مجلة جامعة بابل للعلـــوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com
ISSN: 2312-8135 | Print ISSN: 1992-0652

Automating Twitter Data Annotation Process for Sentiment Analysis

Hasanein Alharbi
College of Information Technology, University of Babylon, hasanein.alharbi@uobabylon.edu.iq, Hilla , Iraq.
*Corresponding author email: hasanein.alharbi@uobabylon.edu.iq; mobile: 07730439040

أتمتة عملية تصنيف بيانات تويتر لأغراض تحليل المشاعر

حسنين يعرب محمد الحربي
كلية تكنولوجيا المعلومات، جامعة بابل، hasanein.alharbi@uobabylon.edu.iq ، بابل، العراق

Accepted: 27/10/2025 **Published:** 31/12/2025

ABSTRACT

Background:
Sentiment analysis algorithms require high-quality annotated data during the training phase. However, this requirement has led to complex, time-consuming and costly manual data annotation process. To address these challenges, this research proposes an automatic data annotation process for sentiment analysis.

Materials and Methods:
Three semantic orientation measures (Pointwise Mutual Information, latent Semantic Analysis, and Word2Vec), five classification algorithms (K-Nearest Neighbors, Logistic Regression, naïve Bayes, Random Forest, Support Vector Machine) and NRC lexicon thesaurus are used to automate the process of tweet annotation for sentiment analysis.

Results:
Tweets were annotated using five classifiers and three semantic measures, forming fifteen combinations. The Inter-Annotator Agreement (IAA) among these combinations was evaluated using Cohen's Kappa statistic. The obtained results show that (Pointwise Mutual Information + Logistic Regression) and (Pointwise Mutual Information + Naïve Bayes) achieved the highest agreement score of 0.7008.

Conclusion:
These results have shown that the corpus-based semantic orientation measures have provided substantive results. However, it can still be enhanced through the use of a broader vocabulary, the application of contextual information and the implementation of the newest deep learning algorithms.

Key words:
Sentiment Analysis, Machine Learning, Semantic Similarity, NRC Lexicon Thesaurus.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

INTRODUCTION

Twitter has been described by social scientists as a social telescope or a huge digital antenna due to its ability to gather huge data with a view to a variety of issues. As of October 2022, it was estimated that the Twitter platform had gathered a substantial volume of approximately 900 billion public tweets [1].

The vast volume of Twitter data presents an exciting opportunity for sentiment analysis (SA). SA involves techniques to extract users' sentiments, opinions, and perspectives. Among the various SA techniques, supervised machine learning algorithms have gained significant interest. However, the application of supervised machine learning algorithms to classify tweets into different sentiment categories required high quality annotated data during the training phase. Therefore, challenges are raised due to the labour-intensive manual data annotation process and the informal nature of the ever-evolving tweet data [2].

The process of annotating words with their appropriate emotion category has been heavily investigated. For example, [3] developed the NRC word-emotion association lexicon. The NRC lexicon is comprised of approximately 42,200 word-emotion pairs. Each NRC word was associated with one of the eight elected emotion categories using Amazon Mechanical Turk [3], [4].

Although the NRC lexicon thesaurus presents numerous advantages, practical analysis reveals that it covers only a restricted portion of the terms used in social media language [5]. Therefore, this article explores the potential advantages of integrating the NRC lexicon with various semantic similarity measures [6] and machine learning algorithms to automate the labelling process. The main goal of this article is to discuss the drawbacks of the existing method of manually annotating the tweets and possibly make the sentiment-analysis algorithms more accurate. The research of this paper is as follows:

1. It suggests a completely automated data-annotation approach, which is lexical-based.
2. It combines machine-learning algorithms, NRC lexicon and semantic similarity measures.

LITERATURE REVIEW

Sentiment analysis of Twitter data has been annotated by many scholars. An example is the case of [7] where the annotation of tweet data was done manually. The aim of the process was to give a designated set of classes to each tweet based on its subjectivity. A static tool called Cup of Statics was used to accomplish this objective to support the annotation process. First, there were 2

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

annotators who were expected to annotate 200 tweets independently. Their annotations were compared to agree on and agreement between the annotations was assessed with the widely used inter-annotator agreement measure that indicated a high level of agreement at 0.87. Also a third annotator was hired to overcome any disagreement on annotation. The third annotator was given a subset of the tweets, and the cases of conflict were to be solved by means of a coefficient-based method. Finally, all the tweets were categorized as positive or negative.

The authors of [8] suggested a semi-automatic process of emotional annotation in their study. The suggested approach has two stages. The former involves an automated stage in which a subset of emotion categories are used to pre-label unlabeled sentences in the first stage. They introduce two pre-annotation strategies, which include an unsupervised, which reduces the human factor, and a supervised which makes use of simple emotion models based on existing corpora. In phase two, a manual refinement is applied to a process where the human annotators are asked to identify the prevailing emotion out of a set of possible categories of emotion per sentence. Through the findings of the article, it was found that the two phase approach had a 20 per cent reduction in the annotation time, which led to improvement in the efficiency and effectiveness of the annotation process.

A very recent deep-learning-based automatic text annotation system is presented in [9]. The suggested approach uses attention-based neural network, which uses semantic regularization; the neural network constructed is aimed at imitating the user reading and annotation process, whereby the created neural network provides a better representation of the document based on the semantic relationship among labels. The integration of two semantic based regularizations, similarity and subsumption are brought to underline the correlation between labels, thus making the output of the network to be very close to label semantics. Finally, the paper brings on board large-scale experiments with four real-world social media datasets, the findings of which showed considerable improvements in terms of accuracy and F1-score, and the training time was also decreased.

A semi-supervised learning system of annotating unlabeled tweets is outlined in [10]. The process is structured into different stages. The first step was to first train six annotators on specific methods of annotation and group them into three. At the same time, every annotator checked and both labeled all tweets, and the inter-annotator agreement of this task was 0.63. After that, the manually marked tweets were divided into training and test sets. Data normalization and cleansing were done during preprocessing, followed by vectors representation of texts through TF-IDF, word2vec

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

and document2vec. One thousand manually labeled instances were used as a baseline classifier and its performance was compared with that of bidirectional encoder representations of Transformers (BERT) feature vectors. A self-training method was used to increase the training corpus, but a lookup list was used to maximize the selection of pseudo-labeled data. Lastly, three deep-learning models, namely CNN, LSTM, and BiLSTM, were tested on the enriched training dataset and their relative efficacy in classifying three million posts on social-media was evaluated.

