2024 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2024

Code-Mixed Sentiment Corpus of Customer Reviews in Malayalam and English for Enhanced Local Market Analysis in India

1st Sree S Bhagya
Research Scholar
Dept. of Computer Application
TKM College of Engineering, Kollam
APJ Abdul Kalam Technological University
Kerala,India
212011@tkmce.ac.in

2nd Hrishiraj G L
PG Scholar
Dept. of Computer Application
TKM College of Engineering,Kollam
Kerala, India
2208@tkmce.ac.in

3rd Dr.NaderaBeevi S
Professor
Dept. of Computer Application
TKM College of Engineering, Kollam
Kerala, India
nadera@tkmce.ac.in

***Abstract*—The popularity of online shopping has boosted the volume of customer reviews, which is crucial for business growth and decision-making as they influence consumers' purchasing decisions. Indians frequently engage in code-mixing, a social media practice where multilingual users publish comments that give information about consumer preferences, cultural relevance, and product or service improvement. Businesses with an Indian market emphasis should consider this factor to improve the consumer experience. Indian languages often face resource constraints, making it difficult to create precise language models and carry out operations like sentiment analysis, spam filtering, offensive text recognition, etc. Through this work, we draw the researcher's attention to the relevance of processing code-mixed customer reviews for e-commerce sites in linguistically diverse regions like India and how it helps to serve customers better, improve goods and services, and stay competitive in a globalised and multicultural market. This work aims to expand Malayalam-English (Manglish) code-mixed literature in the commercial domain by constructing a sentiment corpus featuring customer reviews in Manglish related to commercial products annotated by voluntary annotators. The performance of different machine learning (ML) and deep learning (DL) models for the sentiment identification task is then assessed using the new corpus, and a comprehensive analysis of the results is also presented.**

***Index Terms*—code mixing, customer reviews, sentiment analysis**

I. Introduction

The prominence of online word-of-mouth (WOM) marketing emphasizes the importance of mining online reviews for vendor companies' business insight operations to understand customer preferences and compete in such a more transparent era. Due to the increased popularity of Internet shopping, customers frequently post reviews of the shop/company or the item they purchased. Online retailers and service providers may frequently ask their consumers for feedback regarding their experiences with the goods or services they purchased and whether they were satisfied with developing and enhancing their company. According to survey [1], customers may be more likely to post reviews for a product or service if they had a particularly positive or negative experience with it.

Reviews have the power to influence consumers across a wide range of businesses. Still, they are especially crucial in online shopping, where consumers frequently find that reading comments and reviews about items and services is the easiest way to decide whether or not to purchase them. Meanwhile, sellers and manufacturers are also carrying out investigations of online reviews for decision-making [2]. According to surveys [3], [4], 88% of respondents looked up internet reviews of local businesses to assess their quality, and 90% of respondents said that positive online reviews influenced their purchasing decisions. Thus, it is evident that today's consumers rely heavily on internet reviews to advise them in making decisions on a wide range of goods and services [5].

As reviewers are free to share their thoughts and opinions on online WOM platforms, reviews vary in quality, style, and usefulness [6]. There may even be reviews for the same product in many languages. It is challenging to extract relevant information on product quality, customer sentiment, etc., from this enormous wealth of content. If it were to be done manually, retrieving this information and analyzing this content would be very hard. The decoding of quality aspects from internet reviews can now be automated via developments in machine learning and natural language processing [7]. For retailers and manufacturers to reduce expenses connected with productive and effective business intelligence activities for their future product designs and marketing, it has become essential to filter out valuable and quality evaluations from the abundance of online reviews [6].

The multilingual community and social media interactions typically involve code-mixing (CM) or code-switching (CS) [8]. People who are fluent in several languages have been observed switching between them to make up for the lack of expressions in one language [9], [10]. This frequent switching of languages within the same context is referred to as Code-Mixing [11], [12]. There are over 1600 languages spoken in India, all of which belong to one of the four prominent language families: the Indo-Aryan language family(Hindi, Bangla, Gu-

