Automatic Sentiment Annotation of Idiomatic Expressions for Sentiment Analysis Task

BASHAR M. A. TAHAYNA, RAMESH KUMAR AYYASAMY, AND REHAN AKBAR

1Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman (UTAR), Kampar, Perak 31900, Malaysia
2Department of Computer and Information Sciences, Universiti Teknologi PETRONAS (UTP), Seri Iskandar, Perak 32610, Malaysia

Corresponding author: Ramesh Kumar Ayyasamy (rameshkumar@utar.edu.my)

This work was supported in part by Universiti Tunku Abdul Rahman under the University Tunku Abdul Rahman (UTAR) Research Fund under Grant IPSR/RMC/UTARRF/2020-C1/R01.

ABSTRACT Social media users use words and phrases to convey their views or opinions. However, some people use idioms or proverbs that are implicit and indirect to make a stronger impression on the audience or perhaps catch their attention by utilizing funny, sarcastic, or metaphorical phrases. Idioms and proverbs are figurative expressions with a thematically coherent totality that cannot be understood literally. In previous work, the extension of IBM’s Sentiment Lexicon of Idiomatic Expressions was proposed to include around 9,000 idioms; a crowdsourcing service manually annotates both lexicons. Therefore, in this research, we provide a knowledge-based expansion approach to avoid human annotation of idioms. For sentiment classification, the proposed method has the advantage that it does not require any fine-tuning for the BERT model. Experimental comparisons show that automated idiom enrichment and annotation are very beneficial for the performance of the sentiment classifier. The expanded annotated lexicon will be made available to the general public.

INDEX TERMS Data augmentation, deep learning, fine-tuning, idiomatic expressions, lexicon, sentiment analysis.

I. INTRODUCTION

One of the most popular online activities is the use of social media. Statista website states that “globally, more than 4.26 billion individuals used social media in 2021. This figure is expected to rise to approximately six billion by 2027” [1]. Social media platforms are among the essential modern means of communication and have become a de facto publication medium for businesses, organizations, and governments. They provide a speak-and-listen model of interactive bilateral communication between institutions and their existing or potential clients. The abundance of social media platforms has created an enormous information space that has given rise to various new applications and the growth and development of natural language processing (NLP) specialties like sentiment analysis. Sentiment analysis is an NLP effort focusing on extracting features of textual data and representing them in the proper format, then classifying them into distinct polarities or emotional classes. The manual processing of social media data to determine the concealed sentiments they masked is tedious and time-consuming. Therefore, it is necessary and demanding to find ways to enable computers to process, analyze and understand this volume of data efficiently.

Social media has lowered this barrier and brought us closer together, but comprehending culturally distinct terminology still requires a solid awareness of that culture. Social media users frequently converse and convey their ideas in written, everyday language. But it can be a challenging chore to extract and analyze opinion or sentiment from a text. For instance, because each message on Twitter is limited to a certain number of characters, users frequently employ shorthand and figurative expressions whenever possible to replace lengthy statements or concepts with shorter ones. Users often use informal writing styles and metaphorical language in their tweets to communicate their thoughts or beliefs. The
shorthand approach considers character constraints and condenses the intended message into a small number of phonetically related words or symbols. Idioms, euphemisms, and slang expressions can bring a touch of kindness, civility, or humor to a contentious subject. Another justification might be to prevent a criminal conviction or legal action if inappropriate language was used. In [2], researchers work to show how idioms are essential in the Twitter platform. They assert that it is astounding since idiomatic expressions are used and debated by millions of users on social media platforms like Twitter. They found that during ten months in 2014, idioms accounted for about 10% of Twitter trends [2].

The primary techniques utilized in sentiment analysis are machine learning and deep learning approaches. From a machine learning standpoint, a corpus is required to train a classifier to perform the sentiment classification task [3]. Text and sentiment classification tasks have advanced significantly in recent years because of the utilization of deep learning methods. The goal of deep learning is to discover the underlying concepts and layers of sample data representation [4]. The knowledge gained throughout these learning processes is quite helpful. The ultimate objective is to give machines the capacity to analyze, understand, and identify information like textual data. Deep learning is a sophisticated neural network(s) that performs significantly better at speech and image recognition than earlier (un-) supervised machine learning algorithms.

The neural network structure has achieved outstanding results in sentiment analysis. The neural network design is used to create a variety of neural network algorithms, including bidirectional LSTM and CNN. After [5] successfully applied CNN to the sentiment classification task and reported positive results, the authors of [6] presented a VDCNN model based on the deep convolutional network method [6]. Traditionally, neural network methods begin their model parameters randomly before training them using optimization methods like backpropagation and gradient descent [7]. However, using neural network-based deep learning for natural language processing encountered the following issues before introducing pre-training techniques: First, complex models needed deep learning models that were too primitive at that time. Second, manual annotation is too expensive to run large models, while data-hungry deep learning models lack a substantial amount of annotated data. As a result, researchers are steadily paying more attention to pre-training strategies. The general agreement nowadays is that Deep Learning methods outperform other machine learning approaches in terms of accuracy for most NLP tasks [8], [9], and [10]. The authors of [11] state that when there are few labeled data sets, deep learning models, however, have the propensity to overfit [11]. Deep learning models can only be utilized effectively in a few situations since its time and resource-consuming to gather sufficient-labeled data [9].

