ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

SENTIMENT ANALYSIS OF TWITTER DATA ON THEHINDI LANGUAGE

Ms. Madhuri Thorat, Sanket Bhadale, Santosh Biradar, Mansur Mujawar, Faizan Pathan

Dept. of Information Technology AISSMS's Institute of Information Technology Pune, Maharashtra, India

Abstract

Brands yearn to decode customer whispers and shouts on the bustling streets of Twitter. Sentiment analysis acts as a magic decoder ring, unlocking the emotions behind every tweet. This powerful tool reveals areas of delight, discontent, and apathy, guiding brands towards data-driven decisions. It's a swift shield against online storms, helping brands steer clear of negativity and amplify positive buzz. Embrace the whispers,understand the roar, and watch your brand soar.

Keywords: Sentiment Analysis(SA), Natural Language Processing(NLP), Lexicon- based Approach (LBA) , Hybrid-based Approach(HBA), and Machine learning Approach (MLA)

Introduction

The age of opinion mining dawns, where hidden gems of sentiment glitter within texts. Sentiment Analysis (SA) polishes these gems, revealing the emotional truth within. Twitter, a bustling bazaar of voices, hums with opinions about brands, ideas, and the world. Understanding these whispers and roars becomes crucial. LBA and MLA, two knights in the sentiment analysis quest, offer their services. LBA, wielding its lexicon dictionary, swiftly classifies tweets as positive or negative. MLA, trained on mountains of data, promises nuanced understanding. Yet, both have their chinks in the armor. LBA lacks depth, while MLA craves vast training grounds. Thus, we ventured beyond English, into the vibrant realm of Hindi tweets. We built a bridge, collecting and analyzing tweets, not just classifying them, but discerning the subtle shades of positive, negative, and neutral. This exploration delves deeper, unveiling not just

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

opinions, but the very soul of Hindi social media. So, join us on this linguistic voyage, where tweets become whispers to understand, and Twitter, a treasure trove of emotions to unlock. Let's mine the depths of Hindi sentiment, revealing the hidden gems that shape our digital world.

Literature review

The quest to capture the elusive essence of sentiment has seen researchers embark on diverse paths. Some wield dictionaries like swords, their LBA approach slicing through text to decipher positive and negative opinions. Others forge a delicate blend, wielding both LBA and MLA in a hybrid dance to uncover deeper emotional nuances. Their battleground stretches across varied datasets, from product reviews gleaming with star ratings and scathing comments, to movie critiques dancing between praise and scathing critiques, and even the pulse of elections thrumming through news and blog posts. Anjum Madan [1] In the realm of Hindi sentiment analysis, Anjum Madan et al. (2023) locked horns in a battle between two titans: Lexicon-Based Approach (LBA) and Hybrid-Based Approach (HBA). LBA, wielding its trusty SentiWordNet dictionary, valiantly sliced through text, classifying positive and negative sentiments with accuracies of 60.31% and 62.78%, respectively. HBA, on the other hand, fought with a two-pronged attack. First, LBA scouted the terrain, identifying positive and negative sentiments. Then, supervised machine learning algorithms, like the mighty Decision

Tree (DT) classifier, delivered the final blow, achieving a staggering accuracy of 92.97%! This study highlights the potential of HBA in Hindi sentiment analysis, showcasing its ability to outperform LBA alone. So, when it comes to deciphering the emotions behind Hindi tweets, HBA emerges as the clear victor.

Delving into the depths of Twitter sentiment, researchers in [3] harnessed the power of Python and Twitter APIs to craft a compelling study. Their analysis pitted diverse approaches against each other, revealing that while the Lexicon method offered a quick glimpse into emotional landscapes, its accuracy paled in comparison to the nuanced understanding gleaned from machine learning techniques. This suggests that for truly

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

deciphering the whispers and roars of Twitter sentiment, machine learning reigns supreme. In the ever-evolving landscape of sentiment analysis, authors of [4] paint a clear picture. Their in-depth survey reveals that supervised machine learning algorithms hold the reins for accuracy, outperforming both lexicon-based methods and even the burgeoning frontier of deep learning. This suggests that, at least for now, data meticulously labeled by human hands reigns supreme in unlocking the true emotional landscape of text. Hasan et al. Peering into the Twitterverse, researchers in [11] analyzed public perception of a product with laser-sharp precision. Their weapon of choice? A potent combo of Bag-of-Words (BoW) and TF-IDF techniques, slicing through tweet data to categorize positive and negative sentiments with remarkable accuracy. The TF-IDF vectorizer, the study argued, acted like a sentiment analysis supercharger, boosting accuracy and validating the proposed framework's effectiveness. In short, this research showed that when it comes to deciphering Twitter's emotional heartbeat, the BoW-TFIDF duo reigns supreme.

System Architecture

The diagram illustrates the system architecture for sentiment analysis of twitter on hindi languange. The diagram has several components that work together to analysis the sentiment of tweets.