979-8-3503-0661-3/24/$31.00 ©2024 IEEE

jarati, Punjabi, Marathi, Konkani, Sindhi, Assamese, Maithili, and Oriya), the Dravidian language family (Kannada, Malayalam, Tamil and Telugu), the Tibeto-Burman language family (Bodo, Koch, Dhimal, Kuki, Lepcha, Burmese, Naga), and the Austro-Asiatic language family (Munda and Khasi languages) [13]. Code-mixing is a widespread practice widely used by India's multilingual population. Indians frequently use a mix of languages while posting comments on social networking sites, including English, Hindi, Malayalam, Tamil, etc. A lot of research on analyzing the code-mixed reviews in Indian languages on social media platforms like Twitter, Facebook, YouTube, etc. has been conducted in recent years. Events like FIRE, EACL, EMNLP, and others held workshops, conferences, and shared tasks on this topic, including tasks like sentiment detection, hate speech and offensive content identification, and anaphora resolution on code-mixed social media texts in Indo-Aryan and Dravidian languages. From this trend, it is evident that researchers are currently placing more importance on code-mixed text processing across various fields, including social networking, e-commerce, education, voice assistants, food delivery platforms, etc.

This work addresses the scarcity of code-mixed datasets in the commercial domain for Dravidian languages, particularly the Malayalam-English mix. This work contributes to the new standard code-mixed dataset for Malayalam-English commercial product reviews annotated for sentiment analysis. We also offer comprehensive results on widely used classification techniques.

II. Literature Review

A. Sentiment Analysis on Manglish Text

Sentiments in the digital era are people's feelings, opinions, or reviews on social media sites like Facebook, Instagram, Twitter, YouTube, and WhatsApp. Sentiment analysis (SA) is the automatic analysis of sentiments, viewpoints, or emotions, including happiness, sadness, anger, etc. It is becoming more important because these texts can be used to justify the adoption or rejection of a product [14]–[16]. Since no restrictions exist on the language, subject matter, or procedures used to express ideas, comments, or reviews on social media, users can freely express their sentiments in any language without concerns [17]. However, owing to technological restrictions, users typically use Roman script to express their thoughts or ideas in their language along with English words instead of using their native or local language script [18]. One reason is that Roman letters are readily available and may be typed immediately, unlike most Indian languages, which need a combination of keys to generate a character. These compositions have become more complex as more words from many languages are employed to describe sentiments or reviews, making their analysis more difficult. SA is challenging due to the vastness of these texts and various types of code-mixing. Researchers are still working to improve the effectiveness of sentiment analysis systems for code-mixed text to interpret better the sentiments represented in multilingual, multicultural online materials.

Datasets for sentiment analysis of social media materials in the context of Manglish text processing are still being determined. Properly annotated datasets on this field may be found in the Dravidian-CodeMix-2020 repository, as referenced in sources [1] & [2] and the dataset available in source [3]. Movie reviews are included in datasets 1 & 2, while reviews from two cooking channels are collected in dataset 3.

B. Code-Mixing in e-commerce

In the context of e-commerce, most popular commercial websites might not disclose code-mixed customer feedback to the general public. The company did not publish customer reviews in code-mixed form because of their review moderation process or website policies. Look at the guidelines for Amazon Community Participation[4], where it is stated that content produced in a combination of languages is not permitted. However, some online shopping sites like *meesho*[5] support code-mixed customer reviews, particularly the Hindi-English mix.

Businesses should give customers the best experience possible and pay attention to their needs and preferences. Companies that target the Indian market, such as Tata Tea, Milma, Nirapara, Kitchen Treasures, Aachi, Idhayam Oil, Mysore Sandal Soap from KSDL, Tourist places and hotels, etc., will primarily serve native Indians who commonly prefer to interact with a code-mixed pattern. It would be in the companies' best interests to consider this as a feature to improve the customer experience on their e-commerce platforms if a sizable section of their consumer base chooses code-mixed reviews. For example, suppose customers have expressed a preference for code-mixed reviews. In that case, businesses must adjust their policies to reflect this since these customers are important, and they cannot disregard their concerns about their services or products. Additionally, since code-mixed text is more practical, it is simpler for customers who purchase by reading customer reviews of related products or services. It's important for e-commerce websites to carefully consider the linguistic diversity of their user base and ensure that they have the necessary tools and resources for effective moderation, translation, and support to maintain a high-quality user experience. In Table I, code-mixed customer reviews from various platforms for commercial products are given.

