5th IEEE International Conference on
Emerging Smart Computing & Informatics
(ESCI 2023)

Analysing the Sentiments of Marathi-English Code-Mixed Social Media Data using Machine Learning Techniques

Presented by

Varad Patwardhan
Gauri Takawane
Nirmayi Kelkar
Rutwik Saraf
Omkar Gaikwad
Sheetal Sonawane

Computational Linguistics Lab, SCTR's Pune Institute of Computer Technology, Pune

AGENDA

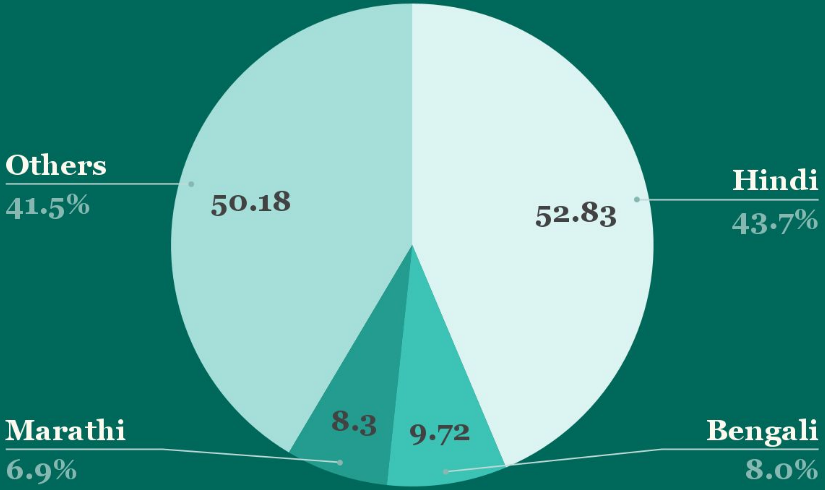
1. Abstract
2. Motivation
3. Applications
4. Literature Survey and Gap Analysis
5. Methodology
6. Explanation with example
7. Results
8. Conclusion
9. Future Scope
10. References

ABSTRACT

- Code-mixed data-Data that combines components of different languages or linguistic subgroups.
- Empirical study on the problem of word-level language identification and text normalisation for Marathi-English data.
- Models used were Naive Bayes and SVM.
- Analysis showed that SVM showed a better accuracy than Naive Bayes.

MOTIVATION

- Marathi is an Indo-Aryan language that is spoken by the Marathi people.
- It has the third largest number of speakers in India, after Hindi and Bengali.[1]
- Increase in multilingual population.
- Necessity of efficient techniques to handle this code-mixed data.
- Applications of code-mixed data:-
  - Machine Translation
  - Text Categorization
  - Speech Recognition



[1] https://censusindia.gov.in/census.website/data/census-tables

LITERATURE SURVEY

Sr No	Title & Authors	Year	Description	Future Scope
1	Multilingual Hate speech and Offensive language detection in English, Hindi, and Marathi languages[1] Kalaivani Adaikkan and Thenmozhi Durairaj	2021	Detection of hate speech and further classifying it as profanity, hate speech and offensive. Creation of machine learning, transfer learning and multilingual pre-trained models such as BERT, MBERT.	For future work, handling of sarcastic feature and imbalanced dataset may avoid misclassification and this work can be extended into other low-resource languages.
2	A Review on Sentiment Analysis for Code-Mix Chinese and English Text on Social Media[2] Ts Lim Kong Hua and Dr Lim Tong Ming	2020	Examines studies on codemix Sentiment analysis in English and Hindi to offer some insights for use in code-mix Chinese and English. Survey paper that provides insights into the research work done in this area.	Formal and informal languages used in the Internet need to be identified for more accurate sentiment analysis classification. Proper word segmentation in Chinese text can be further improved using lexicon approach or machine learning approach or combination of both.

