Sentiment Analysis for Social Media Telugu Language Reviews

Padmaja Tallu[1], Dr.A.V.Krishna Prasad[2], Venkataramana Battula[3]

[1,2,3] *Computer Science and Engineer, Department of Information Technology, MVSR Engineering College, Hyderabad, Telangana, India*

Abstract— **The process of identifying positive, negative, or neutral sentiment in text is referred to as sentiment analysis. Businesses use this technology to assess brand reputation, gain a better understanding of customer concerns, and detect sentiment in social data. Sentiment analysis in low-resource languages and regional languages has recently emerged as an emerging area in natural language processing as industries encourage customers to express their opinions in regional languages. Telugu, also known as the "Italian of the East" is a Dravidian language spoken by over 82 million people worldwide. While sentiment analysis in English has become extremely common and relatively simple to perform due to the availability of a large volume of annotated data and tools, very little work is done in Telegu. This work primarily focuses on creating annotated data sets from various web sources and evaluating different Deep Learning models using pre-trained word embeddings trained on Telugu data, as well as evaluating state-of-the-art models such as BERT using the Transfer Learning concept. Different Machine Learning models with TFIDF Features were also evaluated on the same Telegu own corpus for comparison purposes. The combined CNN and BiLSTM model performed well, with 88 percent model accuracy.**

Index Terms: **Telugu Sentiment Analysis, Deep Learning, CNN, LSTM, Bidirectional -LSTM, BERT.**

1.INTRODUCTION

Sentiment analysis (also known as opinion mining) is the use of text analysis and computational linguistic techniques in natural language processing (NLP) to detect, extract, and classify subjective information from unstructured text. It aims to detect the polarity of sentences using word clues extracted from sentence context. As a result, sentiment analysis is recognised as a significant technique for obtaining useful information from unstructured data sources such as tweets or reviews. In business, companies use sentiment analysis to understand their customers' reactions to their products or services. Sentiment analysis is used in politics as a decision-making tool to analyse the general public's reaction to political events. The fourth rank for Telugu among the languages with the best number of native speakers in India and all over the world, with nearly 82 million speakers. Introduction of Unicode (UTF-8) standards for Indian languages and with Telegu typing supportive keyboards there has been a rapid increase of sites written in Telugu. Nowadays all social media platforms and other product selling plot forms encourage people to travel with regional languages. There are many websites that give information in Telugu are movie review sites like 123telugu.com, telugumirchi.com, entertainment websiteslike telugu.oneindia.com, teluguone.com and news websites like eenadu.net,andhrajyothy.com,etc. and social media informationlike tweets, Koos, comments, reviews, blog posts, etc. with the much data accessible on the web, analyzing the sentiment is beneficial to both customers and producers. With SA producers can gain better customer insights and improve accuracy.

Most of the research done in this area for English Language due to availability of resources and tools available but for telugu very less research done due to less resources. Sentiment analysis of Telugu Language is equally important as other Languages. For sentiment analysis in Telugu, a some of research work has been carried out which ranges from lexicon based approach to machine learning based approach. In recent days deep learning and neural networks plays a major role in sentiment analysis and it is considered as a state-of-the art method for analyzing various languages. There no annotated dataset available in telugu which contain all kind of reviews .

Project work flow includes creating annotated Telugu corpus which contain different kind of reviews from different web sources, With the Fasttext Pre word embeddings evaluated different deep learning models.evaluated the state of art model like BERT .For comparison purpose also evaluated different Machine Learning models with TFIDF Features on same telugu own corpus.

2. BACK GROUND

As compared to English Language very less research done Telugu language due to less availabity of resources and tools.

In 2017,Naidu et al. [1] used subjectivity classification by using Telugu SentiWordNet and got 70% accuracy.

In 2018,Sandeep Sricharan Mukku and Radhika Mamidi et al. [2]created annotated corpus for Telugu sentiment analysis containing telugu news sentences and set standard guidelines for annotating reviews. Mukku, Sandeep, et al. [3] carry the sentiment analysis task with above corpus for binary sentiment analysis using word2vec and Machine learning algorithms got 76% accuracy. In 2020,Srikanth Tammina, et al. [5] applied Hybrid learning approach for telugu sentiment analysis. This approach they combined Lexicon based approach and Mechine Learning approch. they achieved a highest accuracy of 85% for Naïve Bayes classifier for binary Sentiment Analysis.

