DOI: 10.15439/2024R28
ISSN 2300-5963 ACSIS, Vol. 42

A Survey on Sentiment Analysis in Tamil: Critical Analysis

S. Manoj
Alliance College of Engineering and Design
Bangalore, India
manoj632004s@gmail.com

Moumita Pal
Stanley College
Hyderabad, India
moumitafdp@gmail.com

***Abstract*—The review paper delves into the methodologies, techniques, and challenges specific to sentiment analysis and opinion extraction within the Tamil language. As the digital landscape continues to expand, the ability to comprehend sentiments and opinions expressed in Tamil across diverse online platforms has grown increasingly vital. The paper traces the evolution of sentiment analysis techniques tailored for Tamil, covering essential components such as feature extraction, lexicon creation, and the applications of various algorithms. Special attention is given to the distinct details of the Tamil language, encompassing its linguistic complexities, codeswitching, and the expression of sentiment in informal contexts. A critical analysis has been conducted to compare different models. Moreover, the review explores strategies for opinion extraction and provides insightful suggestions for potential areas for future research and development.**

***Index Terms*—Lexicon, Sentiment Analysis, Opinion Extraction, Transformers**

I. Introduction

SENTIMENT analysis, one of the most popular and significant subsets of Natural Language Processing (NLP), requires the analysis of textual data to find and classify sentiments as neutral, positive, or negative. Sentiment analysis tasks have significant implications across various domains, from market research and brand management to social and political analysis. Sentiment analysis in various languages is important as the amount of data and media in these languages continues to grow. This review paper aims to provide a comprehensive survey of methodologies, dataset preprocessing steps, challenges, and recent advancements in sentiment analysis and opinion extraction specific to the Tamil language. Tamil, being a Dravidian language, has its own distinctive syntactic and semantic complexities, posing unique challenges and opportunities in the field of NLP.

Opinion extraction in Tamil requires a nuanced understanding of subjective language and sentiment that depends on context. The review explores strategies for identifying and consolidating opinions from various of sources, including social media, online reviews, and forums, within the linguistic context of Tamil.

Furthermore, this review seeks to highlight new development and how they impact on the accuracy and precision of sentiment analysis task in Tamil. By integrating insights from existing research, this survey aims to provide a valuable resource for researchers, practitioners, and enthusiasts engaged in sentiment analysis and opinion extraction in the Tamil language. Different methodologies used across languages are compared which can serve as a guide for future developments in sentiment analysis tasks for Tamil. The analysis also explores potential areas for future advancements and aims to encourage further research and innovation in this rapidly evolving field, contributing to computational linguistics and natural language processing field of study.

II. Review of Literature

The survey discusses three fundamental sentiment analysis techniques: supervised learning, lexicon-based methods, and transformer-based approaches. Supervised learning involves training a model with labelled text data. Lexicon-based methods use a sentiment-scored word dictionary to assess text sentiment, while transformer-based approaches utilize pre-trained neural networks to understand context and deduce sentiment.

Sentiment analysis is well-established in English but poses challenges in Tamil and other resource-poor languages like Bengali, Malayalam, Kannada, and Telugu due to the scarcity of datasets and lexicons, compounded by the complexity and English code-mixing in these languages. Efforts are ongoing to enhance sentiment analysis capabilities in these languages.

The survey highlights specific studies aimed at improving sentiment analysis. In study [62], Sentiwordnet was enhanced for tweet classification by expanding the lexicon and training a classifier for polarity estimation. In study [29], domain-specific ontology was used to refine sentiment analysis, while [38] focused on lexicon expansion for Tamil through lexical similarity and rule-based analysis.

The Forum for Information Retrieval Evaluation (FIRE) hosts workshops to advance multilingual information access research, including sentiment analysis in code-mixed languages. These workshops have revealed insights into the effectiveness of various algorithms for sentiment analysis tasks. Researchers are employing diverse machine learning algorithms, such as Recurrent Neural Networks(RNN), transformer models, and genetic algorithms, for tasks like aspect-based sentiment analysis and word-level Natural language understanding, showcasing the potential of these algorithms to improve sentiment analysis system performance.

III. Dataset

The survey highlights the utilization of diverse datasets for sentiment analysis, each with unique preprocessing techniques and sources. Romanized Bangla and Bangla text samples, cited in studies [1] and [12], include 9,337 social media posts from platforms like YouTube, Twitter, and Facebook, with preprocessing that removes emoticons, hashtags, and proper nouns, and includes Part-Of-Speech(POS) tagging and manual sentiment categorization. The Bengali Cricket Commentary dataset, referenced in [2] and [25], comprises 2,489 Facebook comments, processed

©PTI 2024

by removing punctuation, digits, and stop words, and then tokenized and stemmed for a Bag of Words representation. An English corpus from Amazon Reviews, used in [12], contains over 68,356 reviews, with preprocessing that eliminates stop words and non-Bangla words, and adjusts for negation. The IMDB and Polarity Detection dataset, mentioned in [28], undergoes stop words, special characters, and URL removal, with additional lowercasing, stemming, and 10-fold cross-validation. Tamil SentiWordNet, discussed in [29] and [30], evolves from the English SentiWordNet 3.0 using various lexicons, classified into positive and negative sentiments. Code-mixed datasets for Dravidian languages, from studies [24], [60], [64], and [68], include YouTube comments processed to remove extraneous char- acters and symbols. FIRE datasets from 2020, 2021, and 2022, cited in [55], [56], [58], and [61], offer a rich source of bilingual and native texts in Malayalam-English, Tamil-English, and Kannada-English, with comprehensive preprocessing for analysis readiness. Each dataset's preprocessing is meticulously tailored to its linguistic features and format, ranging from simple cleaning to advanced tokenization and sentiment classification, providing a foundational basis for sentiment analysis research across various languages and contexts.

Table 1 Summary of Sentiment Analysis investigated on different Languages

Title	**Publication Year**	**Datasets used**	**Pre-processed steps**
Sentiment Analysis on Bangla and Romanized Text using deep recurrent model [1]	2016	Bangla and Romanized Bangla text samples (Written in English) collected from product review pages, YouTube, Twitter, Facebook, and online news portals.	Remove emoticons, hashtags, proper nouns and applied POS tagging. Manually categorized into positive, negative and ambiguous samples
A Sentiment Classification in Bengali and Machine Translated English Corpus [2]	2019	Two Bengali datasets from cricket commentary and other from Drama review (scrapped from YouTube)	English comments were removed leaving only Bengali. Bengali corpora are converted to English using google machine translation. Class balancing using SMOTE. The dataset is stemmed and tokenized, Tokenized and applied term frequeny-inverse document frequeny(tf-idf) method. Manual rating is given to each translation representing the accuracy from 1 to 5
Performing Sentiment Analysis in Bangla Microblog Posts [4]	2014	Bangla tweets using twitter API	Bangla lexicon translated to English and applied tokenization, normalization and POS tagging.
Multilingual Sentiment Analysis: An RNN-Based Frame- work for Limited Data [6]	2018	English reviews and restaurant reviews from Spanish, Dutch, Russian and Turkish	SentiWordNet lexicon used to obtain a positive and negative sentiment score and they're aggregated to classify review as positive or negative
Evaluation of Naıve Bayes and Support Vector Machines on Bangla Textual Movie Reviews [7]	2018	Texts from Bangla movie reviews sites, Facebook, websites, tweets and Movie Database (IMDb).	Punctuation characters, URLs, stop words emoticons were removed. Applied tokenization, stemming and vectorization.
fully Automatic Lexi- con Expansion for Do- main oriented Senti- ment Analysis dataset used in this paper [9]	2006	Japanese text reviews from the following domains Consumer electronics, Travel, Food, Books, Movies	Applied tokenization, stemming, stop words removal and POS tagging
Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments [11]	2018	Comments (Bangla and Romanized Bangla) obtained from YouTube video	Applied tokenization and Reduction of stop words, links, URLs, user tags and mentions from YouTube
Sentiment Mining from Bangla Data using Mutual Information [12]	2016	Product reviews collected from Amazon.	Removed stop words and Non-Bangla alphabetic words from translation. For sentences having presence of a negation word, a negation word is added before each word in the sentence
Opinion-Polarity Identification in Bengali [13]	2010	Bengali news corpus	Applied tokenization, stemming and POS tagging
Supervised Approach of Sentimentality Extraction from Bengali Facebook Status [15]	2016	User's status comments from Facebook	Manually tagging the Facebook data into positive and negative. Removed symbols like hashtags, websites URLs and applied stemming.
Opinion Mining and Analysis for Arabic Language [16]	2014	Arabic reviews and comments collected from different social media resources	Removed digits, punctuations, special symbols and non-letters. Applied normalization and tokenization