The author of [11] came up with a hybrid rule-based emotion annotation algorithm that annotates tweets to the eight basic emotion classes proposed by Plutchik. Each tweet is analyzed with emoji, words of NRC Emotion Lexicon, and lexical relations, in case an emoji corresponds to one of the eight categories, the tweet is annotated, otherwise, it is substituted with a text description.

The algorithm also examines the tweet and determines whether it contains words in NRC lexicon; the recognition of a word in the NRC lexicon initiates the annotation of the respective emotion. Besides that, the idea of lexical relations, such as synonyms, hyponym, and hypernym, is also discussed with each tweet to improve the perspective on the emotional meaning.

In [12], another technique of automatic annotation of data at the aspect level is introduced. It is based on the probability of the word sequences using an N-gram language model to evaluate the probability of two words, Wx (aspect word) and Wy (any other word); the occurrence of a text (positive, negative, or neutral) is judged by the conditional probability P(Wx|Wy). When Wy is related to positive sentiment, then the combination of Wx and Wy is considered to have positive polarity. On the other hand, when Wy has negative sentiment, the pair is called to have a negative polarity. Where Wy is neither positive nor negative, then the combination is said to be neutral. A set of positive and negative words, also known as a bag of positive and negative words, has been developed in order to enhance the performance of the suggested strategy.

The framework that was proposed by the authors of [13] is named Topic2labels (T2L), and it employs a different method of automating the annotation procedure. This model employs the Latent Dirichlet Allocation (LDA) methods to extract topics out of the data and Bidirectional Encoder Representation for Transformers (BERT) to build the feature vector. The three layers of the proposed architecture are structured. The first layer involves the execution of LDA, which tries to generate topics which are ranked based on a newly developed algorithm; the rank that is

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

dominating is then used to annotate the data. BERT is used in the second layer to encode the feature representations of the labelled text. The third layer consists of deep-learning algorithms that are used to categorize the text into several themes.

The experiment in reference [14] defines a method of annotation which combines manual labeling and semi-supervised learning method. First, a collection of tweets is filtered, pruned, and later annotated by human participants to produce six-dimensional vectors of emotive substance. These carefully annotated tweets serve as examples of seeds to label the rest of the unlabeled ones. This process of semi-supervised labeling uses hybrid word-embedding representation (word2vec) and Distance (WMD) of Word Mover. WMD is used to test the semantic dissimilarity of textual units and it is also used to determine the maximum similarity of each unlabeled tweet to the seed tweets. A similarity score is then generated and each unlabeled tweet is annotated.

The article in [15] suggests a data annotation methodology that makes use of rules. The suggested methodology is made up of three stages. To start with, unstructured data are first purged through a pipeline of extensive pre-processing. The data are then assigned three categories of affective values, namely positive, neutral and negative, through the calculation of sentiment scores through a rule-based classifier. Lastly, the classification is done with BERT and zero-shot learning algorithms.

The authors present ALANNO, which is an open-source system, in reference [16] and which uses Active Learning (AL) approaches to streamline data sampling. The site has two different user roles project managers and annotators. Project managers manage annotation campaigns and have the mandate to instantiate three classes of projects, i.e.,, single-label classification, multi-label classification, and sequence-labelling task. Annotators, in their turn, have a role of labeling the data that have not been labeled before in the specified projects.

A summary of the articles that have been reviewed is presented in Table 1 below. The literature review was based on data annotation methods, with the particular emphasis on the strengths and weaknesses of each method.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 1: Overview of Selected Research Articles

Reference	Annotation Method	Key Findings	Limitation
[7]	Three annotators manually assigned polarity (positive, negative) to 600 reviews.	The inter-annotator agreement was 0.87, and the classification algorithms achieved 75% accuracy.	Fully depend on manual annotation, Limited dataset size.
[8]	A semi-automatic approach consists of automatic pre-annotation followed by manual human refinement.	Annotation time reduced by 20% compare to manual annotation without compromising inter-annotator agreement (IAA)	Human intervention is required in the second phase, and lexicon limitations of CountWordEmo and EmoLexicon reduce performance on informal social media text.
[9]	Deep learning-based automatic annotation approach using user generated tags as labels.	The approach achieves improvements in accuracy and F1 score and reduce training time.	The model requires computational and memory resources and its performance varies based on dataset characteristics.
[10]	The model combines manual annotation with NLP and DL then enhanced by semi-supervised learning.	BiLSTM and BERT embedding's have proven high efficiency for classifying patient discussions.	Manual annotation was implemented in the first phase. However, the guidelines for data annotation were not explained.
[11]	A hybrid rule-based approach maps tweets to predefined emotions using emojis, NRC lexicon and lexical relations.	The annotated dataset was used to train LSTM classifier achieving an accuracy of 91.0%.	Informal language poses challenged for the proposed rule-based keyword-centric approach. Only (7%) of the emojis can be mapped to emotions, and there is an overlap between emotion expression
[12]	A fully automated approach combines N-Gram with bags of positive/negative words then the polarity (positive, negative, neutral) is estimated based on lexical matching.	A novel auto-annotation technique is presented, offering annotation quality comparable to manual methods while reducing time and cost.	Narrow focus on music review domain suggests that its implementation on other contexts may require further testing and adaptation.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

[13]	Topics extracted from social media data using Latent Dirichlet Allocation (LDA), and novel T-TF-IDF algorithms assign topic to each document.	Compared to baseline models, the proposed method achieved the highest accuracy with LSTM (83.51%),CNN (83.1%) and ANN (80.73%)	LDA is a bag-of-words model and ignores words semantic relation. T-TF-IDF relies on frequency statistics and does not consider semantic relationships.
[14]	Word2Vec and Word Movers Distance are used to expand the manually annotated dataset.	The Extra Trees Classifier achieved 95% accuracy on 10,000 tweets annotated using the proposed method.	Human intervention is required in the first stage. Word Movers Distance disregards word order, treating text as bag or words.
[15]	The model consists of data preparation, rule-based annotation, and deep learning-based classification.	Using BERT for sentiment analysis, the proposed annotation model achieved 93.21% accuracy.	The rule-based annotation approach in the second phase has not been described.

MATERIALS AND METHODS

In this part, a detailed discussion on the basic techniques used in the implementation of the proposed model is given. The techniques will be explained in more detail in the following sub-sections.