In this paper, we propose a solution to tackle the manual annotation and the classification of idiomatic expressions found in tweets’ datasets. The framework uses state-of-art deep learning for language modeling (pre-trained BERT model). It utilizes an external knowledge base to enrich or expand the context of a given tweet. We assume that words used in idioms affect one another and interact to produce different meanings; we also believe that words in sentences have varying importance to the automatic creation of the idiomatic sentiment lexicon; and finally, we assume that words in sentences have different meanings when used alone. The usual overhead associated with standard data augmentation techniques, which calls for BERT or another pre-trained model to be fine-tuned, is also unnecessary.

II. RELATED WORK

A. IDIOMATIC LEXICON-BASED SENTIMENT ANALYSIS

An idiomatic lexicon for sentiment classification, such as the work of [12], has been applied in the early days of sentiment analysis research. The authors developed a sentiment classification system for customer reviews using a lexicon-based methodology. They use sentimental words, idiomatic lexicon, and other linguistic features to identify and extract sentiment patterns from text. For their dictionary, they painstakingly gathered more than 1000 English idioms. They pointed out that even though collecting and annotating idioms takes time, most of them communicate potent sentiments. The authors in [13] have developed a comprehensive lexicon of Japanese multivord expressions composed of idioms, clichés, and quasi-idioms. The dictionary’s coverage of alternative notations and derived forms makes it applicable in various contexts. However, it is a dictionary that describes a semantic conceptual system. Therefore, sentiment analysis cannot be performed on it directly.

The authors in [14] introduce (PSenti) for sentiment analysis. To identify sentiment polarity and gauge the intensity of polarity in online reviews, they proposed a hybrid strategy combining lexicon-based and machine-learning approaches. A set of 40 English idioms, 116 emoticons, and a feeling word lexicon are used in the hybrid technique to attain excellent accuracy. They determine the polarity strength of the text by giving each emotion pattern a score: \([-1, +1]\) for a sentiment word, \([-2, +2]\) for an emoji, and \([-3, +3]\) for an idiom, which is thought to be the most powerful kind of feeling. In [15], Chinese idioms were identified and extracted from text using an unsupervised sentiment classifier. To create an idiomatic sentiment lexicon, they gathered more than 24000 idioms and chose about 8000 examples, and annotated them with positive and negative orientations. The lexicon that the classifier was trained on. Using three publicly accessible Chinese reviews annotated corpora, they assess the performance of their classifier and the scope of their lexicon (book, hotel, and notebook PC).

Ibrahim et al. proposed an idiomatic sentiment lexicon of contemporary standard Arabic and Egyptian dialects AIPSeLEX [16]. They describe their manual efforts to create the AIPSeLEX vocabulary, which contains 3632 idioms and proverbs. Their experiments have shown that utilizing idioms
as a feature enhances the process of emotion classification even though they use a simplified cosine similarity only and Levenshtein distance [16]. Williams et al. have illustrated the importance of using idioms in the sentiment classification task. They demonstrate that these idiomatic features greatly enhance the overall sentiment analysis performance [17]. They produced a Lexico-semantic resource consisting of 580 idioms annotated with sentiment polarity. They also put into practice a collection of local grammar to identify these idioms when they appear in the text. They show how it is difficult to rigorously examine the function of idioms in sentiment analysis due to their relative scarcity. The main disadvantage of this approach is the substantial time investment needed to manually develop lexico-semantic criteria for detecting idioms and their polarity; to annotate the lexico-semantic with the proper polarity, they used crowdsourcing. This approach has been successful, although idiom polarity acquisition is not automatic.

Interesting research was proposed in [18]. The authors provide criteria for idiom detection in the text. In addition, they propose a technique for automatically generating lexical semantics for sentiment polarity classification tasks. Although the preliminary findings show that even this straightforward method (concatenating idiom and sentence polarity) considerably enhances sentiment analysis results, this approach mostly tends to adopt the polarity of the idiom over the polarity of the sentence, and it does not guarantee an optimal performance [18], [19]. Instead of using this straightforward way, we suggest an automated feature integration strategy that preserves the original “positional context” in the embedding of the idiom within the original tweet using an idiom-expansion algorithm. To improve the effectiveness of word disambiguation for sentiment analysis, the authors of [20] suggest a technique to recognize Chinese metaphors using a neural network-based labelling methodology. They conclude that metaphorical utterances often have a stronger emotional impact than literal expressions. By mixing and interacting the meanings of the source and target semantics in metaphors, emotional content may be produced. This method could be helpful for jobs involving sentiment classification, even though the paper doesn’t address this problem.

In [21], the authors outline a method to build a dictionary of idiomatic sentiment expressions and show how to use it to create a sentiment corpus. They describe a method to extend the corpus and calculate the idiomatic phrase sentiment by evaluating them with more than a hundred sample sentences. They set an acceptance threshold to assign the proper sentiment category or class for the idiom. They claim that around 50% of the idioms provide accurate sentiment estimates. The main issue with this work is that a “new idiom” sentiment polarity is estimated based on the surrounding text only while ignoring the polarity strength that can be conveyed by the idiom itself. They conclude that it might be challenging to ascribe a sentiment to an idiom on its own, and therefore, they suggest that it is required to create a dictionary to incorporate the idiom’s context. In contrast, we suggest an expansion method to incorporate the polarity of an idiom itself while computing the overall sentiment of the given tweet. According to a separate research, [22] improved BERT and RoBERTa should be used to build an ensemble model for idiom and literal (literal meaning) recognition. Even though their primary goal is to discover idioms, they show how language models like BERT may be useful in getting semantic properties of idioms through fine-tuning.