Title	Publication Year	Datasets used	Pre-processed steps
Sarcasm Detection followed by Sentiment Analysis for Bengali Language: Neural Network & Supervised Approach [17]	2023	News headlines from twitter containing English, Bengali and Romanised Bengali	Removing punctuations, non-alphabetical characters, stop words. Applied Stemming and Lemmatization, Word tokenization
Cross-Lingual Sentiment Analysis Without (Good) Translation [19]	2017	English reviews on Yelp, Chinese hotel reviews, Spanish billion-word corpus	Performed data cleaning, tokenization, stemming, normalization to remove diacritics.
Sentiment Analysis in Czech Social Media Using Supervised Machine Learning [20]	2013	Czech movie reviews and product review dataset	Used tokenization, POS tagging stemming, lemmatization and removed stop words
Aspect-Based Opinion Mining from Customer Reviews [21]	2016	Customer reviews from amazon	Removal of symbols and performed sentence splitting, lemmatization, POS tagging, dependency parsing and dependency analysis
Analysis and Tracking of Emotions in English and Bengali Texts: A Computational Approach [22]	2011	News stories and blog corpora in Bengali	Applied tokenization, stemming, lemmatization, POS tagging, stop word removal, named entity recognition
Datasets for Aspect- Based Sentiment Analysis in Bangla and Its Baseline Evaluation [23]	2018	Cricket dataset and Restaurant dataset	Punctuations, digits and stop words were removed. Performed manual data annotation and tokenization on both datasets and represented as bag of words(BOW)
Corpus creation for sentiment analysis in code mixed Tamil -English text [24]	2022	Tamil English mixed YouTube comments	If comment is fully written in Tamil or English it is discarded. Removed emoticons and applied sentence length filters (less than 5 words and more than 15-word sentences are removed) and performed data annotation
Tamil English language sentiment analysis system [25]	2016	Tamil user reviews from domains like books, DVDs and music	Opinionated words from dataset are extracted. Used google translate to turn Tamil comments to English.
Cross-Lingual Sentiment Analysis with Machine Translation [26]	2013	Turkish and English product review dataset	Google translate to translate English data to Turkish and performed manual annotation for the sentence polarity dataset.
Sentiment Analysis Using Machine Learning Techniques [28]	2017	IMDB movie review dataset and polarity dataset, tweets collected from Twitter	Removed Stop words, Numeric and special characters; Count Vectorization and tf-idf vectorization; 10-fold cross validation
Sentiment Analysis: An Approach for Analysing Tamil Movie Reviews Using Tamil Tweets [29]	2021	Tamil language movie tweets	Data cleaning – removal of retweets, URLs, special characters and applied stemming and tf-idf.
Towards Building a SentiWordNet for Tamil [30]	2016	English SentiWordNet 3.0, AFINN-111, Subjectivity Lexicon and Opinion Lexicon.	SentiWordNet and subjectivity lexicons are merged and filtered to strongly subjective data and removed duplicate words. AFINN-111 and Opinion Lexicon added to this. removed words with ambiguous sense
Rough Set Based Opinion Mining Tamil [31]	2017	Tamil product review dataset	Performed Sentence extraction, anaphora resolution. Applied tokenization, stemming, lemmatization and removal stop words.
Corpus Based Senti- ment Classification of Tamil Movie Tweets Using Syntactic Patterns [33]	2017	Tamil movie tweets	Removal of any external links, retweets, characters that repeat more than once and applied tokenization
Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study [35]	2020	Mobile phone user Tamil reviews from e-commerce websites	Performed tokenization, stemming, lemmatization, POS tagging, stop words removal.
Sentiment Extraction for Tamil Political Reviews[36]	2016	Political reviews are gathered from social media websites such as twitter and Facebook	Applied tokenization and POS tagging.

Title	Publication Year	Datasets used	Pre-processed steps
Analysing Sentiment in Tamil Tweets using Deep Neural Network [37]	2020	Tamil tweets data	Removal of symbols, special characters, dates, diacritics, punctuations and emoticons. Tweets converted into word embedding using word2vec.
Sentiment Lexicon Ex- pansion using Word2vec and fastText for Sentiment Prediction in Tamil Tweets [38]	2020	Data from movie review websites - Cineulagam, Filmibeat, Maalaimalar, Samayam, IndiaTimes, Behind-woods, as well as from Twitter, Facebook, Noolaham.com, and Ta.wikipedia.ord	removal of html tags, English words, repeated characters, symbols and emoticons. word embedding is created from this corpus using word2vec.
Sentiment Mining: An Approach for Bengali and Tamil Tweets [41]	2016	999 Bengali tweets and 1103 Tamil tweets	Performed Tokenization and texts converted lowercase. Removal of URLs, usernames, punctuations.
Sentiment Analysis of Tamil-English Codeswitched Text on Social Media Using Sub-Word Level LSTM [42]	2020	English and Tamil mixed comments from Facebook	Manual data annotation. splitting into train, validation and test set
Sentiment Analysis of Dravidian Code Mixed Data [43]	2021	Tamil and Malayalam code mixed dataset	Replace emojis with corresponding description in English. non-Tamil and non-Malayalam characters replaced with roman script representation. Applied tf-idf vectorization
Multilingual Senti- ment Analysis in Tamil, Malayalam and Kannada Code Mixed Social Media Posts using MBERT [44]	2021	Posts from YouTube of Tamil, Malayalam, Kannada code mixed languages	Removal of symbols, special characters, hashtags, punctuations, URLs, emojis and numerals. Texts converted to lowercase
Sentiment Analysis on Tamil Code Mixed Text using Bi LSTM [45]	2021	Tamil code mixed data from FIRE 2021	Removal of emojis, special characters, non- ASCII characters. Texts conversion to lowercase. Maximum size of message fixed to 30 to 70 characters. Applied Tokenization
A Study on the Perfor- mance of Supervised Algorithms for Classifi- cation in Sentiment Analysis [48]	2019	Twitter review data and US airline dataset	Conversion of texts to lowercase. Applied Tokenization, stemming, removal of stop words. filter tokens that exceed length of 15 and below 3.
Unsupervised Self Training for Sentiment Analysis of Code-Switched Data [50]	2021	Data set of four different languages - Hinglish(tweets), Spanglish (tweets), Tanglish (YouTube comments) and Malayalam – English (YouTube comments)	Removed URLs, special characters and use data embedding
Transformer based Sentiment Analysis in Dravidian Languages [52]	2021	FIRE-2021 dataset	Removed emojis and punctuation. Applied tokenization. All sequences are padded with same length.
Analyzing Sentiment in Indian Language Micro Text Using Re-current Neural Network [53]	2016	Twitter data in three languages—Tamil, Hindi, and Bengali, given by SAIL in 2015.	Tweet id is removed. labels of the tweet are merged along with the tweet
Sentiment Analysis in Tamil Texts: A study on Machine Learning Techniques and Feature Representation [54]	2019	Tamil texts from twitter, YouTube, Facebook on topics such as movie, product, news, sports and tv shows	Removal of URLs, hashtags and non-Tamil words. Applied Tokenization and vectorization using fastText, tf-idf and BOW.
Sentiment Analysis and Homophobia de- tection of YouTube comments in Code-Mixed Dravidian Lan- guages using machine learning and Trans- former models [55]	2022	FIRE 2022 dataset	Executed tokenization and data cleaning. Removed URLs, numerals, and tags. Data embed- ded using tf-idf, count vectorizer, and XLM, MPNet, BERT.
An ensemble-based model for sentiment analysis of Dravidian code-mixed social media posts [56]	2021	FIRE 2021 dataset	Extracting tf-idf features using 1–6-gram characters.
Sentiment Analysis in Tamil Language Using Hybrid Deep Learning Approach [57]	2022	Ratings and reviews of Tamil movies from Kaggle.	Performed data cleaning, removed stop words, punctuations, and special characters. Transform the given data(multiclass) into binary class data (into positive and negative)

