

https://dspace.org/

Engaging community members in setting priorities for nutrition interventions in rural northern Ghana

M Dalaba; E Nonterah; S Chatio; J Adoctor; Agnes Erzse ...

https://hdl.handle.net/10539/39692
DOAJ

Downloaded from a DSpace repository. https://dspace.org/

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 October 2022, accepted 12 November 2022, date of publication 23 November 2022, date of current version 21 February 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3224136

SURVEY

Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape

KOENA RONNY MABOKELA[1,2], TURGAY CELIK[2,3], AND MPHO RABORIFE[4]

[1]Department of Applied Information Systems, University of Johannesburg, Johannesburg 2006, South Africa
[2]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa
[3]Faculty of Engineering and Science, University of Agder, 4630 Kristiansand, Norway
[4]Institute for Intelligence Systems, University of Johannesburg, Johannesburg 2006, South Africa

Corresponding author: Turgay Celik (celikturgay@gmail.com)

This work was supported by the National Research Foundation (NRF) for the Black Academics Advancement Program (BAAP) under Grant BAAP200225506825.

ABSTRACT Sentiment analysis automatically evaluates people's opinions of products or services. It is an emerging research area with promising advancements in high-resource languages such as Indo-European languages (e.g. English). However, the same cannot be said for languages with limited resources. In this study, we evaluate multilingual sentiment analysis techniques for under-resourced languages and the use of high-resourced languages to develop resources for low-resource languages. The ultimate goal is to identify appropriate strategies for future investigations. We report over 35 studies with different languages demonstrating an interest in developing models for under-resourced languages in a multilingual context. Furthermore, we illustrate the drawbacks of each strategy used for sentiment analysis. Our focus is to critically compare methods, employed datasets and identify research gaps. This study contributes to theoretical literature reviews with complete coverage of multilingual sentiment analysis studies from 2008 to date. Furthermore, we demonstrate how sentiment analysis studies have grown tremendously. Finally, because most studies propose methods based on deep learning approaches, we offer a deep learning framework for multilingual sentiment analysis that does not rely on the machine translation system. According to the meta-analysis protocol of this literature review, we found that, in general, just over 60% of the studies have used deep learning frameworks, which significantly improved the sentiment analysis performance. Therefore, deep learning methods are recommended for the development of multilingual sentiment analysis for under-resourced languages.

INDEX TERMS Multilingual, sentiment analysis, code-switching, deep learning, cross-lingual, under-resourced languages, systematic review.

I. INTRODUCTION

Sentiment analysis is an intensive research activity in Natural Language Processing (NLP). It uses NLP technologies to analyse textual messages and determine deeper contexts as they apply to a topic, brand, or theme [1], [2]. It is used to determine whether comments are subjective or objective and then classify such texts as *positive*, *negative* or *neutral* sentiment. sentiment analysis can be tackled at different classification levels such as document, sentence, aspect or feature [3], [4]. It has garnered considerable research attention, which can be attributed to its numerous essential NLP applications [5]. In recent years, sentiment analysis has gained even more interest owing to the rapid use of social media platforms. Its primary usage has been in businesses and consumer care services [6], [7].

Social media users have made it a modern-day culture to use social media platforms to share their feelings or thoughts

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

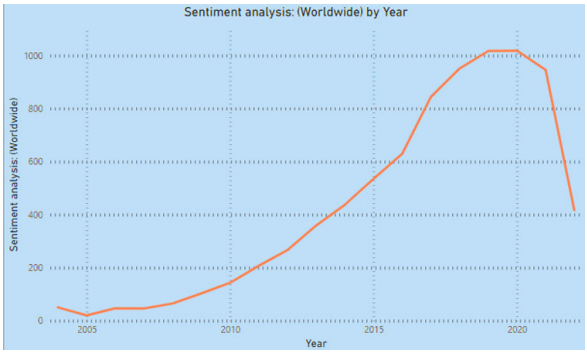


FIGURE 1. Worldwide interest in sentiment analysis over time on Google Trends, 2004 - present.

on various subjects [8]. As a result, social media platforms such as Facebook, Twitter, Instagram, WhatsApp and, most recently, TikTok, generate a large amount of potentially "*rich*" data [9]. Therefore, most sentiment analysis research studies have used social media data, particularly Twitter [10], [11], [12]. Notably, it is common for social media data to be multilingual and multicultural or have many linguistic variations, including mixed languages [13], [14].

Sentiment analysis is a growing research area with promising progress in high-resourced languages [15], [16], [17]. However, in another context, the same cannot be said for under-resourced languages due to a lack of resources to develop NLP technologies. By under-resourced languages, we refer mainly to languages with little or no resources available to create digital language technologies [18]. Our study reviews multilingual sentiment analysis (MSA) methods for under-resourced languages. This paper uses the terms '*under-resourced languages*' and '*low-resource languages*'.

Research on sentiment analysis has focused predominantly on single-language texts, mainly for high-resourced languages such as English [8], [19], [20]. Sentiment analysis research for high-resourced languages was actively studied due to the massive availability of resources such as benchmark datasets, annotated corpora and sentiment lexicons [14], [21]. In addition, sentiment analysis technologies developed for single-language tasks increase the risk of overlooking information in texts written in multiple languages [6], [22]. Deriu et al. [23] report that sentiment analysis methods developed for single-language texts could not be replicated for new or multilingual texts. Therefore, a concerted effort is necessary to create sentiment analysis models that cater for multiple languages. For example, some researchers proposed a cross-lingual sentiment analysis method with the help of a Machine Translation (MT) application and then applied Machine Learning (ML) techniques. The ML techniques such as Support Vector Machine (SVM), Naive Bayes (NB), and Maximum Entropy (ME) are used for sentiments classification [24], [25], [26]. This cross-lingual sentiment analysis method has been successful in languages such as French, Spanish, Italian, Portuguese, Arabic, and Chinese [4], [13], [15].

MSA aims to recognise the sentiment of textual content written in multiple languages. It attempts to address issues presented by various languages, including code-switched comments. The success of monolingual sentiment analysis and MSA technologies mainly depends on the availability of labelled datasets to train computational language models [19], [27]. Recently, some resources and methods, including code-switched datasets, are available on SemEval, the largest workshop on computational semantic evaluations for multiple NLP research [28]. Although some of these MSA methods are already performing well for high-resourced language datasets, they underperform for under-resourced languages, with English as the primary language and other contributing languages [9]. These sentiment analysis methods can only perform very well if labelled datasets are available or if methodologies that address the issues of under-resourced languages can be customised.

Most MSA approaches still rely on MT-based methods [24] or merging of monolingual datasets from different languages to build large-scale multilingual datasets [29]. Then apply ML techniques for sentiment classification. To some degree, MSA approaches that employ training of monolingual datasets from various languages cannot perform well for mixed-language texts. Some of the MSA methods are language-specific and may not be applied across distinct languages [22], [30]. In addition, supervised ML relies on a labelled dataset to produce accurate results [22]. Therefore, previous MSA research used manual data labelling methods, which is, to date the most labour-intensive and expensive process [22], [23], [30].

Code-switched texts originate from the most populous, multicultural societies and culturally diverse countries where more than one official language is spoken [31]. Given this reality, social media users are more comfortable expressing their views in multiple languages [20], [23], [27]. Under-resourced languages are most commonly mixed with English [31], [32], [33]. These mixed-language phenomena pose a significant challenge to existing MSA systems. In this context, we refer to multilingual data as sentences containing monolingual texts or code-switched data —texts written in more than one language. MSA for under-resourced languages is advancing gradually with progressive application of Deep Learning (DL) methods [16], [22], [34]. However, very few generic MSA methods have been developed for under-resourced languages [14]. In this study, we also examine MSA methods that considered mixed-language or code-switched texts intending to address under-resourced language challenges.

Prior studies on MSA explored the use of MT-based systems to transfer knowledge from resource-rich languages to under-resourced languages [18], [24]. These approaches translate text from an under-resourced language to English or vice-versa and then apply ML-based techniques to perform sentiment classification [15], [30]. Moreover, this method generally presents limitations like loss of meaning, and poor translation quality [13], [35]. In addition, [25] say

that MT-based systems should be an obvious baseline system for any new MSA method [16]. In reality, the most recent development in the field of NLP has demonstrated that the effectiveness of MSA is significantly impacted by DL techniques [13], [21], [36]. Thus far, researchers have explored approaches such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Adversarial Neural Networks (ANN), and Generative Adversarial Networks (GAN) [21], [37].

The literature has made several attempts to address sentiment analysis in multilingual environments [13], [34] and others have addressed the problem from the perspective of creating methodologies that can operate on small datasets [12], [38]. However, even though there is a multitude of studies on high-resource languages from the perspective of MSA [20], [39] but none of them focuses specifically on under-resourced languages in a multilingual setting. In this article, we concentrate on analysing the MSA literature survey from the perspective of under-resourced languages. Although there have been several literature surveys for MSA [14], [40], [41], [42], our study is by far the first to cover a mixture of high-resourced languages and under-resourced languages in a multilingual setting. We provide a detailed literature survey, along with the methods, models, mechanisms and performances, with a special focus on rule-based, cross-lingual, machine learning and deep learning techniques. The purpose is to the extent the research field provides scope for future research on under-resourced languages. We used the methods presented in Section III as a guideline to review the relevant studies. We categorised these studies as multilingual, cross-lingual, or code-switched approaches for under-resourced languages. Most significantly, we found that more than 40% of the research presented in this analysis had not been looked at in earlier literature reviews. The following contributions are made by our study:

1) To the best of our knowledge, we provide a detailed systematic review and an overview of MSA techniques for languages with limited resources in multilingual environments.
2) We provide the most recent comprehensive review of the MSA methods and an overview of MSA techniques for languages with limited resources in a multilingual environment.
3) We describe the outcomes of using cross-lingual sentiment analysis approaches to develop MSA methods for resource-constrained languages.
4) We further address the research questions raised in our systematic literature study.
5) Finally, we highlight the areas of research that need more investigation and offer suggestions for using MSA techniques in languages with limited resources in the future.

This paper is organised as follows: Section II presents the research questions. Section III presents the methodology used for this literature study. Section IV offers a summary of previous state-of-the-art studies. Section V describes MSA techniques and shortcomings. The evaluation metrics for MSA will be highlighted in section VI and the results will be discussed in section VII. Section VII will present the limitation of the study. In section IX, we will present the emerging MSA areas. Section X deals with emerging MSA areas. Lastly, we offer a conclusion and future suggestions.

II. RESEARCH QUESTIONS

Our research aims to discover the most recent trends in MSA approaches for under-resourced languages. As a result, the following research questions guided this systematic literature review:

1) What are the existing MSA methods used to generate sentiment classification models, sentiment datasets and sentiment lexicons in a multilingual context?
2) What are MT system applications suitable for developing MSA methods and sentiment resources in multilingual environments?
3) What MSA techniques have been applied to sentiment classification for under-resourced languages in code-switched texts?
4) What are the DL and pre-trained techniques used to perform MSA for under-resourced languages?

III. RESEARCH METHODS

This section outlines the methodology utilised to achieve the purposes of this systematic review and provides a detailed description of the approaches and datasets used in MSA research. To prevent bias in our conclusion, we choose to undertake a thorough systematic literature study using high-quality peer-reviewed articles from 2008 to 2022 mainly because of the following reasons: (i) to bridge the gap between the research methods, which can be relevant and help address under-resourced language challenges, (ii) to provide a detailed overview of the most recent trends in MSA methods and offer an understanding of the research shift from 2008 to 2022, (iii) to recommend suitable research methods for future studies on under-resourced languages. Moreover, we clearly describe the differences across MSA methodologies and datasets. Furthermore, we examine how research has shifted from lexicon-based, cross-lingual methods and statistical ML techniques to more contemporary DL models, concentrating on low-resource languages and incorporating aspect-based sentiment analysis research. Lastly, a meta-analysis of the results from the selected articles is used to produce different summary tables.

A. SEARCH STRATEGIES FOR RELEVANT STUDIES

The literature search was conducted by following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) framework (i.e., Fig. 2). This framework includes the identification phase, screening phase, exclusion and inclusion phase [43]. Research communities widely use PRISMA for conducting a systematic literature review. The PRISMA framework was adopted in this study because it emphasizes the reporting of studies assessing the

intervention's effect. It can be used as a foundation for publishing systematic reviews with goals other than considering interventions [43]. We began this literature review study by identifying the data sources and formulating the search keywords that eventually led to selecting the most relevant studies since the start of MSA research. Several peer-reviewed and published articles relating to sentiment analysis in multiple languages were used from 2008 to 2022. The methodology used in this study is depicted in the schematic diagram presented in Fig. 2.

Note that the methodology used in this study is the same one used in the newly released comprehensive literature review on MSA for deep learning methods [14]. We used research papers produced from the search keywords as: ''*Multilingual Sentiment Analysis*'' OR ''*Cross-lingual AND sentiment analysis*'' OR ''*Multi-language AND NLP*'' OR ''*Multi-language AND sentiment*'' OR '' *Multi-lingual sentiment analysis*'' OR ''*Code mixed AND sentiment analysis*'' OR ''*Code-switching AND sentiment analysis*''.

B. INCLUSION AND EXCLUSION STRATEGIES

We provide a detailed explanation of our inclusion and exclusion criteria in this section.

1) INCLUSION CRITERIA

In the next step, we selected articles based on their abstracts, methods, conclusions, and future directions. We included articles based on MSA studies for under-resourced languages and studies on multilingual sentiment classification where the low-resource language is reported. We had relevant articles that investigated MSA from its infancy to date in the aspect of low-resourced languages. Peer-reviewed journals, book series and conference papers are selected because they are of high quality, and citation counts are also considered. We needed to read the entire article for some articles to determine whether they are to be included. We had a more detailed study version for studies published more than once. Finally, we selected conference proceedings papers that reported complete research during the same period.

2) EXCLUSION CRITERIA

There were few articles that are written in languages other than English. However, this systematic literature is based on peer-reviewed papers written exclusively in English. This approach neglects a requirement in a systematic literature review that discourages language limitations. Excluding these two articles from our analysis did not constitute a bias. Articles not published in computer science, decision science, mathematics and engineering journals and not using the techniques mentioned above were excluded, even if they were related to MSA studies for low-resource languages. Journal articles that did not present a complete or significant portion of their methodology are excluded. We also decided not to consider research studies that reported conceptual papers, work-in-progress, preliminary studies, or unfinished work. Review articles, survey papers and dissertations are excluded from this study.