The authors of [23] proposed an algorithm to extract and classify the sentiment of Persian text with idioms using an extended version of PerSent, a manually labeled sentiment lexicon. They employed different classification methods to compute the overall sentiment for a given Persian text that contains an idiomatic expression. Our method is different as we claim to automatically annotate the lexicon and then utilize it in the deep learning classifier. In [24], the authors discuss a hypothesis on whether idioms are compositional or not for sentiment or semantics. The results of their investigation suggest that idioms are non-compositional for both sentiment and meaning since there is no consistent link between component-wise sentiment polarities and crowdsourced phrase-level classifications. They conclude that idioms are a prime example of a phenomenon where the non-compositionality of sentiment is not specified or immediately clear, and the absence of a link between component words and phrase-level sentiment drives the need for further study of how to deal with idioms in context.

B. PRETRAINED LANGUAGE MODELS

Pre-trained semi-supervised models are revolutionizing existing NLP technology. Semi-supervised learning has drawn much interest since it is one of the paradigms that employs unlabeled data in the most promising ways to complete this classification assignment [25]. The outcomes from these models are noticeably better than earlier models. However, there are hundreds of millions of parameters in these models [25]. For example, BERT is one of these models that has the most representation, but it has a sizable number of parameters, vast in scale, and has a significant latency.

BERT employs an attention mechanism and is considered as one of the most significant advances in natural language processing. The attention mechanism is suggested as a solution to the persistent dependency issues in models that employ a single context vector condensing all data from earlier time steps. Capturing the whole semantics of the texts is made feasible by the attention method; the model may use hidden states from several previous steps and the attention mechanism to determine the relevance of the input concerning the current time step. To capture intricate linguistic patterns, language models have frequently employed the transformer acting as the fundamental design of the attention mechanism [11]. It is true that these pre-trained language models have high requirements for the quality of the training dataset and rely on large-scale data sets, but it is also true that these models excel in a range of NLP tasks [7], [11].
BERT is a pre-trained language model that encodes the context and the meaning of words into densely packed vectors. As shown in Fig. 1, the base BERT model uses \( n = 12 \) layers of transformers block (encoders) with a hidden size of 768 and several self-attention heads and has more than a hundred million trainable parameters [26]. Other BERT variants have a different number of encoders. Re-training BERT requires decent hardware and a massive dataset to reproduce a fine-tuned model to handle a new downstream task, which might have a very limited amount of data. BERT was trained to solve MLM (Masked Language Model) and NSP (Next Sentence Prediction) [26].

The pre-training and fine-tuning paradigm has been successfully used for a number of NLP tasks in recent years, notably in the NLP sector. In this method, a language model should be pre-trained on a sizable corpus and then refined for the target job, according to the [27] proposal. Slanted Triangular Learning Rates and Progressive Unfreezing are two innovative approaches in this model. Researchers are inspired by the pre-trained models’ strong performance and get exceptional results even with sparsely labelled data. To achieve various goals, such as language modelling and masked language modelling, pre-trained models are frequently used. The effectiveness of previously-trained models likewise improves with larger training data sets [28], [29].

In this paper, we provide a solution that does not rely on the literal meaning of the terms constituting the idiom but rather on the expansion form that identifies the actual meaning and the purpose of it. In addition, our proposal provides evidence that it is unnecessary to perform a model fine-tuning which might be extra overhead and unstable in many cases.

III. THE PROPOSED METHOD

Fig. 2 shows the proposed architecture and the process flow diagram. The first phase is designed to manage and prepare the idiomatic lexicon. Unlike the proposed method in [30], this research aims to automatically annotate the idioms in the lexicon and to compare the error rate of the automation process. The second phase involves the data augmentation unit and the connection to the external knowledge bases.

In this research, we have tested different online Thesaurus and dictionaries based on the coverage and the inclusion of the idiom definition availability. We utilized the urban dictionary for the experimental part. The final phase is implementing the BERT model for the sentiment classification task.

A. IDIOMATIC SENTIMENT LEXICON

In our previous research [30], we extended the SliDE lexicon by adding 3930 new idioms. The original Sentiment Lexicon of IDiomatic Expressions (SliDE) of 5000 labelled idioms [31]. We used the eSliDE sentiment lexicon to extract sentiment labels for the 8,930 idioms in that dataset [30]. The sentiment labels were assigned by the lexicon based on a majority vote of at least 10 crowdsourced annotations for each sentence. Table 1 displays the distribution of the idioms. To determine the reliability of the annotated tweet datasets, Krippendorff’s alpha coefficient is used to measure the inter-annotator agreement as in (1).

\[
\alpha = 1 - \left( \frac{D_o}{D_e} \right) 
\]

\( D_o \) denotes the actual disagreement, which is the percentage of items on which both annotators agree. \( D_e \) represents the expected disagreement when annotations are given at random. On the tweet dataset, the agreement was estimated as \( D_e = 0.701, D_o = 0.213, \) and \( \alpha = 0.696\% \). In the experiment section, we compare the accuracy of the proposed automated sentiment annotation with the manual sentiment annotation of eSliDE lexicon.

