

Review Paper on Sentiment Analysis for Hindi Language

Madhuri Thorat¹ and Nuzhat Shaikh²

¹Smt. Kashibai Navale College of Engineering, Pune, India
Email: thoratmadhuri31@gmail.com

²Modern Education Society's College of Engineering, Pune, India
Email: nfshaikh76@gmail.com

Abstract—Sentiment analysis is inevitable in the current era. The Internet is growing day-by-day. Now-a-days everything is online. We can shop, buy, and sell online. People can give feedback / opinions on the internet. Customers can compare among various products by analyzing the product reviews. As more and more people from different age groups and languages are becoming new internet users, we need it in regional languages. Till date most of the work related to sentiment analysis has been done in the English language. But when it comes to Indian languages, not much research has been done except for a few languages. This paper mainly focuses on performing sentiment analysis in one of the Indian languages i.e. Hindi.

Lately, because of the accessibility of voluminous information on the web for Indian dialects, it has become a significant errand to break down this information to recover valuable data. In light of the development of Indian language content, it is helpful to use this blast of information with the end goal of conclusion investigation. This examination portrays a methodical survey in the field of opinion investigation all in all and Indian dialects explicitly. The current status of Indian dialects in estimation examination is grouped by the Indian language families. The periodical development of Indian dialects in the field of supposition examination, wellsprings of chose distributions based on their importance is additionally portrayed. Further, scientific categorization of Indian dialects in estimation examination dependent on strategies, areas, supposition levels and classes has been introduced. This examination work will help specialists in finding the accessible assets, for example, commented on datasets, pre-handling phonetic and lexical assets in Indian dialects for supposition investigation and will likewise uphold in choosing the most appropriate assumption examination strategy in a particular area alongside applicable future exploration bearings. In the event of asset helpless Indian dialects with morphological varieties, one experiences issues of performing estimation examination because of inaccessibility of explained assets, phonetic and lexical apparatuses. Along these lines, to give productive execution utilizing existing feeling examination procedures, the previously mentioned issues ought to be tended to successfully.

Index Terms— Sentiment Analysis, Natural Language Processing (NLP), Convolutional Neural Network (CNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM).

I. INTRODUCTION

In the present life, individuals consistently look for recommendations and assessments of others for their endurance and choices making. These feelings (or conclusions) advantage showcasing and business tasks in improving items or administrations. As of late there is a multiplication of World Wide Web destinations that underlines client produced content as clients are the likely substance benefactors. Likewise, there are a great deal of remarks and blog-entries about the drifting movement via web-based media. The quick development of micro blogging and rising destinations, for example, Twitter has moved online networks to thrive by empowering individuals to make, share and scatter free-streaming messages and data. This blast of stubborn content has formed an energizing zone in text investigation, which is expressed by numerous names like supposition mining, notion examination, evaluation extraction, and subjectivity examination (Pang et al. 2008). The terms slant investigation/feeling mining have been alluded all through this examination. Supposition Analysis (SA) is a characteristic language preparation undertaking to decide the client's mentality toward a specific element by recognizing and characterizing clients' assessments from a bit of text into various conclusions, classes or feelings, for example, cheerful, miserable, irate, or appalled (Rani and Kumar 2017). The fundamental point of SA is to distinguish whether a book is objective or emotional. Objectivity indicates that the content doesn't comprise any assessment while subjectivity expresses that the content bears sentiment content. For instance, the sentence "this celebrity speaks to the objectivity as this sentence is a reality and passes on broad data as opposed to a feeling or a perspective on some person. Furthermore, the sentence "This film by Amir Khan and Kajol is brilliant." speaks to the subjectivity as this sentence speaks to sentiment about the film and the sentiments of essayists. The abstract content can be additionally ordered into the general classes based on the assessments communicated in the content. Consider the sentence "I love to watch Star TV arrangement." suggests the positive slant of the essayist about "star TV arrangement" because of the opinion word "love" and the sentence "The film was terrible." implies the negative feeling about the film due to the supposition word "horrendous". Similarly, the sentence "I typically get eager by early afternoon." is abstract sentence and suggests impartial conclusion as this sentence comprises sentiments and perspectives on client nonetheless, it doesn't comprise of any sure or negative extremity.

