Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.036648
Article

Applying English Idiomatic Expressions to Classify Deep Sentiments in COVID-19 Tweets

Bashar Tahayna and Ramesh Kumar Ayyasamy*

Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, 31900, Malaysia
*Corresponding Author: Ramesh Kumar Ayyasamy. Email: rameshkumar@utar.edu.my
Received: 08 October 2022; Accepted: 29 December 2022; Published: 26 May 2023

Abstract: Millions of people are connecting and exchanging information on social media platforms, where interpersonal interactions are constantly being shared. However, due to inaccurate or misleading information about the COVID-19 pandemic, social media platforms became the scene of tense debates between believers and doubters. Healthcare professionals and public health agencies also use social media to inform the public about COVID-19 news and updates. However, they occasionally have trouble managing massive pandemic-related rumors and frauds. One reason is that people share and engage, regardless of the information source, by assuming the content is unquestionably true. On Twitter, users use words and phrases literally to convey their views or opinion. However, other users choose to utilize idioms or proverbs that are implicit and indirect to make a stronger impression on the audience or perhaps to catch their attention. Idioms and proverbs are figurative expressions with a thematically coherent totality that cannot understand literally. Despite more than 10% of tweets containing idioms or slang, most sentiment analysis research focuses on the accuracy enhancement of various classification algorithms. However, little attention would decipher the hidden sentiments of the expressed idioms in tweets. This paper proposes a novel data expansion strategy for categorizing tweets concerning COVID-19. The following are the benefits of the suggested method: 1) no transformer fine-tuning is necessary, 2) the technique solves the fundamental challenge of the manual data labeling process by automating the construction and annotation of the sentiment lexicon, 3) the method minimizes the error rate in annotating the lexicon, and drastically improves the tweet sentiment classification's accuracy performance.

Keywords: Sentiment analysis; idiomatic lexicon; BERT; COVID-19; deep learning

1 Introduction

As per Statista, "globally, more than 4.26 billion individuals used social media in 2021, and this number is expected to rise by approximately six billion by 2027" [1]. The most critical contemporary communication tools are social media platforms. They offer interactive two-way communication and have supplanted traditional publishing platforms for governments, enterprises, and groups. However, the manual analysis of social media data to unearth concealed emotions takes a lot of time and work. Therefore, it is vital to develop a system that would enable computers to process, analyze, and grasp such a large number of data rapidly and effectively.

Several social media platforms have produced a vast amount of information, which has sparked the establishment of numerous new applications in specialized areas of natural language processing (NLP), like sentiment analysis. The primary goal of NLP is to accomplish human-like language processing. Therefore, it enables computers to comprehend and understand the text to extract meaningful information. NLP is "an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things" [2]. The NLP research encompasses a wide range of disciplines, "The foundations of NLP lie in several disciplines such as computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc." [2]. Several applications can benefit from this research field, for example, machine translation, text summarization, information retrieval, question answering, speech recognition, and sentiment analysis. Sentiment analysis uses machine learning or deep learning methods to mine unstructured data autonomously. It seeks to analyze individuals' opinions, attitudes, and feelings through texts created by users, such as those found on social media platforms. It starts with extracting elements from textual input, representing them correctly, and then classifying them into distinct polarities or sentiments [3].

Although social media has lessened this barrier and increased our interconnectedness, understanding the language that is exclusive to one culture still necessitates understanding that culture. Social media users typically communicate using everyday language. But it is challenging to take out and analyze their opinions or feelings from a text. On Twitter, users typically utilize shorthand and figurative language forms whenever possible to substitute extensive word counts because each message restricts a particular number of characters. Therefore, it is prevalent to find that users use a conversational writing style and metaphorical language to express their views or opinions in their tweets.

The shorthand method reduces the desired message to a few words or symbols that are phonetically related and takes into account character limits. Idioms, euphemisms, and colloquialisms can add a dash of courtesy, etiquette, or humor to a divisive topic. Another possible defense would be to avoid a criminal conviction or legal action if the language used was offensive. In [4], the authors examine a hypothesis about whether or not idioms are compositional in terms of emotion or semantics. Since there is no consistent relationship between component-wise sentiment polarities and crowdsourced phrase-level classifications, the findings of their analysis show that idioms are non-compositional for both sentiment and meaning. They conclude that idioms are a classic example of a situation in which the non-compositionality of emotion does not describe or immediately apparent and that the lack of a relationship between component words and phrase-level sentiment necessitates more research on handling idioms in context. Researchers attempt to demonstrate how essential idioms are to the Twitter platform [5]. The fact that idioms are discussed by millions of people on social media sites like Twitter is astonishing [5]. The authors find that idioms made for roughly 10% of Twitter trends over ten months in 2014.

One way to deal with complex data for analysis is to employ sophisticated lexicons that account for the subjectivity or objectivity of the words, context, and intensity. The lexicon or rule-based sentiment analysis alludes to a study by linguists. As a result, a set of rules—also referred to as a sentiment lexicon—that categorize words as either positive, negative, or neutral, together with the matching intensity measure, have been developed [6–13]. For example, if a word indicates a positive sentiment, we should consider how positive it is; there is essentially a distinction between "*excellent, wonderful, and extraordinary.*"

This study [14] created a VADER lexicon (Valence Aware Dictionary and Sentiment Reasoner) to assess the sentiments. It is frequently employed to evaluate social media posts' tone (sentiment). WordNet is also a lexical database of words identified in the Natural Language Toolkit (NLTK) corpus. WordNet is based on the relationships between words. Another well-known lexicon is SentiWordNet [15,16]. SentiWordNet is an improved version of WordNet [17]; the synsets, which are logical collections of cognate synonyms made up of nouns, verbs, adjectives, and adverbs, are the subject of this lexicon. This aids in determining the sentiment polarity relevant to the specific problem occurrence.

The benefits of lexicon-based approaches are that they are primarily utilized as a tool for the unsupervised approach, which eliminates the requirement for prior training. Additionally, they generally demonstrate speedier execution. The rule-based technique may successfully handle fewer issue cases, unlike the Machine Learning-based approach, which needs an extensive dataset to train. Additionally, if the correct vocabulary is used on the right issue instance, the accuracy of the rule-based technique is typically quite excellent with consistent results. Finally, rule-based approaches exhibit lower risk, have undergone testing, and are widely used. The primary drawback of rule-based approaches is that they are situation-specific; for example, VADER performs better in cases of social media issues. Additionally, various lexicons must be tested, and customization is typically not possible. Because the rules are established in stone by the experts and independent of the provided text, they are applied to the given dataset without considering the dataset itself. The rule-based approaches have the additional drawback that they must initially create the rules with the assistance of an expert. Additionally, adding or amending the rules necessitates subject-matter specialists and requires arduous human labor before being included in the completed product. Rule-based approaches cannot also learn.