B. TWEET DATA COLLECTION

The data collection module extracts a customized dataset from the Twitter platform. We utilize Twitter API to retrieve
TABLE 1. Sentiment polarity distribution of idioms.

| Idiom Polarity   | # of idioms |
|------------------|-------------|
| Positive idioms  | 2612        |
| Negative idioms  | 2892        |
| Neutral idioms   | 3426        |

tweet data from Twitter. There are existing few benchmark Tweets datasets publically available on different websites such as Kaggle and GitHub. However, for this research, these datasets do not include idioms; therefore, our customized tweet collections are domain independent and gathered using queries related to every idiom in the eSliDE idiomatic lexicon. As shown in Fig. 2, we designed the code to retrieve exactly five tweets without redundancy. To keep the sentiment distribution as balanced as possible, we use 1000 idioms per polarity. The total number of retrieved tweets is 15,000 tweets. After preparing our idiom dataset, we launched a query using Twitter API to harvest tweets containing idioms from the eSliDE lexicon dataset.

C. TWEET PREPROCESSING
The typical sentiment analysis solution starts with preprocessing phase. Usually, social media data are mixed with emojis, URLs, hashtags, stop words, numbers, dates, and other features. It’s widespread for researchers to start cleaning the dataset to reduce the processing time. They argue that such noisy data is useless and does not influence the system’s accuracy.

We test this argument and check whether this assumption holds always true. To generate a clean dataset, we manipulate the raw tweets by removing noisy URLs and applying some standard text pre-processing procedures:
1) Stop word removal: Eliminating useless encoding of words missing in any pre-trained word embedding.
2) Case folding: Converting words or phrases to lowercase.
3) Mapping unique values for their type (for example, “9-08-2022” → “DATE”).
4) Special character removal: Removal of hashtags, numbers, punctuation marks, and characters other than letters of the alphabet.
5) Acronym normalization (for example, “UK” → “United Kingdom”) and abbreviation normalization (for example: ‘IDK’ → “I don’t know”).
6) Spelling correction.

D. DATA AUGMENTATION
Data augmentation is a common approach to expanding the training dataset size. In computer vision, speech, and other technologies, many academics are focused on data augmentation. In contrast, there has yet to be as much study on text data augmentation, and there is yet to be a common approach. A dictionary, thesaurus, or database of synonyms is frequently used to substitute words in sentences. The alternative method for locating related terms without a dictionary is to utilize distributed word representation. Synonym augmentation is an approach that fits under this category. Changing the language’s phrasing intentionally is the most excellent way to augment, but this approach is too costly. Therefore, replacing words or phrases with their equivalents is the most practical alternative for data augmentation in most studies. The most well-known open-source lexical database for the English language, for instance, is WordNet. To locate phrases that are semantically related, a new method called semantic similarity augmentation uses word embedding and distributed word representation. You require either pre-trained word embedding models for the relevant language or enough data from the desired application for this approach to function. The advantage of this method is that additional dictionaries are not required to discover synonyms. The last method is known as back-translation. In this method, a phrase or a sentence is translated into another language (forward translation), and then translating the results back into the original language [7].

E. TWEET EXPANSION
To accomplish automatic data expansion, we designed a simple strategy to generate the augmented tweets dataset, and we call it the Non-Compositional Replacement Strategy (NCRS). In this strategy, as shown in Fig. 3, we replace an idiom, using the dom function, with the equivalent meaning/definition sentence from the urban dictionary (or another resource if the idiom definition was not found). For example, a yes-man idiom in this tweet “Fauci is a political yes-man” is replaced by “Fauci is a political weak person who always agrees with their political leader or their superior at work.”

It’s crucial to maintain the original aim of the freshly enriched tweets. To ensure that the learning model is exposed to idiom variants, such as its pertinent definitions or meanings, which might have different sentiment polarity. Certain
Algorithm 2 Idiom Expansion

procedure EXPAND_TWITTERS(F: idiomatic lexicon, a: knowledge base D: set of tweets) 
\[
\begin{align*}
\chi_{temp} & \leftarrow \Phi; \\
\mathcal{F}_{temp} & \leftarrow \Phi; \\
\text{for all } i \in D & \text{ do} \\
& \text{for all } j \in \mathcal{F} \text{ do} \\
& \phi \leftarrow \operatorname{MixUp}([\phi[i], i]); \\
& \chi_{temp} \leftarrow \chi_{temp} \cup \text{WebCrawler}(a, \phi) \\
& \text{dom}(\chi_{temp}) = \{ i \in D \mid \chi_{temp} \neq \chi_{temp} \} \\
& \mathcal{F}_{temp} \leftarrow \mathcal{F}_{temp} \cup \text{dom}(\chi_{temp}) \\
& \text{end for} \\
& \text{end for} \\
& \text{return } \mathcal{F}_{temp} \\
\end{align*}
\]

FIGURE 4. Pseudo code of idiom expansion algorithm.

idioms might hold bipolar sentiments (i.e., there are rare instances when we cannot ensure the same polarity). For instance, depending on the context, the phrase “I had a blast” might mean either “had fun” or “gone bonkers.” Therefore, the model must return all available definitions.