AI for natural language processing and text examination includes utilizing AI calculations and "tight" man-made consciousness (AI) to comprehend the significance of text archives. These records can be just about whatever contains text: web-based media remarks, online audits, study reactions, even budgetary, clinical, lawful and administrative reports. Basically, the part of AI and AI in natural language processing (NLP) and text examination is to improve, quicken and mechanize the fundamental content investigation capacities and NLP includes that transform this unstructured content into useable information and bits of knowledge. AI for NLP and text investigation includes a lot of measurable methods for distinguishing grammatical forms, substances, supposition, and different parts of text. The procedures can be communicated as a model that is then applied to other content, otherwise called directed AI. It additionally could be a lot of calculations that work across enormous arrangements of information to remove meaning, which is known as solo AI. It's imperative to comprehend the distinction among directed and unaided learning, and how you can outwit both in one framework.

II. MOTIVATION

India is a multi-lingual nation with 22 authority dialects. Because of phonetic and social decent varieties of India, it has in every case left open a wide territory for Natural Language Processing (NLP) scientists. The client base of Indian language clients has just developed from 42 million of every 2011 to 234 million out of 2016 and it is relied upon to develop at a pace of 18% (to 536 million) by 2021 as per the KPMG-Google study (Onl 2017). The presentation of Unicode (UTF-8) has lead huge increment of web content. Generally Indian individuals share their feelings on everything including items, most recent films, legislators, occasions, government rules, new patterns, and so forth in their own language as practically all the items, motion pictures, papers and government official sites are presently accessible in Indian dialects. Directly, web indexes, for example, Google likewise uphold Indian dialects including Hindi, Bengali, Punjabi, Tamil, Telugu and Gujarati and so forth. The web-based media sites like Facebook and the miniature blogging locales, for example, Twitter are likewise accessible in a large portion of Indian dialects. Consequently, the acceleration in the accessibility of voluminous Indian dialects information has urged analysts to investigate the exploration here.

III. RELATED WORK

In this paper experimentation is done with deep learning approaches to detect sarcasm in the Hindi-English code mixed dataset. They works on only 5250 tweets dataset, from that only 504 tweets are labelled as sarcastic while remaining 4746 tweets are labelled as not sarcastic.[1].

Table I. Methods and Accuracy

DL Models	Hinglish Data	
	Word2Vec(a)	FastText(b)
1.Series CNN	72.86	72.65
2.Parallel CNN	74.28	73.41
2.1 LSTM	76.19	75.25
2.2 Bi-LSTM	77.12	76.25

Assumption examination is one of the quickest spreading research regions in software engineering, making it trying to monitor all the exercises in the territory. They present client criticism audits on items, where they use supposition mining, text mining and slants, which has influenced the encircled world by changing their feeling on a particular item. Information utilized in this examination is online item surveys gathered from Amazon.com using RNN, LSTM and CNN. We played out a similar notion examination of recovered audits. This examination paper gives you wistful investigation of different advanced cell sentiments on shrewd telephones partitioning them Positive, Negative and Neutral Behaviour. Highest accuracy has been obtained using the CNN algorithm. Ie. 77.63%. [2]

There is only one reported work available for detection of sarcasm in Hindi scripted text. It can not consider natural Hindi tweets. Mapping of tweets and news are based on the keyword so it does not not give full assurance that the news and tweets are of same time stamp.[3]

The author decides to utilize sentiment analysis of twitter tweets to predict the result of the Indian general election. When we define the polarity of a tweet that time emoticons are also an important aspect , but in this work emoticons are not considered. Data was labelled manually so it is not large enough to provide more accuracy. [4].

TABLE II. ALGORITHM AND ACCURACY

Algorithm	Accuracy
Naive Bayes	62.1%
SVM	78.4%
Dictionary Based	34%

In this paper Genetic algorithm discovered the optimal hyper-parameters for the GRU. This approach reduce the time needed for optimization as compared to the traditional grid search approach.65.95% accuracy achieved by using CNN-SVM,using RNN get 72.01% and for proposed GA-GRU model they achieve 88.02% accuracy.[5]

The current decade has seen the surprising advancements in the field of man-made reasoning, and the upset of profound learning has changed the entire man-made consciousness industry. In the long run, profound learning procedures have become basic segments of any model in the present computational world. By the by, profound learning methods guarantee a serious extent of mechanization with summed up rule extraction for both content and assessment arrangement errands. This article plans to give an experimental investigation on different

profound neural systems (DNN) utilized for conclusion characterization and its applications. In the primer advance, the exploration completes an investigation on a few contemporary DNN models and their fundamental hypotheses.[6]