Despite the popularity of lexicon-based sentiment analysis, machine learning algorithms are the mainstay of most sentiment analysis and classification research [18]. The traditional supervised sentiment classification methods have gained popularity due to their impressive results; nevertheless, the major drawback is that they are domain dependent and require human feature engineering before we can use text datasets. While requiring fewer annotated training datasets, unsupervised sentiment classification methods can overcome the issue of domain dependence. Most unsupervised sentiment classification involves generative models [19–22] and lexicon-based algorithms [23–26]. On the other hand, deep learning has become dominant and surpasses conventional machine learning because it can learn from text without needing the manual feature engineering process [27].

2 Related Work

In this section, we discuss the related work of the body of knowledge about sentiment classification. In general, there are three principal methodologies for sentiment analysis: machine learning-based approaches, lexicon-based methods, and a hybrid methodology, where a linguistic and a machine learning method combined to perform a seamless two-phased sentiment classification model [28–30]. Indeed, researchers define two subcategories of the lexicon-based approaches: The dictionary and the corpus-based approaches. The lexicon-based approach has the benefit of being simple in structure (Corpus or dictionaries). Even building an accurate sentiment lexicon necessitates feature selection from a vast and high-quality amount of labeled data, which might need to be more available and take time and resources. The dictionary-based method classifies sentiments using a predefined lexicon of terms; lexicons are compiled from the document and utilize any online thesaurus to find synonyms and antonyms to extend that lexicon further [28]. In contrast, the corpus-based analysis approach analysis a statistical study of the data, such as the k-nearest neighbors clustering algorithm. Therefore, unlike the dictionary-based method, the corpus-based analysis does not need a preset lexicon.

Deep Learning methods outperform conventional machine learning approaches in terms of accuracy for most NLP tasks [31–34]. However, in the early days, the usage of deep learning in NLP ran into different problems [35]: In the first place, deep learning models needed to be more sophisticated to intricate power models. In the second place, data-driven deep learning models lacked a sizable amount of annotated data, and manual annotation is undisputedly tedious and expensive. Researchers have started steadily paying more attention to pre-training strategies to overcome these problems. The current undeniable overlords of NLP are the enormous pre-trained models known as Transformers. Their design aims to handle long-range input and output dependencies with attention and repetition while resolving sequence-to-sequence tasks. Pre-training and self-attention are the two cornerstones of a successful deep learning-based modern NLP system. When self-attention techniques combine with a lot of unsupervised pre-training, the network can make accurate summarised word and phrase vectors while still keeping track of the entire input sequence. In addition to performing better on empirical testing, pre-trained transformer models may be taught much more quickly than architectures built using recurrent or convolutional layers. The core of this research is the employment of a sentiment lexicon in combination with deep learning as a hybrid model. We address the unmet requirement in Twitter sentiment classification using a language of daily sentiments.

2.1 Idiomatic Lexicon for Sentiment Classification

This study [30] classifies “product reviews” using the lexicon-based method. The lexicon consists of annotated sentiments for idioms and other keywords. They apply a pattern matching on their accumulated 1000 English idioms. They emphasized that expressions convey strong sentiments, even though gathering and annotating them takes time. This study [36] presents a glossary of Japanese vocabulary used in idiomatic expressions. It comprises quasi-idioms as well as well-known Japanese idioms and clichés. As a result of addressing alternate notations and derived forms, the authors illustrate their applicability in various contexts. However, this study cannot directly apply sentiment analysis, as it defines only the semantic conceptual framework. Another hybrid model of machine learning and lexicon-based approaches is proposed by [37]. This study experiments on a collection of 40 English idioms, 116 emoticons, and a vocabulary of feeling words. Each emotion pattern was measured using the text’s polarity strength: [−1, +1] for sentiment words, [−2, +2] for emojis, and [−3, +3] for idioms. Idioms express the strongest emotions and prove they enhance the sentiment classification task.

This study [38] uses a Chinese idiomatic lexicon to train an unsupervised sentiment classifier. The lexicon consists of 8,000 labeled idioms to reflect the sentiments as binary positive-negative values. The authors evaluate their classifier’s effectiveness and vocabulary size using three publically available Chinese product reviews and show that utilizing idioms enhances classification accuracy. This study [39] uses an Arabic idiomatic-based sentiment lexicon, the AIPSeLEX lexicon, which contains more than 3,000 idioms and proverbs. The authors demonstrate that adding idioms as a feature significantly improves the sentiment classification task.

According to [40], idioms may be utilized as features in the sentiment classification task. They demonstrate how this may greatly enhance the performance of sentiment analysis as a whole. They construct and annotate 580 idioms in their lexicon. They also recognize these idioms when they appear in the written text using a set of regional grammar. The paper demonstrates how it might be difficult to adequately examine idioms' functions in sentiment analysis due to their relative scarcity. However, this outcome cannot be generalized since their extensively used corpora have unbalanced usage of idioms. The manual annotation strategy's main flaw is the significant work it demands. Although idiom polarity acquisition is not automated, this approach has worked well for sentiment categorization.

The authors [41] performed a highly intriguing study. In addition to describing a method for automatically producing an idiomatic lexicon, the authors added criteria for idiom recognition in the text. Initial studies suggested that even this modest strategy—combining idiom and sentence polarity—significantly improved sentiment analysis results. However, it is essential to note that this approach prefers the idiom's polarity over the sentence's, which could result in suboptimal outcomes [41,42]. Our proposed method solves this issue by injecting the expanded form of an idiom back into its tweet while the original "positional context" is maintained.

This study [43] uses Chinese idioms and other metaphorical expressions for the word disambiguation problem. They deduce that figurative expressions frequently elicit higher sentimental and emotional responses than literal ones. It may create emotional content in metaphors by combining and interacting with the source and destination semantics meanings. Though the study doesn't specifically address the classification problem, this approach may be helpful for sentiment classification tasks. The authors [44] present a way for creating a sentiment corpus and show how to utilize it to develop an idiomatic sentiment lexicon. To provide the idiom with the appropriate sentiment polarity, they set an acceptance level by estimating the sentiment of an idiomatic expression that consistently appears in more than one hundred example sentences. According to their estimates, around 50% of the idioms offered precise sentiment estimates. The primary problem with this study is that the "new idioms" emotion is solely inferred from the language around them, neglecting the polarity strength that the idioms themselves might express. They even conclude that it could be challenging to determine an idiom's emotion on its own, and they advise that creating a dictionary that includes the idiom's context is essential. Our proposed method aligns with this recommendation.

According to [45], BERT (Bidirectional Encoder Representations from Transformers) and roBERTa (Robustly Optimized BERT Pre-training Approach) have been fine-tuned and used to build an orchestrated model to recognize idioms and their literal meaning. They show that, even though their primary goal was to recognize idioms, language models like BERT and roBERTa may help acquire idiom semantic properties through fine-tuning.