F. BERT FINETUNING

Fine-tuning BERT is a customization procedure of re-training the BERT model using your own custom data to solve other specific tasks. For example, to solve downstream tasks like sentiment analysis. The method is to freeze the early layers of the model and modify its architecture by adding some new layers at the end. Thus, in this method, the model can be retrained on a relatively small dataset for a specific domain.

To simplify the experiment, we utilized a fine-tuned version of BERT from Huggingface called Twitter-roBERTa-base. This model was trained on 58M tweets and fine-tuned for sentiment analysis with TweetEval benchmark [32]. Our goal is to fine-tune Twitter-roBERTa-base again to retrain it on the idiomatic tweets since the TweetEval was aimed for sentiment classification of English tweets without considering idioms. Therefore, we retrained the RoBERTa model using eSliDE dataset. While fine-tuning, we freeze only the last 4 layers out of the 12 layers of the original model. The RoBERTa model output layer gives three distinct sentiment percentage values of each estimated percentage that the tweet is sharing. For example, a tweet “I am happy” is classified as {0.002 as negative, 0.019 as neutral, and 0.979 as positive}. We used AdamW optimizer with a batch size= 16, learning rate= 5e-5, and the number of epochs= 3. For training and validation, we split the eSliDE dataset into 80/20 percentages.

IV. EXPERIMENTS

We provided the pre-trained classifier with the idioms in two versions to test the efficiency of the automatic annotation and the precision of the suggested approach. As can be seen in Table 2, the first section, under the column Manual Sentiment Annotation (eSliDE), displays a sample of the original eSliDE annotation and the percentages attained after manually labeling the idioms with their sentiment based on the 10 annotators. As we complete the automated annotation, this column is the reference for computing the error ratio.

With the exception of labels with aquatic annotations and no idiom enrichment, the second section of the table provides the same information as the first part. The emotion polarity with the largest percentage was given the majority label in the last subcolumn. If the labels received an equal number of votes, the negative label takes precedence over the positive and neutral labels. The positive label receives most votes when the positive and neutral labels are similar. We can see from the sample that in Table 2, With Idiom Expansion, the classification and voting percentage are closer to the manually categorized idioms. We have computed the annotation error ratio as in (2). The \( \delta \) represents the percent error, and the \( \upsilon_A \) and \( \upsilon_E \) are the actual observed and the expected values, respectively.

\[
\delta = \left| \frac{\upsilon_A - \upsilon_E}{\upsilon_E} \right| \cdot 100\% \quad (2)
\]

Table 3 shows that the idiom misclassification error rate using the idiom expansion method was dropped for all sentiment polarity classes. However, we noticed that the error rate was the highest in classifying the neutral labels. The interesting about this result is that the uncertainty of the polarity of a bipolar idiom might be mitigated by integrating the multi-definitions of each idiom.

We have conducted the experiment using the different settings as described in the previous section. The precision, recall, and F1-Score results are shown in Table 4. For validation, we use a dataset of 2000 idioms with two forms (raw and expanded forms). In the beginning, we use the RoBERTa model without any fine-tuning using the raw tweets containing the idioms. The raw tweets mean that we perform tweet pre-processing without expanding the idiomatic expressions. The F-score achieved is 81%. We notice that even after fine-tuning the RoBERTa, the F-score moves less than 1.5 points, as shown in the second column of Table 4. In the last part of Table 4, the idiom enrichment method achieves an amazing jump in the F-score even without any fine-tuning of the model.

As was previously indicated, certain idioms are ambiguous by nature and may have a muddled “sentiment” of their own. These idioms change depending on the context they appear in. For instance, the expression “tough as nails” may mean either “manage any difficulty” or “be cruel and unfeeling”. Another example is the expression “well-padded,” which may also mean either “being rich” or “being fat,” and depending on the context in which it appears, can either be taken as praise or as an insult. Even human annotators struggle to identify which polarities to assign for idioms supplied to them in isolation.

We conduct another experiment to find out whether a bipolar idiom can be appropriately categorized based on a definition chosen at random or by combining all definitions.
TABLE 2. Automatic vs. manual annotation of sentiment lexicon.