Title	Publication Year	Datasets used	Pre-processed steps
Sentiment Classifica- tion of Code-Mixed Tweets using Bi-Directional RNN and Language Tags [58]	2021	English-Tamil code-mixed data from FIRE 2020	Eliminating references, Removing the punctua- tions, taking off URLs, removing excess white space, extracting words from hashtags and applied data embedding using fastText.
Sentiment Analysis on Code-Switched Dra- vidian Languages with Kernel Based Extreme Learning Machines [60]	2022	YouTube comments using three codemixed datasets	Removing stop words and emoticons, lemmatizing. the pre written labels for the data is altered. word embedding applied using fastText.
Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages [61]	2021	FIRE2021 dataset	Removal of special Characters, emojis, URLs, and hashtags. Applied tokenization, stop word removal, word embedding and post padding
Hateful Sentiment De- tection in Real-Time Tweets: An LSTM-Based Comparative Approach [65]	2023	Scraped twitter comments	Remove symbol, punctuations. Text conversion to lowercase. Applied tokenization and tf- idf vectorization
PANDAS@TamilNLPAC L2022: Abusive Com- ment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE [66]	2022	Tamil English YouTube comments	Performed text normalisation, removal of punctuations, extra unwanted characters and stop words. Feature extraction using tf-idf and LaBSE.
A Computational Approach to the Analysis and Generation of Emotion in Text [69]	2011	The corpus contains 815,494 blog posts from LiveJournal	Feature Extraction using Bag of words and sentiment orientation

IV. Methodology

A. Supervised algorithms

Supervised algorithms have been majorly used in sentiment classification tasks. Support vector machine (SVM) is opted in by most researchers as shown in Table 2, most of the time, texts are linearly separable which allows for faster processing and fewer parameters to optimize. SVM views the problem as a pattern-matching task that involves learning symbolic patterns that depend on a phrase's lexical and syntactic semantics [13]. SVM with Sequential minimal optimization (SMO) used in [18], [26] and [5] is known for its scalability that is to perform consistently in large datasets. SVM SMO system has been used to develop aspect-based, Word-level and documents-level sentiment analysis systems.

Ensemble models used in the survey are listed in Table 3. Decision tree classifier provides a hierarchical decomposition of the data space in which a condition on the attribute value is used to divide the data [10].

The research that has used naïve bayes model is listed and analyzed in Table 4, the model computes the posterior probability of a class, depending upon the distribution of the words in the dataset. The model works with the BOWs feature extraction which ignores the position of the word in the document [10]. Multinomial naive bayes [12], MNB uses multinomial distribution for all pairs where it uses the word counts and rectify the underlying calculations to act within.

[4], [10], [20], [26], [28] and [34] have used max entropy classifier, The Maximum entropy Classifier transforms labelled feature sets into vectors. Now, by combining the weights that are computed for each feature, the most likely label for a feature set can be found. [10].

[7], [23] and [25] have used K-Nearest Neighbor (KNN) classifier. KNN performs classification by finding k nearest (in Euclidean distance) data objects repetitively with trial and error in classifying, until it the data points are finally classified, and majority vote determines final classification. [3], [19], [24], [43] and [51] used logistic regression, it works by fitting a sigmoid function to the training data that outputs 0 and 1. It is popular for its simplicity and easy interpretation, and it generally achieves good accuracy in various datasets. It is relatively light weight and can be deployed with minimum resources.

[28] Linear Discriminant Analysis(LDA), using a discriminant analysis technique, LDA classifies reviews by representing the dependent variables as a linear combination of the independent variables. This approach focuses on creating a linear combination of the dependent variable based on the independent variables. After that, these linear equations will be processed to produce the necessary categorization outcome [28].

B. Genetic Algorithm Based model

In Genetic Algorithms (GA), the text reviews are classified after being depicted as chromosomes. While applying Neuro GA, GA is employed to select the best features from a large pool of features. Subsequently, neural network is used to classify the reviews based on the selected features. The papers that have opted for GA based model are listed in Table 5.

C. *Fuzzy classifier*

[8], [35] and [40] used fuzzy classifier. Sentiment polarity is vague about its conceptual reach. There is not a clear boundary between the concepts of “+ve”, “-ve” and “neutral”. To better handle such fuzziness in sentiment polarity, fuzzy set classifiers are used.

D. *Roughset based classifier*

Rough set theory-based classification, the fundamental aim of rough set analysis is to derive upper and lower approximations from the available data. This theory assigns a level of affiliation to each object, with a central focus on resolving ambiguity that stems from distinguishing objects within a specific domain [31]. The analysis and comparison of rough set-based classifiers identified in the survey are presented in Table 6.

E. *Lexicon mased models*

The lexicon-based model identified in the survey is presented in Table 7. The Lexicon based approach uses predefined dictionaries and assigned manually annotated sentiment scores. The sentiment is founded from aggregating sentiment scores of all words in each data and determines overall sentiment of the data. Lexicon based approach provides rich linguistic information which helps in improving accuracy and requires less processing time compared to supervised learning method.

F. *Recurrent Neural Network models*

Long short-term memory (LSTM) is an extension of simple RNN which reduces vanishing gradient problem and can remember dependencies over a large gap [1]. A bidirectional long short-term memory (BLSTM) processes the input sequence in both forward and backward directions, with both directions feeding into the same output layer. So, one-layer processes the input in one direction while the other LSTM layer processes the sequence in the opposite direction. [43] used sub word level LSTM model as it accounts for words that have a similar morpheme. For example, in the Tamil dataset, ‘aval’, ‘avanga’ and ‘avala’ have similar meanings

