Sentiment Analysis of Libyan Dialect Using Machine Learning with Stemming and Stop-words Removal

Abdullah Habberrih[1], Mustafa Ali Abuzaraida[1]

[1] Computer Science Department, Faculty of Information Technology, Misurata University, Libya.

***Abstract*— This study evaluates the impact of using Stemming and Stop-words removal techniques on machine learning classifiers, Support Vector Machine (SVM) and Logistic Regression (LR), in detecting sentiment from Libyan dialect poetry. The lack of Arabic Natural Language Process resources has made sentiment analysis for Arabic a challenging task compared to other languages. A secondary dataset was used and two experiments were conducted, with the first exploring the use of Stemming with Stop-words removal techniques and the second investigating the impact of using Stemming alone. Other preprocessing techniques were applied alongside TF-IDF with a combination of Unigrams and Trigrams during feature extraction. The results show that the Stop-words removal technique may have a negative impact on classifier performance. SVM outperformed LR in both experiments, achieving an accuracy of 71.63%, while LR achieved 70.92% in the second experiment. This study's accuracy outperformed previous research on the topic, achieving 71.63%, compared to 69% in earlier studies.**

***Index Terms*—Sentiment Analysis, Arabic Dialects, Machine Learning, TF-IDF, Stemming.**

I. INTRODUCTION

Sentiment Analysis (SA) is one of the Natural Language Processing (NLP) techniques used to determine the polarity of a text, whether it is positive or negative (Omar et al., 2022). SA is extensively utilized to automatically extract people's opinions on various subjects, such as events, services, products, and more, expressed in different forms such as articles, reviews, forum posts, tweets, and short remarks (Altawaier & Tiun, 2016). The significance of sentiment analysis lies in its ability to derive conclusions and make indirect inferences from a vast amount of data (Elgeldawi et al., 2021). Analyzing texts written in a language with complex morphology, such as Arabic, has always been a challenging and multi-level process. The process of Sentiment Analysis (SA) involves several stages, including data collection, data annotation, pre-processing, feature extraction, sentiment detection, and classification. Pre-processing is deemed the most critical step in analyzing sentiment, especially for messages obtained from social networks. This is due to the fact that such messages typically contain informal language, abbreviations, emoticons, elongated words, and inconsistent capitalization, which do not adhere to standard grammatical rules (Nassr et al., 2020). According to (Alshutayri & Atwell, 2017), Arabic is the fifth most commonly spoken language worldwide. The Arabic language comprises three primary variations: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) (Alyami et al., 2022). CA is the language used for writing the holy book of Islam, while MSA is utilized in books, politics, journalism, and education. DA, on the other hand, is an informal form of Arabic extensively used for day-to-day communication (Alyami et al., 2022). SA encounters several challenges when working with the Arabic language and its dialect, which have been discussed in (Oueslati et al., 2020) and a study focused on Saudi dialects, as described in (Alwakid, 2020). One of the primary obstacles in dealing with dialectical Arabic is the lack of standard grammatical rules. Arabic sentences can start with a noun phrase, verb, or nominal phrase, and there are numerous syntactic variations within all sentence types. Moreover, dialect natives may express their opinions using different dialects and employ slang words and abbreviations. Additionally, the repetition of letters may be used to convey emotions and emphasis, resulting in spelling errors. The city of Misurata, also known as the two-beach city and the sand city, is located on the Mediterranean coast of northern Libya and is the country's third-largest city, situated approximately 130.49 miles east of the capital, Tripoli. One of the sub-dialects spoken in Libya is native to this city. The objective of this study is to perform sentiment analysis on poems written in the Misurata sub-dialect of the Libyan language. To achieve this, the study will compare the performance of two machine learning classifiers, namely Support Vector Machine (SVM) and Logistic Regression (LR). Additionally, the study will employ various preprocessing techniques, along with the use of TF-IDF and N-grams as feature extraction techniques. The rest of this paper is structured as follows: Section 2 offers a concise review of the literature, while section 3 details the methodology. Section 4 presents the experimental results and the corresponding discussion. Section 5 discuss how this study can contribute to society and lastly, section 6 concludes the study.