1. Semantic Orientation: An early study [17] observed that words exhibit three semantic factors. Bipolar adjective pairs (good/bad, kind/cruel, and honest/dishonest) are used to reflect the evaluative factor, which is terms as semantic orientation [18]. Motivated by the notion that a word is characterized by the company it keeps [19], Peter d. et. al [20] introduced a statistical association method to quantify word's semantic orientation. A word intensity towards positive or negative polarity evaluated using Pointwise Mutual Information (PMI) [21] and Latent Semantic Analysis (LSA) [22]. The polarity of a word is gauged based on the strength of its association with a set of positive words, subtracted by the strength of its association with a set of negative words. The proposed approach is explained in Eq. (1).

$$SO-A(word)=\sum_{pword\in Pwords} A(word,pword) - \sum_{nword\in Nwords} A(word,nword) \qquad (1)$$

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Where A(word) is the current word of the tweet, pword is a set of positive words, nword is a set of negative words, and A(word1,word2) is a measure of association.

2. Pointwise Mutual Information (PMI): Mutual information is the base of word association measures. According to Fano [23], the PMI between two words, x and y, is defined as:

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (2)$$

PMI allows the comparison of the likelihood of observing both words, x and y, together (the joint probability) with the probabilities of observing them independently (by chance). A large positive value of PMI(x,y) >> 0 indicate strong association between x and y. Conversely, when there is no notable relationship between x and y, the value of PMI(x,y) ≈ 0 [24].

3. Latent Semantic Analysis (LSA): LSA proposed in the late 1980s for information retrieval and is also used to represent word meaning. It basic assumption is that text's meaning reflected by patterns of word occurrence. LSA quantify text using the Vector Space Model (VSM) to create a term-document matrix (A). The matrix (A) is transformed and normalized, and then Singular Value Decomposition (SVD) is applied. SVD reduces the matrix's dimensionality, revealng the underlying semantic structure of the data. The SVD is explained in equation (3) [25].

$$A = U \sum V^T \qquad (3)$$

Where U is the term eigenvectors, V is the document eigenvectors, T denotes transposition, and $\sum$ is the diagonal matrix of singular values.

4. Word2Vec: Word2Vec is used to transform word into vector representations by considering critical attributes such as window size and vector dimensions. Words with similar meaning tend to have vector with similar values [26]. Semantic similarity between words vector is computed using cosine similarity. The resulting similarity values span from -1 to 1, where the value of 1 represents the utmost level of similarity [27].

$$similarity = \cos\theta = \frac{\bar{x} \cdot \bar{y}}{\| \bar{x} \| \| \bar{y} \|} \qquad (4)$$

Where:

$\bar{x} \bullet \bar{y}$: Vector dot product from x and y. $\sum_k^n = 1 \; x_k \, y_k$

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

‖x‖: Long vector x. $\sum_k^n = 1\ x_k^2$

‖y‖: Long vector y. $\sum_k^n = 1\ y_k^2$

Θ : is the angle between the two vectors.

5. NRC Lexicon: The NRC emotion lexicon project presents a large word-emotion association resource created through a massive online annotation project [3]. The annotated words mapped to Plutchik’s eight basic emotion categories [28]. An online annotation platform, Amazon Mechanical Turk, utilized to obtain a substantial volume of human annotations at a low cost. Annotators were presented with four words related to the target term and asked to identify which word is most closely synonymous [29]. Each term is annotated by five individuals, and their responses are aggregated via majority voting. Overall, the lexicon consists of 24,200 word-sense pairs, consolidated into approximately 14,200 word-types [3].

PROPOSED MODEL

This section provides an explanation of the proposed model. Overall, two approaches are investigated. The steps implemented in each approach are discussed in the following subsections. The tweets collection step is identical for both approaches. The Python API executed to collect hashtags that attracted high interactions. High interaction hashtags are those hashtags that have 500+ clicks or replies or retweets. Overall, 33,429 tweets were gathered. Table 2 shows samples of the collected tweets.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 2 tweet samples

No.	Tweet Samples
1	Let's see... @GrabMY services is still on. Insurance protection for Covid-19 to riders but nothing about drivers (like me), food delivery by car for gold and platinum status drivers only.... is this? Not only is demand for rides low but
2	@BCAppelbaum The two situations are completely different. In 2008, what happened was clearly a theft due to deregulations. Now there is a breakdown because many of the stocks are held at relatively high prices. Unfortunately, a complete collapse
3	Not a lot of food left in the supermarket tonight. No eggs, meat, bread, milk, pasta or rice. Plus, of course, no toilet paper or disinfectant. Some frozen food, fruit and plenty of snacks. (Chips, chocolates etc)

1. First Approach: Figure 1 illustrates the steps implemented in the first approach. It shows that the collected tweets are pre-processed, tokenized and the semantic orientation is measured for each tweet based on its relation to the NRC positive/negative word list. The following subsections give a comprehensive detail of the actions taken on each step.

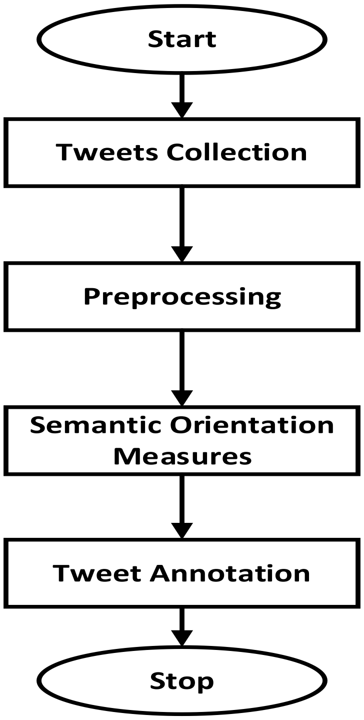


Figure 1 proposed framework (1st approach)

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

a) Tweets Pre-processing: The pre-processing step consists of several tasks that collectively aim to clean the collected tweets and make them suitable for subsequent analysis. It includes the following sub-steps:

- Lowercasing: convert tweets to lowercase.
- Remove unwanted characters: punctuation, URLs, extra white spaces, and Twitter features (hashtag symbols).
- Retweet removal: to enhance the content of the collected tweet, duplicated tweets are excluded.
- Stop-word removal: stop words, do not convey any sentiment. Hence, they have been removed from the collected tweet.
- Spell Correction: It is a common practice that Tweeter users do not adhere to spelling accuracy. Accordingly, an open access dictionary integrated with python are utilized to correct the misspelt words.
- Stemming and lemmatization: this task is aimed at converting words into their roots.
- Tokenization: tweet text is transformed into a sequence of words. This transformation is achieved using Python’s NLTK library. Subsequently, a sequence of words (tokens) is generated.

