IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 7, July 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

Sarcasm Detection in Hindi Text for Comparative Analysis of Three Approaches - Word Embedding, Context Based Approach and Machine Learning Algorithm"

Maryam S. Siddiqui[#1], Sharvari S. Govilkar[#2]

Department of Computer Engineering, Mumbai University, PIIT, New Panvel, India[#1, 2]

ABSTRACT: Sarcasm is defined as a special categorization of figurative language that intents opposite of what is being said, to hurt someone feelings or to criticize someone. Negative thoughts are expressed through positive intensified words. Today with the increase in communication on social media platforms such as Face book, Twitter, Whats App, etc. negativity is expressed indirectly, and this has become a new trend. Hindi is one of the popular Indian languages that is highly used by Indians while communicating on social media. As Hindi language is rich in morphology and complex in structure, sarcasm detection is one of the tedious jobs. An extensive set of annotated training data to detect sarcasm is surely required. Availability of Hindi annotated data set is almost negligible in the domain of sarcasm detection. Small amount of research has been done in the field of sentiment analysis for Hindi Language. Information content in Hindi being a national language is important to be analyzed as it is widely used industries and government(s). The idea is to propose a system where the comparison between three approaches namely - Word Embedding, Context based Approach and Machine learning algorithm for Hindi text. Among three of them which approach works which higher accuracy and efficiency is seen.

KEYWORDS - Sentiment analysis, Sarcasm detection, Pos tagging, Pre processing,stop words, Tweets, Twitter,Word embedding, Context based approach Machine learning approach.

I. INTRODUCTION

Sentiment Analysis is identification and aggregation of attitudes and opinions shown by Internet users for a specific topic. This process involves two steps: a) information mining from various sources of text forms such as blogs, reviews and b) classification as positive, negative or neutral based on polarity.

1.1. Definition of Sarcasm Detection

Sarcasm can be defined as to convey or express feelings where an individual or society says or writes something that is completely opposite of what they meant or said. When one uses sarcasm, the person speaks the contradictory of what the speaker means to express gloomy feelings applying positive words. The feature set used for English scripted tweets might not be applicable efficiently for identifying sarcasm in natural Hindi tweets. To identify sentiment analysis for Hindi Language, a small amount of research has been carried out till date. Researchers face numerous challenges while performing sentiment analysis for Hindi language:
(1) Hindi as a language lacks availability of well annotated standard corpora. This makes application of supervised machine learning algorithms impossible. (2) Absence of efficient parser and tagger for this resource scarce language. (3) Limited numbers of adjectives and adverbs are contained in limited resources like Hindi Senti Word Net (HSWN) available for this language. (4) Also, translation dictionaries do not have all the words for Hindi due to language variations.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

1.2 Motivation

Being rich in morphology and complex in structure, sarcasm detection in Hindi is a tedious job. For the purpose of machine learning analysis for sarcastic Hindi sentences the annotated resources are found to be almost negligible.Sarcasm is an indirect way to negate a sentence to make someone feel insulted or stupid. Positive words or intensified positive words are often used by the people to write or speak sarcastic sentences. Hence, the interesting problems in the text, speech, video, etc.involve detection of sarcasm. from

II. LITERATURE SURVEY

Abulaish et al. (2018) invokes the problem of self deprecating sarcasm detection, a special case of sarcasm detection. Amalgamation of rule-based and machine learning approaches have been proposed for detecting self-deprecating sarcasm. Different categories of figurative language targeted were sarcasm, irony, satire, etc. in Twitter, but self deprecating sarcasm detection was never considered.[1]

Aggarwal et al. (2020) studied a corpus of tweets for training custom word embeddings and a Hinglish dataset labelled for sarcasm detection was used. Deep learning-based approaches (including CNNs, LSTMs, Bi-directional LSTMs (with and without attention) are used to address the problem of sarcasm detection in Hindi-English code-mixed tweets using bilingual word embeddings derived from Fast Text and Word2Vec approaches. Attention based Bi-directional LSTMs gave the best performance exhibiting an accuracy of 78.49%.[2]