II. LITERATURE REVIEW

Elgeldawi et al. conducted two studies related to Arabic sentiment analysis. In their first study published in (Sayed et al., 2020), they proposed a framework to enhance the accuracy of

DOI:http://doi.org/10.24086/cocos2024/paper.1171

sentiment classification in Arabic reviews using nine machine learning classifiers, including Ridge, Gradient Boosting, and Multi-layer Perceptron, and collected 6318 hotel reviews from Booking.com for evaluation. The second study published in (Elgeldawi et al., 2021) investigated the effects of hyperparameter tuning on Arabic sentiment analysis using six machine learning algorithms. They used a hotel review dataset named the Reviews Sentiment Analysis Corpus (RSAC), which contained 7000 reviews in both Modern Standard Arabic (MSA) and Dialectal Arabic (DA). In both studies, the authors used preprocessing techniques such as normalization, stemming, and stop-word removal, and the TF-IDF method for feature extraction. The 2020 study evaluated classifiers based on accuracy, with the Ridge Classifier achieving the best performance. The 2021 study compared five hyperparameter tuning algorithms and found that the SVM classifier with Bayesian Optimization achieved the best accuracy of 95.62%. Another study in 2021, Abugharsa conducted a study to compare the performance of Machine Learning (ML) methods and Deep Learning (DL) techniques on Misurata sub-dialect poems. Four classification algorithms were used in ML, including Logistic Regression classifier, Random Forest classifier, Naïve Bayes (NB) classifier, and Support Vector Machine (SVM) classifier. On the other hand, Mazajak, a tool built on a CNN algorithm and designed to detect sentiment from Modern Standard Arabic (MSA), was used for DL. The results indicated that the ML classifiers were more efficient than the Mazajak classifiers based on their accuracy scores. The ML classifiers achieved a score of 68.0%, while the Mazajak model scored 60.66%. In (Omar et al., 2022) conducted a study that aimed to determine customer opinions of three major Libyan telecommunication companies, namely Libyana, Almadar Aljadid, and Libya Telecom and Technology, using ML approach. The study utilized a dataset collected from Twitter that was annotated into Positive and Negative sentiment. To preprocess the data, several techniques were applied, including data cleaning to remove non-Arabic characters, usernames, hashtags, punctuation marks, and numbers, as well as tokenization, normalization, stop-word removal, and lemmatization. The feature extraction technique used was TF-IDF. To train the model, the dataset was divided into 80% training and 20% testing, and tested on five classifiers - SVM, LR, NB, KNN, and DT. Three experiments were performed, the first two trained the classifiers without and with lemmatization, and the third experiment included under-sampling the corpus to balance between the two classes. The findings of the study indicated that the SVM classifier had the highest accuracy in predicting customer sentiment for Libyana, achieving 80.67%. For Almadar Aljadid, the NB classifier had the highest accuracy at 81.19%, while for Libya Telecom and Technology, the DT classifier achieved the highest accuracy of 75%. In the same year, (Aljwari, 2022) Aljwari published a paper that detects the emotions in Arabic text using five ML algorithms, which are: DT, KNN, NB, Multinomial NB, and SVM. The dataset used is called SemEval-2018 dataset from Twitter and it is publicly avaliable, also it contains 934 tweets with four classes: Joy, Anger, Sadness, and Fear. The tweets were preprocessed by applying Normalization technique and removing the stop-words, punctuation marks, repeating characters, mentions, non-Arabic characters, and Arabic diacritics from the text, however, TF-IDF method were used in the feature extraction step. To build the classifier, the dataset was splitted into two parts: training set and testing set with 80-20 ratios, which mean that the training set contains 747 samples while the testing set contains 187 samples, therefore, the classifier was evaluated by utilizing precision, accuracy recall, and F-score metrics, and the results show that DT, and KNN classifiers scores better performance with 74% accuracy, then NB and Multinomial NB classifiers achieved 69% accuracy, while the SVM classifier obtained 63%.

III. Methodology

While the previous section discussed and presented some previous studies, the current section will focus on outlining the proposed methodology. Figure 1 showed the methodology of this study.