As a business analyst, analyzing code-mixed customer reviews can have several significant benefits and implications, such as:

- Market Insights: Code-mixed reviews offer valuable insights into the preferences, opinions, and behaviours of customers comfortable with mixed languages, thereby enhancing our understanding of the diverse customer base.

[1]https://github.com/bharathichezhiyan/DravidianCodeMix-Dataset
[2]https://dravidian-codemix.github.io/2020/datasets.html
[3] https://doi.org/10.5281/zenodo.3871306
[4]https://www.amazon.in/gp/help/customer/display.html?nodeId=GLHXEX85MENUE4XF
[5]https://www.meesho.com/

TABLE I
CODE-MIXED CUSTOMER REVIEWS FOR COMMERCIAL PRODUCTS.
Language ids- Eng: English, Hin: Hindi, Mal: Malayalam, Tam: Tamil

Review	Lang.Mix	Product	Source
Ghadi bahut achchhi hai ke liye recommend hai 1 ke sath 1free hai bachcho ke liye recommend hai	Hin-Eng	Digital Smart Watch	https://www.meesho.com/square-dial-Hin-Eng&Smart&digital-smart-watches-combo-Watch&for-boys-led-lights-watch-kids-children-pack-of-2/p/3gtlm3
Iphone 13 battery pettanu drain avum .njn ipo purchaseythulu	Mal-Eng	Mobile Phone	https://www.youtube.com/watch?v=8maloCt2-0g&t=3s
fridge backla opena irukanum illana rompa heata irukkum neenka fullcovereda vanki irukinkale problem aakume	Tam-Eng	Refrigerator	https://www.youtube.com/watch?v=MOAyrJknXto

- Cultural Relevance: Code-mixed reviews help identify cultural nuances and language preferences, enabling the customization of marketing strategies, product offerings, and customer engagement to specific cultural and linguistic groups.
- Competitive Analysis: Analyzing code-mixed reviews provides valuable insights into competitors' perceptions and resonated products or services with mixed language customers, thereby guiding competitive strategies.
- SEO and Content Strategy: Including code-mixed content can improve your website's search engine optimization (SEO) for keywords and phrases specific to those languages. Understanding the language mix used by customers can help create more relevant content.
- Product and Service Improvement: Code-mixed reviews often contain candid feedback and suggestions. Analyzing such feedback can guide product development and service improvements, helping you address the unique needs of code-mixing customers.
- User Experience Enhancement: Understanding customer language preferences enhances personalized and user-friendly experiences, boosting customer satisfaction and loyalty.
- Inclusivity and Diversity: Making code-mixed reviews public showcases a commitment to linguistic diversity and inclusivity, potentially improving a brand's reputation and attracting a broader customer base.
- Research Opportunities: Code-mixed data offers valuable linguistic and sociolinguistic research opportunities, enabling an understanding of language trends and the evolving linguistic landscape in your target market.

E-commerce sites dealing with code-mixed customer reviews may need to invest in custom NLP models, collaborate with linguistic experts, and refine their processing methods to address these difficulties. Leveraging state-of-the-art NLP research and adapting it to low-resource languages can improve the accuracy of code-mixed text analysis.

III. METHODOLOGY

A. Corpus creation

The primary objective of this work is to enhance the quantity of Malayalam and English (Manglish) code-mixed literature available on e-commerce. A dataset comprising Manglish customer reviews of commercial products has been constructed to accomplish this. We employed web scraping techniques to gather Manglish customer reviews related to commercial products from YouTube unboxing videos. After the reviews were gathered, they were put for the annotation process; the sentiments of reviews in the corpus were then annotated by voluntary annotators and all the annotated reviews, ensuring high-quality inter-annotator agreement(measured through Krippendorff's alpha score), were compiled to form the standardised corpus. Subsequently, this corpus will serve as the foundation for conducting sentiment analysis tasks and evaluating and showcasing the performance of diverse machine learning (ML) and deep learning (DL) models. This initiative aims to delve deeper into the cultural context embedded within these reviews, facilitating the development of targeted marketing campaigns and deriving enhanced insights to aid in more informed decision-making processes and improved customer support strategies.

