative language in Arabic texts by creating this corpus. In addition, this study created a baseline for future work in annotating the aspect-based level of figurative Arabic texts in terms of hyperbole, sarcasm, metaphor, and simile. Finally, this work increased the dataset by focusing on more annotated words, which was expected to provide better results. Therefore, we plan to extend our work in this direction, based on the proposals in [3].

REFRENCES:

[1] N. S. Elmitwally and S. Alanazi, "Arabic Corpus for Figurative Sentiment Analysis", International Journal of Advanced Science and Technology, Vol. 29, No. 3, 2020, pp. 14.

[2] K. Elshakankery, and M. F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis", Egyptian Informatics Journal, Vol. 20, No. 3, 2019, pp. 163–171.

[3] A. Alsayat and N. Elmitwally, "A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications)", Egyptian Informatics Journal, Vol. 21, No. 1, 2019, pp. 7– 12.

[4] A. T. Cignarella, C. Bosco, V. Patti, and M. Lai, "Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRO", ` Proceedings of the Eleventh International Conference on Language Resources and Evaluation, May 7-12, 2018. 9

[5] D. Zhang, H. Lin, P. Zheng, L. Yang, and S. Zhang, "The Identification of the Emotionality of Metaphorical Expressions Based on a Manually Annotated Chinese Corpus", IEEE Access, Vol. 6, 2018, pp. 71241–71248.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[6] D. I. Hernandez Far ´ ´ıas, V. Patti, and P. Rosso, "ValenTO at SemEval-2018 Task 3: Exploring the Role of Affective Content for Detecting Irony in English Tweets", Proceedings of The 12th International Workshop on Semantic Evaluation, Louisiana, June 5-6, 2018, pp. 643–648.

[7] R. Ahmad, L. Torlakova, D. Liginlal, and R. Meeds, "Figurative Language in Arabic E-Commerce Text", International Journal of Business Communication, Vol. 57, No. 3, 2020, pp. 279–301.

[8] A. S. Manek, P. D. Shenoy, M. C. Mohan, and V. K. R., "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", World Wide Web, Vol. 20, No. 2, 2017, pp. 135–154.

[9] I. Guellil, A. Adeel, F. Azouaou, and A. Hussain, "SentiALG: Automated Corpus Annotation for Algerian Sentiment Analysis", In: Ren J. et al. (eds) Advances in Brain Inspired Cognitive Systems, Vol. 10989, 2018.

[10] J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles, and L. Hadrich-Belguith, "Towards a Contextual Pragmatic Model to Detect Irony in Tweets", in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), China, July, 2015, pp. 644–650.

[11] A. Ghosh et al., "SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Colorado, June, 2015, pp. 470–478.

[12] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, and P. Rosso, "Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA)", Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Italy, December 12-13, 2018, Vol. 2263.

[13] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, "Sentiment Analysis using SVM: A Systematic Literature Review", International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018, pp. 182–188.

[14] G. Imane, D. Kareem, and A. Faical, "A set of parameters for automatically annotating a Sentiment Arabic Corpus", International Journal of Web Information Systems, Vol. 15, No. 5, 2019, pp. 594–615.

[15] N. Malave and S. N. Dhage, "Sarcasm Detection on Twitter: User Behavior Approach", In: Thampi S. et al. (eds) Intelligent Systems, Technologies and Applications. Advances in Intelligent Systems and Computing, Vol. 910, 2020, pp. 65–76.

[16] H. L. Nguyen and J. E. Jung, "Statistical approach for figurative sentiment analysis on Social Networking Services: a case study on Twitter", Multimed Tools Appl, Vol. 76, No. 6, 2017, pp. 8901–8914.

[17] A S. A. AL-Jumaili, and H. K. Tayyeh, "A Hybrid Method of Linguistic and Statistical Features for Arabic Sentiment Analysis", Baghdad Science Journal, Vol. 17, No. 1 Supplement, 2020, pp. 385–390.

[18] O. Al-Harbi, "Classifying Sentiment of Dialectal Arabic Reviews: a Semi-Supervised Approach", The International Arab Journal of Information Technology, Vol. 16, No. 6, 2019, pp. 995–1002.

[19] M. Ahed, B. H. Hammo, and M. A. M. Abushariah, "An Enhanced Twitter Corpus for the Classification of Arabic Speech Acts", International Journal of Advanced Computer Science and Applications, Vol. 11, No. 3, 2020.