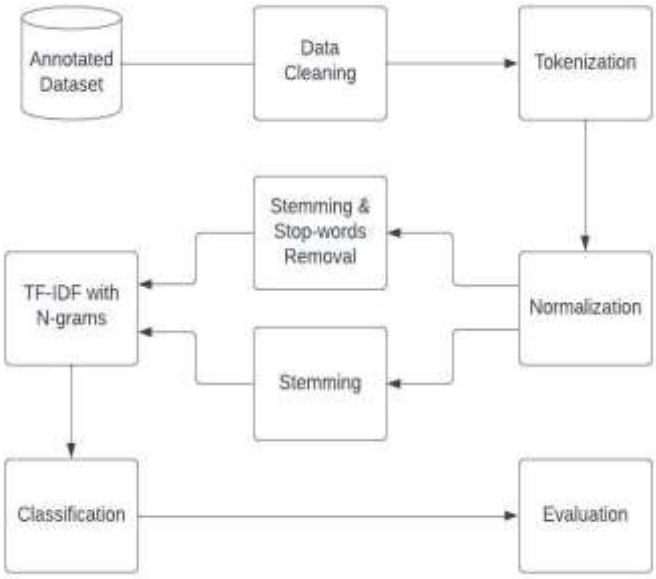

Fig 1 The Methodology of this Study

3.1. Dataset

The study utilizes a secondary dataset, which the author collected and annotated in (Abugharsa, 2021). The dataset concentrates on Misurata sub-dialect poems and comprises 22,762 records that were labeled as positive or negative. Table 1 provides a description of the dataset.

Table 1 Dataset Description

Columns	Sentiment	Sentence
count	22,762	22,762
unique	2	21,387
top	positive	حاجة غريبة
freq	13,216	20

DOI:http://doi.org/10.24086/cocos2024/paper.1171

Moreover, Table 2 shows the number of samples for each class.

Table 2 Number of Samples in The Dataset

Class	Number of Samples
Positive	13,216
Negative	9,546
Total	22,762

Additionally, Table 3 highlights some of the samples of the positive class and the negative class.

Table 3 Examples of The Dataset

Sentiment	Sentence (English)	Sentence
Positive	My heart laughed and my soul fluttered	ضحكت ضحك قلبي ورفرف طاير
Positive	Whoever sees her swears that there are good omens	اللي شوفها نحلف اتقول بشاير
Positive	Her laughter is blessed by God's eyes	ضحكت الله ايعينها
Positive	My whole being has become forever captivated by her	روحي و كلي صرت دوم سجينها
Negative	And I know, I know that it will all disappear	و عارف و اني عارف كلها تغبى
Negative	Just like you and I, my beloved	بحالك و حالي يا عزيز عليا
Negative	Longing overpowers me and stirs up chaos	غالب عليا الشوق هييج شغبه
Negative	My inspiration and news about hidden matters	إلهامي و خبر عن أمور خفية

Data Preprocessing

In order to enhance coherence and facilitate data handling, data preprocessing is a necessary step that involves normalizing data into a consistent form (Abugharsa, 2021). The nature of the language and the analysis objectives determine the various stages of this step. As social media text is often unstructured or noisy, due to a lack of standardization, spelling errors, missing punctuation, non-standard words, and repetitions, pre-processing the text has become increasingly important, especially with the rise of websites generating such text (Alyami et al., 2022). The pre-processing process primarily consists of three steps: normalization, stemming, and stop-words removal (Shoukry & Rafea, 2012). This study employs the following preprocessing techniques:

- **Data cleaning:** In order to cleanse the dataset, empty lines, diacritics, punctuation and were eliminated. As shown in Table 1, there are 1,375 duplicated rows in the dataset that must be eliminated, keeping only the first instance of each duplicated row.
- **Stemming:** It is the process of reducing words to their uninflected base forms, which can be achieved through light and root stemming. Although the stem may differ from the root, stemming is useful as related words can often be mapped to the same stem, even if it is not a valid root (Shoukry & Rafea, 2012).
- **Tokenization:** In this phase, sentences or rows are subdivided into words or tokens using a delimiter, which could be any punctuation character or whitespace (Sayed et al., 2020).
- **Normalization:** Achieving consistency and standardizing the form of the text is the aim of text transformation or Normalization (Shoukry & Rafea, 2012). As an example, (آ إ أ) letters are replaced with (ا), and (ة) at the end of a word is substituted with (ه) during this process.
- **Stop-words Removal:** Stop words are words that frequently occur in a text but hold no significant semantic relation to their context (Shoukry & Rafea, 2012). Examples of such words are "متى" (When), "علاش؟" (Why?), "اني" (Me), "هو" (Him), and similar words.