In 2020,Priya, et al. [8] used Translation Method with less number of tweets implemented BiLSTM got 80.3% accuracy. In 2018,A.K.Goel, et al. [11] performed sentiment analysis on Hindi tweets by using deep learning methods called the RNN and so the machine learning algorithms which they found that RNN model stood out with greater accuracy.

In 2020,S. Anbukkarasi, et al. [6] implemented bidirectional LSTM for Tamil tweets andachieved 86% accuracy.

In 2020,LAL KHAN, et al. [7] implemented various machine learning and deep learning models with n-gram features and pre-trained word embeddings they achieve the best F1 score of 82.05% using LR.The above researches done in other languages motivated to Carrie similar research in Telugu Language due to availability of and pre-trained word embeddings developed by Divyanshu Kakwani, et al. [18] at 2020.

In 2020 ,Kushal Jain, et al. [19] work they created pre prepared transformer models for Hindi, Bengali, and Telugu In request to assess the exhibition on Indian dialects explicitly.In 2021,S. Tam, et al. [13] executed CNN+BiLSTM for English Language and got accuracy 91.13%.

3. BUILDING DATASET

As part of this project one Telugu annotated corpus is created which contains telugu tweets, movie reviews from blogs, product reviews from amazon, koos and news records from ANSTASA dataset.

The data is annotated as per guidelines given in paper [3].

To collect the data from Twitter one program is written with access keys which extract the telugu tweets with key word.Total 3000 reviews collected different web sources and annotated as per ANSTASA guidelines. To get better accuracy and improve the efficiency of Deep learning algorithms 3000 records taken from existing annotated dataset ANSTASA.

Sample Data

Review Type	Review
Positive	చాలాబాగుందిస్వీయరచనా (Chala bagundi swiya Rachana) Very nice self-writing
Negative	చివరకుసినిమానిరాశపరుస్తుంది (Civaraku cinimā nirāśaparustundi) In the end the movieis disappointing .
Neutral	ఎల్జీబీటీవ్యక్తులహక ్కులురక్షణఇత్యాదిఅం శాలపైఈవిభాగంపనిచేస్ తుంది (Eljībīṭī vyaktula hakkulu, rakṣaṇa ityādi anśālapai ī vibhāgaṁ pani cēstundi.) This section works on the rights of LGBT people, protection, etc.

The counts of Data set

Class	Number of Reviews (6000)
Neutral	2000
Positive	1915
Negative	2185

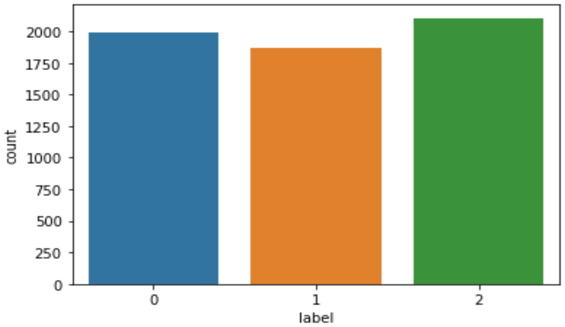

Fig1:The Graph Of Dataset

4. METHODOLOGY

This section mainly focus on the experimental details of machine Learning, Deep Learning and BERT Models. All these models have been implemented on Telugu Corpus created as part of this Project.