1) Data Collection: We aimed to create a code-mixed dataset for Malayalam-English commercial reviews and ensure enough data was available for research. We attempted to gather Manglish reviews from Youtube related to unboxing videos of digital gadgets like smartphones, laptops, tablets, etc. by well-known Malayalam vloggers like CallMeShazzamVINES[6], SarathSNeonTech[7], MrPerfectTech[8], etc., since most e-commercial platforms lack sufficient Manglish customer reviews for research purposes. We used the YouTube Comment-Scraper Tool to extract the comments from YouTube. Initially, we gathered 1,22,5951 reviews from different YouTube channels. A large number of the comments that we downloaded were either entirely in English, Malayalam, or mixed form. Since resources are available for monolingual text processing, we disregarded comments entirely in one language. For that, we used the langdetect library[9] to filter out non-code-mixed corpora based on language identification at the comment level. Then, we manually filtered out some comments unrelated to our research study, like comments related to video quality, a vlogger's presentation, etc. Both native and transliterated scripts of Malayalam text are included

[6]https://www.youtube.com/@CallMeShazzamVINES
[7]https://www.youtube.com/@SarathSNeonTech
[8]https://www.youtube.com/@MrPerfectTechofficial
[9]https://pypi.org/project/langdetect/

in the comments. We preprocessed the comments by removing the URLs, special symbols, etc. Emojis were kept around since they also convey a sense of emotion.

2) Annotation: We used methodology in [19] for annotation, with at least three annotators annotating each review. For our research purpose, we have defined five sentiment class labels: Positive, Negative, Neutral, Mixed Feelings & Not_relevant, and the annotation schema given to the annotators is as follows:

- Positive: There is a clear or hidden signal in the text that the reviewer is feeling well about the product i.e., happy, admiring, relaxed, and satisfied with the product.
- Negative: The reviewer expresses disappointment, dissatisfaction, or criticism regarding the product being assessed. In such reviews, the reviewer highlights the product's shortcomings, flaws, or negative aspects, often suggesting that potential consumers should reconsider their purchase or explore alternative options. E.g., unresponsive touchscreen, poor battery life etc.
- Mixed Feelings: The text contains a verbal or implicit hint that the reviewer feels optimistic and disappointed. These evaluations frequently point out the advantages and disadvantages of the product, enabling prospective customers to consider both sides before making a choice. For example, the ABC mobile phone boasts a sleek design and excellent display, but its battery life and software updates have disappointed me.
- Neutral/Unknown: Deals with cases where the sentiment is ambiguous or uncertain. i.e., it is difficult or impossible to confidently determine the sentiment expressed in a review as positive, negative, etc. For example, What is the battery life of this phone? Here, the comment addresses the product without discussing its benefits, drawbacks, customer satisfaction levels, etc.
- Not_relevant(Not_related_to_product): The comments do not contain any information or opinions regarding the commercial product. For instance, comment texts asking for likes or subscriptions, comments about video sound clarity, etc.

To find volunteers for the annotation process, we contacted PG/Research scholars, software professionals, and academicians from our networks who possess linguistic knowledge in both Malayalam and English. We created Google Forms, in which we collected the annotator's email so the annotator could annotate only once. To determine the diversity of the annotators, we gathered information such as gender, professional status, and linguistic expertise. Each Google form has been set to contain a maximum of 100 reviews. After completing the Google form, we forwarded it to three distinct annotators who consented to annotate.

3) Inter Annotator Agreement: There must be a metric to compare the annotation qualities since there is more than one annotator to label the same data collection. This drives the application of the inter-annotator agreement, which measures the quality of annotation decisions made by several annotators on the same dataset. A high annotation agreement score does not necessarily mean that the annotations are accurate, even though it does show homogeneity of agreement.

TABLE II
CORPUS STATISTICS

Corpus Details	
Number of Manglish reviews	4,158
Number of Sentences	4,423
Number of Word Tokens	62,120
Vocabulary Size	17,320
Average Sentence Length	15

TABLE III
CLASS DISTRIBUTION

Sentiment Class	Quantity
Mixed Feelings	467
Negative	652
Positive	920
Neutral	976
Not_relevant	1,143
Total	4,158

Krippendorff's alpha(α) was employed to measure the annotator agreement score in the present scenario since multiple individuals completed the annotation task, and not all sentences were annotated by the same individuals. We collected all annotated review sets that yielded an agreement score $>= 0.80$ for both the nominal and interval metrics to generate the corpus. The sentiment labels from the annotated file are decided using the majority vote method. After this, we got 4,158 reviews in code-mixed Malayalam and English, containing 62,120 word tokens (17,320 of which are unique) and 4,423 sentence tokens in the corpus. Table II provides the details of corpus statistics, the sentiment class distribution in Table III, and Fig. 1 provides sample reviews from each sentiment class in the corpus.