LITERATURE SURVEY

Sr No	Title & Authors	Year	Description	Future Scope
3	An ensemble model for sentiment classification on code-mixed data in Dravidian Languages.[3]	2018	The paper focuses on ensemble of character-trigrams based LSTM model and word-ngrams based Multinomial Naive Bayes (MNB) model to identify the sentiments. LSTM performed better than MNB for sentences having long length while MNB performed better for words that are rare.	In future, these models can be tried on various language pairs but the accuracy may change according to the choice of languages as well as the preprocessing methods used.
4	Sentiment analysis of mixed code for the transliterated Hindi and Marathi texts[4] Ansari, Mohammed Arshad and Govilkar, Sharvari	2018	The author has designed a system that transliterates Hindi and Marathi text in documents using KNN, SVM, and Naïve Bayes, and ontology based classification. The highest performing algorithm was SVM. The work resulted in creation of Marathi wordnet in Python.	The key area of improvement identified, besides language identification, is in POS tagging of mixed script. This would add a very important tool in the quest of sentiment analysis of transliterated regional languages in English Script.

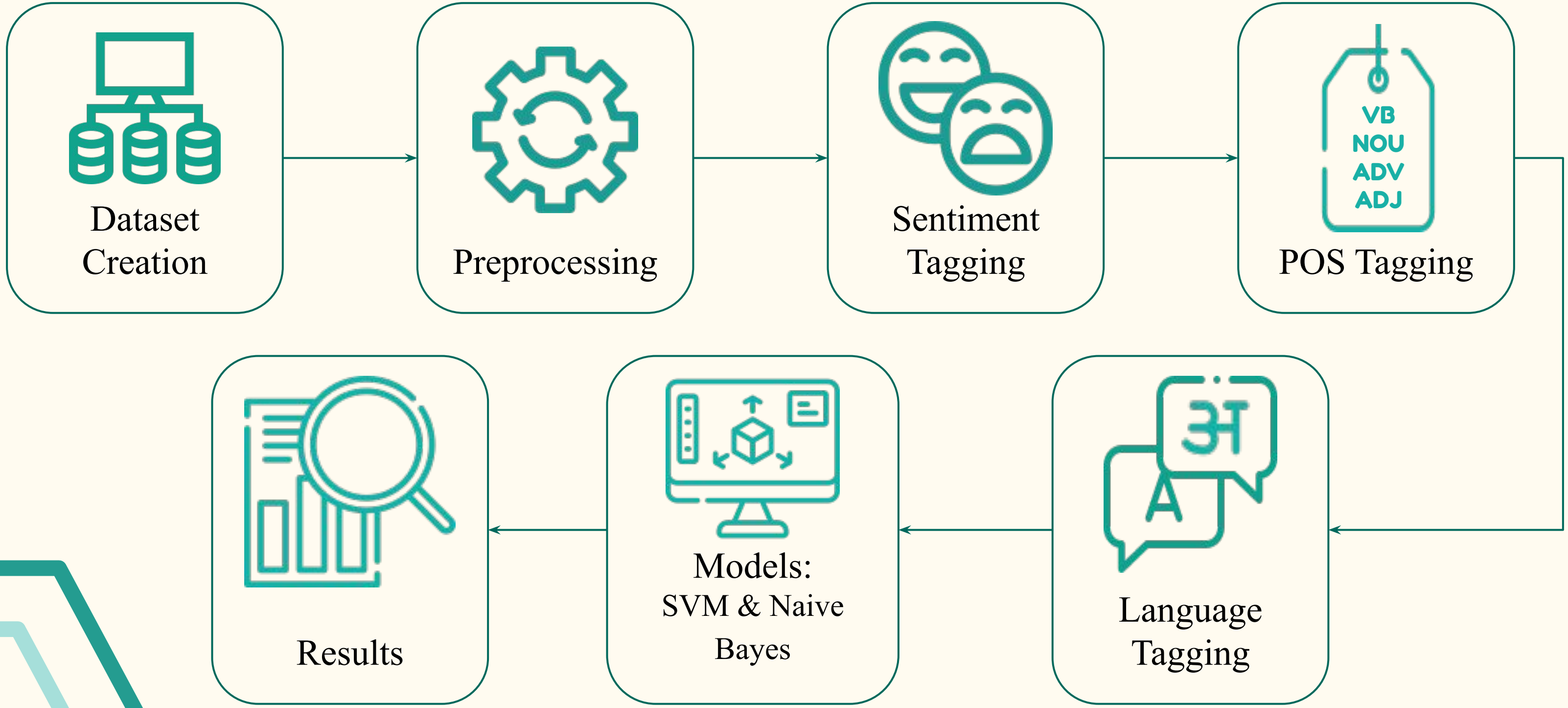
LITERATURE SURVEY

Sr No	Title & Authors	Year	Description	Future Scope
5	Towards Normalising Konkani-English Code-Mixed Social Media Text[5] Akshata Phadte, Gaurish Thakkar	2017	This research work involved presentation of a dataset and development of systems for language identification and text normalization. English Normalizer > Konkani Normalizer	In future, one can to continue creating more annotated code-mixed social media data. This dataset can be used to build tools for codemixed data like POS taggers, morph analysers, chunkers and parsers.
6	Part-of-Speech Tagger for Konkani-English Code-Mixed Social Media Text[6] Akshata Phadte, Radhiya Arsekar	2018	They have focused on building a shallow parser for Konkani-English code-mixed social data using CRF and SVM. This would help in better data-mining and sentiment analysis across the Indian subcontinent.	The creation of a parser and POS tagger for this language pair and, standard dataset of 5088 code-mixed Konkani-English sentences would help various NLP applications such as sentiment analysis.
7	A Twitter Corpus for Hindi-English Code Mixed POS Tagging[7] Kushagra Singh, Indira Sen, Ponnurangam Kumaraguru	2018	They curate a lexically unique dataset to model POS tagging as a sequence labeling task using Conditional Random Field (CRF) and LSTM Recurrent Neural Net-Works.	We note that our model suffers lower performance for POS tag categories like adjectives and adverbs.. In future, this work can used to explore building other downstream NLP tools such as Parsers or Sentiment Analyzers which make use of POS tags using our dataset.

GAP ANALYSIS

- Use of imbalanced data (sentiment tagging) reduces the accuracy of the models.
- For more accuracy, formal and informal texts should be identified independently. Our dataset consists of only general data and not domain specific.
- Many other languages such as Hindi are used in combination with English but not much work is done in Marathi-English code-mixed data.