C. DIGITAL SOURCES AND DATA EXTRACTION

In the identification phase, research papers are searched from the following databases: Elsevier, Scopus, Google Scholar, Science Direct, IEEEXplore, ACM and Springer Link; the databases include studies done across the globe, and therefore geographical bias was not an issue. Furthermore, published books and book chapters are examined. We applied the filtering criteria to 468 papers. Our focus was on MSA for low-resource languages and the use of high-resourced languages to develop low-resource languages from a multilingual perspective. In the end, 38 primary studies were reviewed, and 4 of the 38 articles were later included manually in our research. The primary studies included models used to evaluate MSA systems, bilingual SA, cross-lingual sentiment analysis and code-switched SA, where knowledge/lexicon-based methods, ML-based and DL are reported to execute MSA tasks. Lastly, we derived the results and summary tables in the discussion section from the data extracted from the articles.

We screened and scrutinised the titles, abstracts, keywords, methods, conclusions, and citations and decided on potential eligibility. Studies were eligible if they reported on methods or models related to sentiment analysis in multiple languages. Studies of bilingual sentiment analysis are also considered. Studies with the same techniques used by other researchers were excluded, as well as the sentiment analysis methods developed for a single language. All data sources gathered from social media platforms, and other related data sources are included in the datasets.

This research provides the most relevant and recent systematic literature survey for MSA in under-resourced languages. We aim to set the trend and suggest new approaches for under-resourced languages that can benefit sentiment analysis in a multilingual framework. While there are prior reported literature review studies, they paid more attention to sentiment analysis in a single language [6], [45]. Furthermore, several literature reviews that has been reported on the aspect of MSA [6], [14], [40], [41], [42], [46]. However, very few studies focused on MSA for under-resourced languages. For this, we briefly highlight the objectives of each literature survey and show the significant difference between the available literature review and our literature study. For example, the literature survey study presented by [14] only covered MSA methods for deep learning methods using social media data from 2017 to 2020. They only highlighted a shift of research from cross-lingual to code-switching MSA methods. Abdullah et al. [41] investigated a systematic literature review from 2010 to 2019 that covered the pre-processing methods, methods for sentiment analysis, the evaluation models utilised for MSA and the aspects of common languages supported in sentiment analysis.

Furthermore, Lo et al. [6] reviewed English-based sentiment analysis on social media as well as a few works on MSA

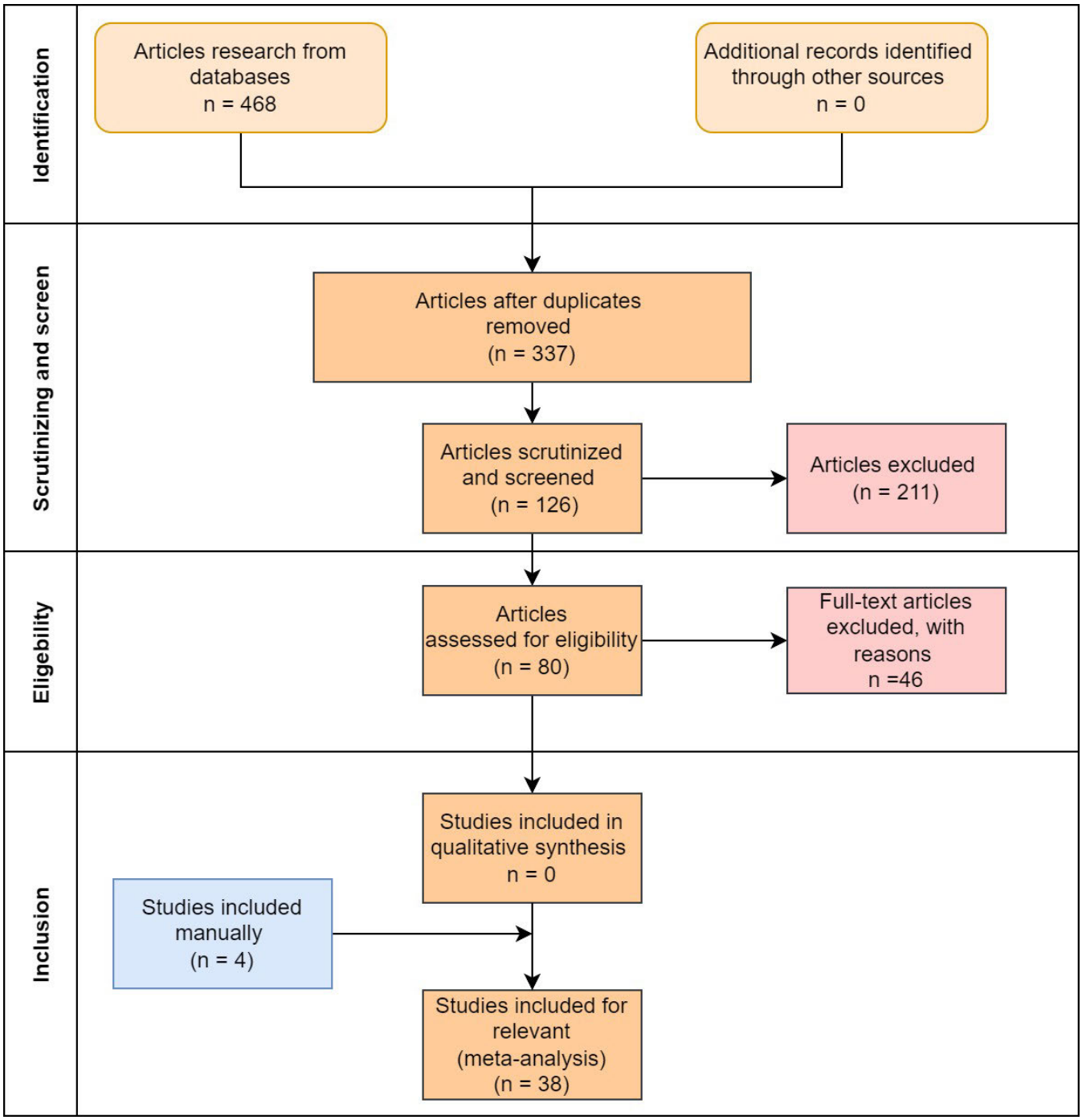


FIGURE 2. PRISMA process flowchart of the systematic literature survey methodology [44].

for social media from 2010 to 2013. Santwana at al. [40] focused only on machine learning techniques for MSA in non-English languages from 2010 to 2018. However, our literature study covers even the most recent MSA methods employed from 2008 to 2022 whilst [42] only focused on the cross-lingual sentiment analysis methods for Chinese languages from 2004 to 2022. Lastly, Xu et al. [46] investigated a systematic literature review for sentiment analysis on social media in single languages from 2018 to 2021. Comparing [6], [14], and [41] with our literature survey, there is an overlap from 2010 to 2018 but [6], [41] provides very little information about recent methods and how the MSA methods work. However, our literature survey includes prior work and the most recent year's work on African languages. According to the best of our knowledge, there is currently a lack of systematic literature survey for

MSA that is published, which covers rule-based/knowledge-based, cross-lingual with machine translation methods, traditional machine learning and deep learning models for under-resourced languages from 2008 to 2022.

D. QUALITY ASSESSMENT

The quality assessment process shown in Fig. 3 was based on the following predefined questions:

- Are the aims of the study clearly stated with objectives and answers our research questions?
- Does the study provide new or unique techniques or contribution in MSA for low-resourced languages?
- Are there any major challenges identified in the study?

We selected several studies after excluding 84 articles. We assessed the quality of the research they presented. We used three quality assessment questions defined to evaluate the quality of the study and provide a quantitative comparison. To determine the quality assessment, we used the scoring procedure: Yes = 1, Partially = 0.5 or No = 0. Each study was given a score between 0 and 3. After the scoring, the points are summed for all the quality assessment questions. If the article received a non-integer total score, it was rounded to the nearest digit. A study is eliminated from our literature review if it receives a score of zero. Thirty-eight (38) articles with a score greater than two (2) are kept because they are considered to meet this study criterion.

IV. STATE-OF-THE-ART STUDIES

Sentiment analysis is an active research field. Previous research has significantly improved the various tasks that makeup sentiment analysis systems. Even commercially produced technologies like Semantria[1] application programming interface (API), IBM Watson[2] and the iFeel tool for SA are accessible to the general public [16], [25]. Many studies have approached sentiment analysis research as a binary classification (i.e., *positive* or *negative*) or ternary classification (i.e., *positive*, *neutral*, and *negative*), and some have even gone as far as to investigate the fine-grained emotions [25]. For fine-grained emotions, researchers have investigated whether a text expresses emotions such as joy, happiness, love, or sadness. Several research studies have examined whether an objective or subjective sentence is positive or negative, as well as the subjective detection of that sentence [18], [47]. Another study for the Persian language focused on developing lexicon-based sentiment analysis [48] to evaluate Persian texts using online Persian language resources [49]. They used available texts from online and used native speakers to manually annotate the texts into positive, negative and neutral [48].

As social media platforms have grown in popularity, interest in studying several powerful sentiment analysis methods has increased. Progress in the field has moved from lexicon-based methods such as AFINN lexicon [50], Valence Aware Dictionary for Sentiment Reasoning (VADER) [51], SentiWordNet, SentiStrength [52], and statistical ML [16], [24] to DL methods like Long Short-Term Memory (LSTM), Bidirectional-LSTM and Bidirectional Encoder Representations from Transformers (BERT) [23], [53]. Twitter data is used mostly for NLP research, especially for sentiment analysis tasks [3], [10]. There are also platforms such as Amazon for sales reviews, music reviews, movie reviews and Twitter, which is the largest source of text datasets so far. The evolution of social media texts or microblogs (e.g. Twitter) has presented new opportunities for language technologies, but it has also posed many new challenges, making it one of the current prime research areas. Some interesting research has emerged using the Twitter dataset for multilingual sentiment classification in SemEval competitions [28], [38], and the introduction of code-switching texts has been studied for SA [54].

V. MULTILINGUAL SENTIMENT ANALYSIS STRATEGIES

Various sentiment analysis methods for multilingual datasets have been explored [13], [16], [17], [22], [35]. There have also been proposed ways for classifying sentiment polarities on a multilingual dataset utilizing lexicon-based techniques along with MT systems and ML approaches [13], [22], [23], [24]. Most SA studies paid more attention to highly resourced languages than those with insufficient resources. However, because English language resources, such as sentiment lexicon, annotated corpora, and benchmark datasets, are easily accessible, most MSA approaches preferred strongly leveraging English language resources [23], [24], [55]. The details of MSA methods will be discussed in the following sections.

A. MULTILINGUAL SUBJECTIVITY DETECTION

Previous sentiment analysis studies introduced the concept of subjectivity detection in multilingual sentiment. Subjectivity detection and sentiment analysis focus on identifying emotional states, such as opinions, emotions, feelings, evaluations, beliefs and speculations [18], [56]. Furthermore, sentiment classification further refines the level of granularity by classifying subjective information as either positive, negative, or neutral.

Although there has been a lot of research on multilingual subjectivity detection, there is still a lot of room for future study in other languages [6], [47]. A lot of the research on the subjectivity detection task was done in English [57], [58]. As a result, most of the gold standard dataset is primarily written in English. Therefore, to create methods for detecting multilingual subjectivity, most studies attempt to use English-language resources [6], [59]. The lexicon and corpus methods dominated early research of multilingual subjectivity analysis [59]. They translated OpinionFinder (i.e., the English subjectivity analysis lexicon) to Romanian using a lexicon-based method and a lemmatized version of the English terminology. This research investigated the effects of corpus-based approaches on Romanian subjectivity-annotated corpora produced by translating English lexicons into Romanian.

[1] https://www.lexalytics.com/semantria
[2] https://www.ibm.com/watson

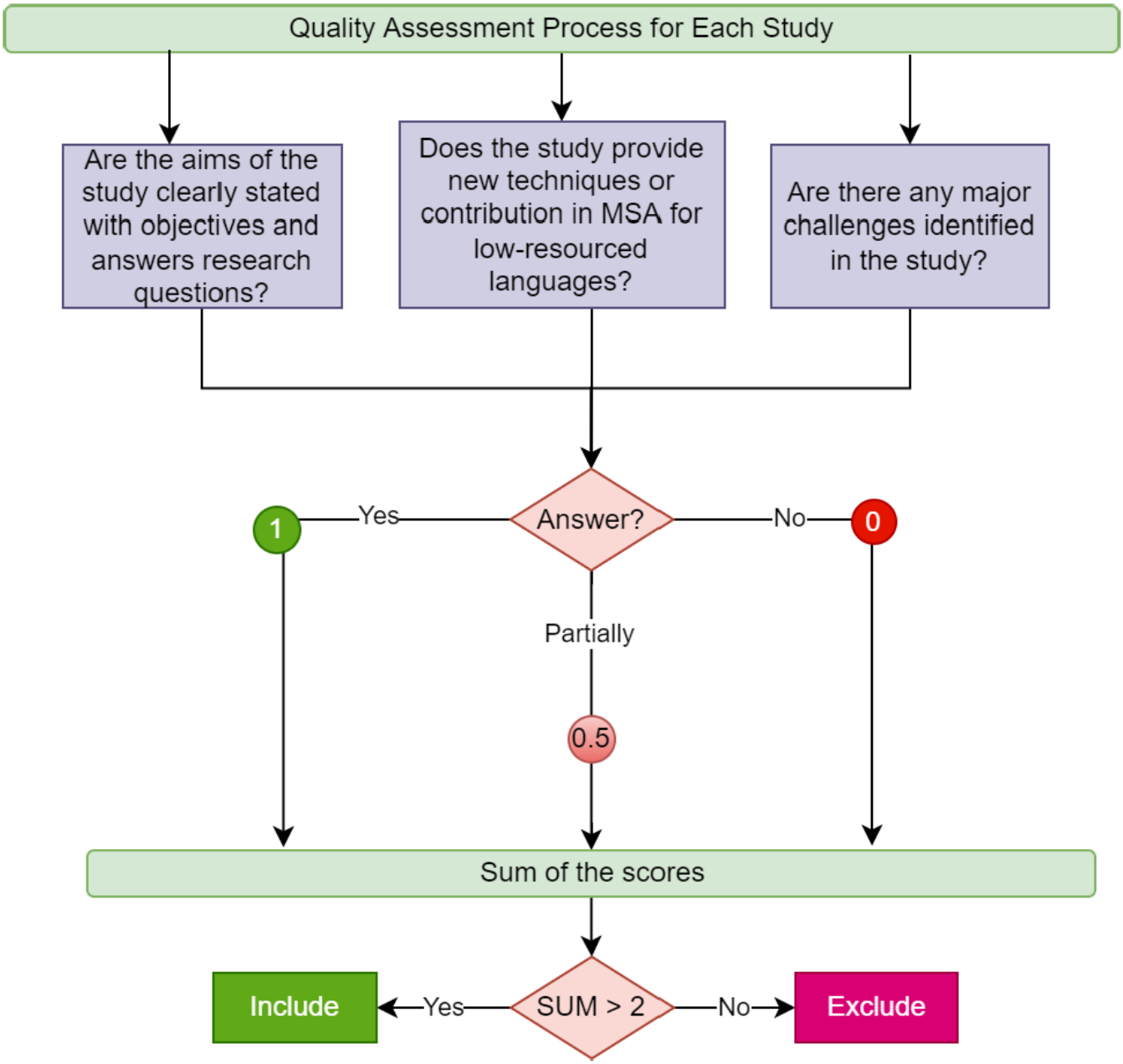


FIGURE 3. Quality Assessment process.