IV. ALGORITHM AND METHODOLOGY

A. NLP

The field of artificial intelligence has always envisioned machines being able to mimic the functioning and abilities of the human mind. Language is considered as one of the most significant achievements of humans that has accelerated the progress of humanity. So, it is not a surprise that there is plenty of work being done to integrate language into the field of artificial intelligence in the form of Natural Language Processing (NLP). NLP primarily comprises of natural language understanding (human to machine) and natural language generation (machine to human). NLP is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

B. CNN

Convolutional Neural networks are designed to process data through multiple layers of arrays. This type of neural networks is used in applications like image recognition or face recognition. The primary difference between CNN and any other ordinary neural network is that CNN takes input as a two-dimensional array and operates directly on the images rather than focusing on feature extraction which other neural networks focus on. The dominant approach of CNN includes solutions for problems of recognition. Top companies like Google and Facebook have invested in research and development towards recognition projects to get activities done with greater speed.

A convolutional neural network uses three basic ideas –

- Local respective fields
- Convolution
- Pooling

To teach an algorithm how to recognize objects in images, we use a specific type of Artificial Neural Network: a Convolutional Neural Network (CNN). It is a special type of Neural Network widely used for recognize the image.

V. CONCLUSIONS

According to the results, the words which are not in the database are considered as neutral though they are sentiment words. Therefore the database should be as rich as possible for efficient results. This concludes that not only the techniques but the resources are also more important for better results. The dataset will be enhanced as part of future work. Also other algorithms and techniques will be implemented to increase the efficiency.

REFERENCES

- [1] Akshita Aggarwal, Anshul Wadhawan “ ‘Did you really mean what you said’ : Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings”, arXiv:2010.00310v3 [cs.CL] 15 Oct 2020
- [2] Rupal Bhargava, Shivangi Arora and Yashvardhan Sharma, “Neural Network-Based Architecture for Sentiment Analysis in Indian Languages,” J. Intell. Syst. 2019.
- [3] Dr. Santosh Bharti and Sanjay Jena, “Harnessing Online News for Sarcasm Detection in Hindi Tweets”, 2017
- [4] Parul Sharma and Teng-Sheng Moh, “Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter”, 2016 IEEE International Conference on Big Data.
- [5] Kush Shrivastava and Shishir Kumar “A Sentiment Analysis System for the Hindi Language by Integrating Gated Recurrent Unit with Genetic Algorithm”, Nov 2020
- [6] Ramesh Wadawadagi, VeerappaPagi, "Sentiment analysis with deep neural networks: comparative study and performance assessment", Artificial Intelligence Review <https://doi.org/10.1007/s10462-020-09845-2>.
- [7] AsadAbdi, SitiMariyamShamsuddin, ShafaatunnurHasan, and JalilPira nMD, "Machine learning -based multi-documents sentiment -oriented summarization using linguistic treatment", Expert Systems with Applications, vol. 109, pp. 66 -85, 1 November 2018.