Table 2 Critical analysis on SVM based model

Title	Model	Result	Critical Analysis
Performing Sentiment Analysis in Bangla Microblog Posts [4]	SVM and Maximum Entropy classifiers	SVM scores highest with accuracy of 93%	SVM has been majorly preferred for sentiment analysis tasks and it mostly scores better accuracy than other supervised algorithms. These experiments that have primarily used the SVM algorithm shows that results of the Tamil dataset didn’t reach high accuracy as much as other languages and had scored poorly in code mixed data[24]. In other cases, SVM has performed very well, and it has the worked best with n-gram features.
Evaluation of Naive Bayes and Support Vector Machines on Bangla Textual Movie Reviews [7]	Naive Bayes (NB) and Support Vector Machines (SVM)	SVM performed slightly better than NB with precision of 0.86.	
Aspect Level Opinion Mining on Customer Reviews using Support Vector Machine [14]	SVM	precision is calculated as 83.34%, recall is calculated as 92.87% and F-measure is calculated as 87.34%.	
Comparative experiments using super- vised learning and machine translation for multilingual sentiment analysis [18]	Translate using translators of google, Moses and Bing to German, French and Spanish and trained using SVM SMO classifier, naive bayes and Random Forest(RF).	The better results were obtained with the SVM SMO classifier in most languages	
Datasets for Aspect- Based Sentiment Analysis in Bangla and Its Baseline Eval- uation [23]	SVM, KNN, Random Forest	SVM obtained the highest precision rate for both of the datasets 0.71 and 0.77 for cricket and restaurant dataset respectively.	
Sentiment Analysis Using Machine Learning Techniques [28]	Naive Bayes, Support Vector Machine, Maximum Entropy (ME), and Stochastic Gradient Descent (SGD)	SVM showed highest accuracy of 88.94% with unigram + bigram + trigram features	
Multilingual Sentiment Analysis using Machine Translation [5]	Classifiers used are SVM SMO, adaboost and bagging classifier on each language with unigram and bigram features	A low quality of the translation led to features extracted being not informative enough thus performing with less accuracy of average accuracy 0.564.	

due to their root word 'aval'. Further evaluation and comparison of the RNN model are provided in Table 8.

G. *Transformer models*

A Transformer based models are state of art models in sentiment analysis tasks were BERT, known for its deep contextual representation, can be expanded by incorporating a classification head to refine the model for downstream NLP tasks. it's primarily used in [24], [46] and [52]. RoBERTA was trained using an approach called Masked Language Modelling (MLM). Through this method, the model can grasp the connections between words in a sentence and comprehend the contextual meaning of the text. [32], [47], [49], [50] and [52] have used XLM ROBERTa, an unsupervised cross- lingual representation approach, was trained on Wikipedia data of 100 languages and fine-tuned on different downstream tasks for evaluation and inference. This involves samples from text sources in a variety of languages extracted, and the model is then trained to predict masked tokens in the input. [44], [49] have used multilingual BERT (MBERT), it has been pretrained in 104 languages with largest Wikipediaes.

Distil BERT is used in [46], [49], [52], which is 60% faster than BERT and includes 40% less parameter. It employs a triple loss language modelling approach, integrating cosine distance loss with the process of knowledge distillation. When compared to MLM loss, the two distillation losses in the triple loss exert a substantial influence on the model's performance [52]. In [49] and [46] character BERT have been used, it reduces the complexity and removes word piece tokenization entirely and instead employs a Character-CNN to represent entire words at the character level rather than at a sub-word level [46]. MuRIL stands as an Indic language model, having undergone extensive training and enhancements to excel in Indian languages. It provides support English and 16 other Indian languages. MuRIL surpassed multilingual BERT on all benchmark data sets of Indic languages in [52]. The papers that have adopted transformer-based models are listed and examined in Table 9.

TABLE 3 CRITICAL ANALYSIS ON ENSEMBLE MODEL

Title	Model	Result	Critical Analysis
Sentiment Analysis Using Machine Learning Techniques [28]	Naive bayes, SVM, random forest and Linear Discriminant Analysis(LDA)	random forest scored highest accuracy in both datasets of 88.88% in IMDB dataset and 95% in polarity dataset	The performance of the tree based model highly varies depending on the features used for the classification task and has surpassed the popular SVM in some cases. Ensemble models have been fairly well tested in Tamil and have shown promising results. The model can be used as a baseline for future research in sentiment analysis tasks
Sentiment Mining an Approach for Bengali and Tamil Tweets [41]	Features extracted: tf-idf, score of unigrams and bigram, tweets specific features - emoticons, hashtags. classifiers used are naive bayes and decision tree classifier	Bengali Tweets: Naive Bayes: (+ve = 0.52, -ve = 0.76, neutral = 0.79); Decision Tree: (+ve = 0.52, -ve = 0.88, neutral = 0.81) Tamil Tweets: Naive Bayes: (+ve = 0.51, -ve = 0.78, neutral = 0.73) Decision Tree: (+ve = 0.50, -ve = 0.82, neutral= 0.77)	
Sentiment Analysis andHomophobia detection of YouTube comments in Code-Mixed Dravidian Languages using machine learning and Transformer models [55]	SVM, Multilayer Perceptron (MLP), random forest classifier, Ada boost, Gradient Boosting, and Extratrees classifiers have been used.	For SA task in Tamil- English Count Vectorizer with the Random Forest model fetched the best F1-score of 0.61.	
Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study [35]	Decision Tree, naïve bayes, NBTree, Rough set, Fuzzy rough set, SVM, Fuzzy SVM, Rough Fuzzy SVM, bagging (random forest), stacking (LDA, KNN, SVM), stacking (C5.0, CART(Classification and Regression Tree), RF)	Bagging and stacking algorithms show accuracy of 91%, Rough Fuzzy SVM show 87% and Decision tree shows 81% in 5 class analysis	
sentiment classification of online consumer reviews using word vector representations [3]	classifiers used are SVM, naive bayes, logistic regression (LR), random forest	random Forest outperforms all the algorithms when used with word2vec representations with 90.21% accuracy	
Corpus creation for sentiment analysis in code mixed tamil-english text [24]	algorithms used for classifying polarities are LR, naïve bayes, decision tree, random forest, SVM, dynamic meta embedding, conv-LSTM and BERT	LR, Random Forest and decision tree performed fairly better with bothscoring the same f-score of 0.68	

TABLE 4 CRITICAL ANALYSIS ON BAYES MODEL

Title	Model	Result	Critical analysis
Cross-Lingual Sentiment Analysis with Machine Translation [26]	Classifiers used were SVM SMO and naive bayes classifier with n-gram features	Naive bayes shows higher accuracy with unigram, bigram features of 75.57%	Naive Bayes algorithms have achieved satisfactory results for opinion extraction across various languages. However, they haven't outclassed SVMs in this domain. Both methods exhibit comparable performance levels. Notably, the observed accuracy for Tamil falls short of that achieved in languages like Bengali and English. This discrepancy might cause from limited resources for Tamil language processing.
Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms [34]	Classifiers used were SVM, naive bayes, Maxent Classifier, decision tree. features: The punctuations and apostrophe, TamilSentiwordnet	SVM gives an accuracy of 75.9% performing better than other methods	
Sentiment Mining from Bangla Data using Mutual Information [12]	Multinomial Naive Bayes (MNB)	For English, using testing data 85.1% accuracy without using negation and 85.8% accuracy with negation. For Bangla dataset, using testing data 84.78% accuracy without using negation and 83.77% accuracy with negation	
Sentiment Mining: An Approach for Bengali and Tamil Tweets [41]	Naive bayes and decision tree classifier	Bengali Tweets: Naive Bayes: (+ve = 0.52, -ve = 0.76, neutral = 0.79) Decision Tree: (+ve = 0.52, -ve = 0.88, neutral= 0.81) Tamil Tweets: Naive Bayes: (+ve = 0.51, -ve = 0.78, neutral= 0.73) Decision Tree: (+ve = 0.50, -ve = 0.82, neutral= 0.77)	
Machine Learning Technique to Detect and Classify Mental Illness on social media Using Lexicon- BasedRecommender System [63]	SVM and naive bayes	Results show that SVM model could better classify the genre of film with 65.73% accuracy	