| Idiom                  | Manual Sentiment Annotation (eSLIDE) | No Idiom Expansion | With Idiom Expansion |
|------------------------|-------------------------------------|--------------------|----------------------|
|                        | Pos   | Neg    | Neu | Maj. Label | Pos   | Neg    | Neu | Label | Pos   | Neg    | Neu | Label |
| against the grain      | 0     | 0.6    | 0.4 | Neg       | 0.019 | 0.410  | 0.571 | Neu    | 0.009 | 0.656  | 0.335 | Neg    |
| Kick the bucket        | 0.1   | 0.5    | 0.4 | Neg       | 0.046 | 0.450  | 0.504 | Neg    | 0.059 | 0.643  | 0.297 | Neg    |
| fresh legs             | 0.5   | 0      | 0.5 | Pos       | 0.283 | 0.066  | 0.651 | Neu    | 0.108 | 0.096  | 0.796 | Neu    |
| kettle of fish         | 0     | 0.5    | 0.5 | Neg       | 0.128 | 0.176  | 0.695 | Neu    | 0.009 | 0.896  | 0.694 | Neg    |
| lost in translation    | 0     | 0.5    | 0.5 | Neg       | 0.029 | 0.494  | 0.477 | Neg    | 0.02  | 0.55   | 0.421 | Neg    |
| at cross-purposes      | 0     | 0.6    | 0.4 | Neg       | 0.099 | 0.101  | 0.799 | Neu    | 0.008 | 0.751  | 0.242 | Neg    |
| back to the wall       | 0     | 0.5    | 0.5 | Neg       | 0.033 | 0.388  | 0.579 | Neu    | 0.021 | 0.644  | 0.335 | Neg    |
| moment in the sun      | 0.5   | 0      | 0.5 | Pos       | 0.396 | 0.005  | 0.599 | Neu    | 0.703 | 0.352  | 0.471 | Pos    |
| bang on                | 0.6   | 0      | 0.4 | Pos       | 0.306 | 0.112  | 0.582 | Neu    | 0.609 | 0.228  | 0.363 | Pos    |
| beg to differ          | 0     | 0.5    | 0.5 | Neg       | 0.075 | 0.201  | 0.724 | Neu    | 0.016 | 0.589  | 0.395 | Neg    |

TABLE 3. Error percentage of the automated sentiment polarity annotation.

| Δ of annotating sentiment class | Without Expansion | With Expansion |
|---------------------------------|-------------------|----------------|
| Positive label                  | 27.55%            | 13.48%         |
| Negative label                  | 36.13%            | 18.58%         |
| Neutral label                   | 41.82%            | 22.31%         |

TABLE 4. F1-score comparison raw tweets and an enriched using the base classifier and a fine-tuned version.

| Sentiment Classification (2000 manually annotated tweets) |
|----------------------------------------------------------|
| Metric                  | Tweets without Expansion | Using Expansion |
|                         | Baseline RoBERTa | Fine-Tuned RoBERTa | Baseline RoBERTa |
| Precision               | 0.794         | 0.814         | 0.94        |
| Recall                  | 0.825         | 0.830         | 0.975       |
| F1-score                | 81.09%        | 82.20%        | 95.72%      |

if there are several meanings available. For this experiment, we have identified 150 tweets with varying idioms having definitions that might have bipolar meanings. Table 5 shows that using random definitions has inconsistent behavior, with some idioms being incorrectly classified while others benefit from being correctly classified. On the other hand, the F-score metric has been enhanced by idiom expansion using the multi-definitions fusion method only once out of the four runs.

In each run, the single-definition selection is done randomly (not the first definition found in the thesaurus or the dictionary). As shown in Table 5, the accuracy of the multi-definition fusion and the no-expansion were kept the same as the setting where no changes in each run of the experiment.

In the last experiment, our goal was to evaluate and validate the accuracy of the sentiment classification of the automated process. To achieve this, we conduct another experiment on the unlabeled tweet dataset collection that we harvested as in described subsection before. We compute the accuracy as shown in formula (3):

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{total predictions}}$$  (3)

We had an assumption that the sentiment polarity of a tweet should hold the same sentiment polarity of the idiom it contains. However, this assumption might only be true if the intention or the meaning of an idiom might be affected by the contextual text of the tweet. Unfortunately, it’s impractical to manually annotate the rest of the unlabeled (13,000) tweets as a benchmark reference to directly compute the accuracy based on our assumption. Therefore, we automatically annotate the whole dataset based on the sentiment polarity using our method and then ask each annotator to manually annotate a subset containing 500 random tweets. In this experiment, we do not use a “majority label” as before based on the inter-annotator agreement, but rather we swap the pair 500-tweets...
TABLE 6. Accuracy and consistency of the tweets classification using the automatic-based annotation method.

| 500 tweets datasets | Annotator ID | Sentiment Classification Accuracy | Difference |
|---------------------|--------------|----------------------------------|------------|
| Dataset 1           | 1            | 89%                              |            |
|                     | 2            | 88%                              | 1%         |
| Dataset 2           | 2            | 92%                              |            |
|                     | 1            | 92%                              | 0%         |
| Dataset 3           | 3            | 94%                              |            |
|                     | 4            | 96%                              |            |
| Dataset 4           | 4            | 95%                              |            |
|                     | 3            | 95%                              | 0%         |
| Dataset 5           | 5            | 88%                              |            |
|                     | 6            | 90%                              | 1%         |
| Dataset 6           | 6            | 87%                              |            |
|                     | 5            | 87%                              | 0%         |
| Dataset 7           | 7            | 79%                              |            |
|                     | 8            | 85%                              | 6%         |
| Dataset 8           | 8            | 83%                              |            |
|                     | 7            | 85%                              | 2%         |
| Dataset 9           | 9            | 93%                              |            |
|                     | 10           | 93%                              | 0%         |
| Dataset 10          | 10           | 94%                              |            |
|                     | 9            | 93%                              | 1%         |

Average accuracy 90%

subsets between every two annotators. This method can give us a clue about how roughly accurate and consistent the automatic annotation is compared to the manual annotation. The accuracy of tweet sentiment classification in each subset of this experiment is illustrated in Table 6.