b) Semantic Orientation Measures: In the preceding steps, each tweet was subject to cleaning and transformed into a sequence of words. Consequently, the computation of the semantic orientation measures is now applicable. Algorithm (1), explained in Table 3 outlines the steps taken to achieve this task.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 3 first approach algorithm

```
for each tweet in the collected corpus
   for each word in the current tweet
      Positive_SO = 0
      Negative_SO = 0
      SO_Measure = 0
      Next_word = GetNextWord(word)
      While Next_word is not the end of the current tweet
         If Next_word in NRC_positive Word_list
             Positive_SO = Positive_SO + Max (SO (word, Next_word),0)
         Elseif Next_word in NRC_negative Word_List
             Negative_SO = Negative_SO + Max (SO (word, Next_word),0)
         Next_word = GetNextWord(word)
        End_if
      End_While
      SO_Measure = Positive_SO - Negative_SO
      If SO_Measure > 0 then print "Positive tweet"
      Elseif SO_Measure < 0 then print "Negative tweet"
      Else print "Neutral tweet"
     End_if
   End_for
End_for
```

The algorithm presented in Table 3 calculates the total semantic orientation for each tweet. The semantic orientation of each word in the current tweet is calculated with its neighbouring words. The proposed algorithm finds whether the neighbouring words belong to the NRC positive/NRC negative word list, and the positive/negative semantic orientation is calculated accordingly. Eventually, the Positive_SO and Negative_SO variables store the total positive and negative semantic orientation measures of the entire tweet. It is important to note that various semantic orientation measures are employed such as PMI, LSA and Word2Vec (Cosine Similarity). As some of these measures may produce negative values, the max function is utilized to convert negative results to zero. The final semantic orientation value is computed by subtracting the total tweet Positive_SO fromm the total tweet Negative_SO, as described in equation (1).

Finally, if the total semantic orientation value is positive, the tweet is annotated as positive. If the value is negative, the tweet is labelled as negative. Otherwise, it is labelled as neutral.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

2. Second Approach: To overcome the limitations addressed in the first approach, a more comprehensive method is investigated in this section. The underlying concept of the second approach is to expand the coverage of the NRC lexicon through the utilization of various machine learning algorithms. The proposed framework is illustrated in figure 2.

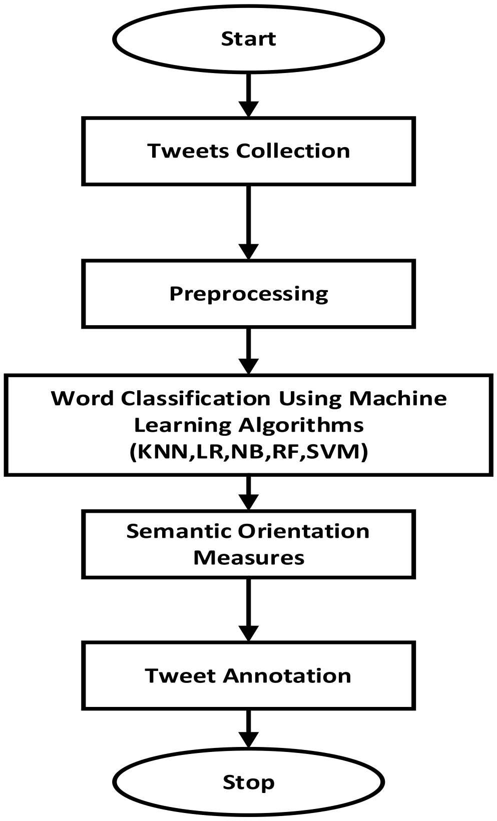

Start
Tweets Collection
Preprocessing
Word Classification Using Machine Learning Algorithms (KNN,LR,NB,RF,SVM)
Semantic Orientation Measures
Tweet Annotation
Stop

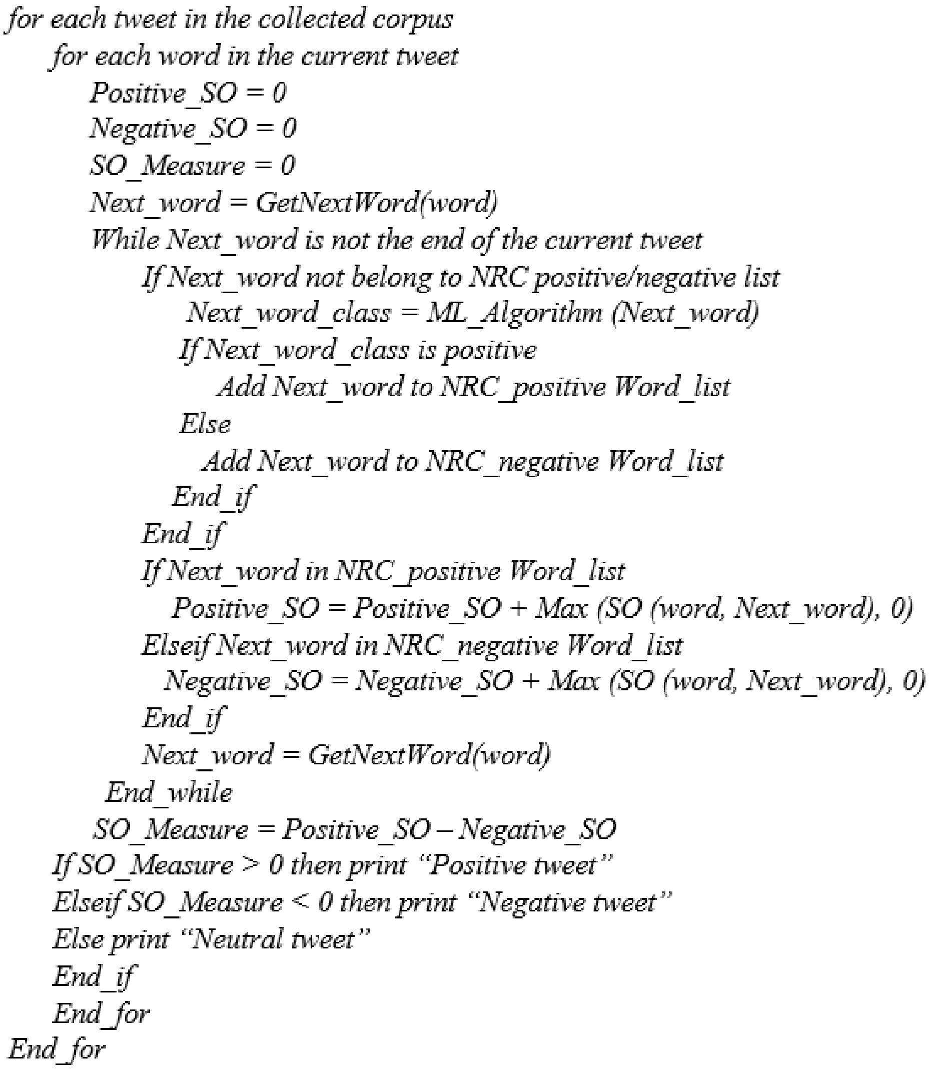
Figure 2 proposed framework (3rd approach)