Using linguistic resources in English, Banea et al. [60] investigated an MT-based method to conduct a subjectivity analysis of Romanian and Spanish. They used the Multi-Perspective Question Answering (MPQA) corpus employed by Balahur and Turchi [15], which contains English-language news articles annotated for subjectivity from various sources. The authors showed that even though the translation system was employed, the results obtained were promising and comparable to those obtained by manually translating the corpora. Furthermore, Banea et al. [56] showed that using multilingual information, subjectivity classification (i.e., objective or subjective) in English could achieve 83% accuracy.

Banea et al. [61] explored the alignment of sense levels in different languages to reflect coherent subjectivity. The researchers claim that it is impossible to map one sense to another across languages because a particular purpose may have additional meanings or uses for a specific language. Additionally, they demonstrated that dual co-occurrence metrics could be used to model multilingual feature spaces, offering a more reliable model when compared to using individual languages as input. These metrics learn from comparable sense definitions. As a result of using a simple SVM classifier trained on multilingual space, the accuracy increased to 73% and 76% for English and Romanian, respectively. With an overall accuracy of >73% across all iterations, they

demonstrated that the multilingual model consistently outperformed its cross-lingual counterpart.

Another approach by [57] used a pre-annotated English corpus (i.e. 10,000 movie reviews) collected and annotated by Pang and Lee [58]. These reviews were obtained from the Rotten Tomatoes website[3] and IMDB[4] for subjective and objective reviews, respectively. They built a model to handle multilingual corpora annotated with opinion labels. Their models used Naïve Bayes (NB) techniques to classify reviews. In addition, they developed a method that can be used between topics and languages with high reliability using novel annotation methods. Parallel corpora of English and Arabic reviews are used for model evaluation. The findings show that the same annotations applied to English sentences in parallel corpora can also be applied to sentences in other languages [57].

B. CROSS-LINGUAL METHODS FOR MSA

Several studies have employed MT to build sentiment analysis corpora for under-resourced languages [24], [25], [26], [29]. They utilised well-known MT applications such as Google Translate to translate a dataset existing in a high-resource language into an under-resourced language. However, translation quality is often affected by missing context information, cultural differences and lack of parallel corpora [9], [22], [26]. Some researchers proposed cross-lingual NLP approaches to solve the problem of low-resource languages by benefiting from high-resource languages like English [16], [26], [35], [62]. Previous sentiment analysis methods usually translate the comments from the original under-resourced language to English. This method allows the sentiment classification task to be performed on well-performing models. However, even though this approach was successful for high-resource languages like Russian, German and Spanish [63], it was reported in [9] that translation from English to German, Urdu, and Hindi had a harmful impact on the sentiment analysis performance. Ghafoor et al. [9] used Arabic social media comments to investigate the impact of MT on sentiment analysis performance. They reported that translation from English into German, Urdu and Hindi revealed poor sentiment analysis performance. According to studies on under-resourced languages, with the help of MT systems, cross-lingual sentiment analysis systems suffer performance degradation [9], [24]. Cross-lingual sentiment classification relies on MT approaches in which a source language is translated into the target language [17]. However, another challenge with approaches that rely on MT is that most APIs are not free of charge. Therefore, the task at hand may be costly when dealing with large text corpora [13]. Several authors have used MT systems to translate information directly from one language into another. However, due to differences in linguistic terms and writing styles, the translated data cannot cover the vital information found in the original data [64]. Some cross-lingual MSA approaches have been developed by training a sentiment polarity classifier in English and then employing MT, translating text written in another language into English and then applying a sentiment classifier. Fig. 4, shows the overview of a cross-lingual MSA method using MT-based techniques.

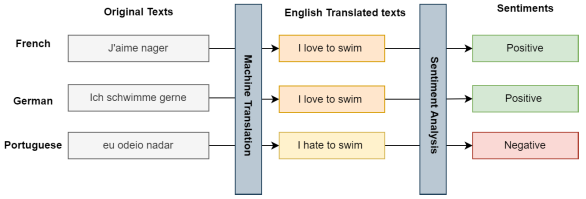


FIGURE 4. Overview of the MT-based approach with a simple example.

One approach uses SentiWordNet, which leverages English lexical resources to perform sentiment analysis [22], [30]. This approach focuses on extracting sentiments in languages other than English and then translating words into English using a standard MT system. Thus, translated documents are classified according to their sentiment, which is either positive or negative. The classification was performed by searching for sentiment-bearing words, such as adjectives, using SentiWordNet. The calculated score determines whether the words are positive or negative. This approach was investigated in German and English languages. The problem with this approach is that MT systems are not always accurate. Moreover, they have many issues, including data sparseness. The drawbacks of automatic MT systems have also been reported and further highlighted in studies [15], [24], [25].

Similarly, [65] proposed a unique way of leveraging reliable English resources to improve Chinese SA. Using MT systems, Chinese reviews were converted into English reviews, and then sentiment polarity in English reviews was identified by directly using English resources.

Balahur and Turchi [15] proposed an MSA approach that uses three distinct MT systems: Google Translate,[5] Bing[6] and Moses[7] translators. The approach used the English dataset from NTCIR 8 Multilingual Opinion Analysis (MOAT[8]). Three MT systems were employed to translate the gold standard dataset into French and German and build the training dataset and some testing datasets. This approach was further extended by using Yahoo systems to translate the dataset for testing into English, French, and German languages [24]. The test dataset translated using Yahoo systems was later corrected and verified by an expert [35]. Sentences containing no sentiments were omitted from the dataset to retain sentences with positive and negative sentiments.

SVM sequential minimal optimisation (SMO) was employed to classify sentiments for all languages to build

[3]www.rottentomatoes.com
[4]https://www.imdb.com/interfaces/

[5]https://translate.google.com/
[6]https://www.bing.com/translator
[7]https://www.statmt.org/moses/
[8]http://research.nii.ac.jp/ntcir/data/data-en.html

a classification model for these three languages. First, the SVM SMO classification models for the three languages were trained separately for each language [24]. Then, the second experiment was executed by combining all separately trained models for each of the three languages and employing the unigram and bigram features extracted from the dataset [6]. The average performance accuracy reported for this approach is <60% for all three MT systems. However, Balahur and Turchi [15] observed that translated data increased features, sparseness and more issues separating positive and negative sentiments during the training phase. This happened due to the low quality of the MT's data, which led to a decrease in performance accuracy. Moreover, the extracted features had insufficient information to allow the classifier to learn.

Balahur and Turchi [15] suggest that the quality of the MT process has implications for the set of features used to build models. According to Becker et al. [13], MT systems are costly, and the results are limited because of the quality of the data translation. In addition, monolingual datasets that are combined to train MSA classification models have shown no impact in improving performance accuracy.

C. IMPROVING MT-BASED METHODS

Balahur and Turchi [24] attempted to improve the performance of an MT system for MSA to obtain the best possible results. They employed MT systems and a supervised ML technique to perform sentiment analysis on a multilingual dataset (i.e. English, Spanish and French). The MT system translated training and test data into a single language, and then a monolingual sentiment classifier was applied. They concluded that the MT system reached a level of maturity and obtained good performance for languages other than English. Although the translated data produced reliable training data, the approach did not address the drawbacks reported in [13] and [23]. They also concluded that the gap in classification performance between the models trained in English and translated data was somewhat in favour of source language data. Nonetheless, with MT systems, there is room for translation errors [15]. Although this technique helps to disambiguate the use of specific words, it does not eliminate translation errors. In addition, adopting this approach requires a more reliable MT system for the accurate performance of MSA models. However, several attempts have been made to improve MT for MSA. Becker et al. [13] argue that even if a perfect MT is readily available, there is always a potential cultural difference between the source and target languages, which may have implications for final classification results. Consequently, the approach mentioned above may not be reliable and will therefore not address the task of MSA, particularly in under-resourced languages.

Similarly, [66] presented a standalone MSA method for English using a gold standard dataset and Google Translate system to translate the dataset from English into four other languages (Italian, Spanish, French and German) to redesign their sentiment analysis system, which caters for data in multilingual settings. The approach employs a supervised learning method (i.e. SVM SMO) which was used previously [15], [24] with a linear kernel on unigrams and bigrams as features. In addition, they used tweet normalisation and MT to obtain high-quality training data for sentiment analysis in the four languages.

In their study, it was further shown that the joint use of training data from different languages, especially a closely related family of languages, can significantly improve the results of the sentiment analysis system. The authors claim that their proposed sentiment analysis approach can perform multilingual sentiment classification with up to 70% performance accuracy. The dataset used in their study was sufficiently small for training and testing. A small dataset allows for easy manual correction of translation errors and eliminates incorrect translations [66]. Furthermore, Balahur and Turchi [66] claim that this approach can be extended to other languages using similar dictionaries created in this work. However, their study focused only on four different languages in which the dataset was presented in a monolingual setting to build a multilingual system. Therefore, we can assume that this method may not be easily adopted for MSA in a mixed-language context.

D. TRANSLATION WITH ML-BASED METHODS

Prior studies indicate that numerous sentiment analysis strategies have explored different methods, but these methods usually rely on lexical resources or ML techniques [22], [25]. Many of these existing methods involve adapting lexical resources without proper comparisons, and validations [16], [25]. Araujo et al. [25] took a different step in the field by evaluating 21 methods for English multi-linear sentence-level SA. These methods are compared with two language-specific strategies based on nine language-specific datasets consisting of Arabic, Dutch, French, German, Italian, Portuguese, Russian, Spanish and Turkish. The two language-specific strategies were a multi-language version of SentiStrength (i.e. ML-SentiStrength) and a commercial API for Semantria.

They investigated these methods to address the problem of multiple languages in SA. First, they used MT (i.e. Google translate) to translate texts from a specific language into English and then employed the existing English-based sentiment classification methods [16], [25] to translate texts from a particular language into English and then employed the current English-based sentiment classification methods in Table 1.

These methods were evaluated and compared across all nine languages [25], [67]. Further details of the classification classes of these methods are presented by [16] and [25]. Although the datasets were small, the researchers concluded that the existing English methods performed better than the two language-specific approaches. In this regard, SentiStrength was shown to be the most accurate method for SA, but language-specific techniques significantly impacted MT-based approaches.

TABLE 1. Overview of MSA methods with classes.

Methods	Classes	L	ML
AFINN	3	✓	
Emolex	2	✓	
Emoticons	2	✓	
Happiness Index	9	✓	
Opinion Finder (MPQA)	3	✓	✓
Opinion Lexicon	3	✓	
PANAS-t	11	✓	
Pattern.en	2	✓	
SO-CAL	3	✓	
NRC Hashtag	8	✓	
Stanford Recursive Deep Model	6		✓
Sentiment140Lexicon	4	✓	
SAsentiment analysis	4		✓
SentiStrength	3		✓
Umigon	3	✓	
Vader	3	✓	

Similarly, Araújo et al. [16] evaluated the performance of 16 English-based methods (including Google prediction API) for multilingual sentence-level sentiment analysis across 14 languages. They added five more languages: Chinese, Greek, Hindi, Czech and Haitian Creole to expand their previous work. They compared these English methods with three language-specific methods, including the IBM Watson API commercial sentiment analysis system developed by IBM. The sentiment analysis approaches employed previously [25] were explored. They investigated how the methods of using MT systems addressed multiple languages and found that the MT strategy should be used as a baseline system for new MSA systems [16]. The goal was to evaluate how effectively non-English texts could be analysed using English sentiment analysis methods, and MT systems [68]. As a final contribution, they developed iFeel 3.0, a web-based framework tool for multilingual sentence-level SA [16], [69]. Their methods, datasets and codes of the research work are freely accessible online to the research community.

E. CO-TRAINING MSA METHODS

Pan et al. [70] investigated a cross-lingual approach that employed an annotated sentiment corpus in English to predict the sentiment polarities in the Chinese language. They used an MT system to create the training dataset. This approach used the co-training of the two models simultaneously and added lexical knowledge to improve model accuracy. Co-training is training two or more monolingual models of the languages involved to build an MSA model [64]. This approach showed that adding lexical knowledge could improve the accuracy of the sentiment classification model.