B. Model Evaluation: Results & Discussions

To provide a simple baseline, we applied several traditional machine learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), K-Nearest Neighbours (KNN), Decision Trees (DT) and Random Forests (RF) separately for sentiment detection on the annotated code-mixed dataset. TF-IDF, or the term frequency-inverse document frequency, is used here to extract the input features. Our approach solely uses this dataset to train the classifier models, not pre-trained embeddings. It is challenging to identify a pattern from the handcrafted characteristics alone because of the dynamic nature of the data. These features can then be used as input into algorithms like logistic regression (LR) and support vector machines (SVM). To provide solid baselines for classification tasks, we also experimented with deep learning-based models such as LSTM, GRU, BiLSTM, and BiGRU on our code-mixed data. To extract features from the code-mixed text, the DL models experimented with FastText (FT) [20] and Word2Vec (W2V)

Sentiment_Class	Review_Text
Mixed Feelings	സുഹൃർത്തെ tesla pi ഫോൺ ഇറങ്ങിയതോടെ apple i phone 14 നേക്കാൾ 10 ടൈംസ് കൂടുതൽ features ഉള്ളപ്പോൾ മറ്റുള്ള ചൈനാ ഫോണുകൾ വട്ടപ്പൂജ്യം
Negative	redmi വേണ്ടേ വേണ്ട മതി ആയി എന്റെ ഫോൺ ബോർഡ് poco x2 അടിച്ചുപോയി ബെറ്ററി camera പോയി pensive face
Neutral	ഇതിന്റെ google dailer and messaging app miui ടേതുപോലെ മാറ്റാൻ വല്ല മാർഗം ഉണ്ടോ
Not_relevant	എവിടെ backgroundil golden playbutton എവിടെ eyes അതും കൂടെ ആവുമ്പോൾ സെറ്റ് അല്ലേ
Positive	i love boxy design ഞാൻ phone വാങ്ങാൻ നോക്കിയപ്പോ opt ചെയ്തത് 12 pro തന്നെ ആയിരുന്നു proplus നേക്കാളും 12 pro ആണ് ഞാൻ prefer ചെയ്തത് വില പോലും reveal ചെയ്യുന്നതിനേക്കാളും മുന്നേ proplus ന്റെ 200 mp യും design ഉം എന്നെ അങ്ങ് ആകർഷിച്ചില്ല

Fig. 1. Snapshot of corpus reviews.

[21] embedding strategies and a new embedding scheme by combining the FT & W2V embeddings.

We evaluated our dataset based on the precision, recall, and F-score of the baselines. The train-test split used for the evaluation is 90-10%. For our test, there are 131 positive examples, 118 negative, 147 neutral, 72 mixed feelings, and 156 not_relevant examples. The classwise best outcomes in terms of accuracy & weighted average F1 scores are provided in Table IV and the source codes are supplied in [10]. The source code for this work, provided here, has the standard values of the hyper-parameters for the deep learning models that were used and adjusted for experimental evaluations.

Based on the data presented in Table IV, the SVM classifier achieved the highest score across all classes concerning weighted average f1 score (51%) and accuracy (52%). Except for MNB, all machine learning algorithms successfully classified all sentiment classes. MNB failed to identify the Mixed Feelings category. All of the deep learning models that have been examined here, however, were successful in identifying every sentiment category. The ensemble of LSTM and GRU with a combined fast text and word2vec embedding achieved the best results among DL models regarding accuracy (43%) and weighted average f1 score(39%).

IV. Conclusion

This work introduced a code-mixed dataset of the under-resourced code-mixed Malayalam-English language pair. This data set comprises customer reviews of commercial products annotated for sentiment analysis. We made an annotation scheme and obtained a high inter-annotator agreement in terms of Krippendorff α from voluntary annotators on their contributions gathered via Google Form. Using gold standard annotated data as a baseline, we produced results for each class regarding precision, recall, f1-Score, and accuracy. Researchers can use this resource to address novel and challenging issues in code-mixed research.