Code-mixed Hinglish tweets are fed as training, validation, and test sets for model training, validation, and testing respectively. NITS-Hinglish - Senti Mix is an ensemble model wherein different models like basic LSTM (Long Short-Term Memory), LSTM + Convolution, a Bi LSTM (Bidirectional LSTM), and a CNN (Convolution Neural Network) model have been amalgamate to improve the general F-Score of the framework. [3]

Ilavarasan (2020) has done a survey about different sarcasm detection works done in past, given a brief about general architecture of sarcasm detection and various approaches used, different types of sarcasm, and some challenges in sarcasm detection. [10]

Sarsam et al. (2020) showed that using lexical, pragmatic, frequency, and POS tagging contribute in improving the performance of SVM, and also lexical and personal features can enhance the performance of CNN-SVM.[16]

Bharti et al. (2018) proposes a pattern-based approach to identify sarcastic Hindi tweets where a set of online news is treated as temporal facts. [4]

Hazarika et al. (2018) adapted a hybrid approach of both context-driven and content modeling for the purpose of sarcasm detection in online social media discussions such as Reddit called CASCADE (Contextual Sarcasm Detector). [9]

Swami et al. (2018)used English-Hindi code-mixed data set of tweets for presence of sarcasm where each token was annotated with a language tag. A supervised classification system was developed using
the same data set that resulted in an average F-score of 78.4 after using random forest classifier and performing 10-fold cross validation.[18]

Bharti et al.(2018) approach followed in this research utilizes Hindi news as the context of a tweet within the same timestamp and obtained an accuracy of 87%.[5]

Sahaet al. (2107) provides classification based on the polarity of tweets as- positive, negative or neutral. In order to find accuracy of tweets Naïve Bayes and SVM classifiers were used .[14]

Machine-learning algorithm and rule-based approach have been discussed in this paper. Sarcasm has
different nature, shape and application in real life, hence it was found that this was a challenging problem and is too wide to be made as a generalized formula. [20]

Dave et al.(2016) various supervised classification techniques mainly used for sarcasm detection and their features have been identified. Results obtained from the classification techniques, on textual data available in various languages on review related sites, Joshi et al. (2016) checked whether prior work could be improved using semantic similarity/discordance between word embedding or not.Four different types of word embedding have been experimented. Irrespective of the word embedding used or the original feature set to which our features are augmented appreciable improvement in sarcasm detection, had been observed. [12]
social media sites and micro-blogging sites have been analyzed.

Bouazizi et al. (2016) detected sarcasm on Twitter . A pattern-based approach has been proposed with four sets of features that cover the different types of sarcasm. The tweets are classified as: Sarcastic and Non-sarcastic. An accuracy of 83.1% with a precision equal to 91.1% is by the proposed approached. [6]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

Zhang et al. (2016) investigates the use of neural network for tweet sarcasm detection, and comparison between the effects of the continuous automatic features with discrete manual features is made. [21]
Bouazizi et al. (2015) has proposed a method to detect sarcasm in Twitter that uses different components of the tweet. Four sets of features containing different types of sarcasm classified tweets into sarcastic and non-sarcastic.[7]

III. METHODOLOGY

Availability of Hindi annotated data set is almost negligible in the domain of sarcasm detection. Small amount of research has been done in the field of sentiment analysis for Hindi Language. Information content in Hindi being a national language is important to be analyzed as it is widely used industries and government(s). The idea is to propose a system where the comparison between three approaches namely-Word Embedding, Context based Approach and Machine learning algorithm for Hindi text. Among three of them, which approach works with higher accuracy and efficiency.

3.1 Hindi Sarcasm detection System Architecture

The proposed methodology compares three different approaches – Word embedding, Context based approach and different machine learning algorithm.