METHODOLOGY



Dataset Creation

Preprocessing

Sentiment Tagging

POS Tagging

Language Tagging

Models: SVM & Naive Bayes

Results

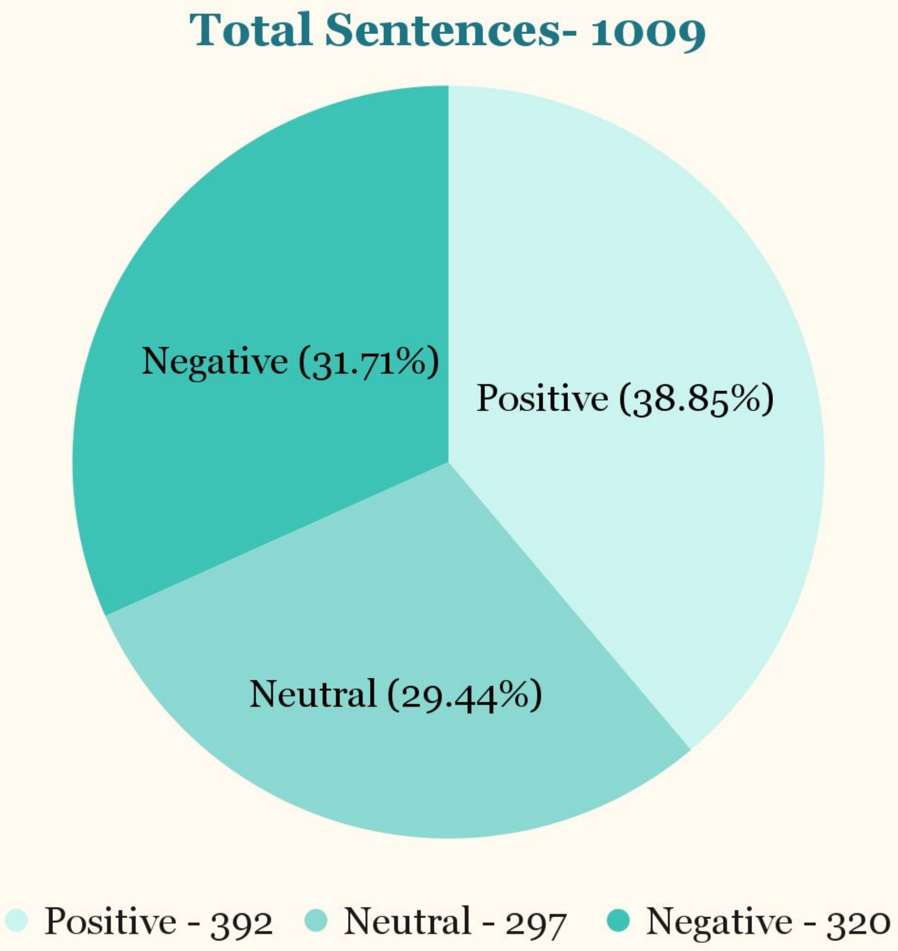
METHODOLOGY

Preprocessing -

- Removing Special Characters
- Removing Punctuation
- Removing Emojis