In another study by [47] and [72], they used the co-training method to overcome the problem of cross-lingual SA. He used the cross-lingual method, which employed a readily available English corpus for Chinese sentiment classification, using the English corpus as training data [47]. He then exploited a bilingual co-training approach to leverage the

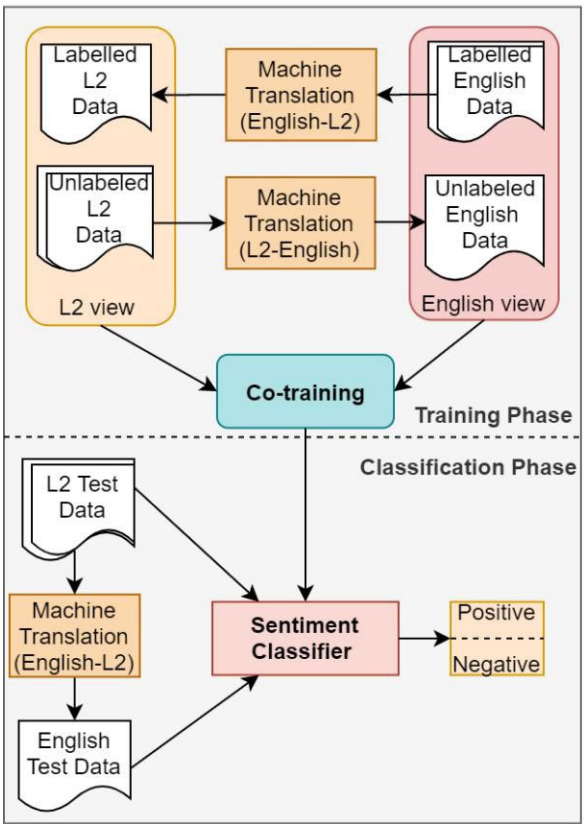


FIGURE 5. A co-training model for cross-lingual sentiment analysis.

annotated English resources to perform sentiment classification in Chinese reviews. An MT system translated English labelled documents into Chinese, and a similar approach was used to translate unlabelled Chinese documents into English. Fig. 5 shows a co-training model used for cross-lingual sentiment analysis where English labelled data is transferred to another language, then apply sentiment classification on English data. In this work, an SVM-based classifier was adopted for sentiment classification. The co-training models are designed to select the high-confidence samples suited for training data. However, the classifiers in each language view will increase the probability of adding incorrect labels to the training set. Furthermore, adding such samples increased the accuracy of the learning model but gradually decreased the performance of the initial classifiers. Nevertheless, the co-training approach can evaluate the different languages when the datasets are readily available.

Another approach incorporated an ensemble of English and Chinese sentiment classifications. Peng et al. [4] used an MT system, conducted an analysis of English and Chinese reviews, and the results were combined to improve the overall performance of the sentiment classification system. However, this approach is not reliable due to the poor results of the translation system when the domain knowledge is different from the target language. Furthermore, the technique used for the Chinese language cannot easily be adopted without modifying the language models. Thus, a sentiment lexicon in the

target language is required for this method to work optimally on other languages and cannot be applied to other languages with no lexical resources. In addition, the MT system led to the accumulation of errors and reduced the accuracy of the translation. Despite the use of the MT system in this approach, a structural correspondence learning technique was applied to find a low-dimensional representation shared by the two languages at the feature level. This technique was done to reduce translation errors [72].

Although the most apparent solution to multilingual sentiment classification is by employing MT and using existing English methods to deal with multiple languages, which may not be an easy and reliable solution to MSA referred to in our study (i.e., mixed-language texts in a sentence). Secondly, the earlier approaches have studied how sentiment analysis can be done for languages other than English using MT. In our view, these methods are "*short-cut*" solutions to address issues presented by multiple languages. Some of these methods have explored the MSA task as a 2-class polarity classification task, while others tackle MSA methods as a 3-class polarity detection problem (i.e., positive, neutral, and negative) for mixed-language comments [67]. However, an extra class is added in some methods to transform the MSA task into a 3-class polarity detection task.

F. MULTILINGUAL CORPUS-BASED SENTIMENT ANALYSIS

The following section deals with studies focusing on MSA methods for low-resource languages rather than multilingual subjectivity analysis in multiple languages. Texts written in different languages pose a considerable challenge to sentiment polarity classification. However, [17] and [29] proposed a multilingual approach that addresses the problem of sentiment polarity classification using Twitter data from different languages. They employed and compared three techniques in English and Spanish and used three ML models to address the issues presented in other languages. As a result, the following models have been developed [29], [62]:

- **Multilingual approach model**: This approach is achieved by training a multilingual dataset that does not require prior language identification or recognition phases. To accomplish a multilingual model, they merged two or more monolingual datasets to train and develop a single-pass multilingual sentiment classifier.
- **Dual monolingual approach models**: These two or more monolingual models know the origin of the text's language. Each model is trained and adjusted by using a monolingual corpus. In this case, the correct monolingual model is executed for sentiment classification once the language of the text is known.
- **Monolingual pipeline with language detection model (pipe model)**: This model acts based on the decision provided by the language identification tool. This approach identifies a language given unknown text using a language identification tool. The training is similar to the monolingual system, as the language of the texts is known before using the correct sentiment classifier.

Vilares et al. [17] evaluated sentiment analysis approaches using Spanish and English datasets. They created a Spanish-English multilingual and monolingual English and Spanish models with language detection tools [29], [62]. They used monolingual corpora from the SemEval 2014 task-B corpus and the TASS 2014 corpus for English and Spanish, respectively. These monolingual corpora were combined to create a multilingual corpus for training and testing the classification model. L2-regularised logistic regression was employed for sentiment classification, which was then compared with a super-supervised model based on the bag-of-words. Four features were considered: words, lemmas, psychometric properties and part of speech tags. The word features are obtained using a simple statistical model for counting word frequencies in texts; psychometric properties refer to emotions such as anger or topics (e.g. job) that commonly appear in messages [17].

Three approaches are evaluated using monolingual, synthetic multilingual and code-switching corpora of English, and Spanish tweets [29]. First, code-switching for the testing set was obtained by filtering tweets containing Spanish and English words. Then, three annotators labelled these tweets manually using the SentiStrength strategy, which uses a dual score to indicate positive or negative sentiments [25]. The conclusion was that the multilingual model approach was the best option when Spanish was the majority language. It was due to the high number of English words in Spanish tweets. Furthermore, monolingual models with language detection performed well only when English was the dominant language. Again, this was because of the lower number of Spanish words in the English corpus. Therefore, the monolingual approach cannot be used for multilingual settings. They also reported that the monolingual (pipeline) model with a language identification tool performed worse on the code-switching test set for most of the features used. Finally, the multilingual model approach obtained the best performance of 59.34%, using features such as lemmas and psychometrics. The lemmas are simply terms labelled using set rules to reduce data sparsity. However, in general, the atomic set of features, such as words, psychometric properties, or lemmatisation and their combinations, performed better under the proposed multilingual model approach [17].

The proposed multilingual model approach appears more robust in environments containing code-switched tweets and tweets written in multiple languages. However, again, it was concluded that neither dual monolingual nor multilingual strategies based on language detection are optimal for addressing code-switching texts. Notably, the performance accuracy of these systems on the experimented features was still $<70\%$, even after the improvements reported by [62]. In addition, they felt it would be interesting to explore the performance of MSA using neural deep network methods. Finally, the authors suggested that using DL architectures can help deal with code-switching texts [19], [62].

Tho et al. [73] investigated a code-mixed sentiment analysis of the Indonesian language and Javanese language using

a lexicon-based approach. The authors compared two translated lexicon models such as SentiNetWord and VADER. They collected 3,963 tweets from two accounts that provide code-mixed tweets. The results of the manual labelling with the lexicons mentioned above showed that SentiNetWord outperformed the VADER lexicon. However, the overall performance showed that the VADER lexicon performed better than SentiNetWord.

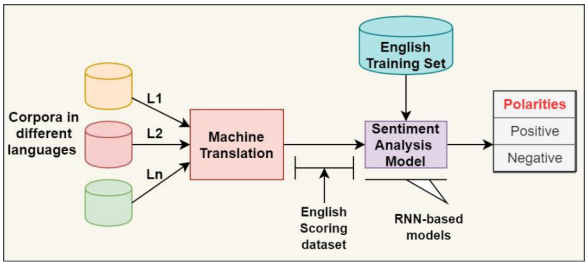


FIGURE 6. Multilingual sentiment analysis approach using RNN models [26].

G. DEEP LEARNING METHODS FOR MSA

Recently, [26] presented an approach for MSA based on an RNN framework which aimed to answer the question: *"Can a sentiment analysis model trained on a language be reused for sentiment analysis in other languages, Russian, Spanish, Turkish, and Dutch, where the data is more limited?"* A single multilingual sentiment model utilising English data was developed for four different languages. The approach was built by training sentiment models using RNN methods with English reviews. An MT strategy was employed, translating Russian, Turkish, Dutch and Spanish studies into English and then reusing the English-based RNN model to classify sentiments. Fig. 6 shows an RNN-based structure for the MSA task where an MT-based system is utilised to translate the text in non-English language to English and apply deep learning sentiment classification on English data.

The MSA approach in this study was developed to eliminate the need to train language-dependent models and sentiment word embeddings in four languages. Can et al. [26] claim that other languages with low resources can utilise this multilingual model. In their study, the method was compared with a lexicon-based method that uses SentiWordNet to obtain positive and negative sentiment scores considerably better, with an accuracy of >80% for the three languages and 74% for Turkish. The RNN-based model eliminates the feature extraction process. Can et al. [26] concluded that the RNN-based model performed significantly better than language-specific models in all four languages, despite the misclassification encountered during translation. Furthermore, their study offers a solution that employs a single multilingual model, but it does not consider mixed-language texts.

Previous MSA methods have utilised English methods for which robust classifiers are readily available [16], [62]. Another work by [13] proposed an approach for the multilingual sentiment classification of Twitter data (i.e. English, German, Portuguese and Spanish), namely an efficient translation-free DL architecture to perform MSA on tweets in multiple languages. This approach was implemented by employing cost-effective character-level embeddings and Adhoc convolutions to learn a different language. The MSA model could learn hidden features from the four languages used during the training phase in their study. The authors compared their work with three different neural network architectures [74]. Each neural network was trained using different embedding strategies.

Their study concluded that the proposed multilingual approach achieved the best performance accuracy with the LSTM-based model. The LSTM models also performed optimally on all four languages when the F_1-*score* was evaluated. However, this classification model did not undergo the pre-processing phase, affecting performance accuracy [13]. The authors claim that these models can be extended to other languages and handle texts written in multiple languages. Furthermore, they suggest that employing a fully supervised CNN model will increase performance accuracy. However, it is not worth the time manually labelling thousands of tweets; therefore, a different approach can be investigated. Finally, the authors argued that a multilingual strategy offers several advantages over a language-specific sentiment model.

Deriu et al. [23] proposed a novel approach for multilingual sentiment classification of short texts in four languages (i.e. English, Italian, German and French) to enhance the system's ability to deal with mixed languages. They used weakly supervised data trained only on a CNN method for up to three layers. This approach trains multi-layer CNN where word embeddings (i.e. word2vec) are created on a large corpus of unlabelled tweets. Word embeddings are generally numerical representations of words input into DL-based methods. These are used for language modelling and feature learning (i.e., Word2Vec, GloVe and FastText) [14], [53].

The CNN model was trained in an unsupervised phase, where word embeddings are created on a large corpus, a distantly supervised step trained on the weakly labelled dataset, and a supervised stage, where the network was fully trained manually annotated tweets [74]. They evaluated the performance of the sentiment model with different datasets, including the benchmark sentiment prediction dataset from SemEval-2016 Task 4. They demonstrated that a single-CNN model could be trained successfully for MSA tasks rather than separate classification models for each language. However, the performance of the model can be improved by training a large number of convolutional layers. This method can be easily extended to new languages, and multilingual texts [23]. Deriu et al. [23] concluded that CNN models require a large amount of training data for the model to perform well, as well as the labelled dataset.

Similarly, [75] followed an approach that achieved the best results in the SemEval 2017 task. Even the system by

TABLE 2. Summary of the methods, corpus, and techniques for MSA.

Sources	Methods	Languages	Sentiment Corpus
[30]	**MT, SentiWordNet**	English, German	2,000 reviews
[56], [60]	**MT, NB, SVM**	English, Romanian, Arabic and Spanish , French, German	MPQA MPQA corpus - 9,700 labelled sentences
[57]	**NB**	English and Arabic	Movie reviews, news corpora TED talks corpus - 183,000 manually annotated sentences
[77]	**SVM, NB, ME**	English, Dutch, French	2,500 blogs reviews and news sites.
[47], [71]	**SVM**	Chinese, English	Amazon product reviews (8,000)
[18]	**MT, NB, DT, SVM**	Chinese, English, Japanese	MPQA corpus - 535 for English, newspaper headlines - 1,200 movie reviews-12K (objective , subjective)
[15], [24]	**MT, SVM**	English, Spanish, French	Annotated NTCIR 8 Multilingual Opinion Analysis Task (MOAT - 6,200)
[35], [66]	**SVM SMO**	Italian, Spanish, French, German	Annotated NTCIR 8 Multilingual Opinion Analysis Task (MOAT - 6,200)
[68]	**AFINN, SentiStrength, Vader MPQA, SO-CAL, Opinion Lexicon SASA, Umigon, Sent140lex Emoticons, NRC, Pattern.en PANAS-t, Opinion Finder, Emolex Stanford Recursive Deep Model**	Arabic, Dutch, French, German, Italian, Portuguese Russian, Spanish, Turkish	Human labelled products, food reviews and tweets - 23,000
[16], [25]	**AFINN, ML-SentiStrength, Vader MPQA, SO-CAL, Opinion Lexicon SASA, Umigon, Sent140lex Emoticons, NRC, Pattern.en PANAS-t, Opinion Finder, Emolex Stanford, Semantria, IBM Watson**	Arabic, Dutch, French, German, Italian, Portuguese, Russian, Spanish, Turkish English, Croatian, Hindi, Haitian Creole, Chinese	Human labelled products, food reviews and tweets - 23,000
[17], [29], [63]	**L2-RLR**	Spanish and English (mixed)	SemEval 2015 Task B (8,200 tweets) TASS 2014 - (7,200 tweets)
[78]	**LSTM and MT**	Chinese and English	NLP& CC 2013 (91,600 unlabelled reviews)
[13]	**CNN, LSTM**	English,German Portuguese, Spanish	Annotated tweets (128,200, subset of 1.6M)
[79]	**CNN**	English, French, Greek	Labelled restaurant reviews (62,600)
[23]	**CNN + word2vec**	English, Italian, German, French	Unlabeled tweets (300M), weakly-labelled data (40-60M), & annotated tweets (71,000)
[38]	**CNN**	English, Spanish, French, Russian, Arabic, Dutch, Turkish, Chinese	SemEval 2016 Task 5 corpus
[80]	**CNN**	English, German, Portuguese, Spanish	Annotated tweets (128,000, subset of 1.6M)
[12]	**BiLSTM**	English, Hindi	English annotated tweets (114,000) Hindi-English labelled Facebook posts (3,800)
[26]	**MT, RNN**	Russian, Turkish, Dutch, Spanish	Yelp Dataset Challenge (8,000), Amazon reviews (8M) Kaggle competition (68,000) Restaurant reviews

TABLE 2. *(Continued.)* **Summary of the methods, corpus, and techniques for MSA.**