Figure 2 shows that the initial steps (tweet collection and pre-processing) are similar to those of the first approach. However, a distinct feature emerges as machine learning algorithms are integrated to classify tokens that are not presented in the NRC word list. This led to a more comprehensive evaluation of the approach. The following subsections explain the steps of the proposed approach.

a) Tweet Word Classification: In order to classify tweets with words that do not belong to NRC positive/negative list, various machine learning algorithms are employed. K Nearest Neighbor (KNN), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) classification algorithms are developed to achieve this task. To train the classification algorithms, the NRC positive/negative word list served as seed data. For each NRC word, its definition (meaning) is obtained using an open access dictionary. NRC word definition data is used to train the classification algorithms. Likewise, the definition of the unclassified tweet words is gathered. Then the trained algorithms are used to predict the class of the tweet words that do not belong to NRC list.

b) Semantic Orientation Measures: The second approach calculated the semantic orientation of each tweet. Algorithm (2) depicted in Table 4 explains the executed steps.

Table 4 second approach algorithm



```
for each tweet in the collected corpus
    for each word in the current tweet
        Positive_SO = 0
        Negative_SO = 0
        SO_Measure = 0
        Next_word = GetNextWord(word)
        While Next_word is not the end of the current tweet
            If Next_word not belong to NRC positive/negative list
                Next_word_class = ML_Algorithm (Next_word)
                If Next_word_class is positive
                    Add Next_word to NRC_positive Word_list
                Else
                    Add Next_word to NRC_negative Word_list
                End_if
            End_if
            If Next_word in NRC_positive Word_list
                Positive_SO = Positive_SO + Max (SO (word, Next_word), 0)
            Elseif Next_word in NRC_negative Word_list
                Negative_SO = Negative_SO + Max (SO (word, Next_word), 0)
            End_if
            Next_word = GetNextWord(word)
        End_while
        SO_Measure = Positive_SO – Negative_SO
    If SO_Measure > 0 then print "Positive tweet"
    Elseif SO_Measure < 0 then print "Negative tweet"
    Else print "Neutral tweet"
    End_if
    End_for
End_for
```

Table 3 illustrates the process of measuring the semantic orientation of each word in the current tweet with respect to its neighbouring words. Unlike the first approach, the second approach utilizes a trained ML algorithm to classify words that do not belong to the NRC positive/negative lexicon. Definitions for unidentified words are sourced from an open-source dictionary and fed to a trained classification algorithm for precise classification.

Similar to the first approach, Positive_SO and Negative_SO variables are employed to accumulate the positive and negative semantic orientation values. Consequently, the final semantic orientation measure is computed by subtracting the values of Positive_SO and the Negative_SO then tweets are labelled accordingly.

RESULTS AND DISCUSSION

In this section, an analysis of the obtained results is presented. The results of both the first and second approaches are examined in detail.

First Approach Results: the frequency distribution of NRC positive and negative lexicon within the collected tweets has been analyzed. It has been observed that only a small fraction of the collected tweets contain more than 5 NRC positive/negative lexicons. Furthermore, there are a notable number of the collected tweets that do not contain any NRC lexicon, which could lead to false annotations. Figure 3 illustrates the distributions of the NRC lexicon among the collected tweets.

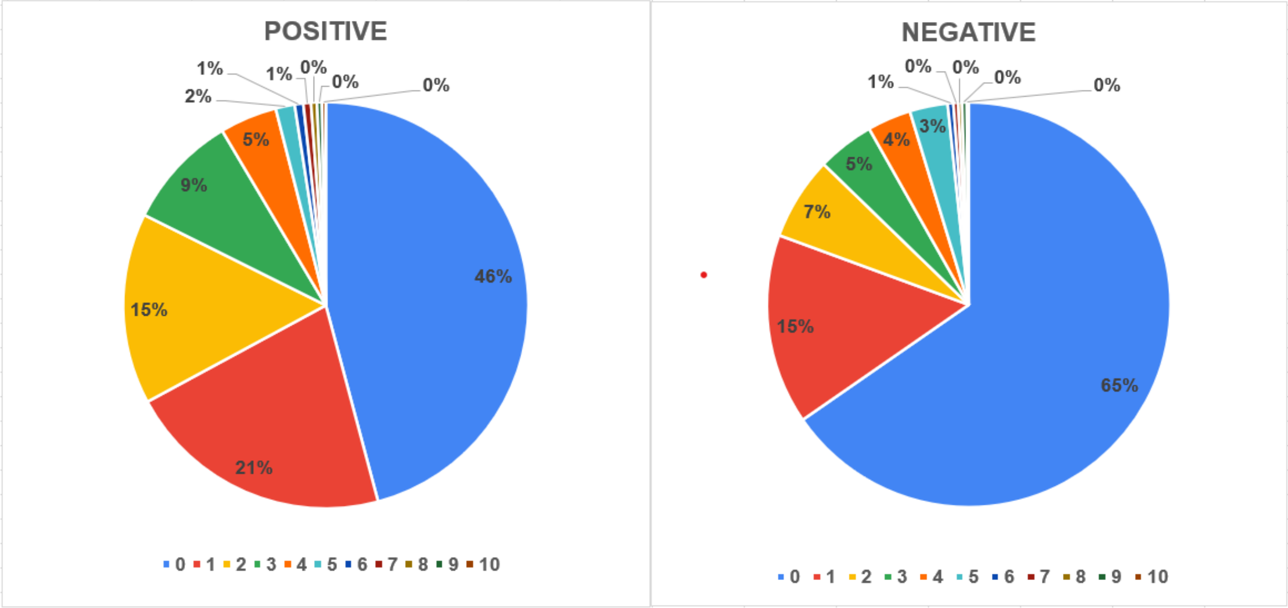


Figure 3 NRC lexicon/tweets distribution

Figure 3 categorizes the collected tweets based on the number of NRC words. The first group shows the number of tweets that contain zero NRC positive/negative tokens, while the last group shows the number of tweets containing ten NRC words. Notably, only a few tweets contain more than five NRC tokens, with 2% and 3% of tweets having positive and negative NRC tokens, respectively. 46% of the collected tweets contain zero positive NRC tokens, while 65% contain zero negative NRC lexicons. Furthermore, the number of tweets that have one or more NRC positive tokens, or fewer than or equal to five NRC positive tokens, is almost 53%, while negative tokens account for only 34%. Overall, the percentage of tweets that contain zero NRC lexicons is considerably high, with almost 50% of tweets lacking both positive and negative NRC lexicons. This may result in inaccurate tweet annotation.