2.2 Transformer-Based Sentiment Classification

Transformers have recently excelled in sentiment analysis tasks due to their complex structure and powerful functions [46]. In Transfer learning, a procedure is followed to allow knowledge transfer. We create and train a neural network language model for a particular task in transfer learning. Then, with little effort, we can optimize the model with fresh data using a fine-tuning approach, enabling us to utilize the same model for the new downstream task. The benefit of transfer learning is that the training takes far less time than starting training a new model from scratch.

Neural networks start their model parameters randomly before training them with optimization algorithms to reduce losses and provide the most accurate possible results. The main idea behind the optimization is to alter weights and rectify the vanishing learning rate. The general agreement nowadays is that Deep Learning methods outperform other machine learning approaches in terms of accuracy for most NLP tasks [32–34].

The current undeniable overlords of the NLP are the enormous pre-trained models known as Transformers. Their design aims to handle long-range input and output dependencies with attention and repetition while resolving sequence-to-sequence tasks. Pre-training and self-attention are the two cornerstones of a successful deep learning-based modern NLP system. The network can generate reliable contextualized word and phrase vectors while keeping track of the whole input sequence thanks to a mix of self-attention methods and thorough unsupervised pre-training. In addition to performing better on empirical testing, pre-trained transformer models may be taught much more quickly than architectures built using recurrent or convolutional layers. Google first introduced Transformers in 2017. Language models such as recurrent neural networks (RNN) and convolutional neural networks (CNN) were primarily used to do NLP tasks. Even though both RNNs and CNNs are capable models, The Transformer is preferred since it does not demand that data sequences be processed in a specific order. Transformers make it possible to train on larger data sets than were previously possible since they can accommodate any order of data. As a result, it became simpler to construct pre-trained models like BERT, which was trained on massive amounts of linguistic data before its release.

For classification, the model is trained in an environment where annotated data is readily available or straightforward to collect. After then, it is fine-tuned and tested in a field where obtaining training data is challenging. Transformers stand out from other AI systems because they can easily be modified (fine-tuned) to operate admirably, even when learning with little or no data. The bulk of transformers is still functional when used off-the-shelf since they have been well-optimized and trained on a lot of data. Deep learning fine-tuning involves using weights from an older model to train a more similar deep learning process to achieve the desired output or increase performance on the downstream task.

Although there is no question about the efficiency of transformers' strategies, there are very few formal comparisons and controlled sandbox studies because of various factors [45]. For example, knowledge bases, ontologies, grammatical characteristics, reasoning, and databases are just a few of the technologies frequently used when employing transformer models. The history of NLP has seen significant investment in developing competitions and cooperative projects to use unlabeled datasets to find answers to particular challenges. Even though this has been crucial for developing NLP research, the experiments prove that even minor changes to the initial random seed can significantly affect model comparison [47]. Therefore, comparing and contrasting a single pre-trained transformer with alternative approaches is challenging.

2.3 BERT Transformer

BERT is a "bidirectional" semi-supervised model that was pre-trained with unlabeled information gathered from the Books Corpus and the English Wikipedia (2,500 M words) (800 M words). BERT produced ground-breaking results throughout its development phases in 11 natural language comprehension tasks, including sentiment analysis. Prior language models like word2vec had trouble and constraints in reading context and polysemous words. However, BERT was able to overcome such limitations. According to research experts in the field, the most significant challenge to understanding natural language is ambiguity, which BERT effectively solves. It can parse words using common sense, essentially human-like.

Google declared in 2019 that they would start using BERT in their American-based production search algorithms. BERT impacts Google search searches by 10%. Users should not optimize search content for BERT, as this search engine strives to provide a superior search experience. Users should make their material and inquiries relevant to the natural user experience and topic. As of 2019, more than 70 distinct languages utilized BERT. However, researchers can use their data to optimize (i.e., fine-tune) these models for downstream tasks (such as classification, entity identification, question answering, etc.) to provide cutting-edge predictions. It can extract excellent linguistic features from text data [48–50]. Various downstream tasks, including named entity recognition, categorization, and question-answering, can be carried out using BERT architecture. The phrase "black box" is often used to describe pre-trained BERTs since they produce H = 768 shaped vectors for each input token (word) in a sequence. Depending on the situation, the arrangement may consist of one or two phrases divided by the separator [SEP] and start with a token [CLS]. As a result, throughout the training phase, BERT learns about both the left and right sides of a token's context (small units of the surrounding text). BERT can forecast concealed words by analyzing the words that come before and after a particular phrase.

3 Method

In Fig. 1, we show the suggested framework composed of modules. The illustration displays the Idiomatic Lexicon module done in our earlier work [51]. The Twitter API modifies the query (idiom: theme) and posts the results to the raw tweet collection. On the other hand, idioms expand by consulting external information sources (the urban thesaurus). Expanded idioms are employed in an experiment to categorize idioms based on their emotional connotations (as shown in Experiment I). The tweet enrichment model adds the expanded idiom to the original tweet, which the classifier uses as test data. As mentioned in Section 3.3, the pre-processing model employs some data cleaning and normalization. The following subsections will explain each module's actions and experiments in more depth.

3.1 Idiomatic Sentiment Lexicon

In our earlier research [51], we added 3,930 different idioms to the 5,000 idioms in the Sentiment Lexicon of IDiomatic Expressions (SliDE). We manually annotated the idioms by a sentiment value based on a majority vote of at least ten crowdsourced annotations for each idiom. Table 1 shows the distribution of the idioms in the extended lexicon (eSliDE). The inter-annotator agreement is measured using Krippendorff's alpha coefficient to assess the dependability of the annotated tweet datasets, as in (1).

$$\alpha = 1 - \left(\frac{D_o}{D_e}\right) \tag{1}$$

D_o stands for the percentage of things on which both annotators disagree, which represents the actual dispute. D_e stands for the anticipated discrepancy when annotations are distributed randomly. The calculated inter-annotator agreement $\alpha = 0.696\%$, whereas $D_e = 0.701$, and $D_o = 0.213$.