Data availability is the primary concern when building language models for interpreting code-mixed texts, particularly in low-resource language mixes. The corpus developed in this work has limited sentiment data, implying that a model's performance measures must be improved. It is expected that strategies like data augmentation, the use of synthetic data, data collaboration, crowd-sourced and community-involved annotations, etc., can be adopted to deal with data scarcity, which can further reduce the misclassification rate [19], [22]–[25]. The design section can also examine the ensemble and transfer learning approaches to determine their applicability and effectiveness. Our future research will also examine whether this corpus can be used to develop corpora for other Dravidian languages with limited resources and to handle multilingual societies where people speak more than one language.

[10]https://github.com/SreeBhagya-S/COM_Senti

References

[1] Crawford Michael, Khoshgoftaar Taghi M, Prusa Joseph D, Richter Aaron N and Al Najada Hamzah, "Survey of review spam detection using machine learning techniques," Journal of Big Data, vol. 2(1), 2015, pp.1–24,

[2] Peng Qingxi and Ming Zhong. "Detecting Spam Review through Sentiment Analysis." J. Softw. 9, no. 8 (2014): 2065-2072.

[3] Anderson M and Anderson M. 88% of consumers trust online reviews as much as personal recommendations. Retrieved November. 2014 Jul;18:2015.

[4] Gesenhues A. Survey: 90% of customers say buying decisions are influenced by online reviews. Marketing Land. 2013 Apr.

[5] Kaemingk Diana, "Online reviews statistics to know in 2021," Qualtrics 30 (2020): 39-58.

[6] Zhang Ying, and Zhijie Lin "Predicting the helpfulness of online product reviews: A multilingual approach." Electronic Commerce Research and Applications 27 (2018): 1-10.

[7] Zhang, Qi, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, and Xuanjing Huang. "Mining product reviews based on shallow dependency parsing." In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 726-727. 2009.

[8] Jose Navya, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae, "A survey of current datasets for code-switching research." In 2020 6th international conference on advanced computing and communication systems (ICACCS), pp. 136-141. IEEE, 2020.

[9] Tay, Mary WJ. "Code switching and code mixing as a communicative strategy in multilingual discourse." World Englishes 8, no. 3 (1989): 407-417.

[10] Nilep, Chad. ""Code switching" in sociocultural linguistics." Colorado research in linguistics (2006).

[11] Milroy, Lesley, and Li Wei, "A social network approach to code-switching: The example of a bilingual community in Britain." One speaker, two languages: Cross-disciplinary perspectives on code-switching 136157 (1995).

[12] Moodley, Visvaganthie. "Codeswitching in the multilingual English first language classroom." International journal of bilingual education and bilingualism 10, no. 6 (2007): 707-722.

[13] Bhagya, Sree S., and Beevi S. Nadera. "Analyzing Performance of Pre-trained Models in Detecting Sentiments of Code Mixed Manglish Text Reviews from Social Media." In 2022 International Conference on

.

TABLE IV
RESULTS OF SA IN TERMS OF ACCURACY, PRECISION(P), RECALL(R), AND F1-SCORES(F1).