V. CONCLUSION AND FUTURE WORK
In order to understand the task of classifying the sentiment of tweets using an idiomatic sentiment lexicon, this research suggests an augmentation technique using the idiom expansion method. It employs a full replacement of an idiom with its factual meaning or definition retrieved from an external knowledge source. The objective was to assess the validity and accuracy of using raw data vs. the expansion approach employing a fine-tuned BERT embedding model. The findings of this study demonstrate that data expansion is quite beneficial and that it may be employed even if the embedding model does not need to be fine-tuned.

Future research will examine the impact of altering training and testing datasets, the embedding model, and the expansion method on the outcomes. The last experiment raised another unanswered question: how can we ensure that the classification model will function well when the expansion procedure is done at random on all English idioms? We think it will be highly worthwhile to go further in order to address this query by modifying the expansion process to take into account the general sentiment of the tweet itself before choosing the appropriate meaning/definition of the idiom.

REFERENCES
[1] Statista. (Sep. 2022). Number of Worldwide Social Network Users 2022[Statista]. Accessed: Sep. 16, 2022. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
[2] K. Rudra, A. Chakraborty, N. Ganguly, and S. Ghosh, “Understanding the usage of idioms in the Twitter social network.” in Pattern Recognition and Big Data, Singapore: World Scientific, 2017, pp. 767–788.
[3] P. Lin, X. Luo, and Y. Fan, “A survey of sentiment analysis based on deep learning,” Int. J. Comput. Inf. Eng., vol. 14, no. 12, pp. 473–485, 2020.
[4] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 1–9.
[5] Y. Chen, “Convolutional neural network for sentence classification,” M. Sc dissertation, School Comput. Sci., Univ. Waterloo, Waterloo, ON, Canada 2015.
[6] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” 2016, arXiv:1606.01781.
[7] X. Gong, W. Ying, S. Zhong, and S. Gong, “Text sentiment analysis based on transformer and augmentation,” Frontiers Psychol., vol. 13, 2022, Art. no. 90661.
[8] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, Aug. 2018.
[9] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaria, M. A. Padhel, M. Al-Amidie, and L. Farhan, “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions;” J. Big Data, vol. 8, no. 1, pp. 1–74, Dec. 2021.
[10] O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, “Combining lexicon and learning based approaches for concept-level sentiment analysis,” in Proc. Int. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM), New York, NY, USA, 2012, pp. 1–8.
[11] S.-X. Xie and T. Wang, “Construction of unsupervised sentiment classifier for modern standard Arabic and colloquial sentiment analysis,” 2015, arXiv:1506.01906.
[12] H. S. Ibrahim, S. M. Abdou, and M. Gheith, “Idioms-proverbs lexicon for modern standard Arabic and colloquial sentiment analysis,” 2015, arXiv:1506.01906.
[13] L. Williams, C. Bannister, M. Arribas-Ayllon, A. Preece, and I. Spasic, “The role of idioms in sentiment analysis,” Exp. Syst. Appl., vol. 42, no. 21, pp. 7375–7385, Nov. 2015.
[14] I. Spasic, L. Williams, and A. Baerki, “Idiom-based features in sentiment analysis: Cutting the gordian knot,” IEEE Trans. Affect. Comput., vol. 11, no. 2, pp. 189–199, Apr. 2020.
[15] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments and Emotions. Cambridge, U.K.: Cambridge Univ. Press, 2015.
BASHAR M. A. TAHAYNA received the B.S. degree in computer science from An-Najah National University, Nablus, Palestine, in 1998, and the M.Sc. degree in media informatics from RWTH Aachen University, Aachen, Germany, in 2006. He is currently pursuing the Ph.D. degree in computer science with Universiti Tunku Abdul Rahman, Perak, Malaysia. From 2007 to 2019, he was a Research and Teaching Assistant at different universities in Palestine and Malaysia. He has published high-quality papers on natural language processing, image processing, and computer vision. His research interests include the development of multimedia processing and retrieval systems. He is serving as a reviewer and a PC chair for different journals and conferences.

REHAN AKBAR (Senior Member, IEEE) received the Ph.D. degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2013. From 2001 to 2008, he was working as a Lecturer with GC University, Lahore, Pakistan. Later, he joined IT Industry as a Project Manager. In 2012, he joined the Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia, where he was an Associate Professor and the Head of the Department until 2021. Since 2022, he has been working as an Associate Professor with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS. He is associated with the Positive Computing Research Centre under Institute of Autonomous Systems. His research interests include software development processes and methodologies, process tailoring, agile methodologies, big data, and cybersecurity. He is a member of the Association for Information Systems. He received the Professional Technologist (T.S./P.Tech.) status from the Malaysian Board of Technologists and Teaching Excellence Award, in 2016. He is serving as the editor, a reviewer for different journals and conferences and a member for editorial board.