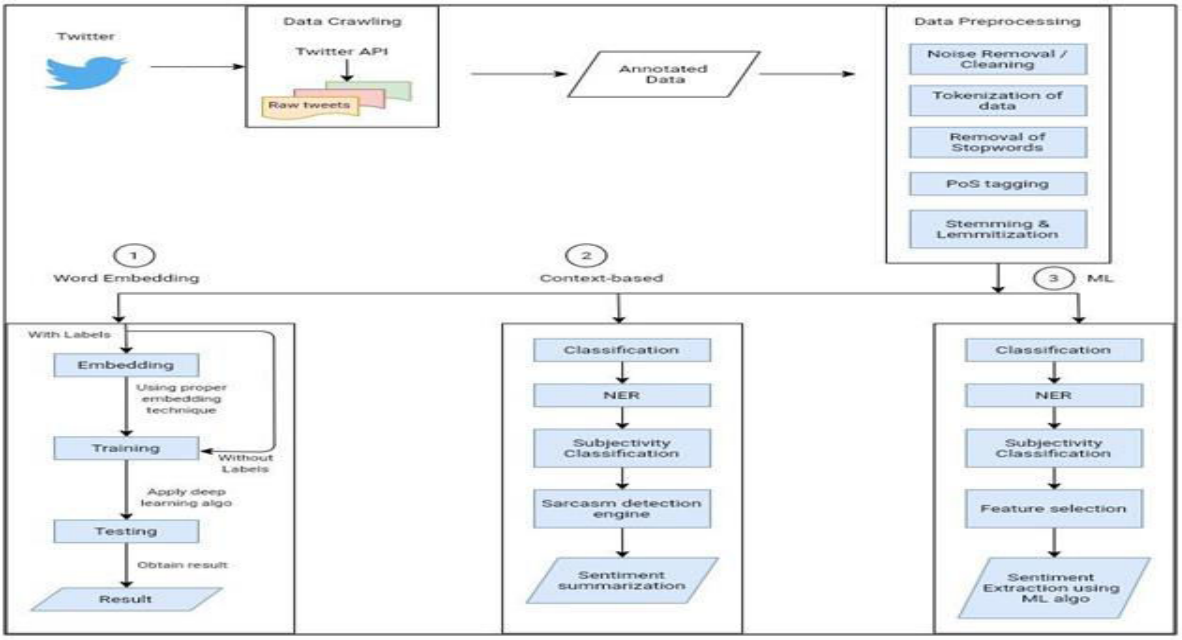

Figure 3.1: Hindi Sarcasm Detection System Architecture

3.2 Input for the system:

In this approach the system takes Hindi tweets as an input. The sentences then pre - process through various steps which includes data crawling, annotated data , data pre – processing**.** After which the pre- processed data is fetched to the different algorithm.

Step I: Data Crawling Step **II:** Annotated data **Step III**: Data pre – processing **Step IV:** Embedding

The two types of embedding which we can try upon for Hindi as language are:

1) Fast Text: Few steps involved in performing Fast Text algorithm are:
a) Obtaining sentence/document vector by averaging the word/n-gram embeddings**. b)** Classification task is done by, multinomial logistic regression, where the sentence/document vector corresponds to the features. **c)** Hierarchical

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

SoftMax is used for applying Fast Text on problems with a large number of classes in order to speed-up the computation.

2) Bidirectional Encoder Representations from Transformer (BERT)
Consider two sentences: "The man was accused of robbing a bank." "The man went fishing by the bank of the river." BERT would produce different word embedding for the word "bank" in both the cases. BERT is the first *deeply bidirectional*, *unsupervised* language representation, pre-trained using only a plain text corpus.

Step V: Using deep learning models for testing and training followed by obtaining of results

Deep learning models use various algorithms. One network cannot be considered perfect, some or the other algorithm is best suited for a specific task as per the requirement.