In the annotation process, three semantic orientation measurement techniques: PMI, LSA and Word2Vec (Cosine Similarity) are applied. Accordingly, equation (1) is calculated to classify tweets into positive, negative, and neutral. Figure 4 depicts the classification of the collected tweets into three categories (positive, negative, and neutral).

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

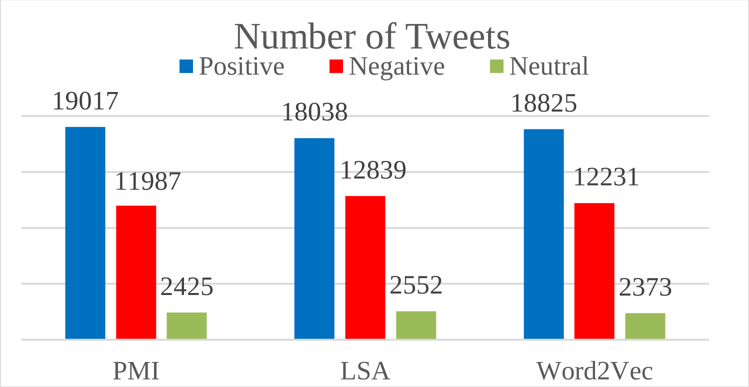


Figure 4 Number of tweets per semantic label

Figure 4 shows that all three semantic orientation measurement techniques predominantly classify tweets as positive. PMI achieves the highest number of positive tweets at 19,017. Word2Vec is coming next at 18,825 and finally LSA with 18,038 positive tweets. On the other hand, LSA seems to be more sensitive to negative tweets. It identifies the largest number of negative tweets, followed by Word2Vec and PMI respectively. Surprisingly, the smallest number of tweets fall into the neutral category. The number of neutral tweets starts at 2,373 using Word2Vec measure, 2,552 using LSA measure and 2425 using PMI measure. Overall, figure 4 shows a general agreement in the tweet annotation process among all three semantic measure techniques. To obtain a more accurate evaluation of the classification decision agreements among all three approaches, Kappa statistics [30] are calculated.

Table 5. Kappa agreement measure (1st approach)

Semantic Orientation Methods	Kappa Measure
PMI & LSA	0.7504
PMI & Word2Vec	0.8099
LSA & Word2Vec	0.7728

Table 5 indicates that PMI and Word2Vec (cosine similarity) exhibit the strongest classification agreement. Nevertheless, all approaches achieved an agreement value exceeding 0.75, which shows a general consensus among all approaches.

Second Approach Results: to overcome the limitations of the first approach, tweet words that do not belong to the NRC lexicon are classified using various machine-learning algorithms. The definitions of non NRC words are sourced from open dictionaries and fed to well-trained machine learning algorithms. During the construction of these algorithms, attention is given to evaluating

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

their accuracy. Figure 5 provides a comprehensive explanation of the obtained accuracy across various models.

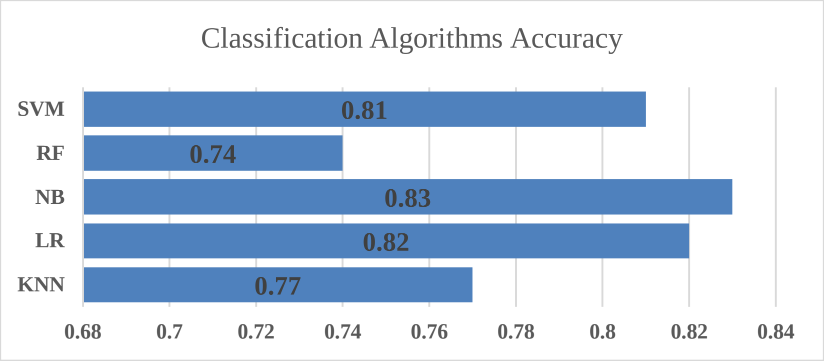


Figure 5 Classification algorithms accuracy

Results in Figure 5 demonstrate that Naïve Bayes (NB) achieved the highest classification accuracy of 83%. The significant performance of NB algorithm could be linked to its probabilistic nature, which is suitable for relatively independent features, a common assumption in text based datasets such as tweets. Logistic Regression (LR) and Support Vector Machine (SVM) closely followed, with accuracies of 82% and 81%, respectively. LR competitive performance indicate that linear relationship between features and labels is probably established. Finally, K-nearest Neighbors (KNN) and Random Forest (RF) generated the lower accuracies of 77% and 74% respectively. Overall, the results suggest that probabilistic and linear models (NB and LR) are more suitable for the targeted task, while distance-based (KNN) and tree-based (RF) algorithms may required more steps to improve performance.

In the annotation process, a total of fifteen combinations, five classification algorithms and three semantic orientation measures are executed. Table 6 gives a comprehensive breakdown of the number of the labelled tweets per sentiment category (positive, negative and neutral).

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلـــوم الصـرفة والتطـــبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 6. Labelled tweets breakdown (2nd approach)

Combination	Positive	Negative	Neutral
PMI+KNN	27387	5986	56
PMI+LR	29852	3521	56
PMI+NB	30979	2394	56
PMI+RF	28265	5108	56
PMI+SVM	19017	11987	2425
LSA+KNN	24639	8709	81
LSA+LR	27487	5861	81
LSA+NB	28834	4511	84
LSA+RF	25317	8030	82
LSA+SVM	26309	7035	85
Word2Vec+KNN	27485	5891	53
Word2Vec+LR	29906	3470	53
Word2Vec+NB	31122	2254	53
Word2Vec+RF	28362	5014	53
Word2Vec+SVM	28449	4927	53

In contrast to the first approach, it can be observed that there is a substantial disparity between the number of tweets categorized as positive compared to those classified as negative. Moreover, a considerable reduction is observed in the number of tweets categorized as neutral. However, an exception to this trend is observed in the case of combining PMI with SVM algorithm. In this particular instance, this discrepancy between the numbers of positive and negative tweets is mitigated, and the number of tweets labelled as neutral is increased. To increase the understanding of the annotation results, kappa agreement measures were calculated. The results for all possible combinations of the proposed techniques are shown in table 7.