RAMESH KUMAR AYYASAMY (Senior Member, IEEE) received the Ph.D. degree in information technology from Monash University, Australia, in 2013. From 2003 to 2008 and 2013 to 2014, he worked as a Lecturer at Tamil Nadu, India, and Monash University, Malaysia, consecutively. Since 2015, he has been working as an Assistant Professor at the Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia. His research interests include artificial intelligence, big data analytics, cyberbullying, deep learning, machine learning, and text mining. He is serving as a reviewer for different journals and conferences and a member for editorial board.

[20] X. Chen, Z. Hai, S. Wang, D. Li, C. Wang, and H. Luan, “Metaphor identification: A contextual inconsistency based neural sequence labeling approach,” *Neurocomputing*, vol. 428, pp. 268–279, Mar. 2021.

[21] K. Matsumoto, S. Tsuchiya, M. Yoshida, and K. Kita, “Construction and expansion of dictionary of idiomatic emotional expressions and idiomatic emotional expression corpus,” *Int. J. Comput. Softw. Eng.*, vol. 6, no. 2, 174, 2021.

[22] J. Brikisal and C. N. Subalalitha, “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa,” *Inf. Process. Manage.*, vol. 59, no. 1, Jan. 2022. Art. no. 102756.

[23] K. Dashtipour, M. Gogate, A. Gelbukh, and A. Hussain, “Extending Persian sentiment lexicon with idiomatic expressions for sentiment analysis,” *Social Netw. Anal. Mining*, vol. 12, no. 1, pp. 1–3, Dec. 2022.

[24] A. Hwang and C. Hidey, “Confirming the non-compositionality of idioms for sentiment analysis,” in *Proc. Joint Workshop Multitask Word Sense Disambiguation*, 2018, pp. 114–119.

[25] N. V. Chawla and G. Karakoulas, “Learning from labeled and unlabeled data: An empirical study across techniques and domains,” *J. Artif. Intell. Res.*, vol. 23, pp. 331–366, Mar. 2005.

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2019, arXiv:1810.04805.

[27] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” 2018, arXiv:1801.06146.

[28] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” 2019, arXiv:1908.10063.

[29] S. M. Rezaeinia, R. Rahmani, H. Veisi, and A. Ghodsi, “Sentiment analysis based on improved pre-trained word embeddings,” *Exp. Syst. Appl.*, vol. 117, pp. 139–147, Mar. 2019.

[30] J. Briskilal and C. N. Subalalitha, “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa,” in *Proc. Joint Workshop Multitask Word Sense Disambiguation*, 2018, pp. 1–6.

[31] C. Jochim, F. Bonin, R. Bar-Haim, and N. Slonim, “SLIDE—A sentiment cardinalnlp/twitter-roberta-base-sentiment Hugging Face (2022).

[32] B. Tahayna, R. K. Ayyasamy, R. Akbar, N. F. B. Subri, and A. Sangodiah, “Lexicon-based non-compositional multiword augmentation enriching tweet sentiment analysis,” in *Proc. Joint Workshop Multitask Word Sense Disambiguation*, 2018, pp. 114–119.

[33] S. Tsuchiya, M. Yoshida, K. Matsumoto, and K. Kita, “Construction and expansion of dictionary of idiomatic emotional expressions and idiomatic emotional expression corpus,” *Int. J. Comput. Softw. Eng.*, vol. 6, no. 2, 174, 2021.

[34] J. Brikisal and C. N. Subalalitha, “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa,” *Inf. Process. Manage.*, vol. 59, no. 1, Jan. 2022. Art. no. 102756.

[35] K. Dashtipour, M. Gogate, A. Gelbukh, and A. Hussain, “Extending Persian sentiment lexicon with idiomatic expressions for sentiment analysis,” *Social Netw. Anal. Mining*, vol. 12, no. 1, pp. 1–3, Dec. 2022.

[36] A. Hwang and C. Hidey, “Confirming the non-compositionality of idioms for sentiment analysis,” in *Proc. Joint Workshop Multitask Word Sense Disambiguation*, 2018, pp. 114–119.

[37] N. V. Chawla and G. Karakoulas, “Learning from labeled and unlabeled data: An empirical study across techniques and domains,” *J. Artif. Intell. Res.*, vol. 23, pp. 331–366, Mar. 2005.

[38] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2019, arXiv:1810.04805.

[39] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” 2018, arXiv:1801.06146.

[40] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” 2019, arXiv:1908.10063.

[41] S. M. Rezaeinia, R. Rahmani, H. Veisi, and A. Ghodsi, “Sentiment analysis based on improved pre-trained word embeddings,” *Exp. Syst. Appl.*, vol. 117, pp. 139–147, Mar. 2019.

[42] B. Tahayna, R. K. Ayyasamy, R. Akbar, N. F. B. Subri, and A. Sangodiah, “Lexicon-based non-compositional multiword augmentation enriching tweet sentiment analysis,” in *Proc. Joint Workshop Multitask Word Sense Disambiguation*, 2018, pp. 1–6.

[43] C. Jochim, F. Bonin, R. Bar-Haim, and N. Slonim, “SLIDE—A sentiment lexicon of common idioms,” in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2012, pp. 1–6.

[44] B. M. A. Tahayna et al.: Automatic Sentiment Annotation of Idiomatic Expressions for Sentiment Analysis Task

[45] B. M. A. Tahayna: Automatic Sentiment Annotation of Idiomatic Expressions for Sentiment Analysis Task