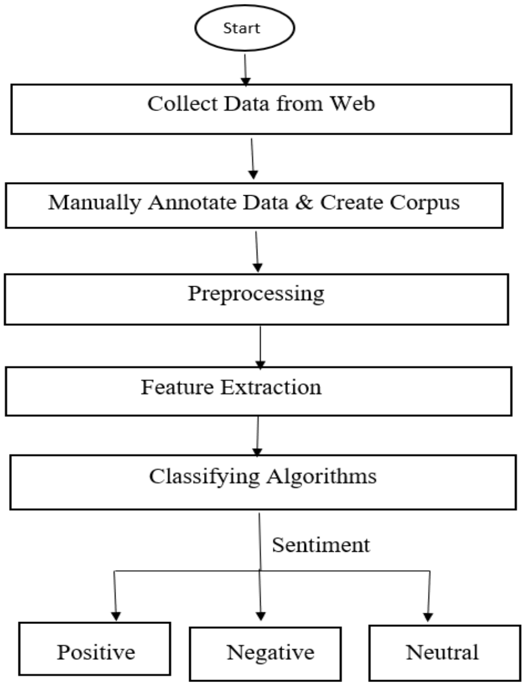

Fig:1.High Level architecture for Telugu sentiment analysis

4.1 Pre Processing

Because the raw dataset always contains words or symbols that computers do not understand, data cleaning is the most important step in NLP. As a result, data cleaning was performed on the datasets to remove punctuation, special characters, and stopping words.

Due Less availability of tools in Language only Removing Stop wordsRemoving special characters. Tokenization are performed.

To remove stop words in telugu one List is created with 150 telugu stop words in telugu then one function is written to remove stop words from entire telugu corpus. Some example for Stopwords .

"ఎక్కడ","తన”,” రెండు”

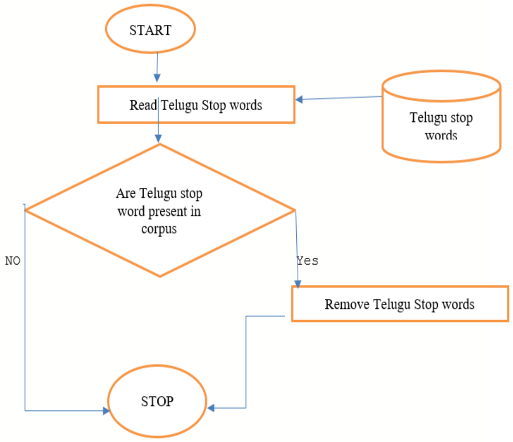

Fig2: Flow Chart for Stopwords removal

4.2 Pre Trained word embeddings

In this phase, the network takes the input of raw text and segments into word or token one by one. Each token is converted into a vector of numeric values. pre-trained word vector models recently applied in many natural processing tasks and have shown better results. The concept behind these pre-trained models is to train these models on very large corpora and fine tune these models for specific tasks.

fastText [17] is a word vector model trained on Wikipedia and common crawl datasets. This model is trained for a total of 157 languages, including Telugu. The INDIC NLP [18] developed the fastText model from scratch with large amount of Telugu data which shown better performance than fastText model. This is the main motive using these pretrained embbeding for Deep learning telugu.

The Fasttext model was trained using skip-gram and continuous bag of words (CBOW) [20]. This model breaks down the unigram into bags of character n-grams. Each single word is represented by the summation of its related n- gram vectors.

Each word of n words is represented as T = {w1, w2, .. ., wn}. Each word is converted into word vector of

d dimension. Text which has a length longer than the predefined l will be truncated. But, if the text which has length shorter than l, zero padding will be added.
This model returns 300d vector for each word.by using all word vectors embbeding weighted matrix created. the maximum sequence length for this work is 128.

4.3. Classification Models

After generating the embbeding layer with embbeding matrix it is passed to different deep learning algorithms to classify the sentiment.
Initially Simple Perceptron model is used then Lstm, BiLSTM and hybrid architecturesCNN+LSTM, CNN+BiLSTM are applied.
With the Transfer Learning concept Pretrained Bert model also applied.
All models hyper parameters are tuned to get best accuracy and to avoid fitting problems.
For the comparison purpose Machine Learning algorithms SVM, Navibias, Decision Tree, Logistic Regression are run with TFIDF vector features on same dataset.

4.4EvaluationMeasures.

We evaluate the effectiveness of our sentiment analysis models using Recall (R), Precision (P), and F1-measure. The mathematical equations are as follows:

Precision = TP /TP + FP ,

Recall = TP /TP + FN ,

F1 = 2 × P × R/ P + R ,

where TP and FP stand for true positive and false positive, and FN stands for false negative.