Table 7 Kappa agreement measure (2nd approach)

Kappa Measure		PMI					LSA					Word2Vec				
		KNN	LR	NB	RF	SVM	KNN	LR	NB	RF	SVM	KNN	LR	NB	RF	SVM
PMI	KNN		0.4858	0.4226	0.4763	0.4763	0.5595	0.3633	0.3451	0.3428	0.3354	0.6848	0.3852	0.3168	0.3847	0.3897
	LR	0.4858		0.7008	0.5972	0.2105	0.2905	0.5236	0.4452	0.333	0.4015	0.3871	0.6737	0.4888	0.4549	0.5089
	NB	0.4226	0.7006		0.517	0.157	0.233	0.3693	0.4909	0.2667	0.3058	0.3344	0.5055	0.65	0.393	0.4076
	RF	0.4763	0.5972	0.517		0.2567	0.3155	0.3967	0.3857	0.542	0.3817	0.3762	0.4539	0.3774	0.6812	0.4506
	SVM	0.2749	0.2105	0.157	0.2567		0.2642	0.2303	0.2059	0.2563	0.2485	0.2514	0.1942	0.1371	0.2365	0.2466
LSA	KNN	0.5595	0.2905	0.233	0.3155	0.2642		0.4864	0.4347	0.4747	0.4589	0.551	0.2925	0.2221	0.3105	0.3121
	LR	0.3633	0.5236	0.3693	0.3967	0.2303	0.4864		0.6745	0.5566	0.6574	0.3704	0.5135	0.3464	0.3953	0.4388
	NB	0.3451	0.4452	0.4909	0.3857	0.2059	0.4347	0.6745		0.5044	0.5705	0.3546	0.4466	0.4721	0.388	0.3986
	RF	0.3428	0.333	0.2667	0.542	0.2563	0.4747	0.5566	0.5044		0.5454	0.34	0.3378	0.258	0.5322	0.3631
	SVM	0.3354	0.4015	0.3058	0.3817	0.2485	0.4589	0.6574	0.5705	0.5454		0.3548	0.3981	0.292	0.3869	0.555
Word2Vec	KNN	0.6848	0.3871	0.3344	0.3762	0.2514	0.551	0.3704	0.3546	0.34	0.3548		0.4875	0.4108	0.4779	0.4841
	LR	0.3852	0.6737	0.5055	0.4539	0.1942	0.2925	0.5135	0.4466	0.3378	0.3981	0.4875		0.6745	0.601	0.6529
	NB	0.3168	0.4888	0.65	0.3774	0.1371	0.2221	0.3464	0.4721	0.258	0.292	0.4108	0.6745		0.5037	0.5121
	RF	0.3847	0.4549	0.393	0.6812	0.2365	0.3105	0.3953	0.388	0.5322	0.3869	0.4779	0.601	0.5037		0.5689
	SVM	0.3897	0.5089	0.4076	0.4506	0.2466	0.3121	0.4388	0.3986	0.3631	0.555	0.4841	0.6529	0.5121	0.5689	

Table 7 records a strong variation of obtained results, ranging from 0.1372 (minimum value) to 0.7008 (maximum value). Of particular interest, the hybrid models PMI+LR and PMI+NB achieve the highest concordance; an agreement measure of 0.7008. Furthermore there is a strong tendency for semantic orientation measures based on Word2Vec to converge with other alternative techniques. In eight different comparison tests, Word2Vec had values for agreement greater than or equal to 0.65, whereas the methodology using PMI had a value greater than 0.65 in only four tests.

CONCLUSION

Within the current study, measures of semantic orientation were investigated covering a range from the usage of the NRC lexicon to the use of machine-learning algorithms for automatic annotation of tweet data for sentiment analysis. The analytical techniques used in this article provided interesting insights and highlighted future research directions. The presented results show that corpus-driven semantic orientation measures have yielded significant results. Nevertheless, further enhancement is possible through the incorporation of a more comprehensive lexicon, the integration of contextual knowledge, and the adoption of advanced deep learning techniques. These potential approaches will be the focus of the coming research works.

Conflict of interests:

There are non-conflicts of interest.

References

[1] J. Pfeffer, D. Matter, K. Jaidka, O. Varol, A. Mashhadi, J. Lasser, D. Assenmacher, S. Wu, D. Yang, C. Brantner, D. M. Romero, J. Otterbacher, C. Schwemmer, K. Joseph, D. Garcia, and F. Morstatter, "Just Another Day on Twitter: A Complete 24 Hours of Twitter Data," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 1073–1081, 2023. https://doi.org/10.1609/icwsm.v17i1.22215

[2] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp.5731-5780, 2022. http://dx.doi.org/10.1007/s10462-022-10144-1

[3] S. M. Mohammad and P. D. Turney, "NRC Emotion Lexicon. NRC Technical Report," *National Research Council Canada*, pp. 1–234, 2013.

[4] C. ÇILGIN, M. BAŞ, H. BİLGEHAN, and C. ÜNAL, "Twitter Sentiment Analysis During Covid-19 Outbreak with VADER," *AJIT-e: Academic Journal of Information Technology*, vol. 13, no. 49, pp. 72–89, 2022. http://dx.doi.org/10.5824/ajite.2022.02.001.x

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

[5] J. Miazga and T. Hachaj, “Evaluation of most popular sentiment lexicons coverage on various datasets,” *Proceedings of the 2019 2nd International Conference on Sensors, Signal and Image Processing*, pp. 86–90, 2019. http://dx.doi.org/10.1145/3365245.3365251

[6] P. Sunilkumar and A. P. Shaji, “A Survey on Semantic Similarity,” *2019 6th IEEE International Conference on Advances in Computing, Communication and Control, ICAC3*, pp. 1-8, 2019. https://doi.org/10.1109/ICAC347590.2019.9036843