[81]	**CNN**	English, Spanish, Dutch, German, Russian, Italian, Czech, Japanese, French	Movie reviews (12,200), labeled reviews (20,800, TripAdvisor, Amazon Fine Food & labeled tweets (labelled
[82]	**LSTM**	English, French, Greek	Labeled restaurant and hotel reviews (91,800)
[37]	**GAN + DAN**	English, Chinese, Arabic	Annotated tweets (48,100) & Weibo posts (53,600)
[34]	**BiLSTM**	English, Bengali, Hindi, Kannada (English mixed)	Labelled Facebook comments (22,500)
[19]	**CNN**	Bambara & French (mixed)	Facebook comments (17,000, subset of 74,000)
[83]	**SVM_$Linear$, SVM_RBF**	English, Greek	(5,300 reviews)
[16]	**AFINN, SO-CAL**	Arabic, Dutch, French, German, Italian, Portuguese, Russian	Human labelled products, food reviews and tweets (23,000)
[84]	**CNN, Bi-LSTM**	English, Hindi	Facebook posts (3,800)
[20]	**Double LSTM**	English & Hindi (mixed)	Hindi-English labeled sentences of Facebook posts (3,800)
[27]	**Bi-LSTM**	English, Hindi, Bengali	Annotated tweets: English (9,200), Bengali-English (5,500) English-Hindi (18,400) labeled Facebook posts English-Hindi (3,800)
[85]	**mBERT XLM-RoBERTa**	Hindi-English, Spanish-English, Tamil-English, Malayalam-English. Malayalam-English	Hinglish Tweets (14,000), Tanglish - YouTube comments (9,600) Spanglish - tweets (12,000) YouTube comments (3,900)
[11]	**XLM-RoBERTa**	Malayalam-English, Tamil-English YouTube Tamil-English	6,740 YouTube comments comments (15,740)
[73]	**SentiNetWord VADER**	Indonesian, Javanese (mix)	3,963 tweets
[86]	**BERT**	Persian-English	3,640 labeled tweets
[87]	**LR, NB, DT, RF, SVM, BERT, DistilBERT, ALBERT RoBERTa,XLM, XLM-R Character BERT**	Tamil-English, Kannada-English, Malayalam-English	60,000 YouTube comments 44,000 - Tamil-English, 20,000 Malayalam-English
[88], [89]	**NB, SVM TeluguSentiWordNet**	Telugu–English	15,744 YouTube comments
[90]	**CNN, LSTM, GRU, BiGRU BiLSTM, BiLSTM+CNN, LSTM+CNN BiGRU+CNN, GRU+CNN, XLM-R**	Malayalam– English	7,000 comments (FIRE 2020) 20,000 comments (EACL 2021)

Nguyen and Nguyen [76] that employs deep CNN and Bi-LSTM has shown that word2vec strategies can significantly improve classification accuracy. Additionally, recent improvements in DL techniques, especially the combination of CNN and LSTM techniques, have produced greater accuracy than per-language models [21], [75].

Medrouk et al. [79] proposed an approach which employs a deep neural network for sentiment analysis in a multilingual corpus. The deep neural networks used in their study employed CNNs (i.e. feed-forward). The authors constructed their multilingual opinion corpus in three languages (English, French and Greek). The CNN exploited n-gram level information, and the system achieved high accuracy for sentiment polarity prediction. They concluded that the model used for feature extraction was language-independent. Another study by [84] presented a hybrid architecture for sentiment analysis of English-Hindi code-mixed data. They trained sub-word-level representations for sentences using the CNN model and employed a dual-encoder network consisting of two Bi-LSTMs. The model combined a network of

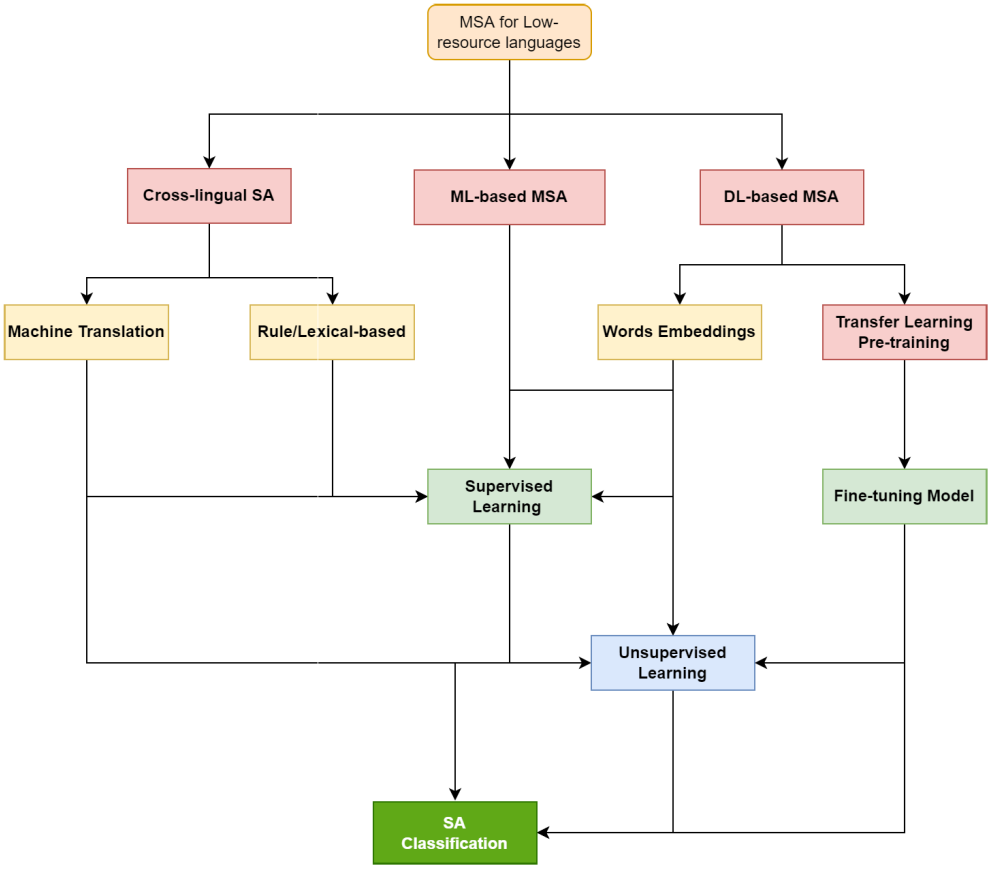


FIGURE 7. A general taxonomy of the MSA for low-resource languages.

orthographic features and word embeddings and achieved the best results with an accuracy of 83.54%.

Zhou et al. [78] proposed a cross-lingual sentiment classification approach using attention-based bilingual LSTM networks. The attention-based bilingual representation learning model was used to learn document distributed semantics for both the source and target languages. This approach was implemented for languages like Chinese and English. The authors used Google Translate MT to translate the training data into the desired target languages and then employed a bidirectional LSTM network to model the documents for both the source and target languages. A dataset from the cross-language sentiment evaluation of NLP&CC 2013 was used [78]. Reviews are divided into three categories: books, DVDs and music. On average, the bilingual model yielded an accuracy of 82.4% across all domains. They concluded that LSTM could capture the compositional semantics of bilingual texts and that the proposed model achieved promising results on the dataset used [14], [78]. It also outperformed the best results in NLP&CC cross-language systems. An interesting part of their study is that the attention model could find the key sentences in a document, and the sentiment signals were captured with the help of word-level attention [23], [78].

Another work by [91] proposed a Character-to-Sentence CNN (CharSCNN) method that exploits characters to sentence-level information to perform sentiment analysis of short texts (i.e. Twitter data). They used the CharSCNN model with two convolutional layers to extract semantic information from word and sentence features to improve the performance of the sentiment analysis system. Irsoy and Cardie [92] improved sentiment classification accuracy using an RNN model on time-series information to obtain sentence representations. Socher et al. [36] improved sentiment analysis by using a recursive neural tensor network model, which synthesises the semantics of the syntactic tree of binary sentiment polarity. A good sentiment classification accuracy was also obtained using a tree-structured LSTM model with semantic association [93]. Furthermore, Baziotis et al. [94] presented an attention strategy for the LSTM model to achieve good sentiment analysis results on the SemEval-2017 Task-4 dataset.

H. CODE-SWITCHED MSA METHODS

To some extent, most of the research on the code-mixed text has focused on the English-Hindi setting [12], [20], [84]. However, Code-mixed challenges can be addressed by learning the sentiment feature space and preserving the similarity of the sentences in which the sentiment is portrayed [12]. It allows a straightforward measure of the relatedness between code-switched content and labelled data from

a resource-rich corpus. Choudhary et al. [12] demonstrated this using Siamese Bi-LSTM with tri-gram embeddings and a fully connected layer. Additionally, they compared the model, which was trained with a pair of Hindi-English texts, one with pairs of English sentences and another with code-mixed texts, yielding a lower *F-score* by 8.7% to the other with 75.9%. Their results suggest that adding more resource-rich data (i.e. the English dataset) is beneficial, as it increases model performance.

In other cases, most researchers have used DL techniques to model sentiment analysis for code-switched datasets. For example, Konate and Du [19] used Facebook comments of Bambara-French with different DL architectures such as LSTM, Bi-LSTM, CNN and Bi-LSTM-CNN, together with embeddings as input at the word or character level. Their proposed model can learn multilingual embeddings from input characters and words to mitigate the embeddings from the language model (Elmo) lack of pre-trained embeddings in the code-mixed corpus. They obtained the best results using a single-layer CNN model with an accuracy of $>80\%$. In addition, they compared LSTM and CNN, where the latter showed the best results in such a domain.

Furthermore, Kusampudi et al. [89], proposed a sentiment analysis in code-mixed Telugu-English text with Unsupervised Data Normalization. They reported accuracy of an 80.22% on this dataset using novel unsupervised data normalisation with an MLP model, which is an increase of 2.53% accuracy due to this data normalisation. According to Saikrishna et al. [88], combining Telugu and English or Tamil and English in the same sentence is commonly observed. Saikrishna et al. [88] developed a sentiment analysis system for Telugu–English code-mixed sentences. They classified the polarity of the code-mixed sentences collected from Youtube comments into positive and negative sentiments using lexicon-based approaches and ML approaches such as NB and SVM classifiers. They achieved an accuracy of 82% and 85%, respectively, outperforming the lexicon-based method.

I. PRE-TRAINED METHODS FOR MSA

The BERT model has been applied to several NLP tasks. BERT has demonstrated outstanding performance in state-of-the-art text classification, including MSA tasks [53], [95]. BERT is a Google model created and reported in [53]. It was created using a significant amount of plain text data freely available on the Internet, and the model was trained unsupervised. Before BERT, a few other pre-trained language models used bidirectional unsupervised learning.

The ELMo is one such model that focuses on contextualised word representations [96]. It constructs word embeddings by utilising LSTM, which separately trains left-to-right and right-to-left word representations and then concatenates these embeddings [96]. On the other hand, BERT does not use LSTM to obtain word context characteristics; instead, it employs attention-based transformers [53]. These models are beneficial for low-resource languages when there is a large amount of unlabelled data, but not much task-specific labelled data. Tomohiro et al. [97] presented a novel model that learns from sentences by labelling them with emojis, utilising English and Japanese tweets to create the corpus. The authors validated and evaluated many models based on attention LSTM, CNN and BERT. In addition, they compared the BERT model with the standard models CNN, FastText and attention Bi-LSTM, all of which received good results in prior studies. Compared to the traditional models, the authors performed better using the BERT model.

Gupta et al. [85] maintain that their unsupervised model understood code-switched languages or learnt only their representations. They introduced an unsupervised self-training method as a generic framework and demonstrated its applicability to the specific use of code-switched data. They exploited the power of pre-trained BERT models to initialise and fine-tune them using only pseudo-labels generated via zero-shot transfer. Their study was conducted in four code-mixed languages: Hinglish (Hindi-English), Spanglish (Spanish-English), Tanglish (Tamil-English) and Malayalam-English. They concluded that their unsupervised models outperformed their supervised counterparts, with performance ranging from 1% to 7%. Another study by [86] used a pre-trained multilingual BERT model to learn the polarity scores of these tweets for code-mixed Persian-English sentiment analysis. They collected tweets and employed two annotators to label the code-mixed tweets. Their Multilingual BERT (mBERT) model outperformed the baseline models that use NB and random forest (RF).

Ou and Li [11] proposed a system to identify the sentiment polarity of the code-mixed dataset of the Dravidian dataset. They built on a pre-trained multi-language model such as the Cross-lingual Language Model RoBERTa (XLM-RoBERTa) [98], and their system employed a k-folding approach to the ensemble and addressed the sentiment analysis problem of multilingual code mixed across language models. They took part in two code-mixed language challenges (Malay-English and Tamil-English). Their system had the highest *F-Score* of > 0.7 in Malayalam-English and ranked third in Tamil-English with an *F-score* > 0.6.

Chakravarthi et al. [87] introduced a code-mixed dataset of the under-resourced Dravidian languages. They manually annotated a dataset from social media comments for three under-resourced Dravidian languages. For over 60,000 YouTube comments, the dataset was annotated for sentiment analysis and identifying offensive language. The collection includes roughly 20,000 comments in Malayalam and English, 7,000 comments in Kannada, and 44,000 comments in Tamil [99]. Unpaid volunteers manually annotated the data, and Krippendorff's alpha indicates a high level of inter-annotator agreement. Utilising machine learning and deep learning techniques, they provided baseline studies to create benchmarks on the dataset with the highest accuracy of 71% with the XLM technique. Notably, traditional machine learning methods have suffered low performance.

Thara et al. [90] investigated two major aspects of the code-mixed text: offensive language identification and sentiment analysis for Malayalathe m–English code-mixed data set. Their framework utilises different word embedding methods, such as Word2vec and FastText. They evaluated different deep learning methods (CNN, LSTM, Gated Recurrent Unit (GRU), BiLSTM, and Bidirectional GRU (BiGRU)) on Forum for Information Retrieval Evaluation (FIRE [9]) 2020 and European Chapter of the Association for Computational Linguistics (EACL [10]) 2021 dataset. Among the hybrid models, GRU+CNN and BiLSTM+CNN turned in the highest F1-score of 0.9969. The challenge with this study is that the training dataset for sentiment analysis was minimal. They obtained the best performance accuracy of 99% using the transformer-based model XLM-R. Next, we will outline the evaluation metrics for MSA.