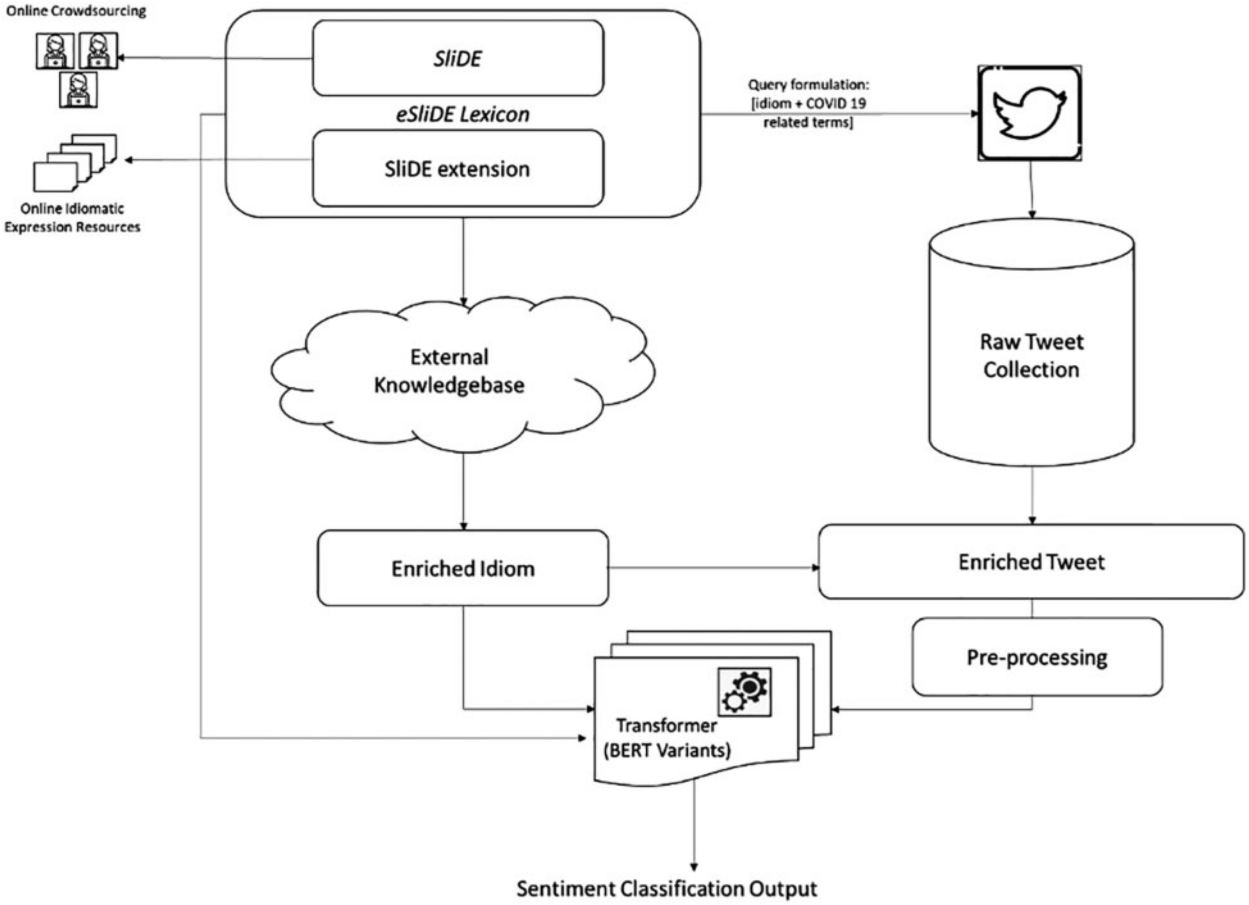

Online Crowdsourcing
SliDE
eSliDE Lexicon
SliDE extension
Online Idiomatic Expression Resources
Query formulation: [idiom + COVID 19 related terms]
Raw Tweet Collection
External Knowledgebase
Enriched Idiom
Enriched Tweet
Pre-processing
Transformer (BERT Variants)
Sentiment Classification Output

Figure 1: The proposed framework architecture

Table 1: eSliDE idioms' polarity distribution

Idiom polarity	# of idioms
Positive idioms	2612
Negative idioms	2892
Neutral idioms	3426

3.2 Data Collection

On various platforms like Kaggle and GitHub, there are now a few benchmark COVID-19 Tweets datasets that are freely accessible. These datasets do not, however, contain idioms for this study. As a result, we leverage the Twitter API, as shown in Fig. 2, to find tweets that include idioms by running searches that are connected to all of the idioms in the eSliDE idiomatic lexicon and include COVID-19-related phrases. For experimental purposes, we set COVID-19-related terms are COVID-19, COVID-19 virus, COVID-19 vaccine, Movement Control Order, COVID-19 MOC, Corona Virus, etc. The WebMD website shows the full glossary of terms.

```
Algorithm 1 Tweet DB Creation
Output← $\mathcal{D}$: $Final\ Tweets\ Dataset$
Input← $\mathcal{F}$ : $Lexicon\ of\ Idiomatic\ Lexicon$; $\sigma$ :$max\ \#\ tweets\ per\ idiom$
Step 1: Initialization
$\mathcal{F}_{temp}$← ϕ; $\mathcal{D}$ ← ϕ;
Step 2: Iterate over the idioms & retrieve relevant tweets
For all $i \in \mathcal{F}$ do
    $count$ ← 0;
    While $count < \sigma$
        $\mathcal{F}_{temp}$← $\mathcal{F}_{temp} \cup$ API. Query$(i)_{fulltext}$
        $count++$;
    End while
    $\mathcal{D}$ ← $\mathcal{D} \cup \mathcal{F}_{temp}$;
End for
Output ≔ $\mathcal{D}$← Tweet collection containing idiomatic expressions
End procedure
```

Figure 2: Pseudo code of tweet collection algorithm

We created the code to extract precisely five tweets without duplication. We employ 1000 idioms from each polarity class to maintain the emotion distribution as evenly distributed as feasible. We retrieved around 15,000 tweets for our experiment.

3.3 Idiom Expansion

We devised a straightforward plan to include the idiom's enlarged version to implement tweet augmentation. In this method, as seen in Fig. 3, we substitute an idiom with a similar meaning or definition discovered in the urban dictionary function (or any other resources when we did not find the idiom definition). For example, the "*kick the bucket*" idiom in this tweet "*I think you are going to kick the bucket after the booster shot*" is replaced by "*I think you are going to die after the booster shot*".

```
Algorithm 2 Idiom Expansion
Output← $\mathcal{F}_{temp}$: $Expanded\ Idioms$
Input← $\mathcal{F}$ : $Lexicon\ of\ Idiomatic\ Lexicon$ ; $\alpha$: $External\ KB$; $\mathcal{D}$: $tweets\ dataset$
Step 1: Initialization
$\mathcal{F}_{temp}$← ϕ; $x_{temp}$ ← ϕ;
Step 2: Connect to the external knowledge base to retrieve the proper definition of idioms
For all $i \in \mathcal{D}$ do
    For all $j \in \mathcal{F}$ do
        $\varphi$ ← MIXUP(EMBED$[j, i]$);
        $x_{temp}$ ← $x_{temp} \cup$ WEB. CRAWLER$(\alpha, \varphi)$;
        DOM$(x_{temp})$:= $\{ i \in \mathcal{D} \mid x^{i}{}_{temp} \neq x_{temp} \}$
        $\mathcal{F}_{temp}$← $\mathcal{F}_{temp} \cup$ DOM$(x_{temp})$
    End for
End for
Output:=$\mathcal{F}_{temp}$
End procedure
```

Figure 3: Pseudo code of idiom expansion algorithm

3.4 Tweet Data Pre-Processing

Pre-processing is where most sentiment analysis solutions begin. Social media data typically blend with emoticons, URLs, hashtags, stop words, numbers, dates, and other elements. It is usual for researchers to begin cleaning the dataset to speed up processing. They contend that such noisy data is meaningless and does not affect the system's accuracy. We perform this step immediately after the idiom expansion step.