- [8] ParamitaRay, and AmlanChakrabarti, "A Mixed approach of Deep Learning method and Rule -Based method to improve Aspect Level Sentiment Analysis", *Applied Computing and Informatics*, Available online 4 March 2019.
- [9] ZiyuanZhao, HuiyingZhu, ZehaoXue, ZhaoLiu, JingTian, Matthew Chin HengChua, and MaofuLiu, "An image text consistency drivenmultimodal sentiment analysis approach for social media", *Information Processing & Management*, vol. 56, no. 6, November 2019.
- [10] SaeromPark, JaewookLee, and KyoungokKim, "Semi -supervised distributed representations of documents for sentiment analysis", *Neural Networks*, vol. 119, pp. 139-150, November 2019.
- [11] SrishtiVashishtha, and SebaSusan, "Fuzzy rule based unsupervised sentiment analysis from social media posts", *Expert Systems with Applications*, vol. 138, 30 December 2019.
- [12] AbdallahYousif, ZhendongNiu, Jame sChambua, and Zahid YounasKhan, "Multi - task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification", *Neurocomputing*, vol. 335, pp. 195-205, 28 March 2019.
- [13] Mohammad A.Hassonah, RizikAl-Sayyed, AliRodan, Ala' M.Al-Zoubi, IbrahimAljarah, and HossamFaris, "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter", *Knowledge -Based Systems*, vol. 192, 15 March 2020.
- [14] ha, Vandana, N. Manjunath, P. Deepa Shenoy, K. R. Venugopal, and Lalit M. Patnaik. "Homs: Hindi opinion mining system." In *Recent Trends in Information Systems (ReTIS)*, 2015 IEEE 2nd International Conference on, pp. 366-371. IEEE, 2015.
- [15] Sitaram, Dinkar, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. "Sentiment analysis of mixed language employing Hindi-English code switching." In *Machine Learning and Cybernetics (ICMLC)*, 2015 International Conference on, vol. 1, pp. 271-276. IEEE, 2015.
- [16] Sharma, Richa, Shweta Nigam, and Rekha Jain. "Opinion mining in Hindi language: a survey." *arXiv preprint arXiv:1404.4935* (2014).
- [17] Hegde, Yashaswini, and S. K. Padma. "Sentiment Analysis for Kannada using mobile product reviews: A case study." In *Advance Computing Conference (IACC)*, 2015 IEEE International, pp. 822-827. IEEE, 2015.
- [18] Nair, Deepu S., Jisha P. Jayan, and Elizabeth Sherly. "SentiMa-sentiment extraction for Malayalam." In *Advances in Computing, Communications and Informatics (ICACCI 2014 International Conference on*, pp. 1719-1723. IEEE, 2014.
- [19] Nair, Deepu S., Jisha P. Jayan, R. R. Rajeev, and Elizabeth Sherly. "Sentiment Analysis of Malayalam film review using machine learning techniques." In *Advances in Computing, Communications and Informatics (ICACCI)*, 2015 International Conference on, pp. 2381-2384. IEEE, 2015.
- [20] Hasan, KM Azharul, and Mosiur Rahman. "Sentiment detection from Bangla text using contextual valency analysis." In *Computer and Information Technology (ICCIT)*, 2014 17th International Conference on, pp. 292-295. IEEE, 2014.
- [21] Ghosal, Tirthankar, Sajal K. Das, and Saprativa Bhattacharjee. "Sentiment analysis on (Bengali horoscope) corpus." In *India Conference (INDICON)*, 2015 Annual IEEE, pp. 1-6. IEEE, 2015.
- [22] Michael R. Berthold "Mixed fuzzy rule formation", South San Francisco, April 2002.
- [23] Hajime Watanabe, MondherBouazizi, TomoakiOhtsuki"Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection", *IEEE Access* Volume: 6,February 2018.
- [24] Nicholas Cummins,Shahin Amiriparian,Sandra Ottl, Maurice Gerczuk Maximilian Schmitt, Björn Schuller"MULTIMODAL BAG-OF-WORDS FOR CROSS DOMAINS SENTIMENT ANALYSIS" *Journal of Computer Applications* (0975 – 8887) Volume 154 – No.11, November 2016
- [25] Wafa Alorainy, Pete Burnap, Han Liu, Matthew L. Williams,"The Enemy Among Us": "Detecting Hate Speech with Threats Based Othering Language Embeddings" *International Journal of Scientific & Engineering Research* Volume 9, Issue 3, March-2018.
- [26] Jianguo Jiang1, Jiuming Chen, Kim-Kwang Raymond Choo, Kunying Liu1, Chao Liu1, Min Yu1, PrasantMohapatra "Prediction and Detection of Malicious Insiders' Motivation based on Sentiment Profile on Webpages and Emails ", *IEEE* 2018.
- [27] Chris Jefferson, Han Liu ,"Fuzzy Approach for Sentiment Analysis", *IEEE School of Computing* July 2017.
- [28] Dr. RanjanaRajnish ,"Fuzzy Aspects in Sentiment Analysis and Opinion Mining".
- [29] U. Sree Krishna, Hima Shree, K. Jayadeep, P.LakshmiPrasanna,"Text Classification Using Fuzzy Neural Network",*International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-7, Issue-6S4, April 2019
- [30] Orestes Appela, Francisco Chiclanaa, Jenny Cartera, HamidoFujitab ,"A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level", *Preprint submitted to Knowledge-Based Systems* May 13, 2016.