VI. EVALUATION METRICS FOR MSA

In addition, we looked at evaluation measures for sentiment classification model performance. Several evaluation metrics have been identified from the systematic literature review [19], [23], [100]. These are reported in the SemEval 2016 challenge [38], averaging the macro F1-score of the positive and negative classes. The confusion matrix is also used as evaluationon parameter to measure sentiment analysis performance [13], [14], [23]. Researchers use four metrics such as: *True positive*, *True negative*, *False negative* and *False positive*. These metrics are described as in [13] and [23]. Metrics such as *accuracy, precision, recall and F-score* are generally used to evaluate the performance of the sentiment analysis classifiers [13], [25].

VII. RESULTS AND DISCUSSION

In this section, we describe the results and then further discuss them.

A. RESULTS

We explore how we answer our research questions as we present the findings. Research question 1 aims to identify MSA datasets and resources for under-resourced languages. Research question 2 aims to determine if MT methods are suitable for building MSA systems. To achieve this, we used the information from the literature review. A summary of the methods from the selected studies is presented in Table 2. Additionally, Table 3 shows a quantitative summary of the results of our research questions 1 to 4. Our results show that DL methods at 61% were the most common MSA techniques for multiples languages, followed by ML methods at 40% and lexicon methods at 37%. Lastly, 29% of these studies used MT systems to help build their MSA resources. From the results in Table 3, we can conclude that DL methods are the leading techniques for MSA, including those where the English language is mixed with other languages, followed by ML and lexicon methods. In this case, lexicon-based, ML and MT methods were almost equally adopted by some of the studies presented in Table 2. Additionally, our research reveals that 63% of the publications studied ternary classification, while 37% and 13% looked into binary and five-category classification. Furthermore, only 31% of the DL methods have explored the binary classification, while 45% of the DL methods have concentrated on ternary classification.

Furthermore, we identified languages studied for MSA and the methods focused on mixed languages. Table 5 presents a summary of results for the languages involved. 42% of the selected studies used two languages in their proposed methods, 18% of the selected studies used three languages, and about 21% used four languages. Four studies with 3% used five, six, eight and fourteen languages in their studies, 34% of these studies have focused on code-switched datasets and only 8% used nine languages. Moreover, 63% of these studies have tackled MSA as a 3-class problem followed by a 2-class problem at 37%. Furthermore, out of the 34% of the studies which focused on a mixed-language dataset, English is the most commonly mixed language except for studies where the French language was mixed with the Bambara language, and Indonesian was combined with the Javanese language. Furthermore, we have noticed that the Persian language is mixed with English. In table 4, we presented our top thirty-five articles, which are highly cited. We also indicate their FWCI score which shows how well cited the article is compared to other similar articles. It is suggested that a value greater than 1.00 means the document is more cited than expected according to the average as in Table 4. Although some studies used a mixture of English, French, Hindi, and Bambara, their classification techniques are mainly based on DL methods.

B. DISCUSSION

Table 2 and Table 3 are the summarised versions of the different methods and techniques used for sentiment analysis in multiple languages. Although most of the research studies utilised English methods, it is still a challenge for languages that do not have sufficient resources. Even methods that use MT systems are unreliable in tackling the task of MSA, owing to the limitations of MT systems [24], [35], [66]. Furthermore, many state-of-the-art sentiment analysis classification methods are based primarily on supervised learning algorithms. This means that a large amount of manually labelled data is required. Therefore, there is currently an immense need for techniques that require less human intervention [13], [23], [104] and even for data annotation of mixed-language texts. In this study, we found that research on MSA has shifted from lexicon or corpus-based and MT-based methods to a multilingual approach using DL techniques, which currently show incredibly encouraging results. Research on low-resource languages is gradually gaining momentum, and studies are paying more attention to code-mixed, and code-switched texts [27], [85]. It can also be noted from Table 2 that over 30% of the MSA studies preferred

[9] https://dravidian-codemix.github.io/2020/
[10] https://competitions.codalab.org/competitions/27654

TABLE 3. Summary of the methods and techniques for MSA with their classification categories where L = lexicon, B = binary, T = ternary, F = five categories.

Sources	L	MT	ML	DL	B	T	F
[30]	✓	✓	✓		✓		
[60]	✓	✓	✓		✓		
[57]	✓		✓		✓		
[77]		✓	✓			✓	
[47], [71]		✓	✓		✓		
[18]	✓	✓	✓		✓		✓
[15], [24]	✓	✓	✓			✓	
[35], [66]	✓	✓	✓			✓	
[17]	✓		✓		✓		
[29], [62]	✓		✓		✓		
[25], [68], [69]	✓	✓	✓			✓	✓
[13]				✓	✓		
[79]	✓			✓	✓		
[83]	✓		✓			✓	
[79]				✓	✓		
[23]				✓		✓	
[78]		✓		✓	✓		
[38]				✓		✓	
[80]				✓	✓		
[26]		✓		✓		✓	
[12]				✓		✓	
[81]				✓	✓		
[82]				✓	✓		
[37]				✓		✓	✓
[34]				✓		✓	
[19]				✓		✓	
[84]				✓		✓	
[16]	✓	✓	✓			✓	✓
[27]				✓		✓	
[39]				✓		✓	
[97]				✓		✓	
[85]				✓		✓	
[11]				✓		✓	
[73]	✓					✓	
[86]				✓		✓	
[87]			✓	✓		✓	
[88]	✓		✓			✓	
[90]				✓		✓	
Total	14	11	15	23	14	24	5
%	37%	29%	40%	61%	37%	63%	13.2%

to use data collected from the Twitter platform. Moreover, most of these datasets were hand labelled by hand annotators despite efforts to build auto-labelling methods [52].

We are guided by the methods in the systematic literature survey to draw a general taxonomy of MSA for low-resource languages, as shown in Fig 7. In Table 5, we can deduce that the number of languages studied increases yearly. The selected studies show that an increasing number of languages are gaining traction in the context of MSA. Several studies have used Twitter to develop sentiment analysis datasets. The fact that more languages are studied means that there are still more under-resourced languages to consider since many different languages are used in social media. Additionally, although researchers are using MT-based methods to generate language resources for under-resourced languages, the universal approach to cater for many languages is still far from being achieved. Therefore, there is an excellent opportunity for future sentiment analysis studies to focus on developing versatile techniques [14].

We noticed a shift in how MT systems were used for MSA research in languages with limited resources. Many studies used the monolingual dataset to address the issues of multiple languages [105]. In contrast, other studies looked at utilising a cross-lingual approach using MT systems [81]. The evidence is presented in Table 2, which demonstrates that DL methods account for 62% of all methods employed,

TABLE 4. Top 35 and leading articles for MSA studies with citations and field-weighted citation impact (FWCI).

No.	Sources	Year	No. of Languages	No. of Citations	FWCI
1	[30]	2008	2	233	7.82
2	[47]	2009	2	356	11.4
3	[77]	2009	3	291	5
4	[56]	2010	4	96	7.01
5	[18]	2011	3	175	14.42
6	[70]	2011	2	29	3.19
7	[35]	2014	4	123	7.94
8	[66]	2013	4	24	3.12
9	[17]	2015	2	36	1.17
10	[79]	2016	2	166	14.74
11	[101]	2016	2	100	3.44
12	[22]	2016	2	131	5.18
13	[38]	2016	8	54	7.45
14	[25]	2016	9	42	3.46
15	[29]	2016	2	21	2.13
16	[83]	2017	2	194	15.58
17	[6]	2017	3	106	12.67
18	[23]	2017	4	124	9.8
19	[62]	2017	2	48	4.14
20	[79]	2017	3	16	3.18
21	[34]	2017	5	17	2.53
22	[13]	2017	4	12	1.2
23	[12]	2018	2	33	4.75
24	[26]	2018	4	31	2.12
25	[19]	2018	2	15	0.69
26	[102]	2018	4	14	1.16
27	[84]	2019	2	28	4.6
28	[103]	2019	9	17	4.51
29	[20]	2019	2	5	1.88
30	[16]	2020	14	32	4.75
31	[86]	2021	2	4	-
34	[89]	2021	2	6	-
32	[87]	2022	4	24	3.94
33	[88]	2022	2	-	-
35	[90]	2022	2	-	-

TABLE 5. Number of languages used for MSA models including mixed languages.

No.	No. of Languages	Studies	Total	%
1	2	[17], [22], [47], [78], [83], [84], [101] [12], [19], [30], [57], [70], [83], [86], [88], [90]	16	**42%**
2	3	[11], [18], [27], [37], [77], [79], [82]	7	**18%**
3	4	[13], [23], [26], [34], [66], [82], [87], [102]	8	**21%**
4	5	[85],	1	**3%**
5	6	[56]	1	**3%**
6	8	[38]	1	**3%**
7	9	[25], [69], [103]	3	**8%**
8	14	[16]	1	**3%**
9	Mixed	[11], [13], [19], [23], [27], [29], [34], [62], [85] [88], [90]	13	**34%**

followed by ML-based methods at 41%, lexicon-based methods (38%), and MT methods (35%). The MT-based techniques are used mainly because of their advantage of reproducing the language resources from English to other languages where MT APIs are supported. It is, therefore, difficult for languages with limited resources and not supported by various MT APIs. It is also clear that only a few studies have utilised pre-trained models to fine-tune the MSA task in low-resource languages, although there is a significant increase in DL methods.

This is mainly because the methods are still new in the NLP community. In addition, pre-trained models for low-resource languages still need to be explored. Annotated dataset remains a challenge for low-resource languages. For this, another approach used BERT methods to fine-tune the downstream tasks for low-resource languages [53], but these have been unable to outperform the existing DL methods [14].

Furthermore, multilingual aspect-based sentiment analysis is still in its infancy [106]. Despite its first study in SemEval 2016 Task 5, it produced the highest accuracy of 88.13% for English and the lowest of 73.35% for Chinese [38]. Efforts were devoted to compare DL-based methods such as CNN, LSTM, or Bi-LSTM performance and improving performance by adding attention mechanisms. However, few methods focus on self-learning sentiment analysis classification, with less attention paid to multilingual contexts. Recently, a study by [5] used Twitter to develop NaijaSenti corpus (i.e. languages such as Hausa, Igbo, Pidgin, and Yorùbá), and they evaluated their corpus using mBERT, XLM-R and Roberta. This study demonstrated that model fine-tuning on pre-trained models performs well.

On a different note, while conducting this research study, we were able to identify several challenges with some MSA studies: (i) some of the research methodologies applied are difficult to follow to build baseline studies, (ii) some research cannot be easily replicated to obtain the exact results reported in the published research papers [6], (iii) some of the MSA research studies have not yet been released or published their datasets, tools and other resources for easy rebuilding or reproduction, (iv) although some studies provided Internet links for their resources used, some were not available for use or were not updated, and some resources and datasets are available only on request [16], [25].

The practical implications of this research are to identify the gaps that can be filled in the future and to set the trend of research shift concerning sentiment analysis of under-resourced languages. For diverse MSA datasets, our systematic literature review offers a variety of tools and techniques. This systematic literature review aims to provide several contributions, covering different application methods used for MSA for under-resourced languages. We focused on illustrating the contributions of each research work and observing the type of language-specific methods, transfer learning methods with MT systems, machine learning and deep learning algorithms used. Our investigations also focus on identifying the type of dataset used, how it was gathered, and how these datasets were annotated. Additionally, they used the environment and the performance measures covered in each study, evaluating them and concluding with apparent research gaps and obstacles, which aids in identifying the non-saturated applications for which the MSA is most required in future research. For instance, aspect-based MSA deep learning systems need more sophisticated learning algorithms to produce better results.

Our findings indicate that deep learning methods are used in more than 60% of the studies, which is where significant research innovation can occur. This comprehensive literature review results bring alternative study directions for languages with limited resources and shed insight into current MSA research trends. Last but not least, this research will assist in identifying current, significant challenges in MSA for low-resource languages. Using computational models created for English or other rich languages with plenty of resources by many NLP systems puts technology developed for languages with limited resources at risk. It is beneficial to develop strategies that support languages with limited resources. We also support the development of platforms from languages with few resources accessible to everyone in other societies and using new NLP technologies for under-resourced languages. This comprehensive literature review, which examines MSA studies from the past and recent years, is anticipated to be helpful to other researchers in the future. The most popular datasets for MSA study are provided in Table 2. Next, we will describe our research methods and limitations.

VIII. STUDY LIMITATIONS

The systematic literature study may have a few limitations. There may have been published papers we missed due to our selection criteria or search keywords, even though those studies examined the MSA approaches per the period specified in the methodology section. The review was conducted by only a few researchers, meaning there may be bias in selecting studies for comparison. Only original and unique studies published from 2008 to 2022 were included, with MSA studies with knowledge-based/lexicon techniques, multilingual subjectivity and MT-based methods for optimal comparison. The intention is to provide a complete picture of the origin of the MSA methods and the direction of progress, including ML and DL methods. Comparing techniques that appear to be out-of-date may also be a drawback. However, we believe they could help develop baseline systems for other languages and dialects.

IX. EMERGING MSA AREAS

The findings from this research are discussed in this part, along with some recent development in the field that may necessitate further investigation.

A. TEXTUAL REPRESENTATIONS STRATEGIES

Text representation methods have been explored for low-research languages [96]. Research is still needed, especially for mixed-language contexts, on the desire to progress the text representation problem for under-resourced languages. Word2vec has known limitations in handling similar words. The ELMo approach was used to lessen these limitations [96]. ELMo extracts context-based word representations. For mixed languages, research on cross-lingual word embedding has drawn considerable attention. Cross-lingual word embeddings are vector representations of words from multiple languages in the same vector space, allowing

words with the same meaning but different languages to have the same vector representation. However, this technique is determined by the nature of the data requirements rather than the structure of the model. Despite these efforts, the question of which type of embedding captures better text features in the MSA task remains unanswered.

Recent research to address word representation in multilingual texts will increase as researchers are trying to study the effect of different languages and closely related language families. An interest in using DL models and pre-trained language models in under-resourced languages has grown in the most recent MSA models. However, many under-resourced languages continue to struggle with annotated datasets. Despite the advances in NLP, many under-resourced languages are still not covered by pre-trained models like BERT, RoBERTa, and XLM-RoBERTa. The use of fine-tuning language models is one strategy for addressing this problem [5].