Feature Extraction

To improve the efficiency and effectiveness of the analysis, feature extraction and selection is the next step after data preprocessing. During this phase, the most effective features for the sentiment analysis process are identified, and irrelevant, redundant, and noisy data is eliminated. This step serves to reduce the dimensionality of the feature space and processing time (Sayed et al., 2020). Surface features, such as n-grams and syntactic features, are the most commonly used features in machine learning sentiment analysis (Alwakid, 2020). The feature extraction techniques that utilized in this study as following:

- **Term Frequency-Inverse Document Frequency (TF-IDF):** It is a technique that consists of two main components Term Frequency (TF) and Inverse Document Frequency (IDF). It can be defined as the following equations:

$$tfidf(w,d,D)=tf(w,d)\times idf(w,D) \quad (1)$$

$$tf(w,d)=fd(w):frequency\ of\ w\ in\ document\ d \quad (2)$$

$$idf(w,D)=log(1+|D|1+df(d,w)) \quad (3)$$

In equation (2), TF (w, d) signifies the frequency of specific words (w) within a given document (d). The total value of TF-IDF (1) is calculated by dividing the number of documents in the corpus by the number of times w is repeated within the corpus for IDF (w, D). IDF (3) is a numerical statistical identifier used to represent the significance of a word in a document within a collection or corpus of work (Alwakid, 2020).

- **N-grams:** It involves subdividing text into contiguous sequences of n items, where n can be any positive integer. The three most common types of N-grams are Unigrams, Bigrams, and Trigrams. Unigrams, also known as Bag of Words (BOW), are the simplest N-gram features to extract, providing good coverage of

DOI:http://doi.org/10.24086/cocos2024/paper.1171

the data. On the other hand, bigrams and trigrams enable the capture of negation or sentiment expression patterns, allowing for a more nuanced analysis of the text sentiment (Shoukry & Rafea, 2012).

IV. Result & Discussion

In this study, two experiments were conducted to evaluate the effects of various preprocessing techniques on the performance of two machine learning classifiers, Support Vector Machine (SVM) and Logistic Regression (LR). The first experiment examined the impact of using both Stemming and Stop-words removal techniques, while the second experiment focused solely on the use of Stemming technique without Stop-words removal. Other preprocessing techniques were applied consistently across both experiments, and both utilized TF-IDF with a combination of Unigrams and Trigrams. The performance of both classifiers was evaluated using four metrics: Accuracy, Precision, Recall, and F1-score. During the training of the classifiers, a 10-fold cross-validation was implemented, which involved dividing the data into training and testing sets. This approach was utilized to avoid overfitting or underfitting issues that may arise from the conventional 80% train and 20% test split. Table 4 presents the results of the first experiment, which focused on evaluating the impact of using both Stemming and Stop-words removal techniques.

Table 4 First Experiment Results

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	69.19	69.65	83.64	76.00
LR	68.61	69.72	81.67	75.22

The findings presented in Table 4 indicate that SVM outperformed LR across all evaluation metrics, with an accuracy of 69.19%, precision of 69.65%, recall of 83.64%, and f1-score of 76.00%, while LR achieved an accuracy of 68.61%, precision of 69.72%, recall of 81.67%, and f1-score of 75.22%. It's worth noting that SVM's f1-score showed a noticeable improvement of +0.78% compared to LR. Figures 2, and 3 show the confusion metrics of the first experiment.

Fig 2 Confusion Matrix of SVM in First Experiment.

Fig 3 Confusion Matrix of LR in First Experiment.

Furthermore, Table 5 displays the outcomes of the second experiment, which aimed to investigate the effect of using Stemming technique alone, without Stop-words removal technique, as previously mentioned.