We cleaned up tweets by eliminating distracting URLs and using some standard text pre-processing techniques, such as:

- Stop word removal: Eliminating the needless encoding of words that are absent from any word embedding that has been learned beforehand.
- Case folding is the process of lowering the case of words or sentences.
- Mapping unique values specific to their category (for instance, "09-2022" against "DATE").
- Special character removal: Remove hashtags, numerals, punctuation, and other characters outside alphabetic ones.
- Normalization of acronyms (such as "MYR" for "Malaysian Ringgit") and abbreviations (such as "VAX" for "vaccine").
- Correction of the spelling.

3.5 Fine-Tuning Classifiers for Downstream Task

As mentioned before, BERT uses a fine-tuning strategy to learn the language representations and then applies them to downstream tasks. Throughout the fine-tuning phase, BERT performs better in the bidirectional language model. There are several versions of BERT, and each version train differently. Twitter-roBERTa-base-sentiment trained on 58 M tweets, while Twitter-roBERTa-base-sentiment-latest trained on 124 M tweets [52].

Fine-tuning BERT and any of its variants is a customization procedure of retraining the model using your 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 and then applied to solve the new downstream task.

However, fine-tuning stability still needs to be determined by critics. Some researchers argue that the reasons behind the instability of the fine-tuning process are the gradient vanishing problem and the lack of generalization [53]. To resolve this problem, they propose a set of training hyperparameters. Other scholars have questioned whether these characteristics are the true causes of the instability and have asserted that catastrophic forgetting and the limited size of the datasets are to blame [54–56].

Regardless of the actual reasons, our expansion method avoids the need for the fine-tuning process at folds. First, the BERT (or other versions) does not require a retaining to teach the model because of the "implicit" idioms in the tweet. Second, while the classification performance is degraded to not understanding the context of the idioms, our expansion method provides the model with a self-explained English definition/meaning of the idiom.

As shown in Fig. 4, the BERT model uses $n = 12$ layers (other variants have different layers) of transformers block (aka encoders) with a hidden size of 768 and several self-attention headers, and more than a hundred million trainable parameters [54].

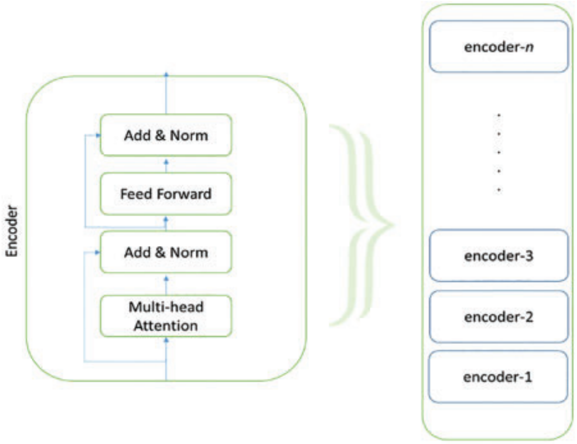
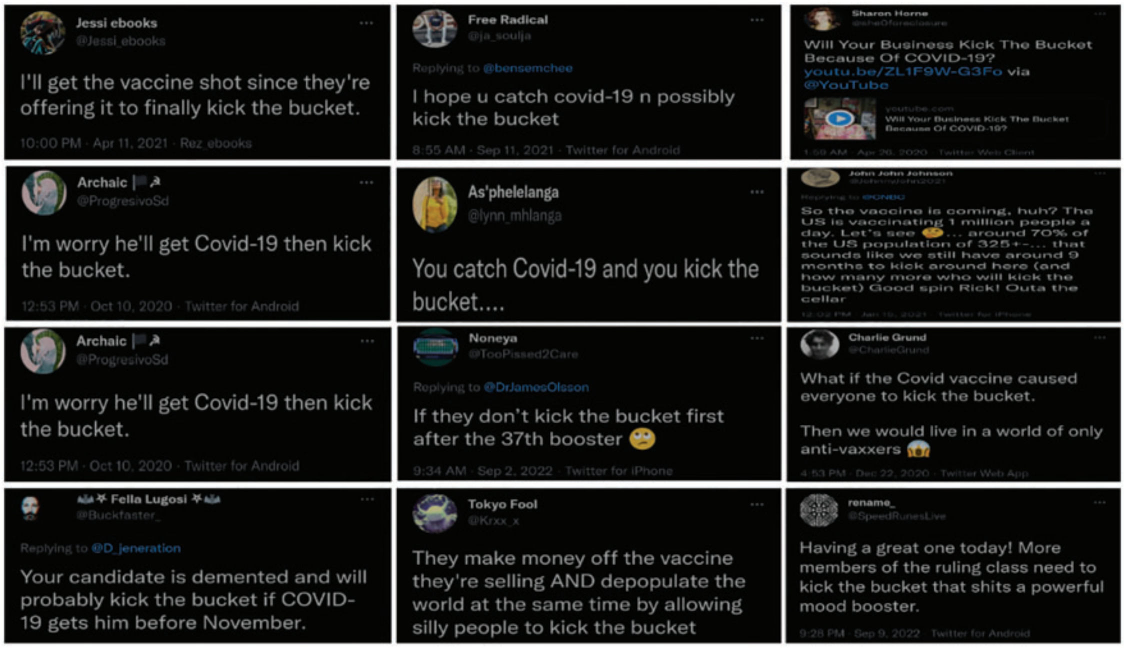


Figure 4: Generic BERT-variant structure

4 Results

In this study, three different tests are carried out to demonstrate the advantages of using idiomatic idioms to assess the sentiment of a tweet. Thanks to the expansion approach, the overhead required for fine-tuning the transformer is no longer necessary. To build ground truth data from the tweets' dataset, we use a tweets dataset that we manually labeled by a group of annotators. We accept only the labels with "all-agree" status. We were able to successfully collect 3000 tweets. The goal of the first experiment is to compare the "idiom only" sentiment classification by using roBERTa with the ground truth data given in the manually annotated lexicon (eSliDE lexicon). The second experiment is similar to the first one except that idioms are augmented and replaced by their definition/meaning phrase retrieved from the external Thesaurus/Dictionary. For tweet sentiment classification, we conduct the last experiment in three parts. The first is by directly assigning the label of the idiom of the tweet to the tweet itself. There is no learning used in this part but rather a simple assignment.