[7] M. Asif, M. A. Qureshi, A. Abid, and A. Kamal, “A Dataset for the Sentiment Analysis of Indo-Pak Music Industry,” *3rd International Conference on Innovative Computing, ICIC 2019*, no. Icic, 2019. https://doi.org/10.1109/ICIC48496.2019.8966720

[8] L. Canales, W. Daelemans, E. Boldrini, and P. Martinez-Barco, “EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text,” *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 579–591, 2022. https://doi.org/10.1109/TAFFC.2019.2927564

[9] H. Dong, W. Wang, K. Huang, and F. Coenen, “Automated Social Text Annotation with Joint Multilabel Attention Networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2224–2238, 2021. https://doi.org/10.1109/TNNLS.2020.3002798

[10] B. Athira, J. Jones, S. M. Idicula, A. Kulanthaivel, and E. Zhang, “Annotating and detecting topics in social media forum and modelling the annotation to derive directions-a case study,” *Journal of Big Data*, vol. 8, no. 1, pp.1-23, 2021. https://doi.org/10.1186/s40537-021-00429-7

[11] M. Krommyda, A. Rigos, K. Bouklas, and A. Amditis, “An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media,” *Informatics*, vol. 8, no. 1, p. 19, 2021. http://dx.doi.org/10.3390/informatics8010019

[12] M. A. Qureshi, M. Asif, M. F. Hassan, G. Mustafa, M. K. Ehsan, A. Ali, and U. Sajid, “A novel auto-annotation technique for aspect level sentiment analysis,” *Computers, Materials and Continua*, vol. 70, no. 3, pp. 4987–5004, 2022. DOI:10.32604/cmc.2022.020544

[13] J. A. Wahid, L. Shi, Y. Gao, B. Yang, L. Wei, Y. Tao, S. Hussain, M. Ayoub, and I. Yagoub, “Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response,” *Expert Systems with Applications*, vol. 195, p. 116562, 2022. https://doi.org/10.1016/j.eswa.2022.116562

[14] B. V. Namrutha Sridhar, K. Mrinalini, and P. Vijayalakshmi, “Data Annotation and Multi-Emotion Classification for Social Media Text,” *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, pp. 1011–1015, 2020. http://dx.doi.org/10.1109/ICCSP48568.2020.9182362

[15] A. Sahar, M. Ayoub, S. Hussain, Y. Yu, and A. Khan, “Transfer Learning-Based Framework for Sentiment Classification of Cosmetics Products Reviews,” *Pakistan Journal of Engineering and Technology*, vol. 5, no. 3, pp. 38–43, 2022. https://doi.org/10.51846/vol5iss3pp38-43

[16] J. Jukić, F. Jelenić, M. Bićanić, and J. Šnajder, “ALANNO: An Active Learning Annotation System for Mortals,” *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pp. 228–235, 2023. http://dx.doi.org/10.18653/v1/2023.eacl-demo.26

[17] P. H. Osgood, Charles Egerton and Suci, George J and Tannenbaum, *The measurement of meaning*. University of Illinois press, no. 47,1957.

[18] K. Hatzivassiloglou, Vasileios and McKeown, "Predicting the semantic orientation of adjectives," *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, pp. 174-181, 1997. http://dx.doi.org/10.3115/976909.979640

[19] J. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, pp.10-32, 1957.

[20] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 315–346, 2003. https://doi.org/10.1145/944012.944013

[21] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1989-June, no. 1, pp. 76–83, 1989.

[22] S. T. Landauer, Thomas K and Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological review*, p. 211, 1997. https://psycnet.apa.org/doi/10.1037/0033-295X.104.2.211

[23] D. Fano, Robert M and Hawkins, "Transmission of information: A statistical theory of communications," *American Journal of Physics*, vol. 29, pp.793-794, 1961. https://doi.org/10.1119/1.1937609

[24] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[25] N. Evangelopoulos, X. Zhang, and V. R. Prybutok, "Latent semantic analysis: Five methodological recommendations," *European Journal of Information Systems*, vol. 21, no. 1, pp. 70–86, 2012. https://doi.org/10.1057/ejis.2010.61

[26] G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," *Journal of Supercomputing*, vol. 77, no. 11, pp. 12320–12335, 2021. http://dx.doi.org/10.1007/s11227-021-03743-2

[27] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2vec model analysis for semantic similarities in English words," *Procedia Computer Science*, vol. 157, pp. 160–167, 2019. http://dx.doi.org/10.1016/j.procs.2019.08.153

[28] R. Plutchik, "The emotions," *University Press of America*, 1991.

[29] S. M. Mohammad and P. D. Turney, "Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon," *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 26–34, 2010.

[30] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Fam Med*, vol. 37, no. 5, pp. 360–363, 2005.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

الخلاصة

المقدمة:

تتطلب خوارزميات تحليل المشاعر بيانات مصنفه عالية الجودة اثناء مرحلة تدريب الخوارزميات. مما أدى الى عملية تصنيف يدوي معقد للبيانات والذي يتطلب وقتا طويلا وجهد وكلفه عالية. ولغرض معالجة هذه التحديات يقترح هذا البحث عملية تصنيف تلقائية للبيانات لغرض استخدامها في خوارزميات تحليل المشاعر.

طرق العمل:

تم استخدام ثلاثة مقاييس للتشابه الدلالي (PMI, LSA, Word2Vec) وخمسة خوارزميات تصنيف (KNN,LR,BN,RF,SVM) ودمجها مع قاموس المرادفات (NRC) لغرض أتمتة عملية تصنيف التغريدات في منصة تويتر لغرض تحليل المشاعر المتضمنة في هذه التغريدات.

النتائج:

تم تصنيف بيانات تويتر لغرض تدريب خوارزميات تحليل المشاعر باستخدام خمسة خوارزميات للتعلم الالي وثلاثة مقاييس للتشابه الدلالي وقد أظهرت نتائج (kappa) لقياس مستوى التوافقية بأن (PMI+LR) و (PMI+NB) قد حققت اعلى درجة اتفاق وبلغت (0.7008)

الاستنتاجات:

تشير النتائج الى ان مقاييس التشابه الدلالي المعتمدة على بيانات تويتر قد حققت نتائج مجدية وبنسبة اتفاقية عالية نسبيا. ومع ذلك هناك مجال لتحسين أداء النماذج المقترحة من خلال دمج قواميس لغوية اكثر شمولا ودمج المعرفة السياقية للكلام واعتماد تقنيات التعلم العميق المتقدمة.

الكلمات المفتاحية: تحليل المشاعر، التعلم الالي، التشابه الدلالي