Table 5 Second Experiment Results

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	71.63	72.00	84.02	77.54
LR	70.92	71.30	70.56	70.92

Table 4.2 demonstrates that SVM surpassed LR in all evaluation metrics, achieving an accuracy of 71.63%, precision of 72.00%, recall of 84.02%, and f1-score of 77.54%, while LR achieved an accuracy of 70.92%, precision of 71.30%, recall of 70.56%, and f1-score of 70.92%. It is notable that SVM's f1-score displayed a considerable improvement of +6.62% compared to LR. Figures 4, and 5 present the confusion metrics of the second experiment.

Fig 4 Confusion Matrix of SVM in Second Experiment.

DOI:http://doi.org/10.24086/cocos2024/paper.1171

TARGET / OUTPUT	Positive	Negative
Positive	631 35.47%	254 14.28%
Negative	263 14.78%	631 35.47%

Fig 5 Confusion Matrix of LR in Second Experiment.

It is noteworthy that the second experiment resulted in an improvement in the classifiers' performance in comparison to the first experiment. This improvement could be attributed to the implementation of the Stop-words removal technique during the preprocessing phase in the first experiment. However, it is worth noting that while this technique is commonly employed to improve classifier performance, its effectiveness is not always guaranteed and may even lead to a decrease in performance, as observed in the results of this study. Furthermore, Figure 6 displays the experimental results of this study.

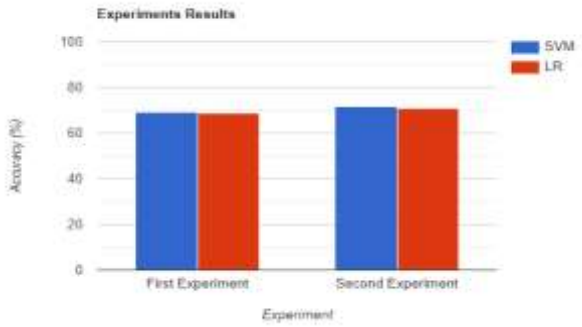

Fig 6 Study Experiments

Furthermore, in this study, a combination of Unigrams and Trigrams was utilized instead of using them separately. This approach was adopted because combining Unigrams and Trigrams can extract a greater number of features in terms of N-grams, allowing for a more comprehensive analysis of the text. By capturing both basic and complex language patterns, this approach leads to a more accurate sentiment analysis than using only Unigrams or Trigrams. Additionally, it is worth noting that Trigrams can sometimes create unfavorable feature dependencies, which makes it challenging to detect sentiment in shorter text samples, such as the dataset used in this study. Additionally, the results of this study demonstrated better performance of the ML classifiers compared to the results reported in (Abugharsa, 2021), with an accuracy of 71.63% for the SVM classifier and 69% for NB in (Abugharsa, 2021), which could be attributed to several factors. Firstly, the previous study (Abugharsa, 2021) did not utilize Stemming techniques during the preprocessing phase. Secondly, (Abugharsa, 2021) applied the Stop-word removal technique without determining its impact on the poems dataset, whereas this study found that the classifier's performance was enhanced when the Stop-word removal technique was not used (see Table 5). Finally, (Abugharsa, 2021) did not employ N-grams techniques, while this study used a combination of Unigrams and Trigrams, which ultimately enhanced the classifier's performance. However, it is important to note that the domain of poems presents unique challenges compared to other domains, such as movies, reviews, politics, and sports. This is supported by the literature review conducted in this study and is reflected in the experimental results, which are consistent with those reported in (Abugharsa, 2021). Moreover, Figure 7 shows the comparison results between this study and study (Abugharsa, 2021).