TABLE 5 CRITICAL ANALYSIS ON GA BASED MODEL

Title	Model	Result	Critical analysis
Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada [47]	Tamil and Malayalam dataset used genetic algorithm technique. Kannada dataset used ensemble of mBERT and XLMRoBERTa models	obtained F1-score of Malayalam dataset is 0.97 obtained F1-score of Tamil dataset is 0.78 obtained F1-score of Kannada dataset is 0.75	The GA-based approach has not performed quite well for Tamil-English mixed data compared to English, Bengali, and Malayalam. The feature selection for this classification task has been a crucial aspect that impacts the results.A notable difference of the GA based classifier from other statistical systems is its ability to encode a whole sentence in GA and use it as a feature whereas in most classifier systems, the n-gram method has been followed[8].
Opinion Extraction and Summarization from Text Docu- ments in Bengali [8]	(i) Rule based approach, CRF (conditional random field) based approach, hybrid approach and GA based technique (ii) SVM classifier with features	(i) the GA based approach gives an accuracy of 90.22 and 90.6 for English and Bengali corpus respectively (ii) The model gives an accuracy of 70.04% with all 7 features	
Sentiment Analysis Using Machine Learning Techniques [28]	Using Genetic Algorithm and classification using KNN algorithm and Neuro-Genetic Algorithm	accuracy of proposed approach GA = 0.93, NeuroGA = 0.963	

H. Clustering models

Clustering functions by grouping similar unlabeled data, eliminating the need for extracting supervised informational features. There are two types of clustering: hierarchical and partition. Models for both types of clustering have been tested in [28] and the analysis is presented in Table 10.

Table 6 critical analysis on rough set theory-based model

Title	Model	result	critical analysis
Rough Set Based Opinion Mining Tamil [31]	Rough set theory-based Classification	The algorithm achieved accuracy of 0.99, 0.80, 0.92, 0.99 and 0.93 for most positive, positive, neutral, most negative, negative respectively	Although the method yielded good results, it has fallen behind comparing with results of other models. Rough set-based feature selection offers a more computationally efficient approach to selecting feature so it can be utilised as a hybrid model alongside other algorithm
Sentiment Analysis on Tamil Reviews as Proucts in Social Media Using Machine Learning Techniques: A Novel Study [35]	Decision Tree, naïve Bayes & NBTree, Rough set, Fuzzy rough set, SVM, Fuzzy SVM, Rough Fuzzy SVM, bagging (random forest), stacking LDA, KNN, SVM), stacking (C5.0, CART, RF)	Bagging and stacking algorithms show accuracy of 91%, Rough Fuzzy SVM show 87% and Decision tree shows 81% in 5 class analysis	

Table 7 Critical analysis on Lexicon based model

Title	Model	Result	Critical analysis
fully Automatic Lexicon Expansion for Domain oriented Sentiment Analysis [9]	Lexicon based approach- sentence delimitation; proposition detection; polarity assignment	The precision of polarity assignment using the automatically acquired lexicon averaged 94%	While lexicon-based approaches offer an alternative to supervised methods, their effectiveness has been found limited in Tamil. where lexical resources are still evolving, the algorithm is fully dependent on quality and variety of words in the lexicon. Dictionary based approach takes less processing time than supervised learning approach, but their accuracy often falls short [10].
A Survey on Sentiment Analysis Algorithms for Opinion Mining [10]	(i) Supervised techniques: Decision tree classifier; SVM; rule-based classifier; naive bayes; max entropy (ii) Dictionary-based approach	Supervised techniques provide better accuracy compared to dictionary based approach	
Opinion extraction from online blogs and public reviews [27]	Lexicon repositories used are Senti- wordnet, wordnet, slang is used for polarity classification	Proposed method achieves an average accuracy of 79% on word level and 81% at sentence level	
Sentiment Analysis: An Approach for Analysing Tamil Movie R views Using Tamil Tweets [29]	Lexicon apparoach - TamilWordNet (TWN) and Tamil SentiWordNet (TSWN).	The proposed model given the best accuracy of 77.89%	

Table 8 critical analysis on RNN based model

Title	Model	Result	Critical analysis
Sentiment Analysis on Bangla and Romanized Text using deep recurrent model [1]	LSTM with word2vec	Bangla dataset attaining highest accuracy of 70%	RNN is specially used for serialized data making it ideal for sentiment analysis tasks. The RNN model, Bi-LSTM which has an advantage of understanding the meaning of the sentence in bi-directional propagation mechanism [57]. from the survey it looks that RNN model performed better than stand alone CNN model and has shown really high accuracy for Tamil than in other results [61]. The hybrid version of CNN- BiLSTM has shown leading results and finds potential for further tests [57].
Multilingual Sentiment Analysis: An RNN- Based Framework for Limited Data [6]	Lexicon based baseline model and RNN in all four datasets	Results showed that the RNN model outperforms in all four datasets	
Detecting Multilabel Sentiment and Emo- tions from Bangla YouTube Comments [11]	LSTM method, CNN method and Baseline using – SVM, NB classifiers	LSTM model performs slightly better than CNN. The highest achievable accuracy for 3 and 5 class sentiment analysis is 65.97% and 54.24%.	
Analyzing Sentiment in Tamil Tweets using Deep Neural Network [37]	Deep bi directional LSTM model with n-gram feature	The model scored an accuracy of 86.2%	
Sentiment Analysis of Online Tamil Con- tents using Recursive Neural Network Models Approach for Tamil Language [39]	RNN model using a binary tree model	The model scored an accuracy of 71.1% in long phrases, 73% in intra sentential negation and 70.8% in inter sentential Negation	

TABLE 8 CRITICAL ANALYSIS ON RNN BASED MODEL (CONTINUATION)

Title	Model	Result	Critical analysis
Sentiment Analysis of Tamil-English Code-Switched Text on Social Media Using Sub-Word Level LSTM [42]	6-layer RNN model including convolutional layer and LSTM layers	the proposed sub word level LSTM model was recorded an accuracy of 75%	
Sentiment Analysis on Tamil Code Mixed Text using Bi LSTM [45]	The datasets are Embedded using GLoVe and passed to bidirectional LSTM model	The framework gives an f1-score of 0.552	
Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach [65]	Long-short term memory (LSTM)	The accuracy score was found to be 97%	
Analyzing Senti- ment in Indian Languages Micro Text Using Recurrent Neural Network [53]	Simple RNN model	F-score obtained by the RNN model in Tamil, Hindi and Bengali are 88.23,72.01 and 65.16 respectively	
Sentiment Analysis in Tamil Language Using Hybrid Deep Learning Approach [57]	feature extraction using fastText models used to classify - CNN-LSTM, CNN-BiLSTM and CNN- BiGRU	CNN-BiLSTM has achieved the higher accuracy of 80.2% and highest f1-score of 0.64	
Sentiment Classifi- cation of Code-Mixed Tweets using Bi-Directional RNN and Language Tags [58]	Bi-Directional LSTM model	The performance of the developed algorithm, garnered precision, re- call, and F1 scores of 0.59, 0.66, and 0.58 respectively	
Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages [61]	Model-1: Convolutional network with LSTM Model-2: Bi-directional LSTM model Model-3: contains an Embedding layer, a Flatten, a hidden and, a Dense layer.	For Malayalam – English the best performance was given by Model-2 of accuracy 0.9482. For Kannada - English the best performance was given by Model-3 of accuracy 0.9896. For Tamil - English the best performance was given by Model-3 of accuracy 0.9905	