APPROACH II: CONTEXT BASED

a) Tweet Collection
An algorithm needs to be implemented to work upon extracting of important keywords from collected and processed tweets. A check needs to be carried on for any proper noun, verb and noun POS tags. If there is a presence of these tags are found, then corresponding keywords are appended with a set of keywords. This algorithm produces tags such as- proper noun (NNP), verb (V), and noun (NN) will work as subject, verb, and object, respectively.

b) Tweet Annotation
Practitioners or teachers who are professionals carry out this process of tweet annotation.

c) Sarcasm Detection Engine
Sentiment identification and Context identification are carried out by Sarcasm detection Engine.

i) Sentiment identification
Classification as either positive, negative or neutral of tweets and sentiment values is done in this step.

ii) Context Identification in Tweets:
The inputs to this algorithm are Tweet Corpus (C1), News Corpus (C2) and Hindi SentiWordNet (HSWN). The output obtained is the context polarity of tweets and news. Also, polarity values are found for each phrase.

d) Sarcasm Detection Algorithm
The tweet will be classified as sarcastic if there occurs a contradiction between the context polarity of the tweet and related news. Also, there might be a case where the context polarity of a tweet and related news may be the same, but there exists a contradiction of sentiment, still the tweets will be labeled as sarcastic. Else, tweets will not be labeled under the category of sarcastic.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

IV. RESULTS AND DISCUSSION

Sarcasm Detection for different algorithms:

The sarcasm detection result for various algorithm is shown in the given figure.

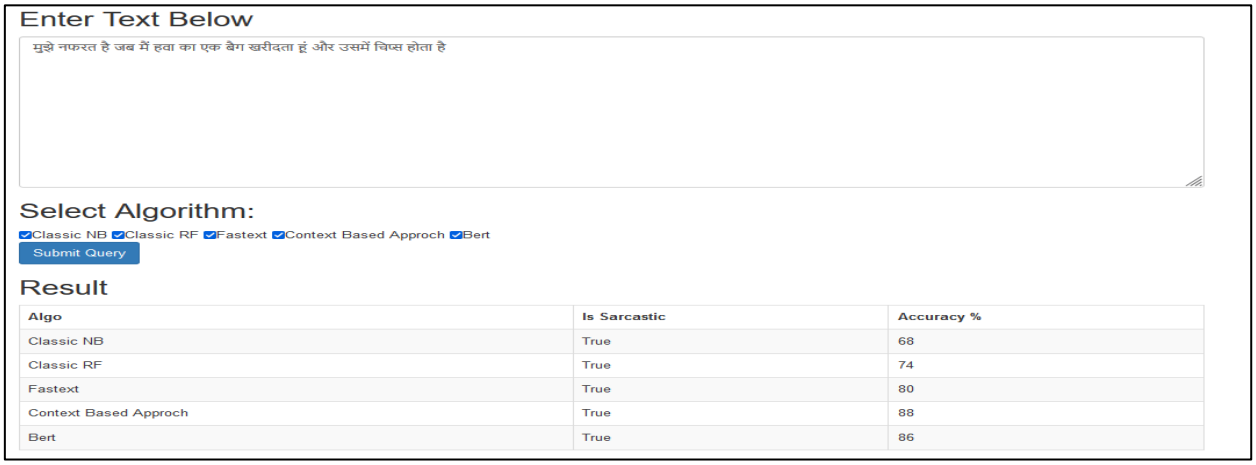

Figure 4.6(a): Sarcasm detection of various algorithm

Figure 4.6(b): Sarcasm detection of various algorithms (Corner Case)

V. TRAINING AND TESTING

5.1 Dataset used:

The dataset used in this work consists of tweets from Twitter. The data set for training has tweets which get classified as Sarcastic and non-sarcastic when given input. The dataset is divided into 80:20 ratio.80 percent of data as used training data and the rest 20 percent as used for testing the algorithm.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

5.2 Performance Evaluation:

Table 5.1: Confusion Matrix for performance evaluation

ACTUAL RESULTS	PREDICTED RESULTS				TOTAL DATASET
	SARCASTIC		NON-SARCASTIC		
SARCASTIC	280	TP	60	FN	**340**
NON- SARCASTIC	55	FP	390	TN	**445**

5.3.1 Analysis of accuracy:

Accuracy measure for each similarity is shown in the following table. The statistical methods were tested to find out the best one.