B. MSA FOR PRE-TRAINED MODELS

The use of transformer-based models like BERT, RoBERTa, and mBERT allows researchers to focus on fine-tuning models for downstream tasks rather than training models from scratch. The recent interest in using pre-trained models for MSA tasks has become more useful for under-resourced languages with promising results for high-resourced languages [11], [53]. The research work of [39] presents extra language-specific pre-training for multilingual contextual word representations in a low-resource situation before usage in a downstream task. They enhanced the current vocabulary with frequent tokens from the low-resourced language (i.e. Irish, Maltese, Vietnamese, and Singlish (Singapore-English) and mimicked better language-specific phrases [39]. Chau et al. [39] suggests that we can improve the performance of the multilingual models on low-resourced language variations similarly by applying additional pre-training on language-specific corpora. They examined dependency parsing in four topologically varied low-resource language varieties with varying degrees of similarity to the pre-training data of a multilingual model. According to their findings, these methods consistently improve performance for each target variety, especially in low-resource conditions.

C. MULTILINGUAL ASPECT-BASED SENTIMENT ANALYSIS

Although there is an interest in continuing research on multilingual aspect-based sentiment analysis, there is a need to standardise the approach to this issue [38]. It is quite clear that this problem has not been widely addressed, especially with using DL-based architectures and considering under-resourced languages. Research on this topic suggests that the use of attention mechanisms and aspect-based embedding may significantly help resolve this problem. Perhaps MSA in an aspect-based context should be adopted for the methods that handle code-switched setups. It is essential to evaluate whether the coupling of processes will impact word embeddings and sentiment analysis classification.

D. UNDER-RESOURCED LANGUAGES

High-resource languages from different language families have been studied extensively [34], [81]. Despite steady interest in under-resourced languages, code-switching and code-mixing remain challenging within multilingual communities, preferably mixing languages with high resources. A general approach to address these challenges is lacking. Although there is promising progress in some Indian languages like Tamil, Urdu [107], [108] and Telugu [109] and Iranian languages like Persian [86], [110], much is still required to develop models that employ deep learning techniques [111]. Also, they attempted to address challenges in the Persian language by applying DL methods which only achieved the *f-score* of 55.5%. Ghasemi et al. [112] developed a sentiment analysis task in Persian language by proposing a cross-lingual deep learning framework to benefit from available training data of the English language. Deep learning models such as CNN and LSTM and their combinations were experimented with to achieve the f-score of 91.8% on LSTM-CNN. Recent work by [113] explored a cross-lingual sentiment analysis approach in the Bengali language. Cross-lingual sentiment classification is another process to handle low-resource language issues. Bengali is considered a low-resourced language due to the scarcity of annotated data and the lack of text processing tools. They created and annotated a comprehensive corpus of around 12,000 Bengali reviews using the MT system and prior sentiment information to generate accurate pseudo-labels from English-based lexicons. The best F1 score of 0.897 is achieved by integrating LR and SVM classifiers as a hybrid method. For the SVM, the best accuracy of 91.5% was achieved. The Decision Tree (DT) based methods, RF and Extra Trees Classifier (ET), achieved the lowest F1 scores. sentiment analysis for monolingual, code-switched and multilingual comments in under-resourced languages has been studied only for a few African languages, e.g. several Nigerian languages [5], [100], Swahili [114] and Bambara [19]. Annotated datasets for MSA are lacking. A concerted effort to build datasets for sentiment analysis is required, especially for under-resourced languages such as African languages [100], including the languages in South Africa [115].

A study by [5] investigated the development of NaijaSenti—an introduction of the first African large-scale human-annotated dataset for sentiment analysis in Nigerian languages (i.e. Igbo, Yoruba, Hausa and Nigerian-Pidgin). The authors evaluated their methods on several pre-trained models such as mBERT [53], XLM-R [98], mDeBERTaV3, mDeBERTaV3 and AfriBERTa [116]. They further evaluated their dataset using language-adaptive fine-tuning methods. To address this under-representation, AfriBERTa was developed –an African version of BERT trained from scratch to accommodate some of the African languages [116]. AfriBERTa has been trained in 11 languages: Afaan Oromoo (also known as Oromo), Amharic, Gahuza (i.e. a hybrid language that includes Kinyarwanda and Kirundi), Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya, and Yoruba [5], [116].

AfriBERTa was tested for named entity recognition and text categorisation in ten languages. In numerous languages, it outperformed mBERT and the Cross-lingual Language Model with RoBERTa (XLM-R) [11], [98], and it was reported to be a competitive model.

Some of the NLP models derived from BERT, such as mBERT [53], RoBERTa [117] or XLM-RoBERTa [98], were trained with many languages and can classify comments straight-forward from those languages. Unfortunately, XLM-RoBERTa [98] and mBERT are not trained with any data containing South African languages [118]. Most of these PLMs models cover 50 to 110 languages with only a few African languages, which are represented due to a lack of large monolingual corpora [116]. AfriBERTa, RoBERTa and XLM-R have not yet been evaluated with any South African languages from the sentiment analysis perspective. IsiZulu, Sesotho and IsiXhosa are only now represented and assessed using a multilingual adaptive fine-tuning model [100] trained on XLM-RoBERTa, and AfriBERTa [100], [116] but for a different NLP task.

In the context of South Africa, a concerted effort is required to create resources for South African under-resourced languages. The research could start by curating the SAfriSenti corpus—a multilingual sentiment corpus for South African languages and then expand to other African languages such as Lingala, Shona and Swahili and other African languages. In South Africa, there are 11 official languages. Languages like Sepedi (i.e. *Northern Sotho*), isiZulu, isiXhosa, Setswana, siSwati, Tshivenda, Xitsonga, and Sesotho (i.e. *Southern Sotho*) and their dialects are spoken by large populations. In addition, the lack of NLP resources for under-resourced languages makes it difficult to develop digital language technologies. Therefore, it is for these reasons that a massive data collection for under-resourced languages, in general, is necessary to address the under-resourced language challenges. Another exciting area for future research is code-switching between English and under-resourced languages.

E. AUTOMATIC DATA ANNOTATION

Supervised algorithms rely on the labelled dataset. Recent studies have investigated models which can be used to reduce human effort during data annotation [119]. For example, Kranjc et al. [104] developed active learning methods using pre-trained BERT language models. These techniques appear to be effective for text labelling vast amounts of data. Another emerging area for further research is the investigation of multi-class sentiment classification using active learning methods [119].

X. CONCLUSION

As many MSA strategies have been proposed and experimented with, many studies have evaluated and contrasted the performances of different statistical ML models and DL-based methods using MT strategies for MSA over the years. However, there is currently no method identified as the best-performing model. In this study, we reviewed the most used methods for MSA that apply traditional ML and DL models, together with those that employ MT-based methods. The performances of the MT-based methods and statistical ML and DL models reported in the literature were compared. Although some studies suggest that MT-based methods should be a baseline system for newly proposed MSA systems, these methods are still not proven for under-resourced languages. The literature results show that most DL architectures have recently spiked research attention compared to traditional ML-based classifiers, including even methods that rely on MT-based systems. Furthermore, the literature has proven that a combination of DL models such as CNN, Bi-LSTM and LSTM can significantly improve the performance of the MSA system. This study also highlighted the limitations of the use of MT-based methods. Furthermore, we propose a DL learning framework for MSA without using an MT-based system. We have also stressed that the BERT or XLM-RoBERTa model can play a pivotal role in the performance of MSA models if it is fine-tuned to handle the downstream task.

Future studies on sentiment analysis must focus more on developing gold-standard datasets suitable for sentiment analysis of multiple languages. Researchers should focus more on developing sentiment lexicons for low-resourced languages so that, in future, they can concentrate on developing advanced models. Investigation of how much sentiment is lost in translation, for instance, when moving between multilingual and a single language versus the sentiment of the original text, should be studied to report MT challenges. This study did not consider single-language sentiment analysis evaluation; future research should focus on multilingual SA, particularly in under-resourced languages. Furthermore, future research should focus more on developing DL MSA models for multiple languages or robust languages with independent MSA techniques that can be used to analyse monolingual, numerous and mixed languages. An exciting research direction is to focus on methods that address code-switching sentiment analysis and multilingual aspect-based sentiment analysis in a multilingual setting. In addition, the significant challenges and current research gaps in MSA were reviewed. Finally, future directions for research in MSA will be investigated.

REFERENCES

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[2] M. Wankhade, A. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, pp. 5731–5780, Feb. 2022.

[3] B. Pang, L. Lillian, and V. Shivakumar, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, vol. 10, Jul. 2002, pp. 79–86.

[4] H. Peng, E. Cambria, and A. Hussain, "A review of sentiment analysis research in Chinese language," *Cognit. Comput.*, vol. 9, no. 4, pp. 423–435, Aug. 2017.

[5] S. H. Muhammad, D. I. Adelani, S. Ruder, I. S. Ahmad, I. Abdulmumin, B. S. Bello, M. Choudhury, C. Chinenye Emezue, S. S. Abdullahi, A. Aremu, A. Jeorge, and P. Brazdil, "NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis," 2022, *arXiv:2201.08277*.

[6] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual sentiment analysis: From formal to informal and scarce resource languages," *Artif. Intell. Rev.*, vol. 48, no. 4, pp. 499–527, Dec. 2017.

[7] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 417–424.

[8] R. Bhargava and Y. Sharma, "MSATS: Multilingual sentiment analysis via text summarization," in *Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng.-Confluence*, Jan. 2017, pp. 71–76.

[9] A. Ghafoor, A. S. Imran, S. M. Daudpota, Z. Kastrati, Abdullah, R. Batra, and M. A. Wani, "The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing," *IEEE Access*, vol. 9, pp. 124478–124490, 2021.

[10] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. LREc*, 2010, pp. 1320–1326.

[11] X. Ou and H. Li, "YNU@Dravidian-CodeMix-FIRE2020: XLM-RoBERTa for multi-language sentiment analysis," in *Proc. FIRE*, 2020, pp. 1–6.

[12] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava, "Sentiment analysis of code-mixed languages leveraging resource rich languages," 2018, *arXiv:1804.00806*.

[13] W. Becker, J. Wehrmann, H. E. Cagnini, and R. C. Barros, "An efficient deep neural architecture for multilingual sentiment analysis in Twitter," in *Proc. 30th Int. Flairs Conf.*, 2017, pp. 246–251.

[14] M. M. Agüero-Torales, J. I. A. Salas, and A. G. López-Herrera, "Deep learning and multilingual sentiment analysis on social media data: An overview," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107373.

[15] A. Balahur and M. Turchi, "Multilingual sentiment analysis using machine translation?" in *Proc. 3rd Workshop Comput. Approaches Subjectivity Sentiment Anal.*, 2012, pp. 52–60.

[16] M. Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," *Inf. Sci.*, vol. 512, pp. 1078–1102, Feb. 2020.

[17] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora," in *Proc. 6th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2015, pp. 2–8.

[18] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual sentiment and subjectivity analysis," *Multilingual Natural Lang. Process.*, vol. 6, pp. 1–19, May 2011.

[19] A. Konate and R. Du, "Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques," *Wuhan Univ. J. Natural Sci.*, vol. 23, no. 3, pp. 237–243, Jun. 2018.

[20] S. Mukherjee, "Deep learning technique for sentiment analysis of Hindi–English code-mixed text using late fusion of character and word features," in *Proc. IEEE 16th India Council Int. Conf. (INDICON)*, Dec. 2019, pp. 1–4.

[21] S. Seo, C. Kim, H. Kim, K. Mo, and P. Kang, "Comparative study of deep learning-based sentiment classification," *IEEE Access*, vol. 8, pp. 6861–6875, 2020.

[22] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: State of the art and independent comparison of techniques," *Cognit. Comput.*, vol. 8, no. 4, pp. 757–771, 2016.

[23] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, and M. Jaggi, "Leveraging large amounts of weakly supervised data for multi-language sentiment classification," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1045–1052.

[24] A. Balahur and M. Turchi, "Comparative experiments for multilingual sentiment analysis using machine translation," in *Proc. SDAD@ ECML/PKDD*, 2012, pp. 75–86.

[25] M. Araujo, J. Reis, A. Pereira, and F. Benevenuto, "An evaluation of machine translation for multilingual sentence-level sentiment analysis," in *Proc. 31st Annu. ACM Symp. Appl. Comput.*, Apr. 2016, pp. 1140–1145.

[26] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual sentiment analysis: An RNN-based framework for limited data," 2018, *arXiv:1806.04511*.

[27] A. Jamatia, S. D. Swamy, B. Gamback, A. Das, and S. Debbarma, "Deep learning based sentiment analysis in a code-mixed English–Hindi and English–Bengali social media corpus," *Int. J. Artif. Intell. Tools*, vol. 29, pp. 399–408, Jan. 2020.

[28] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," 2019, *arXiv:1912.01973*.

[29] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "En-ES-CS: An English–Spanish code-switching Twitter corpus for multilingual sentiment analysis," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, 2016, pp. 4149–4153.

[30] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *Proc. IEEE 24th Int. Conf. Data Eng. Workshop*, Apr. 2008, pp. 507–512.

[31] K. R. Mabokela and M. J. Manamela, "An integrated language identification for code- switched speech using decoded-phonemes and support vector machine," in *Proc. 7th Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2013, pp. 1–6.

[32] K. R. Mabokela, M. J. Manamela, and M. Manaileng, "Modeling code-switching speech on under-resourced languages for language identification," in *Proc. Int. Workshop Spoken Lang. Technol. Under-Resourced Lang.*, 2014, pp. 1–6.

[33] R. Mabokela, "Phone clustering methods for multilingual language identification," in *Proc. Comput. Sci. Inf. Technol.*, Nov. 2020, pp. 285–298.

[34] K. Shalini, H. B. Ganesh, M. A. Kumar, and K. P. Soman, "Sentiment analysis for code-mixed Indian social media text with distributed representation," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 1126–1131.

[35] A. Balahur and M. Turchi, "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 56–75, Jan. 2014.

[36] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.

[37] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 557–570, Dec. 2018.

[38] S. Ruder, P. Ghaffari, and J. G. Breslin, "INSIGHT-1 at SemEval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis," 2016, *arXiv:1609.02748*.

[39] E. C. Chau, L. H. Lin, and N. A. Smith, "Parsing with multilingual bert, a small corpus, and a small treebank," 2020, *arXiv:2009.14124*.

[40] S. Sagnika, A. Pattanaik, B. S. P. Mishra, and S. K. Meher, "A review on multi-lingual sentiment analysis by machine learning methods," *J. Eng. Sci. Technol. Rev.*, vol. 13, no. 2, pp. 154–166, Apr. 2020.