TABLE 9 CRITICAL ANALYSIS ON TRANSFORMER MODEL

Title	Model	Result	Critical Analysis
Multilingual Senti- ment Analysis in Tamil, Malayalam and Kannada Code Mixed Social Media Posts using MBERT [44]	MBERT model	the precision, recall and weighted f1 score for Tamil are 0.59, 0.60 and 0.60 respectively. The precision, recall and weighted f1 score for Kannada is 0.61, 0.61 and 0.61 respectively. The precision, recall and weighted f1 score for Malayalam are 0.72, 0.72and 0.72 respectively.	Transformers, a pretrained unsupervised approach, Numerous studies have been conducted to integrate deep learning and machine learning models to achieve optimal sentiment analysis. These models now are being widely used and models such as BERT, Distil- BERT, and fast-Text show very decent performance, yet there remains room for improvement and fine-tuning for each language. Transformer-based models have shown competitive performance compared to supervised methods[50]. Class imbalance significantly impacts the model's performance in low-support classes[52].
Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text [47]	XLM−RoBERTa	Tamil English, Malayalam-English and Kannada-English scored weighted average F1−score of 0.711, 0.804, and 0.630, respectively	
Unsupervised Self- Training for Senti- ment Analysis of Code-Switched Data [50]	RoBERTa	Hinglish showed f1 score of 0.32 and accuracy 0.36; Spanglish showed f1 score of 0.31 and accuracy 0.32; Tanglish showed f1 score of 0.15 and accuracy 0.16; Malayalam- English showed f1 score of 0.17 and accuracy 0.14	

TABLE 9 CRITICAL ANALYSIS ON TRANSFORMER MODEL (CONTINUATION)

Title	Model	Result	Critical Analysis
Transformer based Sentiment Analysis in Dravidian Languages [52]	MuRIL, vBERT, XLM-RoBERTa, DistilBERT	Using soft voting technique the average F1- Score are 0.708, 0.626, and 0.609 in Malayalam, Tamil, and Kannada respectively	
sentiment Analysis on Dravidian Code- Mixed YouTube Comments using Paraphrase XLM- RoBERTa Model [59]	XLM-RoBERTa model	the model on Tamil, Malayalam, and Kannada code-Mixed language datasets, and achieve F1-scores of 71.1, 75.3, and 62.5 respectively.	
Overview of Abusive Comment Detection in Tamil - ACL 2022 [64]	ML algorithms - Logistic Regression, Linear Support Vector Machines, Gradient Boost classifier, and KNN classifier. Deep learning algorithm - Multilayered perceptron, Vanilla LSTM, Recurrent Neural Networks (RNN) Transformers - mBERT, MuRIL BERT, XLM RoBERTa, and ULMFit.	MuRIL BERT model have shown the best performance the highest F-score of 0.41.	
NLP-CUET @DravidianLang- Tech-EACL2021: Offensive Language Detection from Multilingual Code- Mixed Text using Transformers [68]	SVM, LR, ensemble, LSTM with fastText, LSTM with word2vec and LSTM with attention, MBERT, indic-BERT, XLM-R	Transformer based models show best results for all languages: Tamil F1-score 0.76 by XLM-R, Malayalam F1-score 0.93 by XLM-R and Kannada F1-score 0.71 by M-BERT	

TABLE 10 CRITICAL ANALYSIS ON CLUSTERING MODEL

Title	Model	Result	Critical analysis
Sentiment Analysis Using Machine Learning Techniques [28]	Clustering method(unsupervised) – K means, mini batch K means, Affinity Propagation, and DBSCAN	DBSCAN performed the best with 0.95 adjusted rand index	Due to the semantic complexities involved, unsupervised methods are not widely used in sentiment analysis. For resource-poor languages like Tamil, it is early to expect unsupervised algorithms to perform well.

V. CHALLENGES IN SENTIMENT ANALYSIS

Sentiment analysis faces several challenges related to dataset and model development. The breakdown of some key challenges is the following

A. Dataset Oriented Challenges

Sentiment analysis models rely heavily on training data. However, language evolves, and the sentiment associated with words can shift over time. Training data from a specific period might not accurately reflect current sentiment. This can lead to models misinterpreting the emotions expressed in newer data.

The meaning of a word can be highly contextual and domain dependent. A model trained on general language data might struggle with domain-specific sentiment. For example, the word ‘rock’ in context of mu- sic refers to a genre but when used in casual sentence like ‘she rocks’, the word means to be amazing or great.

Sentiment analysis models need to effectively handle negation (e.g., "not," "no"). Negation can completely flip the sentiment of a sentence. Models require careful feature selection and processing techniques to accurately identify negation and its impact on sentiment.

Sentiment analysis across languages presents unique challenges. Sarcasm, humour, and even positive/negative words can vary significantly due to cultural differences. Furthermore, languages like Tamil, which often mix with English, require understanding the lexical nuances of both languages for accurate sentiment analysis. This increases computational complexity and can lead to ambiguity in some cases.

Sentiment can be conveyed through non-linguistic cues like emojis, hashtags, and capitalization. Integrating these cues into sentiment analysis models requires additional processing power and learning capacity. Models need to be efficient in handling both linguistic and non-linguistic features to provide a comprehensive sentiment analysis.

B. Algorithm Oriented Challenges

Text data needs to be converted into vectors for machine learning algorithms to process it. Choosing the right vectorization technique and feature selection methods significantly impacts the model's ability to capture sentiment-bearing information from the text.

For languages with limited sentiment analysis resources like the Tamil, translation to a well-resourced language might be necessary. Translation accuracy is crucial. Any loss of meaning or sentiment during translation can negatively impact the model's performance.

Sentiment in a sentence can be influenced by words far apart from each other. Traditional sentiment analysis models

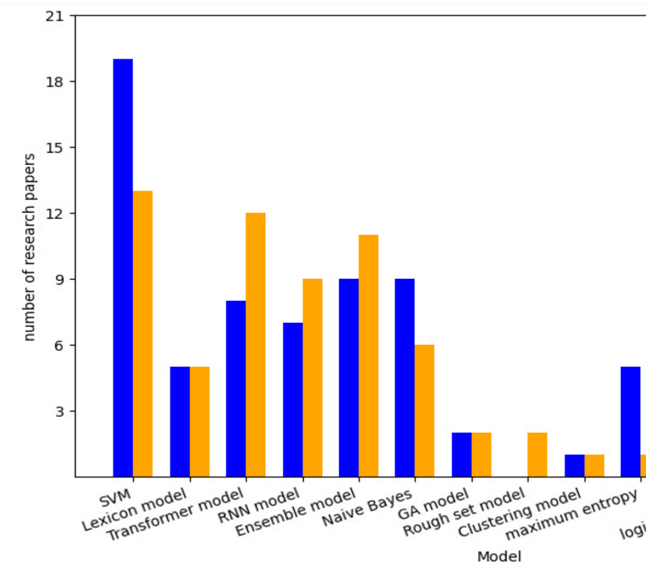

Fig. 1 comparison graph of number of papers that have used the respective model to perform SA task for Tamil and other languages

may struggle to capture these long-range dependencies. Techniques like recurrent neural net- works (RNNs) and transformers are more effective at modelling these relationships.

Sentiment data often exhibits class imbalance to a majority class outweighing the other. This imbalance can bias the model towards the majority class, leading to inaccurate classification of neutral sentiment.

VI. Conclusion

This survey has analysed sentiment analysis research in Tamil and other languages, comparing their performance. The findings reveal that supervised methods, LSTM, and transformer-based models generally outperform other approaches as discussed in the critical analysis. However, their results in Tamil lag behind those in other languages due to the imbalanced distribution of Tamil datasets and the complexities of the language, including prevalent code-mixed data.

Algorithms like lexicon-based and clustering methods are highly dependent on corpus quality. Existing datasets for Tamil sentiment analysis are outdated and lack the sophistication required for effective benchmarking. Figure 1 highlights the areas explored in Tamil sentiment analysis so far.