Sentiment Tagging -

- Positive: 1
- Neutral: 0
- Negative: -1

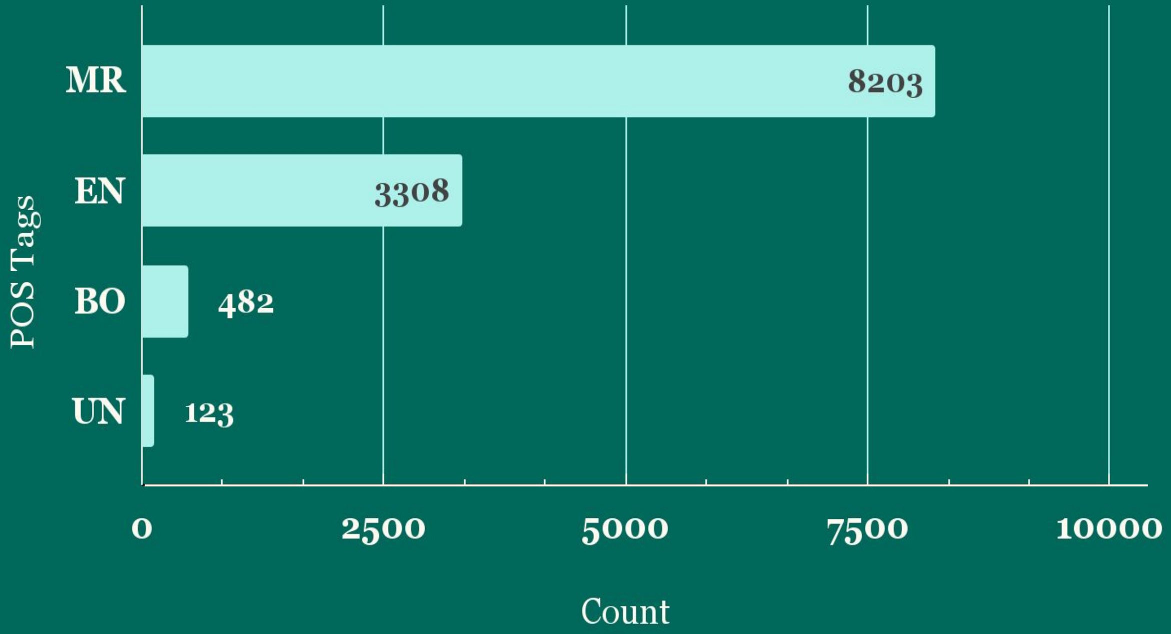


METHODOLOGY

Language Tagging -

- EN:- e.g: friend
- MR:- e.g: mitra
- BO:- e.g: Simran
- UN:- e.g: aankhein

Tokens	12116

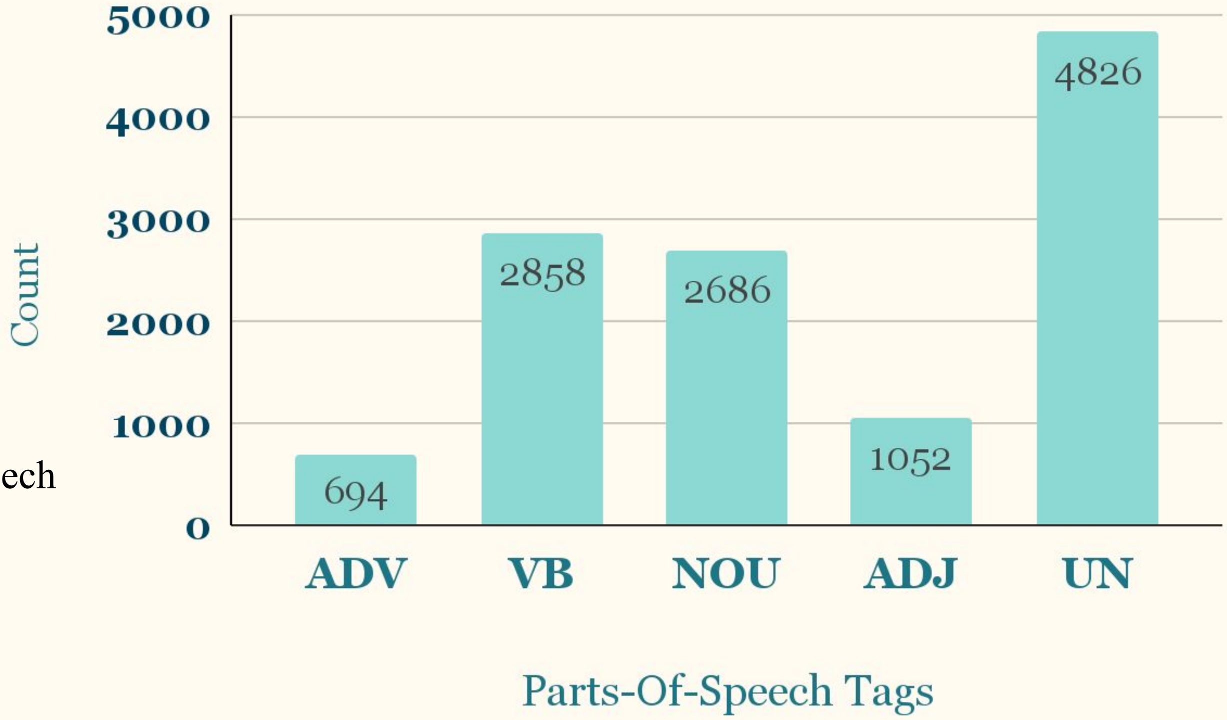