[20] M. Al-Smadi, M. Al-Ayyoub, and Y. Jararweh, "Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features", Information Processing & Management, Vol. 56, No. 2, 2019, pp. 308–319.

[21] H. M. Al-Barhamtoshy, H. T. Hemdi, M. M. Khamis, and T. F. Himdi, "Semantic and Sentiment Analysis for Arabic Texts Using Intelligent Model", Biosc. Biotech. Res. Comm., Vol. 12, No. 2, 2019.

[22] M. M. Eldefrawi, D. S. Elzanfaly, M. S. Farhan, and A. S. Eldin, "Sentiment analysis of Arabic comparative opinions", SN Applied Sciences, Vol. 1, No. 5, 2019, pp. 411.

[23] M. A. H. Madhfar, and M. A. H. Al-Hagery, "Arabic Text Classification: A Comparative Approach Using a Big Dataset", 2019 International Conference on Computer and Information Sciences, Saudi Arabia, 2019, pp. 1–5. 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[24] M. Jarrar, F. Zaraket, R. Asia, and H. Amayreh, "Diacritic-Based Matching of Arabic Words", ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 18, No. 2, 2018.

[25] J. O. Atoum, and M. Nouman, "Sentiment Analysis of Arabic Jordanian Dialect Tweets", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019, pp. 256–262.

[26] I. Abu Farha, and W. Magdy, "From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset", Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, France, May, 2020, pp. 32–39.

[27] M. E. M. Abo, R. G. Raj and A. Qazi, "A Review on Arabic Sentiment Analysis: State-ofthe-Art, Taxonomy and Open Research Challenges", IEEE Access, Vol. 7, 2019, pp. 162008–162024.

[28] I. Guellil, F. Azouaou, and M. Mendoza, "Arabic sentiment analysis: studies, resources, and tools", Social Network Analysis and Mining, Vol. 9, No, 1, 2019, pp. 56.

[29] A. Ghallab, A. Mohsen, and Y. Ali. "Arabic Sentiment Analysis: A Systematic Literature Review", Applied Computational Intelligence and Soft Computing, Vol. 2020, 2020.

[30] L. A. Qadi, H. E. Rifai, S. Obaid, and A. Elnagar, "Arabic Text Classification of News Articles Using Classical Supervised Classifiers", 2nd International Conference on new Trends in Computing Sciences, Amman, Jordan, 2019, pp. 1–6.

[31] J. Zahir, Y. M. Oukaja, and H. Mousannif, "Author Gender Identification from Arabic Youtube Comments", 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems, Sorrento, Italy, 2019, pp. 672–676.

[32] C. Qwaider, S. Chatzikyriakidis, and S. Dobnik, "Can Modern Standard Arabic Approaches be used for Arabic Dialects? Sentiment Analysis as a Case Study", Proceedings of the 3rd Workshop on Arabic Corpus Linguistics, Cardiff, United Kingdom, 2019, pp. 40–50.

[33] A. Omar, T. M. Mahmoud, and T. Abd-ElHafeez, "Comparative Performance of Machine Learning and Deep Learning Algorithms for Arabic Hate Speech Detection in OSNs", In: Hassanien AE., Azar A., Gaber T., Oliva D., Tolba F. (eds) Proceedings of the International Conference on Artificial Intelligence and Computer Vision, Vol. 1153, 2020, pp. 247–257.

[34] B. Sadik and R. Aberkane, "Subjective Sentiment Analysis for Arabic Newswire Comments", Journal of Digital Information Management, Vol. 17, No.5, 2019, pp. 289–295.

[35] N. A. AlSomaikhi and Z. A. Alzamil, "Twitter Users' Classification Based on Interest: Case Study on Arabic Tweets", International Journal of Information Retrieval Research, Vol. 10, No. 1, 2020, pp. 1–12.

[36] M. Lichouri, M. Abbas, A. A. Freihat, and D. E. H. Megtouf, "Word-Level vs Sentence-Level Language Identification: Application to Algerian and Arabic Dialects", Procedia Computer Science, Vol. 142, 2018, pp. 246–253.

[37] A. Al-Tamimi, A. Shatnawi, and E. Bani-Issa, "Arabic sentiment analysis of YouTube comments", IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, Aqaba, Jordan, 2017, pp. 1–6.

