\section{Review of Literature}

In \cite{ref2}, authors propose to provide an effective neural network based technique for the hostility detection in the hindi text. Authors have evaluated different models like SVM, RF, BiLSTM, and also pre-trained language models that are variants of BERT. They have curated the dataset on hostile and non-hostile hindi text from social media platforms like twitter, facebook and whatsapp, etc., which is further annotated into fine grained labels like fake, hate, defamation and offensive. The datasets have been processed into the four fine grained labels using two different methods, OnevsRest(OvR) and Direct Multi-Label Multi-Classification(DMLMC) and both of them are used for the training and testing and the results obtained are compared. Authors in \cite{ref2} received 91.63\% and 89.76\% accuracies on individual mBERT and XLM-R with their set parameters and the hybrid received accuracy of 92.6\% that is the best performance among all the models employed for the coarse-grained evaluations.

\cite{ref3} proposes a system for identification of low-toxic statements used by users on Educational and specialized web resources,which are characterized by a different type of user. The people using these sites are characterized by good manners, restraint in statements and expressions of emotion. Despite this fact, heated discussions also arise on these web resources, characterized not by highly toxic, but by low-toxic statements, ridicule, sharp jokes, provocative statements and hidden injections.The authors of this paper propose to annotate these low-toxic texts. Datasets are trained on XLM-RoBERTa by the authors in [3] because of its better performance for detection of low-toxic texts as compared to other models. Government agencies can detect low-toxic texts on educational and other related platforms helping them take any corrective actions if necessary.

The research done in \cite{ref4} shows the effect of translation on the sentiment classification task from resource-rich language to a low-resource language. It identifies and enlists words causing polarity shifts into five distinct categories. It further finds the correlation between the languages with similar roots. Our study shows 2-3 percentage points performance degradation in sentiment classification due to polarity shift as a result of translation from resource-rich languages to low-resource languages. To explore the translation approach to develop a sentiment analysis dataset for low-resource languages. To study the effect of translating the English reviews into German, Urdu, and Hindi and compare the classification results of all languages. Authors have studied the importance of handling words affected by Negation. Authors have shown that google translator translated “Faultless Production” into Urdu which means “Bad Production”. This translation is incorrect, and this is another proof that Negation affects translation.

In this paper \cite{ref5}, authors have shown that pre-training multilingual language models at scale leads to significant performance gains for a wide range of cross lingual transfer tasks.Authors have trained a Transformer based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data.Authors model, dubbed XLM-R, significantly outperforms multilingual BERT (mBERT) on a variety of cross-lingual benchmarks.Authors have shown that XLM-R performs particularly well on low-resource languages, improving 15.7\% in XNLI accuracy for Swahili and 11.4\% for Urdu over previous XLM models. Authors have introduced XLM-R, a new state of the art multilingual masked language model,they show that it provides strong gains over previous multilingual models like mBERT and XLM on classification and sequence labeling.

\cite{ref6} has provided an effective neural network based technique for the classification task of SemEval 2020 for two code mixed languages: Hindi-English and Spanish-English. They have used the dataset of SemEval of these two mixed languages. The datasets have been processed and then employed under various models like BiLSTM, mBERT and XLM-R. The Hindi-English dataset consists of 17000 labelled texts from social media while the Spanish-English dataset consists of 15000 labelled texts. All text either labelled positive, negative or neutral.Authors have shown that proper word embeddings can boost performances by a large margin, considering the fact that it already offers the model an insight into that language. The problem becomes more complicated here as the authors deal with two languages instead of one.

The sentiment analysis of low resource language Hindi is still lacking in richly populated linguistic resources, owing to the challenges involved in dealing with the Hindi language is displayed in the research done by authors in \cite{ref7}. Hindi, is the fourth-most popular language, In this article authors first explore the machine learning-based approaches—Naïve Bayes, Support Vector Machine, Decision Tree, and Logistic Regression—to analyze the sentiment contained in Hindi language text derived from Twitter. The dataset employed by authors for sentiment analysis has been fetched from Twitter. Authors have downloaded tweets for movie and product reviews from Twitter, selecting the language “Hindi” in the search filters.They have manually labelled 23,767 tweets into positive or negative. Authors removed the tweets with ironic content, slang language, non-Hindi language, and English words written in Hindi. The tweets without subjectivity were also dropped from the dataset by the authors. After removing these tweets, 16,901 subjective tweets were left.The availability of easy translation provided on the Web, netizens find it interesting to write in their native languages. This pushes for the requirement to perform sentiment analysis in other languages also. Large amounts of content in different languages are available on the Web, which needs to be analyzed to determine the opinion of non-English speaking masses. The proposed CNN approach by the authors gives an accuracy of 85%.

Authors in \cite{ref8} perform sentiment analysis for Manipuri language where orientation of the text is classified into either negative, positive or neutral sentiment. Manipuri is the lingua franca of Manipur, a northeastern state of India. It is not only the official language of Manipur but also included in the 8th Schedule of Indian Constitution. Pre-processing methods used by authors include white space removal, stemming, removal of stop words, removal of numbers, removal of URL links, negation handling, replacing negative mentions, reverting words that contain repeated letters in their original form. Authors have collected and prepared a goal standard dataset for Manipuri sentiment analysis from a local daily newspaper. Transliteration systems are implemented to transliterate Bengali script text to Roman script text and Meetei Mayek script text to Roman script text.Limited availability of good language-specific toolkits for Manipuri language acted as a major constraint for the authors. The transliterated gold standard dataset prepared by the authors could be of use in extending the work on the dataset collected from social media with proper normalization.

\cite{ref9} presents a detailed description of the feature-based TSA system (incorporated with an improved corpus-based negation modeling approach), which classifies tweets based on syntactic and semantic features extracted from them.This work contributes in presenting a feature extraction system that would help in generation of varieties of feature sets, which can be used as an input to classifiers. Authors provide an algorithm for implementing a set of rules for handling those tweets where negation occurrence does not necessarily mean negation.This article by the authors contributes in presenting a comprehensive research in the field of TSA by looking into the critical aspects of NLP that are tweet normalization and negation

All the previous work done in this domain has been done for the languages with plenty resources. Marathi is one language where the research done is still way behind other languages like Hindi in the field of opinion mining. Hence we propose to create a framework where we will be deploying XLM-RoBERTa based model fine-tuned over Marathi tweets dataset. This will achieve better accuracies than other models where there is a need of translation before the task of Opinion Mining and also perform better than the models where Language Models are trained over news headings and articles.