METHODOLOGY

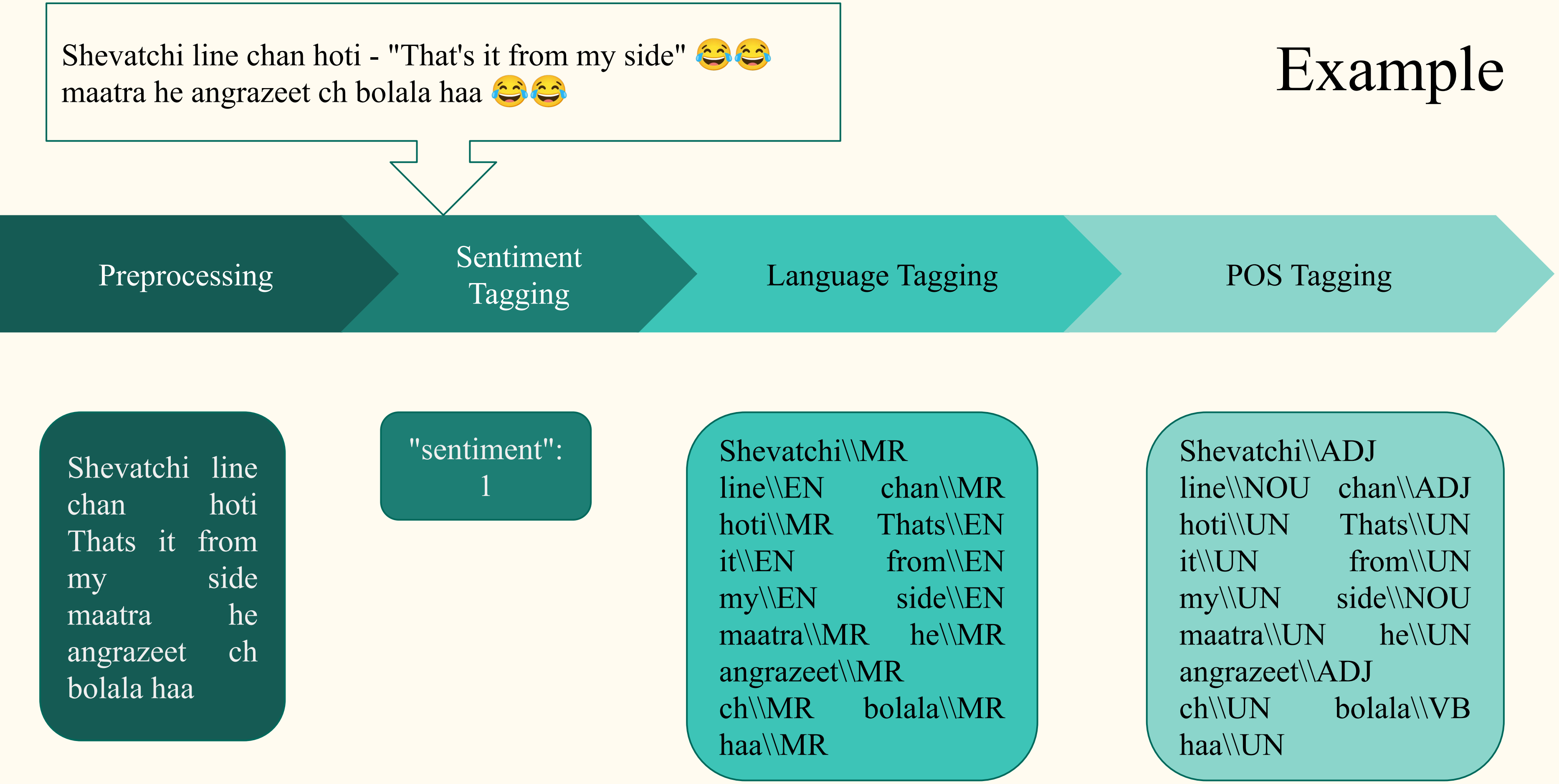
Parts of Speech Tagging -

- NOU: for Noun
- VB: for Verb
- ADV: for Adverb
- ADJ: for Adjective
- UN: for other parts of speech like pronouns, conjunctions, etc

Tokens	12116



Example