1. Data Collection Program & API: Using the Data Collection program we get the Hinditweets from Twitter through the Twitter API.
2. Data Set: Dataset contains the data on which we'll train the model. This Data will contain tweets from various/specific twitter handles that we will analyze.
3. Data Cleaning Program: In data cleaning, we will clean the data for the model to train. This process will remove the stop words and redundant data. Here also stemming of the word will be carried out. And word will be converted.
4. Classification Program: Here we use machine we use machine learning algorithms to classify the sentences/tweets. Algorithms that perform best on the data set are selected. Here also we'll try to tune the algorithms to

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

obtain higher efficiency. Here also the grouping of significant words would be done.

5. Result: After training model on the model on the dataset we analyze to obtain results. Finally, the model has successfully classified the tweets as positive, negative, or neutral Sentiments.

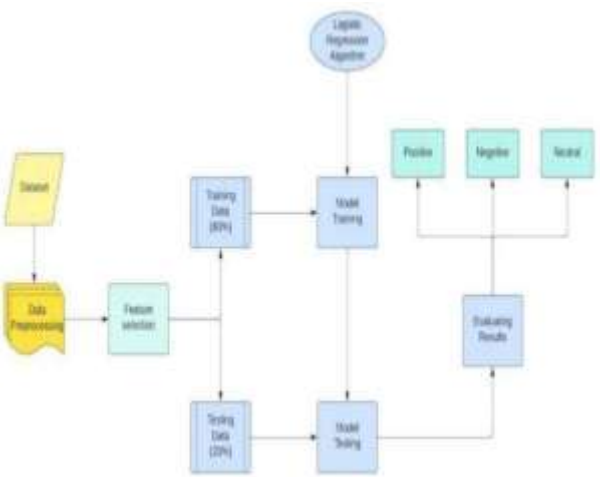

Fig.1- System Architecture Diagram

Methodology and model specifications

Step 1: Reading the dataset for tweets and its sentiments Social media creates a large quantity of sentiment data in many formats, such as tweet id, status updates, reviews, author, content, tweet type, and tweet status update.

Step 2: Pre-processing of data for filtration of redundant data The Twitter datasets utilized in this reviews work is now partitioned into two classes, negative and positive extremity, empowering feeling examination of the information clear and permitting the impact of numerous boundaries to be assessed. Irregularity and excess are normal in crude information with extremity. The accompanying focuses are remembered for tweet preprocessing:

- Removes all URLs (e. g. www.abc.com), hashtags (e. g. #name), targets(@username)

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

- Make the spellings; repeating characters must be handled.
- All of the emoticons should be replaced with their relevant feelings.
- Remove all symbols, signs, or digits.
- Remove Stop Words

Step 3: Use a machine translation tool or library to convert the Hindi tweets into English. There are several options available, such as the Google Translate API, Microsoft Translator API, or open-source libraries like TranslatePy. We here using TranslatePy Library.

Step 4: Extracting the features from data using principal component analysis We extract elements of processed datasets using the feature extraction approach.

Step 5: Creating training and testing data Generating randomly sample of 70% training data and 30% of testing data.

Step 6: Training of classifier using data Supervised learning is a useful approach for dealing with classification challenges. Training the classifier facilitates subsequent predictions for unknown data.

Step 7: Testing of data using the designed network model. Calculation of results In this step I will be graphically evaluating the results by showing which model is best suitable for this data.

Model Testing

	Algorithm	R2 Score (Train)	R2 Score (Test)	Mean Squared Error (Train)	Mean Squared Error (Test)	Root Mean Squared Error (Train)	Root Mean Squared Error (Test)	Mean Absolute Error (Train)	Mean Absolute Error (Test)	Explained Variance (Train)	Explained Variance (Test)	Max Error (Train)	Max Error (Test)
0	Logistic Regression	0.284909	0.184452	0.046681	0.053060	0.216059	0.230347	0.046681	0.053060	0.303855	0.202523	1	1
1	Naive Bayes	0.093835	0.011341	0.059155	0.064322	0.243218	0.253619	0.059155	0.064322	0.094102	0.011388	1	1
2	SVM	0.437641	0.199840	0.036711	0.052059	0.191601	0.228163	0.036711	0.052059	0.454363	0.225421	1	1
3	XGBoost	0.280436	0.203687	0.046973	0.051808	0.216734	0.227614	0.046973	0.051808	0.304903	0.225932	1	1

Fig. 2: Model Testing Result

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

Metrics Details:

1. **R2 score train :** The R2 score train in machine learning is a measure that tells us how well a trained model fits or predicts the data it was trained on. It assesses the performance of the model specifically on the training data.
2. R2 score test : The R2 score test in machine learning is a way to measure how well a trained model can predict new, unseen data. It helps us understand if the model's performance on the training data also holds up when applied to new data.
3. **Mean Squared Error (MSE) train** : Mean Squared Error (MSE) train is a common metric used in machine learning to measure the average of difference between square of the predicted values and the actual value in the training dataset. It quantifies how well a model fits the training data by calculating the average of the squared errors.
4. **Mean Squared Error (MSE) test :** Mean Squared Error (MSE) test in machine learning is a metric used to assess the performance of a trained model on new, unseen data. The average accuracy of a model's predictions on the test dataset by calculating the squared difference between its guesses and the true values, then taking the average of those squared differences.
5. **Root Mean Squared Error (RMSE) train :** This (RMSE) train in machine learning is a metric used to measure the average magnitude of the errors made by a model when predicting the values in the training dataset. It is derived from the Mean Squared Error (MSE) and provides a more interpretable value.
6. **Root Mean Squared Error (RMSE) test:** This (RMSE) test in machine learning is a metric used to measure the average magnitude of the errors made by a trained model when predicting values in new, unseen data. It is derived from the Mean Squared Error (MSE) and provides a more interpretable value.
7. **Mean Absolute Error (MAE) train:** It is (MAE) train in machine learning is a metric used to measure the average absolute difference between the predicted values and the actual values in the training dataset. It quantifies the average magnitude of the errors made by a

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

model during training.

8. **Mean Absolute Error (MAE) test:** It is test in machine learning is a metric used to measure the average absolute difference between the predicted values and the actual values in new, unseen data. It quantifies the average magnitude of the errors made by a trained model when predicting values in the test dataset.
9. **Variance Score train:** Variance Score train in machine learning, also known as explained variance score train, is a metric that indicates the proportion of variance in the training data that is explained by the model. It measures how well the model captures and accounts for the variability in the training dataset.
10. **Variance Score test:** Variance Score test in machine learning, also known as explained variance score test, is a metric that indicates the proportion of variance in new, unseen data that is explained by the trained model. It measures how well the model can generalize and explain the variability in unseen examples.
11. **Max Error train:** Max error in machine learning is like the teacher marking the model's worst mistake on a training exercise. It captures the biggest gap between the model's predicted answer and the true one, highlighting the largest misunderstanding it has about the data. While good scores focus on average performance, max error shines a light on the potential pitfalls, urging the model to improve its grasp of even the most challenging examples.

Conclusion

In this project we classifying tweets in Hindi language as Positive negative and Neutral. In this to classify tweets we are converting Hindi Tweets in English Language and then by using different Algorithm We are performing sentiment analysis of twitter data. By doing performance analysis of different algorithm we here using Logistic Regression algorithm to developed sentiment analysis Model. Following Graphs Shows sentiment score of words. Sentiment analysis of tweets helps in emotions and thoughts of users on twitter. Sentiment analysis- based tweets is also helps in businesses to monitor brand and product sentiment

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

in customer feedback and customer needs. This research dives into Twitter's emotional depths, exploring sentiment analysis through the lens of machine learning. Two distinct approaches take center stage: the first, a stalwart champion used in many classifiers, and the second, a rising star. In a thrilling turn of events, the newcomer outshines its predecessor, demonstrating superior classification prowess. This study sheds light on the power of machine learning, revealing its ability to surpass purely lexical methods in unlocking the emotional nuances of Hindi language tweets. Our team delved into various algorithms, ultimately crowning the Logistic Regression classifier as the king of Twitter sentiment analysis. Buckle up, for this research unveils a potent weapon in the battle to understand the digital heartbeat of Hindi Twitter!

References

[1] Anjum Madan and Udayan Ghose, "Sentiment Analysis For Twitter Data In the Hindi Language",2021 11th international conference on cloud computing, Data Science and Engineering, 2021.

[2] Nikita Kolambe, Yashashri Belkhede, Nikhil Wagh, "A Review on Sentiment Analysis on Hindi Language using Neural Network", international conference on cloud computing, Data Science and Engineering,2020, ISSN:-0886-9367.

[3] Kusrini and M. Madhuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019.

[4] N. F. Alshammari and A. A. Al Mansour, "State-of-the-art review on Twitter Sentiment Analysis," in 2nd International Conference on Computer Applications & Information Security (ICCAIS)., 2019.

[5] A. A. Anees, H. P. Gupta, A. P. Dalvi, S. Gopinath and B. R. Mohan, "Performance Analysis of Multiple Classifiers using different Term Weighting Schemes for Sentiment Analysis," in Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019), 2019.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

[6] Charu Nanda, Mohit Dua, and Garima Nanda "Sentiment Analysis of Movies Reviews in Hindi Language using Machine Learning" International Conference on Communication and Signal Processing,2018, India.

[7] Hasan, M. R., Maliha, M., & Arifuzzaman, M., "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication,Chemical, Materials and Electronic Engineering, 2019.

[8] Dinkar Sitaram, Savita Murthy, and Devansh Sharma, "Sentiment Analysis of Mixed language employing Hindi English code-switching" international conference on machine learning and cybernetics, 2018.

[9] Jie Li and Lirong Qui, "Sentiment Analysis Method of Short Texts in Microblog", 2017 IEEE conference on computational Science.An Analysis Method for Flight Delays based on Bayesian Network Li Qianya, Wang Lei, Fei Rong, Wang Bin, Hei Xinhong.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024