While transformer models have shown moderate performance in Tamil [50], they remain state-of-the-art and, hold potential for further exploration. Research has primarily focused on a few popular datasets, so future studies should venture into less explored areas. Hybrid models combining transformers with algorithms like GA could improve accuracy.

Improvements in Tamil lexicons, such as adding more words and focusing on adverbs, could enhance lexicon-based methods. Further research on linguistic models is crucial to better capture contextual and domain-specific nuances, paving the way for advancements in Tamil sentiment analysis.

References

[1] Hassan, Asif, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Moham- med. ”Sentiment analysis on bangla and romanized bangla text using deep recurrent models.” In 2016 International Workshop on Computational Intelligence (IWCI), pp. 51- 56. IEEE, 2016.

[2] Sazzed, Salim, and Sampath Jayarathna. ”A sentiment classification in bengali and machine translated english corpus.” In 2019 IEEE 20th international conference on infor- mation reuse and integration for data science (IRI), pp. 107-114. IEEE, 2019.

[3] Bansal, Barkha, and Sangeet Srivastava. ”Sentiment classification of online consumer reviews using word vector representations.” Procedia computer science 132 (2018): 1147-1153.

[4] Chowdhury, Shaika, and Wasifa Chowdhury. ”Performing sentiment analysis in Bangla microblog posts.” In 2014 International Conference on Informatics, Electronics & Vision (ICIEV), pp. 1-6. IEEE, 2014.

[5] Balahur, Alexandra, and Marco Turchi. “Multilingual sentiment analysis using ma- chine translation?.” In Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis, pp. 52-60. 2012.

[6] Can, E. F., A. Ezen-Can, and F. Can. “Multilingual sentiment analysis: an RNN-based framework for limited data (2018).” arXiv preprint arXiv:1806.04511.

[7] Banik, Nayan, and Md Hasan Hafizur Rahman. “Evaluation of naive bayes and support vector machines on bangla textual movie reviews.” In 2018 international conference on Bangla speech and language processing (ICBSLP), pp. 1-6. IEEE, 2018.

[8] Das, A. M. I. T. A. V. A. “Opinion Extraction and Summarization from Text Documents in Bengali.” Kolkata, India (2011).

[9] Kanayama, Hiroshi, and Tetsuya Nasukawa. ”Fully automatic lexicon expansion for domainoriented sentiment analysis.” In Proceedings of the 2006 conference on empirical methods in natural language processing, pp. 355-363. 2006.

[10] Pradhan, Vidisha M., Jay Vala, and Prem Balani. ”A survey on sentiment analysis al- gorithms for opinion mining.” International Journal of Computer Applications 133, no. 9 (2016): 7-11.

[11] Tripto, Nafis Irtiza, and Mohammed Eunus Ali. ”Detecting multilabel sentiment and emotions from bangla youtube comments.” In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1-6. IEEE, 2018.

[12] Paul, Animesh Kumar, and Pintu Chandra Shill. ”Sentiment mining from bangla data using mutual information.” In 2016 2nd international conference on electrical, computer & telecommunication engineering (ICECTE), pp. 1-4. IEEE, 2016.

[13] Das, Amitava, and Sivaji Bandyopadhyay. ”Opinion-polarity identification in bengali.” In International conference on computer processing of oriental languages, pp. 169-182. California, USA: Chinese and Oriental Languages Computer Society, 2010.

[14] Joshi, Anju, and Anubhooti Papola. ”Aspect Level Opinion Mining on Customer Re- views using Support Vector Machine.” International

Journal of Advanced Research in Computer and Communication Engineering (2017).

[15] Islam, Md Saiful, Md Ashiqul Islam, Md Afjal Hossain, and Jagoth Jyoti Dey. ”Super- vised approach of sentimentality extraction from bengali facebook status.” In 2016 19th international conference on computer and information technology (ICCIT), pp. 383-387. IEEE, 2016.

[16] Al-Kabi, Mohammed N., Amal H. Gigieh, Izzat M. Alsmadi, Heider A. Wahsheh, and Mohamad M. Haidar. ”Opinion mining and analysis for Arabic language.” IJACSA) Inter- national Journal of Advanced Computer Science and Applications 5, no. 5 2014: 181-195.

[17] Pal, Moumita, and Rajesh Prasad. “Sarcasm Detection followed by Sentiment Analysis for Bengali Language: Neural Network & Supervised Approach.” In 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS), pp. 1-7. IEEE, 2023.

[18] Balahur, Alexandra, and Marco Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language 28, no. 1 2014.

[19] Abdalla, Mohamed, and Graeme Hirst. “Cross-lingual sentiment analysis without (good) translation.” arXiv preprint arXiv:1707.01626 (2017).

[20] Habernal, Ivan, Tom´aˇs Pt´aˇcek, and Josef Steinberger. “Sentiment analysis in czech social media using supervised machine learning.” In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 65- 74. 2013.

[21] Samha, Amani Khalaf. “Aspect-based opinion mining from customer reviews.” PhD diss., Queensland University of Technology, 2016.

[22] Das, Dipankar. “Analysis and tracking of emotions in english and bengali texts: a com- putational approach.” In Proceedings of the 20th international conference companion on World wide web, pp. 343-348. 2011.

[23] Rahman, Md Atikur, and Emon Kumar Dey. “Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation.” Data 3, no. 2 (2018): 15.

[24] Chakravarthi, Bharathi Raja, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. "Corpus creation for sentiment analysis in code-mixed Tamil-English text." arXiv preprint arXiv:2006.00206 (2020).

[25] Thilagavathi, R., and K. Krishnakumari. ”Tamil english language sentiment analysis system.” International Journal of Engineering Research & Technology (IJERT) 4, no. 16 (2016).

[26] Demirtas, Erkin. ”Cross-lingual sentiment analysis with machine translation.” (2013).

[27] Asghar, Muhammad Zubair. ”Opinion Extraction From Online Blogs And Public Re- views.” PhD diss., GOMAL UNIVERSITY DI KHAN, 2014.

[28] Tripathy, Abinash. ”Sentiment Analysis Using Machine Learning Techniques.” PhD diss., 2017.

[29] Ramanathan, Vallikannu, T. Meyyappan, and S. M. Thamarai. ”Sentiment analysis: an approach for analysing tamil movie reviews using Tamil tweets.” Recent Advances in Mathematical Research and Computer Science 3 (2021): 28-39.

[30] Kannan, Abishek, Gaurav Mohanty, and Radhika Mamidi. “Towards building a Senti- WordNet for Tamil.” In Proceedings of the 13th International Conference on Natural Lan- guage Processing, pp. 30-35. 2016.

[31] Sharmista, Ramaswami, and M. Ramaswami. “Rough set based opinion mining in Tamil.” International Journal of Engineering Research and Development (2017).

[32] Sean, Benhur. “Findings of the shared task on Emotion Analysis in Tamil.” In Proceed- ings of the Second Workshop on Speech and Language Technologies for Dravidian Lan- guages, pp. 279-285. 2022.

[33] Ravishankar, Nadana, and Shriram Raghunathan. “Corpus based sentiment classifica- tion of tamil movie tweets using syntactic patterns.” IIOAB Journal: A Journal of Multi- disciplinary Science and Technology 8, no. 2 (2017): 172-178.

[34] Se, Shriya, R. Vinayakumar, M. Anand Kumar, and K. P. Soman. “Predicting the senti- mental reviews in tamil movie using machine learning algorithms.” Indian journal of sci- ence and technology 9, no. 45 (2016): 1-5.

[35] Sharmista, A., and Dr M. Ramaswami. “Sentiment Analysis on Tamil Reviews as Prod- ucts in Social Media Using Machine Learning Techniques: A Novel Study.” Madurai Kama- raj University Madurai-625 21 (2020).