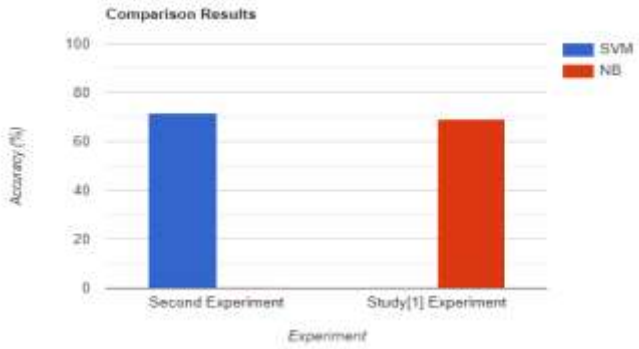

Fig 7 Comparison of this Study and Study (Abugharsa, 2021)

Contribution to Society

This study on sentiment analysis of Libyan dialect using machine learning with stemming and stop-words removal holds great promise for society. By leveraging Machine Learning techniques, this research can provide valuable insights into the sentiments and attitudes of the Libyan population. This understanding of public sentiment can significantly benefit policymakers, enabling them to make informed decisions and develop targeted strategies that align with the desires and concerns of the people. Moreover, in a region like Libya, where conflicts and tensions persist, this study can aid in effective crisis management and conflict resolution by identifying underlying sources of dissatisfaction and facilitating dialogue between different factions. Additionally, businesses operating in Libya can utilize the findings to better understand customer feedback, improve their products or services, and ultimately enhance customer satisfaction. Overall, this study has the potential to contribute to a more informed, harmonious, and prosperous society in Libya.

V. Conclusion

Sentiment analysis has gained significant popularity and become a speculative industry over the last decade. However, analyzing sentiment in Arabic language and its dialects still poses several challenges. In this study, the impact of Stemming

DOI:http://doi.org/10.24086/cocos2024/paper.1171

and Stop-words removal techniques on the performance of machine learning classifiers in detecting sentiment from dialect poetry was explored. Two experiments were conducted on SVM and LR classifiers. In the first experiment, the impact of using both Stemming and Stop-words removal techniques in the preprocessing step was investigated, while the second experiment explored the impact of using Stemming without the Stop-words removal technique. During the feature extraction step, TF-IDF with a combination of Unigrams and Trigrams was applied. The results of the experiments showed that the second experiment outperformed the first experiment in terms of the Stop-words removal technique. Furthermore, the SVM classifier achieved the highest accuracy compared to LR in both experiments, with 71.63%. Interestingly, the results of this study outperformed those reported in (Abugharsa, 2021), suggesting that the use of the Stop-words removal technique in the Libyan poetry domain can have a negative impact on machine learning classifiers. For future work, the authors aim to explore the impact of deep learning techniques in detecting sentiment from Misurata sub-dialect poetry and compare the results with those of machine learning classifiers.

REFERENCES

Abugharsa, A. (2021). Sentiment Analysis in Poems in Misurata Sub-dialect--A Sentiment Detection in an Arabic Sub-dialect. *ArXiv Preprint ArXiv:2109.07203*.

Aljwari, F. (2022). Emotion Detection in Arabic Text Using Machine Learning Methods. *IJISCS (International Journal of Information System and Computer Science)*, *6*(3), 175–185.

Alshutayri, A. O. O., & Atwell, E. (2017). Exploring Twitter as a source of an Arabic dialect corpus. *International* Journal *of Computational Linguistics (IJCL)*, *8*(2), 37–44.

Altawaier, M., & Tiun, S. (2016). Comparison of machine learning approaches on arabic twitter sentiment analysis. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 1067–1073.

Alwakid, G. (2020). *Sentiment analysis of dialectical Arabic social media content* using *a hybrid linguistic-machine learning approach*. Nottingham Trent University (United Kingdom).

Alyami, S., Alhothali, A., & Jamal, A. (2022). Systematic literature review of arabic aspect-based sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, *34*(9), 6524–6551.

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. Informatics, *8*(4), 79.

Nassr, Z., Sael, N., & Benabbou, F. (2020). Preprocessing arabic dialect for sentiment mining: State of art. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *44*, 323–330.

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

Oueslati, O., Cambria, E., HajHmida, M. Ben, & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, *112*, 408–430.

Sayed, A. A., Elgeldawi, E., Zaki, A. M., & Galal, A. R. (2020). Sentiment analysis for arabic reviews using machine learning classification algorithms. *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, 56–63.

Shoukry, A., & Rafea, A. (2012). Preprocessing Egyptian dialect tweets for sentiment mining. *Fourth Workshop on Computational Approaches to Arabic-Script-Based Languages*, 47–56.

DOI:http://doi.org/10.24086/cocos2024/paper.1171