5.EXPERIMENTAL SETTINGS AND RESULTS

The telugu corpus contains total 6000 reviews. The dataset split into training which contains 80% of reviews and testing which contain 20% of reviews.
For all MachineLearning we use default parameters. For Deep Learning we used Categorical cross entropy as loss function and sgd as optimizer. we set epochs as 50.
Remaing hyper parametersare tuned as per model to avoid fitting problems. For BERT model pretrained indic-trnsformers-te-bert [19] is used which is available in Hugging Face Library.

5.1 Result and Discussion

From Tabel1of results we can observe deep Learning models are working better.
Combination of CNN and BiLSTM classifier model achieved the highest accuracy 88.0 compared to other models due CNN extracts local features and BiLSTM look for sequential features from both directions. Performance of Bert is less due to limited amount of data available in web for Telugu Language. Fig4 the precision for positive class CNN+BiLSTM worked well with 0.94. Compared to Machine Learning Deep Learning models achieved better accuracy.

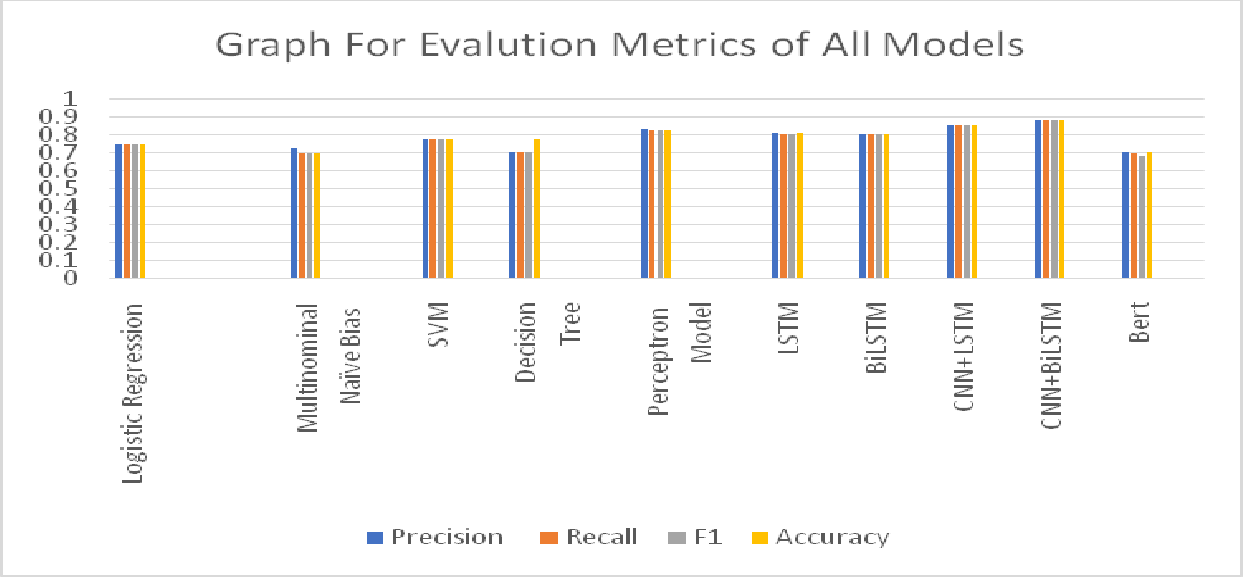

Fig:3: Graph of Telugu Sentiment Analysis Results with deep Learning, Machine Learning models

6.CONCLUSION AND FUTURE WORK

In comparison to Machine Learning Deep Learning algorithms are performing admirably. After evaluating all classifiers, the hybrid model of CNN+BiLSTM performed well with an accuracy of 88 percent because CNN extracts local features and BiLSTM looks for sequential features in both directions. Bert's performance is limited due to the scarcity of Telugu language data on the internet.

Future work will include expanding the corpus and developing cutting-edge classifiers such as UMLFiT and Elmo sentence embeddings. Aspect-based Sentiment Analysis in Telugu and multilingual Sentiment Analysis are also in the works for the future.