In Fig. 5, different examples were retrieved on different aspects related to COVID-19 tweets and the idiom "*kick the bucket*". This idiom is negative as it refers to "death". By looking at some sample tweets, the tweet might look neutral or positive as the context has some humor or positive keywords. For example, the tweet "*You catch the covid-19 and you then kick the bucket*" was classified as neutral without idiom enrichment. However, after enrichment, it was classified as negative, which matches with annotators voting.



Jessi ebooks
@Jessi_ebooks

I'll get the vaccine shot since they're offering it to finally kick the bucket.

10:00 PM · Apr 11, 2021 · Rez_ebooks

Free Radical
@ja_soulja

Replying to @bensemchee

I hope u catch covid-19 n possibly kick the bucket

8:55 AM · Sep 11, 2021 · Twitter for Android

Sharon Horne
@she0foreclosure

Will Your Business Kick The Bucket Because Of COVID-19? youtu.be/ZL1F9W-G3Fo via @YouTube

youtube.com
Will Your Business Kick The Bucket Because Of COVID-19?

1:59 AM · Apr 26, 2020 · Twitter Web Client

Archaic | 🏴 ☭
@ProgresivoSd

I'm worry he'll get Covid-19 then kick the bucket.

12:53 PM · Oct 10, 2020 · Twitter for Android

As'phelelanga
@lynn_mhlanga

You catch Covid-19 and you kick the bucket....

John John Johnson
@JohnnyJohn2021

Replying to @CNBC

So the vaccine is coming, huh? The US is vaccinating 1 million people a day. Let's see 🤔... around 70% of the US population of 325+-... that sounds like we still have around 9 months to kick around here (and how many more who will kick the bucket) Good spin Rick! Outa the cellar

12:02 PM · Jan 19, 2021 · Twitter for iPhone

Archaic | 🏴 ☭
@ProgresivoSd

I'm worry he'll get Covid-19 then kick the bucket.

12:53 PM · Oct 10, 2020 · Twitter for Android

Noneya
@TooPissed2Care

Replying to @DrJamesOlsson

If they don't kick the bucket first after the 37th booster 🙄

9:34 AM · Sep 2, 2022 · Twitter for iPhone

Charlie Grund
@CharlieGrund

What if the Covid vaccine caused everyone to kick the bucket.

Then we would live in a world of only anti-vaxxers 😱

4:53 PM · Dec 22, 2020 · Twitter Web App

Fella Lugosi
@Buckfaster_

Replying to @D_jeneration

Your candidate is demented and will probably kick the bucket if COVID-19 gets him before November.

Tokyo Fool
@Krxx_x

They make money off the vaccine they're selling AND depopulate the world at the same time by allowing silly people to kick the bucket

rename_
@SpeedRunesLive

Having a great one today! More members of the ruling class need to kick the bucket that shits a powerful mood booster.

9:28 PM · Sep 9, 2022 · Twitter for Android

Figure 5: Sample of retrieved tweets (*idiom: kick the bucket*, *theme: COVID*, *booster, vaccine*)

Experiment I:

Table 2 compares the idiom annotation results of the raw idiomatic expressions and compares them with the ground truth data in the eSliDE lexicon. The roBERTa classifier achieves a horrible result with a high error ratio. The classifier fails with a 78% error ratio in detecting the positive labels. We have computed the error rate as in (2). The δ represents the percent error, and the υ_A and υ_E are the actual observed and the expected values, respectively. The best error rate is 0, whereas the worst is 1. Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1, whereas the worst is 0. The accuracy percentage is computed as [1-error rate]. The error rate and the accuracy for each sentiment class are shown in Table 3.

$$\delta = \left| \frac{\upsilon_A - \upsilon_E}{\upsilon_E} \right| \cdot 100\% \qquad (2)$$

Experiment II

This experiment aims to put the idiom extension approach to the test for idioms in the eSliDE lexicon. When we compare the annotation result to the manually annotated idioms, we see that the classifier obtains an intriguing result for the positive sentiment, with a significant decrease in the error ratio. Table 4 shows a sample of the obtained results after using the expanded form of the idioms. Table 5 displays the error ratio and the accuracy for each annotation class.

Table 2: Idioms sentiment classification comparison (manual *vs.* roBERTa classifier)

Idioms sentiment classification comparison (Manual *vs.* roBERTa classifier)									
	Manual sentiment classification (eSliDE)				*roBERTA sentiment classification for isolated idioms*				Label Matching
Idiom	Pos	Neg	Neu	Maj. Label	Pos	Neg	Neu	Label	
Kindred spirit	0.5	0	0.5	positive	0.163	0.081	0.754	Neu	**False**
Killer instinct	0.3	0.2	0.5	neutral	0.104	0.209	0.686	Neu	True
Kettle of fish	0	0.5	0.5	negative	0.095	0.226	0.678	Neu	**False**
Jump through hoops	0.3	0.2	0.5	neutral	0.073	0.184	0.743	Neu	True
Jump on the bandwagon	0.5	0	0.5	positive	0.207	0.135	0.659	Neg	**False**
Jack of all trades	0.4	0.1	0.5	neutral	0.077	0.247	0.676	Neu	True
Inside track	0.5	0	0.5	positive	0.129	0.131	0.740	Neu	**False**
In the dark	0	0.5	0.5	negative	0.107	0.172	0.721	Neu	**False**
In the bag	0.5	0	0.5	positive	0.228	0.111	0.661	Neu	**False**

Table 3: Error ratio and accuracy of idiom annotation without expansion method

Annotation class	*Error rate* ***δ***	*Accuracy*
Positive label	78%	22%
Negative label	64%	36%
Neutral label	51%	49%

Table 4: Idioms sentiment annotation comparison (manual *vs.* idiomexpansion using roBERTa classifier)

Idioms sentiment classification comparison (Manual *vs.* idiom expansion using roBERTa classifier)									
	Manual sentiment annotation (eSliDE)				*roBERTA sentiment classification using idioms expansion*				Label matching
Idiom	Pos	Neg	Neu	Maj. Label	Pos	Neg	Neu	Label	
Kindred spirit	0.5	0	0.5	positive	0.083	0.040	0.878	Neu	**False**
Killer instinct	0.3	0.2	0.5	neutral	0.453	0.028	0.519	Neu	True
Kettle of fish	0	0.5	0.5	negative	0.016	0.690	0.294	Neg	True
Jump through hoops	0.3	0.2	0.5	neutral	0.035	0.125	0.841	Neu	True

(Continued)

Table 4: Continued

Idioms sentiment classification comparison (Manual *vs.* idiom expansion using roBERTa classifier)									
	Manual sentiment annotation (eSliDE)				*roBERTA sentiment classification using idioms expansion*				Label matching
Idiom	Pos	Neg	Neu	Maj. Label	Pos	Neg	Neu	Label	
Jump on the bandwagon	0.5	0	0.5	positive	0.712	0.004	0.284	Pos	True
Jack of all trades	0.4	0.1	0.5	neutral	0.573	0.009	0.418	Pos	**False**
Inside track	0.5	0	0.5	positive	0.510	0.017	0.473	Pos	True
In the dark	0	0.5	0.5	negative	0.015	0.796	0.189	Neg	True
In the bag	0.5	0	0.5	positive	0.634	0.044	0.323	Pos	True