[38] D. Gamal, M. Alfonse, E. M. El-horbaty, and A. M.Salem, "Opinion mining for Arabic dialects on twitter", Egyptian Computer Science Journal, Vol. 42, No. 4, 2018, pp. 52–61,

[39] S. Abuelenin, S. Elmougy, and E. Naguib, "Twitter Sentiment Analysis for Arabic Tweets", In: A. E. Hassanien et al. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Vol. 639, 2017, pp. 467–476.

[40] M. Al Omari, M. Al-Hajj, N. Hammami and A. Sabra, "Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon," International Conference on Computer and Information Sciences, Sakaka, Saudi Arabia, 2019, pp. 1–5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

APPENDIX:

Table 1 Summarization Of The Related Studies In Subsection 2.2

Study	Dataset		Feature Extraction	Classes	Best Classifier		Classifiers Compared (only if MNB, LR, or BNB)	
	size	source			Name	performance	Name	performance
[30]	89,189	beINSPORTS.COM, TECH-WD.COM, SKYNEWSARABIC.COM, ARABIC.RT.COM, CNBCARABIA.COM, ARABIC.CNN.COM, YOUM7.COM	TF-IDF, and Count	Business, sports, technology, and Middle East	SVM	Prec= 98% Rec= 98% F1= 98%	LR	Prec= 98% Rec= 98% F1= 98%
							MNB	Prec= 96% Rec= 96% F1= 96%
	46,900	Akhbarona		Finance, sports, culture, technology, polotics, medical, and religion	SVM	Prec= 94% Rec= 94% F1= 94%	LR	Prec= 94% Rec= 94% F1= 94%
							MNB	Prec= 91% Rec= 88% F1= 88%
[31]	27,929	YouTube comments	TF-IDF	Male, and female	NMB	Acc= 75.2%	-	-
[32]	2242	Shami-Senti	TF, Unigram + Bigram, and Feature Engineering	Positive, negative, and neutral	MNB	Acc= 75.2%	-	-
	19,738	LARB3-Balanced		Positive, negative, and neutral	-	-	MNB	Acc= 58.2
[33]	20,000	Facebook, Twitter, Instagram, and YouTube	N/A	Hate, and not hate	Complement NB	Acc= 97.59% Prec= 97.63% F1= 97.59% Rec= 97.59%	MNB	Acc= 97.52% Prec= 97.55% F1= 97.52% Rec= 97.52%
							BNB	Acc= 96.19% Prec= 96.38% F1= 96.18% Rec= 96.19%
							LR	Acc= 97.54% Prec= 97.55% F1= 97.54% Rec=

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

								97.54%
[34]	63,055	USERS COMMENTS FROM ECHOROUK, AN ONLINE ALGERIAN NEWSPAPER	UNIGRAM + BIGRAM	POSITIVE, NEGATIVE, AND NEUTRAL	MNB	ACC= 65.64% PREC= 67% F1= 64% REC= 66%	-	-
[35]	140,841	TWITTER	BIGRAM	SPORT, RELIGION, TECHNOLOGY, HEALTH, ECONOMY, AND LITERATURE	MNB	PREC= 91% F1= 80%	-	-
[36]	793	ALGERIAN DIALECTS	TF-IDF	8 DIFFERENT ALGERIAN CITIES	-	-	BNB	ACC= 25.82%
							MNB	ACC= 33.11%
	6000	PADIC		MODERN STANDARD ARABIC WITH 5 DIFFERENT ARABIC COUNTRIES	BNB	ACC= 73.15%	MNB	ACC= 72.61%
[37]	5986	YOUTUBE COMMENTS	BIGRAM + UNIGRAM + TF-IDF	POSITIVE, NEGATIVE, AND NEUTRAL	SVM WITH UNBALANCED 2-CLASS (NORMALIZED)	AVG. F1= 88% AVG. PREC= 88% AVG. REC= 88%	BNB	AVG. F1= .859 AVG. PREC= .859 AVG. REC= .859
[38]	151,548	TWITTER	TF	POSITIVE AND NEGATIVE	SVM	ACC= 93.57%	BNB	ACC= 85.8%
							MNB	ACC= 89.44%
[39]	1560	TWITTER	TF-IDF, COUNT, HASHING	POSITIVE, NEGATIVE, AND NEUTRAL	SVM	ACC= 92.98%	BNB	ACC= 81.50%
							MNB	ACC= 78.94%
[40]	3916	GOOGLE REVIEWS, AND ZOMATO	TF-IDF	POSITIVE, AND NEGATIVE	LR	MICRO AVG.: PREC= 0.88 REC= 0.88	N/A	N/A