REFERENCES

[1] Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu, and Ramesh Kumar Mohapatra. 2017. Sentiment analysis using telugu sent wordnet.

[2] Sandeep Sricharan Mukku and Radhika Mamidi. 2017. ACTSA: Annotated corpus for Telugu sentiment analysis. In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, pages 54–58, Copenhagen, De

[3] Mukku, Sandeep. (2017). Sentiment Analysis for Telugu Language. 10.13140/RG.2.2.22769.38243.

[4] Rani, Sujata & Bhatia, Parteek. (2018). A journey of Indian languages over Sentiment analysis: a systematic review. Artificial Intelligence Review.

[5] S. Tammina, "A Hybrid Learning approach for Sentiment Classification in Telugu Language," *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, 2020, pp. 1-6, doi: 10.1109/AISP48273.2020.9073109

[6] S. Anbukkarasi and S. Varadhaganapathy, "Analyzing Sentiment in Tamil Tweets using Deep Neural Network," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 449-453, doi: 10.1109/ICCMC48092.2020.ICCMC-00084.

[7] L. Khan, A. Amjad, N. Ashraf, H. -T. Chang and A. Gelbukh, "Urdu Sentiment Analysis With Deep Learning Methods," in *IEEE Access*, vol. 9, pp. 97803-97812, 2021, doi: 10.1109/ACCESS.2021.3093078

[8] Priya, G. & Usha, M. (2020). A Framework for Sentiment Analysis of Telugu Tweets. International Journal of Engineering and Advanced Technology. 9. 2249-8958. 10.35940/ijeat. F1602.089620.

[9] Djatmiko, Fahim & Ferdiana, Ridi & Faris, Muhammad. (2019). A Review of Sentiment Analysis for Non-English Language. 448-451. 10.1109/ICAIIT.2019.8834552.

[10] C. Nanda, M. Due and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," *2018 International Conference on Communication and Signal Processing (ICCSP)*, 2018, pp. 1069-1072, doi: 10.1109/ICCSP.2018.8524223.

[11] A. K. Goel and K. Batra, "A Deep Learning Classification Approach for Short Messages Sentiment Analysis," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2020, pp. 1-3, doi: 10.1109/ICSCAN49426.2020.9262430

[12] S. Seo, C. Kim, H. Kim, K. Mo and P. Kang, "Comparative Study of Deep Learning-Based Sentiment Classification," in *IEEE Access*, vol. 8, pp. 6861-6875, 2020, doi: 10.1109/ACCESS.2019.2963426.

[13] S. Tam, R. B. Said and Ö. Ö. Tanriöver, "A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification," in *IEEE Access*, vol. 9, pp. 41283-41293, 2021, doi: 10.1109/ACCESS.2021.3064830.

[14] Abburi, Harika & Eswar, Sai & Ganga Shetty, Suryakanth & Mamidi, Radhika. (2017). Multimodal Sentiment Analysis of Telugu Songs.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–

4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[16] Z. A. Guven, "Comparison of BERT Models and Machine Learning Methods for Sentiment Analysis on Turkish Tweets," *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 2021, pp. 98-101, doi: 10.1109/UBMK52708.2021.9559014

[17] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

[18] @inproceedings{kakwani2020indicnlpsuite,title ={{IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages}}, author={Divyanshu Kakwani and Anoop Kunchukuttan and Satish Golla and Gokul N.C. and Avik Bhattacharyya and Mitesh M. Khapra and Pratyush Kumar}, year={2020}, booktitle={Findings of EMNLP}, }

[19] jain2020indictransformers, title={Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages}, author={Kushal Jain and Adwait Deshpande and Kumar Shridhar and Felix Laumann and Ayushman Dash}, year={2020}, eprint={2011.02323}, archivePrefix={arXiv}, primaryClass={cs.CL}

[20] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, ‘‘Enriching word vectors with subword information,’’ Trans. Assoc. Comput. Linguistics, vol. 5, pp. 135–146, Dec. 2017.