[36] Anish, D., and V. Sumathy. “Sentiment Extraction for Tamil Political reviews” (2016).

[37] Anbukkarasi, S., and S. Varadhaganapathy. ”Analyzing sentiment in Tamil tweets us- ing deep neural network.” In 2020 Fourth International Conference on Computing Meth- odologies and Communication (ICCMC), pp. 449-453. IEEE, 2020.

[38] Thavareesan, Sajeetha, and Sinnathamby Mahesan. “Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts.” In 2020 Moratuwa engineering research conference (MERCon), pp. 272-276. IEEE, 2020.

[39] Padmamala, R., and V. Prema. “Sentiment analysis of online Tamil contents using re- cursive neural network models approach for Tamil language.” In 2017 IEEE International conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM), pp. 28-31. IEEE, 2017.

[40] Mouthami, K., K. Nirmala Devi, and V. Murali Bhaskaran. “Sentiment analysis and classification based on textual reviews.” In 2013 international conference on Information communication and embedded systems (ICICES), pp. 271-276. IEEE, 2013.

[41] Prasad, Sudha Shanker, Jitendra Kumar, Dinesh Kumar Prabhakar, and Sachin Tripa- thi. ”Sentiment mining: An approach for Bengali and Tamil tweets.” In 2016 Ninth Inter- national Conference on Contemporary Computing (IC3), pp. 1-4. IEEE, 2016.

[42] Raveendirarasa, Vidyapiratha, and C. R. J. Amalraj. ”Sentiment analysis of tamil-eng- lish codeswitched text on social media using subword level lstm.” In 2020 5th Interna- tional Conference on Information Technology Research (ICITR), pp. 1-5. IEEE, 2020.

[43] Mandalam, Asrita Venkata, and Yashvardhan Sharma. ”Sentiment analysis of Dravid- ian code mixed data.” In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 46-54. 2021.

[44] Kalaivani, Adaikkan, and Durairaj Thenmozhi. ”Multilingual Sentiment Analysis in Tamil Malayalam and Kannada code-mixed social media posts using MBERT.” In FIRE (Working Notes), pp. 1020-1028. 2021.

[45] Roy, Pradeep Kumar, and Abhinav Kumar. ”Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM.” In Working Notes of FIRE 2021-Forum for Information Retrieval Eval- uation (Online). CEUR. 2021.

[46] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. ”Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text.” Language Resources and Evaluation 56, no. 3 (2022): 765-806.

[47] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Sajeetha Thavareesan, Dhivya Chin- nappa, Durairaj Thenmozhi, Elizabeth Sherly, John P. McCrae et al. ”Findings of the sen- timent analysis of dravidian languages in code-mixed text.” arXiv preprint arXiv:2111.09811 (2021).

[48] Sunitha, P. B., Shelbi Joseph, and P. V. Akhil. ”A study on the performance of super- vised algorithms for classification in sentiment analysis.” In TENCON 2019-2019 IEEE Re- gion 10 Conference (TENCON), pp. 1351-1356. IEEE, 2019.

[49] Hande, Adeep, Siddhanth U. Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. ”Benchmarking multi-task learning for sentiment analysis and offensive language identi- fication in under-resourced dravidian languages.” arXiv preprint arXiv:2108.03867 (2021).

[50] Gupta, Akshat, Sargam Menghani, Sai Krishna Rallabandi, and Alan W. Black. ”Unsu- pervised self-training for sentiment analysis of code-switched data.” arXiv preprint arXiv:2103.14797 (2021).

[51] Srinivasan, R., and C. N. Subalalitha. ”Sentimental analysis from imbalanced code- mixed data using machine learning approaches.” Distributed and Parallel Databases (2021): 1-16.

[52] Jada, Pawan Kalyan, D. Sashidhar Reddy, Konthala Yasaswini, Arunaggiri Pandian K, Prabakaran Chandran, Anbukkarasi Sampath, and Sathiyaraj Thangasamy. ”Transformer based Sentiment Analysis in Dravidian Languages.” In FIRE (Working Notes), pp. 926-938. 2021

[53] Seshadri, Shriya, Anand Kumar Madasamy, Soman Kotti Padannayil, and M. Anand Kumar. ”Analyzing sentiment in indian languages micro text using recurrent neural net- work.” IIOAB J 7 (2016): 313-318

[54] Thavareesan, Sajeetha, and Sinnathamby Mahesan. ”Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation.” In 2019 14th Conference on industrial and information systems (ICIIS), pp. 320-325. IEEE, 2019.

[55] Varsha, Josephine, B. Bharathi, and A. Meenakshi. ”Sentiment Analysis and Homo- phobia detection of YouTube comments in Code-Mixed Dravidian Languages using ma- chine learning and transformer models.” In Working Notes of FIRE 2022-Forum for Infor- mation Retrieval Evaluation (Hybrid). CEUR. 2022.

[56] Kumar, Abhinav, Sunil Saumya, and Jyoti Prakash Singh. ”An ensemble-based model for sentiment analysis of Dravidian code-mixed social media posts.” In Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR. 2021.

[57] Ramesh Babu, Suba Sri. ”Sentiment Analysis In Tamil Language Using Hybrid Deep Learning Approach.” PhD diss., Dublin, National College of Ireland, 2022.

[58] Mahata, Sainik, Dipankar Das, and Sivaji Bandyopadhyay. ”Sentiment classification of codemixed tweets using bi-directional rnn and language tags.” In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 28-35. 2021.

[59] Babu, Yandrapati Prakash, and Rajagopal Eswari. ”Sentiment Analysis on Dravidian CodeMixed YouTube Comments using Paraphrase XLM-RoBERTa Model.” Working Notes of FIRE (2021).

[60] SR, Mithun Kumar, Lov Kumar, and Aruna Malapati. “Sentiment Analysis on Code- Switched Dravidian Languages with Kernel Based Extreme Learning Machines.” In Pro- ceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pp. 184-190. 2022.

[61] Pavan Kumar, P. H. V., B. Premjith, J. P. Sanjanasri, and K. P. Soman. “Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages.” (2021).

[62] Bravo-Marquez, Felipe. “Acquiring and exploiting lexical knowledge for twitter senti- ment analysis.” PhD diss., University of Waikato, 2017.

[63] Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Ku- mar Shukla, and Stephen Jeswinde Nuagah. ”Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System.” Computational Intelligence and Neuroscience 2022 (2022).

[64] Priyadharshini, Ruba, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U. Hegde, and Prasanna Kumaresan. ”Overview of abusive comment detection in Tamil-ACL 2022.” In Proceed- ings of the Second Workshop on Speech and Language Technologies for Dravidian Lan- guages, pp. 292-298. 2022.

[65] Roy, Sanjiban Sekhar, Akash Roy, Pijush Samui, Mostafa Gandomi, and Amir H. Gan- domi. ”Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach.” IEEE Transactions on Computational Social Systems (2023).

[66] Swaminathan, Krithika, K. Divyasri, G. L. Gayathri, Thenmozhi Durairaj, and B. Bhara- thi. ”PANDAS@ Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE.” In Proceedings of the Second Workshop on Speech and Lan- guage Technologies for Dravidian Languages, pp. 112-119. 2022.

[67] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, R. L. Hariharan, John Philip McCrae, and Elizabeth Sherly. “Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada.” In Proceedings of the first workshop on speech and language technologies for Dravidian languages, pp. 133-145. 2021.

[68] Sharif, Omar, Eftekhar Hossain, and Mohammed Moshiul Hoque. “Nlp-cuet@ dravid- ianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers.” arXiv preprint arXiv:2103.00455 (2021).

[69] Keshtkar, Fazel. A computational approach to the analysis and generation of emotion in text. University of Ottawa (Canada), 2011