Shevatchi line chan hoti - "That's it from my side" 😂😂 maatra he angrazeet ch bolala haa 😂😂

Preprocessing

Sentiment Tagging

Language Tagging

POS Tagging

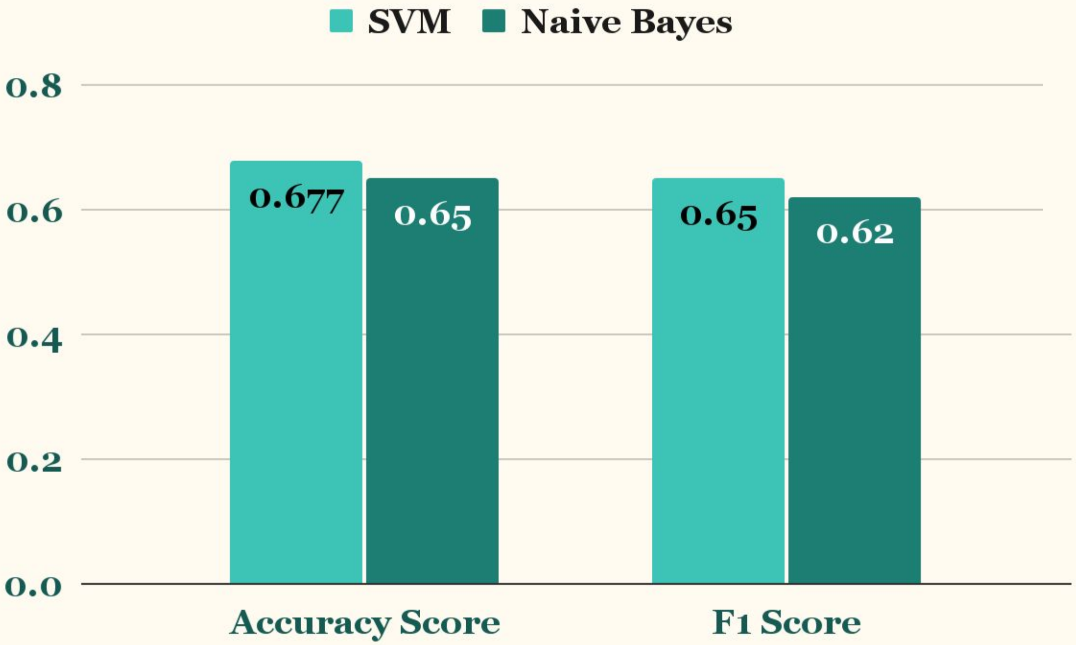
Shevatchi line chan hoti Thats it from my side maatra he angrazeet ch bolala haa