Table 5: Error ratio and accuracy of idiom annotation after expansion

Annotation class	*Error Rate δ*	*Accuracy*
Positive label	13%	87%
Negative label	18%	82%
Neutral label	22%	78%

Experiment III

In this experiment, we used the 3000 tweets to check the accuracy and F1-score the classifier might achieve. Table 6 illustrates the sentiment classification results for tweets without idioms (idioms are intentionally removed from the tweet). The other two parts show the roBERTa classifier accuracy as shown in formula (3). The harmonic mean of precision and recall (F1-score) is computed according to (4).

$$\text{Accuracy} = \frac{\text{correct predictions}}{\text{total predictions}} \tag{3}$$

$$\text{F1} = * \frac{2\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Table 6: Tweet sentiment classification results

Tweet sentiment classification accuracy and F1-score			
Test benchmark dataset 3000 tweets	*roBERTa accuracy using raw tweets without idioms*	*roBERTa classifier accuracy*	
		No expansion	*With idiom expansion*
1470 Neg	69%	71%	90%
860 Pos	72%	70%	89%

(Continued)

Table 6: Continued

Tweet sentiment classification accuracy and F1-score

Test benchmark dataset 3000 tweets	*roBERTa accuracy using raw tweets without idioms*	*roBERTa classifier accuracy*	
		No expansion	*With idiom expansion*
670 Neu	47%	67%	85%
Average F1-score	**70%**	**71%**	**93%**

5 Conclusion

This paper proposed an idiom expansion method to enrich tweets' context to improve sentiment classification accuracy. The expansion method utilizes an external knowledge base to extract the non-literal meaning of idioms. The technique avoids the instability caused by the conventional transformers' fine-tuning to solve a downstream task.

The expansion strategy has lowered the overall error rate associated with the annotation of idiomatic expressions. The error rate decreased by 65 percentage points for positive labels and 46 percentage points for negative labels. However, the approach was less effective at annotating the neutral labels. For the sentiment classification task, the tweet enrichment technique increases the classification accuracy by an average of 18 points across sentiment classes.

It's worth studying the suggested expansion for autonomous production of an idiomatic sentiment lexicon rather than crowdsourcing services. The framework only handles three sentiment classes (positive, negative, and neutral). We plan to enhance this framework to support multi-label classes like "Optimistic, Thankful, Empathetic, Pessimistic, Anxious, Sad, Annoyed, Denial, Surprise, and Joking," which are presented in a Kaggle dataset [57].

Funding Statement: This work was supported in part by the UTAR Research Fund (IPSR/RMC/UTARRF/2020-C1/R01).

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

References

[1] "Number of worldwide social network users 2027 | statista," Statista, [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. [Accessed: 16-Sep-2022].

[2] B. Fatkhurrozi, "Natural language processing (NLP)," *Majalah Ilmiah Dinamika*, vol. 32, no. 2, pp. 1–20, 2013.

[3] T. K. Tran, H. M. Dinh and T. T. Phan, "Building an enhanced sentiment classification framework based on natural language processing," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 1771–1777, 2022.

[4] A. Hwang and C. Hidey, "Confirming the non-compositionality of idioms for sentiment analysis," in *Proc. of the Joint Workshop on Multiword Expressions and WordNet*, Florence, Italy, Association for Computational Linguistics, pp. 125–129, 2019.

[5] K. Rudra, A. Chakraborty, N. Ganguly and S. Ghosh, "Chapter 24: Understanding the usage of idioms in the Twitter social network," *Pattern Recognition & Big Data*, vol. 1, pp. 767–788, 2017.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[7] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,” *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[8] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim and I. A. Khan, “Lexicon-enhanced sentiment analysis framework using rule-based classification scheme,” *PloS One*, vol. 12, no. 2, pp. e0171649, 2017.

[9] I. P. Windasari and D. Eridani, “Sentiment analysis on travel destination in Indonesia,” in *2017 4th Int. Conf. on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, pp. 276–279, 2017.

[10] C. Fellbaum, “WordNet: An electronic lexical resource,” in *The Oxford Handbook of Cognitive Science*, New York: Oxford University Press, pp. 301–314, 2017.

[11] M. d. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodriguez-Garcia *et al.*, “Sentiment analysis on tweets about diabetes: An aspect-level approach,” *Computational and Mathematical Methods in Medicine*, vol. 2017, no. 1, pp. 5140631, 2017.

[12] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha and M. Alkeshr, “Effective method for sentiment lexical dictionary enrichment based on Word2Vec for sentiment analysis,” in *2018 Fourth Int. Conf. on Information Retrieval and Knowledge Management (CAMP’18)*, Kota Kinabalu, Sabah, Malaysia, pp. 1–5, 2018.

[13] A. Bittar, S. Velupillai, A. Roberts and R. Dutta, “Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: Corpus-based analysis,” *JMIR Medical Informatics*, vol. 9, no. 4, pp. e22397, 2021.

[14] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proc. of the Eighth Int. AAAI Conf. on Weblogs and Social Media, AAAI Press*, Ann Arbor, Michigan, USA, vol. 8, no. 1, pp. 216–225, 2014.

[15] A. Esuli and F. Sebastiani, “Determining the semantic orientation of terms through gloss classification,” in *Proc. of the 14th ACM Int. Conf. on Information and Knowledge Management (CIKM’05)*, ACM, New York, NY, pp. 617–624, 2005.

[16] S. Baccianella, A. Esuli and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *Proc. of the European Language Resources Association (LREC’10)*, Valletta, Malta, pp. 2200–2204, 2010.

[17] B. Ohana and B. Tierney, “Sentiment classification of reviews using SentiWordNet,” in *Proc. 9th. IT&T Conf., Technological University Dublin*, Dublin, Ireland, pp. 124–132, 2009.

[18] P. Sudhir and V. D. Suresh, “Comparative study of various approaches, applications and classifiers for sentiment analysis,” *Global Transitions Proceedings*, vol. 2, no. 2, pp. 205–211, 2021.

[19] Y. Jo and A. H. Oh, “Aspect and sentiment unification model for online review analysis,” in *Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining*, ACM, Hong Kong, China, pp. 815–8, 24, 2011.

[20] Q. Mei, X. Ling, M. Wondra, H. Su and C. Zhai, “Topic sentiment mixture: Modeling facets and opinions in weblogs,” in *Proc. of the 16th Int. Conf. on World Wide Web*, ACM, Banff, Alberta Canada, pp. 171–180, 2007.