Table 5.2: Analysis of Accuracy

Measure	Machine learning Approach		Context Based approach		
Algorithm name	Classic Naïve Bayes	Classic Random Forest	Fast text	Bert	Word Embeddings
Accuracy	72%	75%	81%	83%	90%

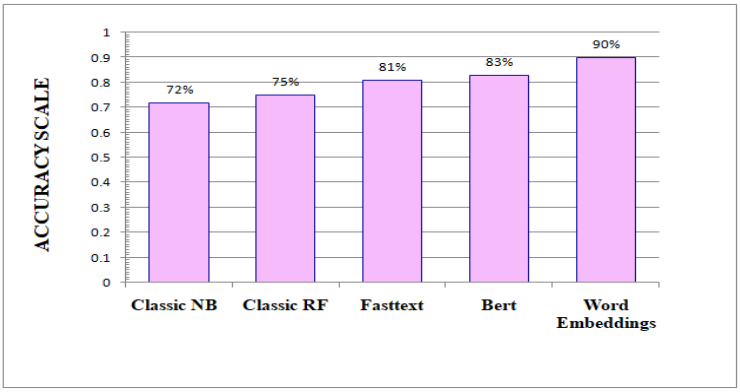

Figure 5.1 Accuracy comparisons of all Sarcasm detection techniques

From table 5.2 and Figure 5.1, Word Embeddings gives the best score among all methods. Hence for finding tweets in Hindi word embeddings approach can be proved effectively accurate.

5.3.2 Analysis of Precision:

As precision is the best measure of prediction i.e., how much our prediction is correct. The precision will be high only when the false positive ratio is very less. The precision analysis is mentioned in the following table 5.2:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

Table 5.3: Analysis of Precision

Measure	Machine learning Approach		Context Based approach		
Algorithm name	Classic Naïve bayes	Classic Random Forest	Fast text	Bert	Word Embeddings
Precision	80%	77%	82%	76%	88%

In figure 5.2, the clear picture of high precision is displayed. This conveyed that the proposed system is expert enough to identify the non-paraphrase sentences. Among all techniques Word Embedding proved best for identifying dissimilar sentences from the copra.

5.3.3 Analysis of Recall:

Recall is also known as true positive rate or sensitivity. Itis the fraction of relevant instances that have been retrieved over the total amount of relevant instances. High value of recall ensures high probability of relevance of instances.

Table 5.4: Analysis of Recall

Measure	Machine learning Approach		Context Based approach		
Algorithm name	Classic Naïve bayes	Classic Random Forest	Fast text	Bert	Word Embeddings
Recall	77%	75%	80%	88%	83%

From the above figure 5.3, it can be stated that Bert is the best method to find sarcasm from the given tweets. The tweets which are actually can be identified by Word embeddings accurately.

5.3.4 Analysis of F-measure:

As alone precision or alone recall cannot decide a system's performance, because precision deals with positive predictive value and recall deals with true positive value. Hence, both precision and recall needs to be combined to get exact positive performance of the system. F-measure is the harmonic mean of precision and recall in which precision and recall is evenly weighted.

Table 5.5: Analysis of F-measure

Measure	Machine learning Approach		Context Based approach		
Algorithm name	Classic Naïve bayes	Classic Random Forest	Fast text	Bert	Word Embeddings
F-Measure	78.4%	75.98%	80.9%	81.5%	85.42%

The figure 5.4 shows the graphical representation of F-measure comparison. Word embeddings gives highest measure score of 85% which outperforms all other techniques.