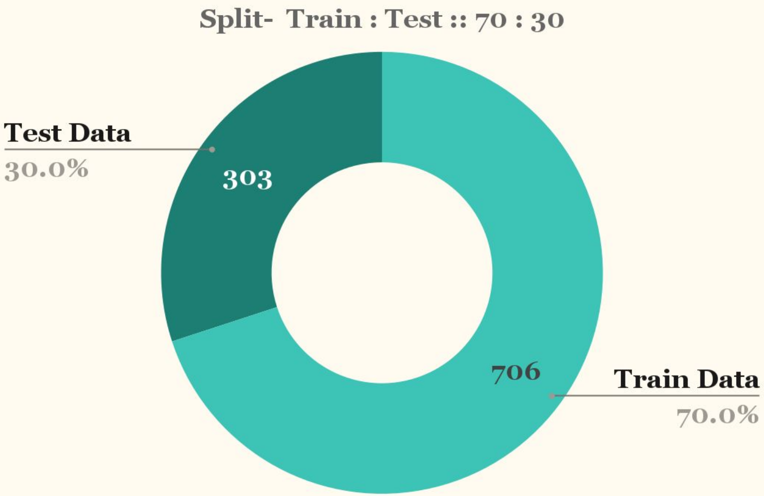
"sentiment": 1

Shevatchi\\MR line\\EN chan\\MR hoti\\MR Thats\\EN it\\EN from\\EN my\\EN side\\EN maatra\\MR he\\MR angrazeet\\MR ch\\MR bolala\\MR haa\\MR

Shevatchi\\ADJ line\\NOU chan\\ADJ hoti\\UN Thats\\UN it\\UN from\\UN my\\UN side\\NOU maatra\\UN he\\UN angrazeet\\ADJ ch\\UN bolala\\VB haa\\UN

RESULTS





Sentences	Count
Total	1009

Dataset Link - https://github.com/Varad0210/Code-Mix-Marathi-English

RESULTS

	Positive	Neutral	Negative
Positive	40	33	16
Neutral	20	39	37
Negative	15	26	77

SVM Confusion Matrix

	Positive	Neutral	Negative
Positive	33	21	25
Neutral	20	36	40
Negative	20	23	75

Naive Bayes Confusion Matrix

CONCLUSION

- The sentiment analysis of code mixed data for Marathi and English language is explored in this research work.
- First attempt to handle sentiments of code mixed Marathi and English languages.
- SVM out performed Naive Bayes for this dataset.

- Feature Extraction
- Expanding Dataset
- Deep Learning Models -
  - AlBERT
  - BERT
  - RoBERTa

FUTURE SCOPE

REFERENCES

[1] Kalaivani Adaikkan and Thenmozhi Durairaj. “Multilingual Hate Speech and Offensive Language Detection in English Hindi and Marathi languages”. In: Fire. 2021.

[2] Kong Lim and Tong Lim. “A Review on Sentiment Analysis for Code-Mix Chinese and English Text on Social Media”. In: Jan. 2020, pp. 53–57. DOI: 10.56453/ icdxa.2020.1001.

[3] SR Mithun Kumar et al. An ensemble model for sentiment classification on code-mixed data in Dravidian Languages. Tech. rep. EasyChair, 2021.

[4] Mohammed Arshad Ansari and Sharvari Govilkar. “Sentiment analysis of mixed code for the transliterated hindi and marathi texts”. In: International Journal on Natural Language Computing (IJNLC) Vol 7 (2018).

[5] Akshata Phadte and Gaurish Thakkar. “Towards normalising Konkani-English code-mixed social media text”. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). 2017, pp. 85–94.

[6] Akshata Phadte and Radhiya Arsekar. “Part-of-Speech Tagger for Konkani-English Code-Mixed Social Media Text”. In: Jan. 2018, pp. 303–307. ISBN: 978-3-319- 91946-1. DOI: 10.1007/978-3-319-91947-8 31.

[7] Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. “A Twitter Corpus for Hindi-English Code Mixed POS Tagging”. In: Jan. 2018, pp. 12–17. DOI: 10.18653/v1/W18-3503.