[41] N. A. S. Abdullah and N. I. A. Rusli, "Multilingual sentiment analysis: A systematic literature review," *Pertanika J. Sci. Technol.*, vol. 29, no. 1, pp. 1–26, 2021.

[42] Y. Xu, H. Cao, W. Du, and W. Wang, "A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations," *Data Sci. Eng.*, vol. 7, pp. 1–21, Jun. 2022.

[43] D. Moher, A. Liberati, J. M. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Int. J. Surg.*, vol. 8, no. 5, pp. 336–341, 2010.

[44] A. Qazi, N. Hasan, O. Abayomi-Alli, G. Hardaker, R. Scherer, Y. Sarker, S. K. Paul, and J. Z. Maitama, "Gender differences in information and communication technology use & skills: A systematic review and meta-analysis," *Educ. Inf. Technol.*, vol. 27, no. 3, pp. 4225–4258, Apr. 2022.

[45] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," *Appl. Comput. Intell. Soft Comput.*, vol. 2020, pp. 1–21, Jan. 2020.

[46] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decis. Anal. J.*, vol. 3, Jun. 2022, Art. no. 100073.

[47] X. Wan, "Co-training for cross-lingual sentiment classification," in *Proc. 4th Int. Joint Conf. Natural Lang. Process.*, vol. 1, Aug. 2009, pp. 235–243.

[48] M. E. Basiri, A. R. N. Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in Persian," *Open Trans. Inf. Process.*, vol. 1, no. 3, pp. 1–14, 2014.

[49] F. Amiri, S. Scerri, and M. H. Khodashahi, "Lexicon-based sentiment analysis for Persian text," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2015, pp. 9–16.

[50] F. A. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," 2011, *arXiv:1103.2903*.

[51] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. Conf. Weblogs Social Media*, 2015, pp. 218–225.

[52] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter events," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 406–418, Feb. 2011.

[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[54] J. Angel, S. T. Aroyehun, A. Tamayo, and A. Gelbukh, "NLP-CIC at SemEval-2020 task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier," in *Proc. 14th Workshop Semantic Eval.*, Barcelona, Spain, Dec. 2020, pp. 957–962.

[55] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Syst. Appl.*, vol. 77, pp. 236–246, Jul. 2017.

[56] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proc. COLING*, 2010, pp. 1–15.

[57] M. Saad, D. Langlois, and K. Smaïli, "Building and modelling multilingual subjective corpora," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, Reykjavik, Iceland, May 2014, pp. 1–7.

[58] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," 2004, *arXiv:0409058*.

[59] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, Prague, Czech Republic, Jun. 2007, pp. 976–983.

[60] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Honolulu, HI, USA, 2008, pp. 127–135.

[61] C. Banea, R. Mihalcea, and J. Wiebe, "Sense-level subjectivity in a multilingual setting," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 7–19, 2014.

[62] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "Supervised sentiment analysis in multilingual environments," *Inf. Process. Manage.*, vol. 53, no. 3, pp. 595–607, May 2017.

[63] G. Shalunts, G. Backfried, and N. Commeignes, "The impact of machine translation on sentiment analysis," in *Proc. 5th Int. Conf. Data Anal.*, 2016, pp. 51–56.

[64] M. S. Hajmohammadi, R. Ibrahim, and A. Selamat, "Bi-view semi-supervised active learning for cross-lingual sentiment classification," *Inf. Process. Manage.*, vol. 50, no. 5, pp. 718–732, Sep. 2014.

[65] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Honolulu, HI, USA, 2008, pp. 553–561.

[66] A. Balahur and M. Turchi, "Improving sentiment analysis in Twitter using multilingual machine translated data," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2013, pp. 49–55.

[67] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto, "SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Sci.*, vol. 5, no. 1, pp. 1–29, 2016.

[68] M. Araújo, P. Gonçalves, M. Cha, and F. Benevenuto, "iFeel: A system that compares and combines sentiment analysis methods," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 75–78.

[69] M. L. D. Araujo, J. P. Diniz, L. Bastos, E. Soares, M. Júnior, M. Ferreira, F. Ribeiro, and F. Benevenuto, "iFeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis," in *Proc. 10th Int. AAAI Conf. Web Social Media*, 2016, pp. 1–2.

[70] J. Pan, G.-R. Xue, Y. Yu, and Y. Wang, "Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2011, pp. 289–300.

[71] X. Wan, "Bilingual co-training for sentiment classification of Chinese product reviews," *Comput. Linguistics*, vol. 37, no. 3, pp. 587–616, Sep. 2011.

[72] B. Wei and C. Pal, "Cross lingual adaptation: An experiment on sentiment classifications," in *Proc. ACL Conf.*, 2010, pp. 258–262.

[73] C. Tho, Y. Heryadi, L. Lukas, and A. Wibowo, "Code-mixed sentiment analysis of Indonesian language and javanese language using lexicon based approach," *J. Phys., Conf. Ser.*, vol. 1869, pp. 12–84, Apr. 2021.

[74] L. Zhang and C. Chen, "Sentiment classification with convolutional neural networks: An experimental study on a large-scale Chinese conversation corpus," in *Proc. 12th Int. Conf. Comput. Intell. Secur. (CIS)*, Dec. 2016, pp. 165–169.

[75] M. Cliche, "BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs," 2017, *arXiv:1704.06125*.

[76] H. Nguyen and M.-L. Nguyen, "A deep neural architecture for sentence-level sentiment classification in Twitter social networking," in *Proc. Int. Conf. Pacific Assoc. Comput. Linguistics*. Singapore: Springer, 2017, pp. 15–27.

[77] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526–558, 2009.

[78] X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proc. 2016 Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 247–256.

[79] L. Medrouk and A. Pappa, "Deep learning model for sentiment analysis in multi-lingual corpus," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2017, pp. 205–212.

[80] J. Wehrmann, W. Becker, H. E. L. Cagnini, and R. C. Barros, "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2384–2391.

[81] X. Dong and G. de Melo, "Cross-lingual propagation for deep sentiment analysis," in *Proc. AAAI*, 2018, pp. 1–15.

[82] L. Medrouk and A. Pappa, "Do deep networks really need complex modules for multilingual sentiment polarity detection and domain classification?" in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6.

[83] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Syst. Appl.*, vol. 69, pp. 214–224, Mar. 2017.

[84] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava, and P. Koehn, "De-mixing sentiment from code-mixed text," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2019, pp. 371–377.

[85] A. Gupta, S. Menghani, S. K. Rallabandi, and A. W. Black, "Unconscious self-training for sentiment analysis of code-linked data," 2021, *arXiv:2103.14797*.

[86] N. Sabri, A. Edalat, and B. Bahrak, "Sentiment analysis of Persian–English code-mixed texts," in *Proc. 26th Int. Comput. Conf., Comput. Soc. Iran (CSICC)*, Mar. 2021, pp. 1–4.

[87] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly, and J. P. McCrae, "DravidianCodeMix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text," *Lang. Resour. Eval.*, vol. 56, no. 3, pp. 765–806, Sep. 2022.

[88] K. S. B. S. Saikrishna and C. N. Subalalitha, "Sentiment analysis on Telugu–English code-mixed data," in *Intelligent Data Engineering and Analytics* S. C. Satapathy, P. Peer, J. Tang, V. Bhateja, and A. Ghosh, Eds. Singapore: Springer, 2022, pp. 151–163.

[89] S. S. V. Kusampudi, P. Sathineni, and R. Mamidi, "Sentiment analysis in code-mixed Telugu–English text with unsupervised data normalization," in *Proc. Conf. Recent Adv. Natural Lang. Process.*, 2021, pp. 753–760.

[90] S. Thara and P. Poornachandran, "Social media text analytics of Malayalam–English code-mixed using deep learning," *J. Big Data*, vol. 9, no. 1, pp. 1–25, Dec. 2022.

[91] C. D. Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. COLING*, 2014, pp. 69–78.

[92] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 720–728.

[93] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," 2015, *arXiv:1503.00075*.

[94] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proc. 11th Int. Workshop Semantic Eval.*, 2017, pp. 747–754.

[95] S. R. Shah and A. Kaushik, "Sentiment analysis on Indian indigenous languages: A review on multilingual opinion mining," 2019, *arXiv:1911.12848*.

[96] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1–20.

[97] T. Tomihira, A. Otsuka, A. Yamashita, and T. Satoh, "Multilingual emoji prediction using BERT for sentiment analysis," *Int. J. Web Inf. Syst.*, vol. 16, no. 3, pp. 265–280, Sep. 2020.

[98] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[99] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae, "Corpus creation for sentiment analysis in code-mixed Tamil–English text," in *Proc. 1st Joint Workshop Spoken Lang. Technol. Under-Resourced Lang. (SLTU) Collaboration Comput. Under-Resourced Lang. (CCURL)*, Marseille, France, May 2020, pp. 202–210.

[100] J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, "Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning," 2022, *arXiv:2204.06487*.

[101] I. Mozetič, M. Grčar, and J. Smailovič, "Multilingual Twitter sentiment classification: The role of human annotators," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0155036.

[102] J. Wehrmann, W. E. Becker, and R. C. Barros, "A multi-task neural network for multilingual sentiment classification and language detection on Twitter," in *Proc. 33rd Annu. ACM Symp. Appl. Comput.*, Apr. 2018, pp. 1805–1812.

[103] X. Dong and G. de Melo, "A robust self-learning framework for cross-lingual text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6306–6310.

[104] J. Kranjc, J. Smailović, V. Podpečan, M. Grčar, M. Žnidaršič, and N. Lavrač, "Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform," *Inf. Process. Manage.*, vol. 51, no. 2, pp. 187–203, Mar. 2015.

[105] A. Balahur and J. M. Perea-Ortega, "Sentiment analysis system adaptation for multilingual processing: The case of tweets," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 547–556, Jul. 2015.

[106] G. Liu, X. Huang, X. Liu, and A. Yang, "A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network," *Comput. J.*, vol. 63, no. 3, pp. 410–424, Mar. 2020.

[107] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, and A. Hussain, "A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks," *Neurocomputing*, vol. 380, pp. 1–10, Mar. 2020.

[108] K. Rakshitha, M. Pavithra, and M. Hegde, "Sentimental analysis of Indian regional languages on social media," *Global Transitions Proc.*, vol. 2, no. 2, pp. 414–420, Nov. 2021.

[109] S. S. Mukku and R. Mamidi, "ACTSA: Annotated corpus for Telugu sentiment analysis," in *Proc. 1st Workshop Building Linguistically Generalizable NLP Syst.*, Copenhagen, Denmark, 2017, pp. 54–58.

[110] M. E. Basiri and A. Kabiri, "Sentence-level sentiment analysis in Persian," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 84–89.

[111] B. Roshanfekr, S. Khadivi, and M. Rahmati, "Sentiment analysis using deep learning on Persian texts," in *Proc. Iranian Conf. Electr. Eng. (ICEE)*, May 2017, pp. 1503–1508.

[112] R. Ghasemi, S. A. Ashrafi Asli, and S. Momtazi, "Deep Persian sentiment analysis: Cross-lingual training for low-resource languages," *J. Inf. Sci.*, vol. 48, no. 4, pp. 449–462, Aug. 2022.

[113] S. Sazzed, "Improving sentiment classification in low-resource Bengali language utilizing cross-lingual self-supervised learning," in *Proc. 26th Int. Conf. Appl. Natural Lang. Inf. Syst.*, Saarbrücken, Germany, Jun. 2021, pp. 218–230.

[114] G. L. Martin, M. E. Mswahili, and Y.-S. Jeong, "Sentiment classification in Swahili language using multilingual BERT," 2021, *arXiv:2104.09006*.

[115] V. Marivate, T. Sefara, V. Chabalala, K. Makhaya, T. B. Mokgonyane, R. Mokoena, and A. Modupe, "Low resource language dataset creation, curation and classification: Setswana and Sepedi—Extended abstract," 2020, *arXiv:2004.13842*.

[116] K. Ogueji, Y. Zhu, and J. Lin, "Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages," in *Proc. 1st Workshop Multilingual Represent. Learn.*, Punta Cana, Dominican Republic, Nov. 2021, pp. 116–126.

[117] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[118] H. Xu, B. Van Durme, and K. Murray, "BERT, mBERT, or BiBERT? A study on contextualized embeddings for neural machine translation," 2021, *arXiv:2109.04588*.

[119] L. Ein-Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim, "Active learning for BERT: An empirical study," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7949–7962.



KOENA RONNY MABOKELA received the bachelor's degree in computer science and mathematics and the M.Sc. degree in computer science from the University of Limpopo. He is currently pursuing the Ph.D. degree with the University of the Witwatersrand. He is currently the Head of the Technopreneurship Centre, School of Consumer Intelligence and Information Systems. He is also a Lecturer with the Department of Applied Information Systems, University of Johannesburg. Prior to joining the University of Johannesburg, in 2019, he worked for over five years in the telecommunications sector, thus acquiring industry experience. He has presented his research work on numerous platforms nationally and internationally and has a keen research interest in NLP, speech technologies, and multilingual sentiment analysis for under-resourced languages, among other areas. He is a member of the South African Institute for Computer Scientists & Information Technologists and also serves on various boards.



TURGAY CELIK received the Ph.D. degree from the University of Warwick, U.K., in 2011. He is currently a Professor in digital transformation engineering with the School of Electrical and Information Engineering and the Director of the Wits Institute of Data Science, University of the Witwatersrand. He actively reviews and publishes in various international journals and conferences. He is also an Associate Editor of IEEE ACCESS, *IET Electronics Letters*, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and *Signal, Image and Video Processing* journal (Springer). His research interests include the areas of signal and image processing, computer vision, machine learning, artificial intelligence, robotics, data science, and remote sensing.



MPHO RABORIFE received the Ph.D. degree in computer science from the University of the Witwatersrand. She is currently an Associate Professor and the Deputy Director of the Institute for Intelligent Systems, University of Johannesburg. An NRF-Rated Researcher, she specializes in theoretical computer science and computational phonology. Her achievements include the L'OREAL-UNESCO sub-Saharan regional Women in Science Fellow, the M&G 200 Award and a DST Women in Science alumni. In addition, she received numerous scholarships and bursaries for her undergraduate through to her post-Ph.D. studies. She has successfully supervised master's and Ph.D. students. She has worked on numerous projects (multidisciplinary) with project partners across three continents. She also contributes to scientific citizenship by being a member of scientific bodies.