Class Label	**LR**			**SVM**			**MNB**			
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	Support
Mixed feelings	0.60	0.08	0.15	0.48	0.18	0.26	0.00	0.00	0.00	72
Negative	0.77	0.14	0.24	0.54	0.32	0.40	0.85	0.14	0.25	118
Neutral	0.37	0.43	0.40	0.53	0.64	0.58	0.46	0.67	0.55	147
Not_Relevant	0.38	0.72	0.50	0.54	0.68	0.60	0.47	0.70	0.56	156
Positive	0.36	0.36	0.36	0.49	0.58	0.53	0.48	0.58	0.53	131
macro avg	0.50	0.35	0.33	0.52	0.48	0.48	0.45	0.42	0.38	
weighted avg	0.47	0.39	**0.36**	0.52	0.52	**0.51**	0.49	0.48	**0.43**	624
Accuracy			**0.39**			**0.52**			**0.48**	
	DT			**KNN**			**RF**			
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	Support
Mixed feelings	0.22	0.21	0.21	0.42	0.15	0.22	0.64	0.10	0.17	72
Negative	0.44	0.32	0.37	0.47	0.24	0.32	0.60	0.25	0.35	118
Neutral	0.41	0.38	0.40	0.38	0.56	0.45	0.47	0.45	0.46	147
Not_Relevant	0.46	0.56	0.51	0.45	0.51	0.48	0.44	0.72	0.55	156
Positive	0.37	0.40	0.38	0.42	0.47	0.45	0.46	0.60	0.52	131
macro avg	0.38	0.37	0.37	0.43	0.39	0.38	0.52	0.42	0.41	
weighted avg	0.40	0.40	**0.39**	0.43	0.42	**0.41**	0.50	0.47	**0.44**	624
Accuracy			**0.40**			**0.42**			**0.47**	
	LSTM (FT+W2V)			**BiLSTM (FT+W2V)**			**GRU (FT)**			
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	Support
Mixed feelings	0.30	0.10	0.15	0.27	0.06	0.09	0.31	0.06	0.09	72
Negative	0.48	0.11	0.18	0.46	0.20	0.28	0.48	0.09	0.16	118
Neutral	0.41	0.51	0.45	0.41	0.55	0.47	0.44	0.52	0.48	147
Not_Relevant	0.42	0.57	0.48	0.46	0.53	0.49	0.43	0.58	0.50	156
Positive	0.38	0.52	0.44	0.39	0.54	0.45	0.38	0.58	0.46	131
macro avg	0.40	0.36	0.34	0.40	0.38	0.36	0.41	0.37	0.34	
weighted avg	0.41	0.40	**0.37**	0.41	0.42	**0.39**	0.42	0.42	**0.37**	624
Accuracy			**0.40**			**0.42**			**0.42**	
	BiGRU (FT+W2V)			**LSTM+GRU (FT+W2V)**			**BiLSTM+GRU (FT+W2V)**			
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	Support
Mixed feelings	0.38	0.07	0.12	0.31	0.06	0.09	0.29	0.64	0.09	72
Negative	0.58	0.09	0.16	0.50	0.15	0.23	0.520	0.09	0.16	118
Neutral	0.42	0.56	0.48	0.43	0.53	0.48	0.43	0.51	0.46	147
Not_Relevant	0.44	0.53	0.48	0.44	0.57	0.50	0.43	0.58	0.50	156
Positive	0.38	0.60	0.46	0.40	0.59	0.48	0.38	0.0.66	0.46	131
macro avg	0.44	0.37	0.34	0.42	0.38	0.36	0.41	0.37	0.33	
weighted avg	0.44	0.42	**0.38**	0.43	0.43	**0.39**	0.42	0.41	**0.37**	624
Accuracy			**0.42**			**0.43**			**0.41**	

Computing, Communication, Security and Intelligent Systems (IC3SIS), pp. 1-6. IEEE, 2022.

[14] Chakravarthi, Bharathi Raja, and Vigneshwaran Muralidaran. "Findings of the shared task on hope speech detection for equality, diversity, and inclusion." In Proceedings of the first workshop on language technology for equality, diversity and inclusion, pp. 61-72. 2021.

[15] 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.

[16] Suryawanshi, Shardul, and Bharathi Raja Chakravarthi. "Findings of the shared task on Troll Meme Classification in Tamil." In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 126-132. 2021.

[17] 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).

[18] Balouchzahi, Fazlourrahman, and H. L. Shashirekha. "MUCS@ Dravidian-CodeMix-FIRE2020: SACO-SentimentsAnalysis for CodeMix Text." In FIRE (Working Notes), pp. 495-502. 2020.

[19] 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.

[20] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, pp.135-146.

[21] Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[22] Barman, Utsab. "Automatic processing of code-mixed social media content." PhD diss., Dublin City University, 2019.

[23] Sree S Bhagya, Dr.Nadera Beevi S. A Systematic Approach for Generating Manglish Code-Mixed Customer Reviews from Amazon's Monolingual English Customer Review Corpus, 16 November 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3603551/v1]

[24] Rizvi, Mohd Sanad Zaki, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. "GCM: A toolkit for generating synthetic code-mixed text." In Proceedings of the 16th Conference of

the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp. 205-211. 2021.

[25] Pratapa, Adithya, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. "Language modeling for code-mixing: The role of linguistic theory based synthetic data." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1543-1553. 2018.