[21] C. Lin and Y. He, “Joint sentiment/topic model for sentiment analysis,” in *Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM’09)*, ACM, Hong Kong, China, pp. 375–384, 2009.

[22] Y. He, C. Lin, W. Gao and K. -F. Wong, “Dynamic joint sentiment-topic model,” *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 1, pp. 1–21, 2013.

[23] P. D. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02)*, Philadelphia, PA, USA, pp. 417–424, 2002.

[24] T. Zagibalov and J. Carroll, “Automatic seed word selection for unsupervised sentiment classification of Chinese text,” in *Proc. of the 22nd Int. Conf. on Computational Linguistics*, Manchester United Kingdom, pp. 1073–1080, 2008.

[25] S. Huang, Z. Niu and C. Shi, "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowledge-Based Systems*, vol. 56, pp. 191–200, 2014.

[26] Y. Lu, M. Castellanos, U. Dayal and C. Zhai, "Automatic construction of a context-aware sentiment lexicon: An optimization approach," in *Proc. of the 20th Int. Conf. on World Wide Web*, Hyderabad India, pp. 347–356, 2011.

[27] N. C. Dang, M. N. Moreno-García and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, pp. 483, 2020.

[28] C. Wu, F. Wu, S. Wu, Z. Yuan and Y. Huang, "A hybrid unsupervised method for aspect term and opinion target extraction," *Knowledge-Based Systems*, vol. 148, no. 2018, pp. 66–73, 2018.

[29] P. Ray and A. Chakrabarti, "A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis," *Applied Computing and Informatics, v*ol. 18, no. 1/2, pp. 163–178, 2020.

[30] X. Ding, B. Liu and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proc. of the Int. Conf. on Web Search (WSDM'08)*, New York, NY, USA, pp. 231–240, 2008.

[31] A. R. W. Sait and M. K. Ishak, "Deep learning with natural language processing enabled sentimental analysis on sarcasm classification," *Computer Systems Science and Engineering*, vol. 44, no. 3, pp. 2553–2567, 2023.

[32] T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[33] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan *et al.,* "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, pp. 53, 2021.

[34] S. Casola, I. Lauriola and A. Lavelli, "Pre-trained transformers: An empirical comparison," *Machine Learning with Applications*, vol. 9, pp. 100334, 2022.

[35] E. Fersini, "Sentiment analysis in social networks: A machine learning perspective," in *Sentiment Analysis in Social Networks*, Boston: Morgan Kaufmann, vol. 22, no. 14, pp. 91–111, 2017.

[36] K. Shudo and T. Tanabe, "JDMWE: A Japanese dictionary of multi-word expressions," *Journal of Natural Language Processing*, vol. 17, pp. 51–74, 2010.

[37] A. Mudinas, D. Zhang and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," in *Proc. of the First Int. Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'12)*, ACM Press, New York, NY, USA, pp. 1–8, 2012.

[38] S. Song and W. Ting, "Construction of unsupervised sentiment classifier on idioms resources," *Journal of Central South University*, vol. 21, pp. 1376–1384, 2014.

[39] H. S. Ibrahim, S. M. Abdou and M. Gheith, "Idioms-proverbs lexicon for modern standard Arabic and colloquial sentiment analysis," *International Journal of Computer Applications*, vol. 118, no. 11, pp. 26–31, 2015.

[40] L. Williams, C. Bannister, M. Arribas-Ayllon, A. Preece and I. Spasić, "The role of idioms in sentiment analysis," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7375–7385, 2015.

[41] I. Spasić, L. Williams and A. Buerki, "Idiom-based features in sentiment analysis: Cutting the Gordian knot.," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 189–199, 2020.

[42] B. Liu, *Sentiment Analysis. Mining Opinions, Sentiments and Emotions*. Cambridge, UK: Cambridge University Press, 2015.

[43] X. Chen, Z. Hai, S. Wang, D. Li, C. Wang *et al.,* "Metaphor identification: A contextual inconsistency based neural sequence labeling approach," *Neurocomputing*, vol. 428, no. 2021, pp. 268–279, 2021.

[44] K. Matsumoto, S. Tsuchiya, M. Yoshida and K. Kita, "Construction and expansion of dictionary of idiomatic emotional expressions and idiomatic emotional expression corpus," *International Journal of Computer & Software Engineering*, vol. 6, no. 2, pp. 174, 2021.

[45] J. Briskilal and C. N. Subalalitha, “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa,” *Information Processing & Management*, vol. 59, no. 1, pp. 1–9, 2022.

[46] H. Li, Y. Ma, Z. Ma and H. Zhu, “Weibo text sentiment analysis based on BERT and deep learning,” *Applied Sciences*, vol. 11, no. 22, pp. 10774, 2021.

[47] J. Lin, D. Campos, N. Craswell, B. Mitra and E. Yilmaz, “Significant improvements over the state of the art? a case study of the ms marco document ranking leaderboard,” in *Proc. of the 44th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Virtual Event, Canada, pp. 2283–2287, 2021.

[48] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.,* “Roberta: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[49] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma *et al.,* “Albert: A lite BERT for self-supervised learning of language representations,” in *Int. Conf. on Learning Representations (ICLR’20)*, Addis Ababa, Ethiopia, pp. 1–17, 2020.

[50] V. Sanh, L. Debut, J. Chaumond and T. Wolf, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” *arXiv: Computation and Language*, vol. 4, pp. 1–5, 2019.

[51] B. Tahayna, R. K. Ayyasamy, R. Akbar, N. F. Subri and A. Sangodiah, “Lexicon-based non-compositional multiword augmentation enriching tweet sentiment analysis,” in *2022 3rd Int. Conf. on Artificial Intelligence and Data Sciences (AiDAS’22)*, Ipoh, Malaysia, IEEE, pp. 19–24, 2022.

[52] Hugging Face, Huggingface.co, [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment. [Accessed: 30-Jul-2022].

[53] M. Mosbach, M. Andriushchenko and D. Klakow, “On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines,” in *Int. Conf. on Learning Representations*, Virtual Conference, USA, pp. 1–19, 2021.

[54] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis*, Minnesota, MN, USA, vol. 1, (Long and Short Papers), pp. 4171–4186, 2018.

[55] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi *et al.,* “Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping,” arXiv preprint arXiv:2002.06305, 2020.

[56] C. Lee, K. Cho and W. Kang, “Mixout: Effective regularization to finetune large-scale pretrained language models,” in *Int. Conf. on Learning Representations (ICLR’20)*, Addis Ababa, Ethiopia, pp. 542, 2020.

[57] “Sentiment Analysis of Covid-19 related Tweets | Kaggle,” Kaggle.com, 2022. [Online]. Available: https://www.kaggle.com/competitions/sentiment-analysis-of-covid-19-related-tweets/. [Accessed: 11-Jun-2022].