From testing of all similarity measures, it is obvious to say that Word embeddings is the best method among all techniques. When considering all sentences in a given dataset, other statistical techniques also performed well for different tweets.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

5.3.5 Analysis of Overall Sarcastic tweets:

The overall score is analyzed with all 4 performance measures to find out the performance level of the proposed system. Overall Similarity is taken as input for performance analysis. Even if there is variation in the similarity values of each method, the overall similarity score provides a better result. Here all the statistical concerns and semantic concerns are considered.

Table 5.6: PERFORMANCE ANALYSIS

MEASURE	VALUE
ACCURACY	85.3%
PRECISION	83.5%
RECALL	82.3%
F-MEASURE	82.8%

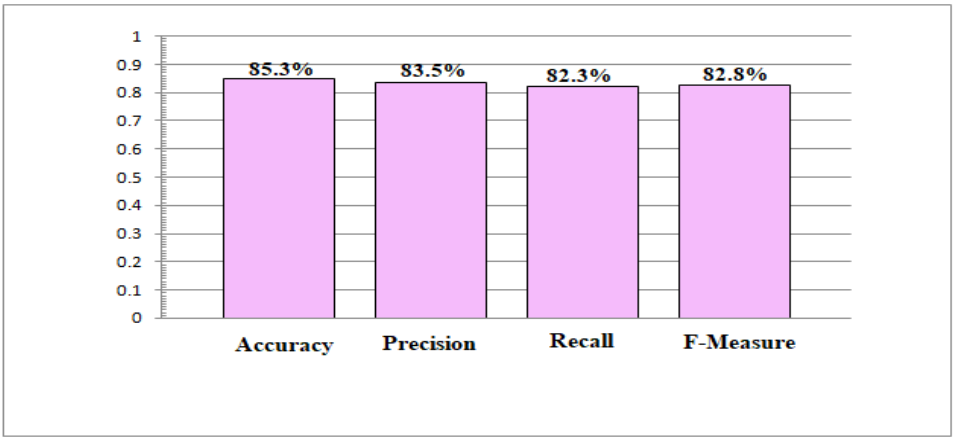

Figure 5.5: Performance Analysis of all Sarcasm detection techniques

The Hindi Sarcasm detection system got accuracy of 90% and F-measure of 85% which is pretty enough to prove that the Hindi Sarcasm Detection System performed very well.

VI. APPLICATIONS

1. Due to rapid increase in usage of the online social media platforms, users can freely give their opinion. This way companies make use of this ecosystem to access major public opinion about aspects associated with products, services, and to provide real-time customer assistance.
2. An active team for marketing and customer assistance purposes is present as a huge volume of information is available on social media websites.
3. Flooding of data results in depending upon tools like Hootsuite which is capable of performing complicated tasks that includes content management, sentiment analysis, and extraction of relevant messages for the company's customer service representatives to respond to. But the disadvantage of using these tools is that they lack the sophistication to decode nuanced forms of language such as sarcasm that carry indirect messages.
4. Detection of sarcasm is of great importance and beneficial to many NLP applications, such as sentiment analysis, opinion mining and advertising.

VII. CONCLUSION AND FUTURE SCOPE

Rich morphology and complex structure of Hindi language makes sarcasm detection one of the tedious jobs. To make this job easy an extensive set of annotated training data is required. Available Hindi annotated data set is almost negligible. This work used 785 tweets. The data set contained 280 positive examples and 60 negative examples. The non-sarcastic data set contained 55 positive examples and 390 negative examples. The Context based approach has

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007083 |

achieved a higher accuracy as compared to all other approaches. The second highest accuracy is achieved by Word Embedding – Bert Algorithm.

In addition to this research, Fake news detection can also be implemented by detecting sarcasm in the tweet as future scope. Detecting the type of news labeled as positive or negative can help to detect whether news is fake or real. Also, Emotion detection can be done wherein we can find the type of tweet on Twitter. The tweet could be sarcastic or non-sarcastic. This may help us to know the trend on Twitter and also can find views.

REFERENCES

[1] Leilei Kong, ZhenyuanHao, Kaisheng Chen, ZhongyuanHan ,Liuyang Tian, Haoliang Qi "HIT2016@DPIL-FIRE2016:Detecting Paraphrases in Indian Languages based on Gradient Tree Boosting", DPIL 2016
[2] Sandip Sarkar, ParthaPakray, Saurav Saha, Dipankar Das, Jereemi Bentham, Alexander Gelbukh "NLP-NITMZ@DPIL-FIRE 2016: Language Independent Paraphrases Detection ", DPIL 2016
[3] Kamal Sarkar "KS_JU@DPIL-FIRE2016: Detecting Paraphrases in Indian Languages Using Multinomial Logistic Regression Model", DPIL 2016
[4] TanikSaikh, Sudip Kumar, Naskar Sivaji, Bandyopadhyay "JU_NLP@DPIL-FIRE2016: Paraphrase Detection in Indian Languages - A Machine Learning Approach", DPIL 2016
[5] Rupal Bhargava, Anushka Baoni, Harshit Jain "BITS_PILANI@DPIL-FIRE 2016: Paraphrase Detection in Hindi Language using Syntactic Features of Phrase", DPIL 2016
[6] Vani K, Deepa Gupta. "ASE@DPIL-FIRE2016: Hindi Paraphrase Detection using Natural Language Processing Techniques & Semantic Similarity Computations" DPIL 2016
[7] Anuj Saini "Anuj@DPIL-FIRE2016: A Novel Paraphrase Detection Method in Hindi Language using Machine Learning", DPIL 2016
[8] Sindhu L. and Sumam Mary Idicula "CUSAT_NLP@DPIL-FIRE2016: Malayalam Paraphrase Detection", DPIL 2016
[9] Manju K. and Sumam Mary Idicula," CUSAT TEAM@DPIL-FIRE2016: Detecting Paraphrase in Indian Languages-Malayalam", DPIL 2016
[10] Ditty Mathew and Dr. Suman Mary Idicula," Paraphrase identification of Malayalam sentences - an experience", IEEE 2013
[11] R. Thangarajan, S. V. Kogilavani, A. Karthic, S. Jawahar, "KEC@DPIL-FIRE2016: Detection of Paraphrases on Indian Languages", DPIL 2016
[12] Joao Cordeiro, Gael Dais and Pravel Brazil, "A Metric for Paraphrase Detection", IEEE 2007
[13] Jun Choi Lee & Yu-N Cheah, "Paraphrase detection using string similarity with synonyms" The Fourth Asian Conference on Information Systems, ACIS 2015
[14] Saha, S., Yadav, J., & Ranjan, P. (2017). Proposed approach for sarcasm detection in twitter. Indian Journalof Science and Technology, 10(25), 1-8.
[15] Sana, P., & Avinash, S. (2017). Opinion Mining in Twitter: How to make use of Sarcasm to EnhanceSentiment Analysis. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 6.
[16] Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machinelearning algorithms in Twitter: A systematic review. International Journal of Market Research,1470785320921779.
[17] Sharma, A. Hindi Text Emotion Recognition based on Deep Learning.
[18] Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A corpus of English-Hindicode-mixed tweets for sarcasm detection. arXiv preprint arXiv:1805.11869.
[19] Wang, Z., Wu, Z., Wang, R., & Ren, Y. (2015, November). Twitter sarcasm detection exploiting a context-based model. In international conference on web information systems engineering (pp. 77-91). Springer, Cham.
[20] Wicana, S. G., İbisoglu, T. Y., & Yavanoglu, U. (2017, January). A review on sarcasm detection frommachine-learning perspective. In 2017 IEEE 11th International Conference on Semantic Computing (ICSC) (pp.469-476). IEEE.
[21] Zhang, M., Zhang, Y., & Fu, G. (2016, December). Tweet sarcasm detection using deep neural network.In Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers (pp. 2449-2460).

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details