International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 12 • Issue 2 pp. 31-35 June - Dec 2020 www.csjournals.com

Sentiment Analysis in Dogri Language Using Machine Learning

Shubhnandan Singh Jamwal
jamwalsnj@gmail.com
PG Department of Computer Science and IT, University of Jammu, J&K

Abstract: Addressing the challenges of low resource languages is crucial for preserving linguistic diversity and promoting inclusion. Efforts are being made by researchers, organizations, and communities to document and revitalize endangered languages and develop language technologies specifically tailored for low resource languages. These efforts involve collecting and digitizing linguistic data, creating language resources, and collaborating with local communities to ensure the preservation and promotion of these languages. In this paper, Naive Bayes, Linear Regression and Support Vector Machine learning models are analyzed for the Dogri language. The precision, recall and F1 score are also presented in the paper.

Keywords: Naive Bayes, Support Vector Machine, Linear Regression, Dogri, Sentiment Analysis, Machine Learning

Introduction

A low resourced language lack comprehensive linguistic documentation, dictionaries, grammars, corpora, linguistic databases tools, and resources making it challenging to develop technologies such as machine translation, speech recognition, and natural language processing systems. Some low resource languages are without standardized writing systems. This absence of a writing system further complicates the process of documentation and resource development.

Dogri is an Indo-Aryan language spoken primarily in the Jammu region of Jammu and Kashmir, a state in northern India. It is also spoken by a significant number of people in the neighboring states of Himachal Pradesh and Punjab. Dogri is recognized as one of the official languages of Jammu and Kashmir. Dogri vocabulary is influenced by various languages, including Sanskrit, Persian, Punjabi, and Hindi. The grammar of Dogri is similar to other Indo-Aryan languages, with noun declensions, verb conjugations, and gender agreement. Dogri language is a low resourced language that has limited linguistic resources available for research, development, and technological support. Dogri language is spoken 11 million people as per the 2011 census. Given the wide applicability of social media platforms, individuals increasingly turn to the web to seek and share information, opinions, comments and suggestions which results in proliferation of user generated large volume of text data available for interpretation. N. Hadiya and N. Nanavati [1] observed that performing Sentiment Analysis on the Indian languages is also challenging task which is undertaken by researchers that extracts opinions from the given text and classifies them as a negative or positive. Research works in sentiment analysis is mostly conducted for English language. Nowadays, the web indexes and other websites related to reviews also support non-English languages. It is therefore necessary to perform sentiment analysis for these languages as well. There are numerous works found in the literature for sentiment analysis in other languages worldwide. However, sentiment analysis for Indian languages needs exploration.

Review of Literature

G. I. Ahmad, J. Singla and N. Nikita [2] observed that a large number of users in India write their feelings or emotions in more than one language; thereby a large volume of text data is made available for Natural Language Processing (NLP) researchers. Sentiment Analysis (SA) of code-mixed text provides useful information in the field of politics, marketing, business, health, sports and what not. During the past decade the work on Sentiment Analysis of Indian language textual data, particularly in Hindi has got momentum in contrast to code-mixed Indian language text. However, due to non-availability of language and vocabulary (linguistic and lexical) tools and annotated resources, the task of Sentiment Analysis of Indian Languages becomes somehow difficult. They studied a detailed summary of Sentiment Analysis of Indian languages with a special focus on code mixed Indian Languages.

P. Impana and J. S. Kallimani [3] observed that sentiment analysis involves computational identification of opinion on given dataset which is also referred as the extraction of the opinion. Cross lingual Sentiment Analysis refers to the generation of the opinions in two languages. One language is highly rich in its resources providing sufficient dataset required for the opinion extraction known as Resource Rich Language. The other languages such as Kannada, Hindi, Marati which are poor in its resources and lacks in the data Wordnet and seeks the help of resource rich languages for the opinion extraction known as Resource Poor Language. They used architecture of auto encoder which helps in the generation of the sentiment analysis in two languages. Sentiment Analysis of two languages can be performed by using the Bilingually Constrained Recursive Auto-encoder (BRAE) model and also with the help of linked Wordnet datasets.

R. Naidu, S. K. Bharti and K. Sathya Babu[4] observed that the challenges of sentiment analysis in Indian languages are due to rich morphology and little availability of the annotated datasets. They developed SentiWordNets (SWNets) to tag the sentiment of each word for languages like Hindi, Telugu, Tamil, Bengali, Malayalam and observed that some unigram words of the existing Telugu SWNet are classified as ambiguous and are not sufficient to analyze the sentiment. In such situations, bigram and trigram phrases can be used to resolve the problem of ambiguity in sentiment prediction and proposed an algorithm to build the Telugu SentiPhraseNet (SPNet) for the sentiment analysis in Telugu. To build SPNet, they collected the data from various sources namely, Telugu e-Newspapers, Twitter and NLTK Indian Telugu data which resolve the problems with existing SWNet. With the proposed SPNet, they performed the sentiment analysis and it is compared with SWNet, various existing Machine Learning approaches namely, Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Multilayer Perceptron Neural Network (MLPNN), Decision Tree (DT) and Random Forest (RF) and observed that the performance of the proposed system outperformed the other existing approaches and attains an accuracy of 85.6%. K. Chakraborty, R. Bag and S. Bhattacharyya [5] presented the works that has already been done regarding analysis of sentiments in most available yet underrated Indian languages. Other than this, the probable problem domains are explored which can be solved in the area of multilingual sentiment analysis.

P. Sharma and T. -S. Moh [6] used Twitter Archiver tool to get tweets in Hindi language and performed data (text) mining on 42,235 tweets collected over a period of a month that referenced five national political parties in India, during the campaigning period for general state elections in 2016. They made use of both supervised and unsupervised approaches and utilized Dictionary Based, Naive Bayes and SVM algorithm to build our classifier and classified the test data as positive, negative and neutral. They identified the sentiment of Twitter users towards each of the considered Indian political parties. The results of the analysis for Naive Bayes was the BJP (Bhartiya Janta Party), for SVM it was the BJP

(Bhartiya Janta Party) and for the Dictionary Approach it was the Indian Nathional Congress. SVM predicted a 78.4% chance that the BJP would win more elections in the general election due to the positive sentiment they received in tweets. As it turned out, BJP won 60 out of 126 constituencies in the 2016 general election, far more than any other political party as the next party (the Indian National Congress) only won 26 out of 126 constituencies. S. Tammina [7] in their research illustrated a methodical approach which leverages lexicon based approach and machine learning in the field of sentiment analysis to classify the opinions in Telugu language. Firstly, by employing Lexicon based approach - Telugu SentiWordNet we identified the subjective sentences from the Telugu corpus. Secondly, by utilizing machine learning algorithms - SVM, Naïve Bayes and Random Forest they categorized the sentiment in the corpus. Our proposed methodology achieved highest accuracy of 85%. K. Sarkar [8] presented a sentiment polarity detection approach that detects sentiment polarity of Bengali tweets using character n-gram features and a supervised machine learning algorithm called Multinomial Naïve Bayes. The proposed approach has been tested on the SAIL 2015 dataset and the experimental results show that character n-gram features are more effective than the traditional word n-gram features. L. G. Singh and S. R. Singh [9] proposed various syllabic features by exploiting syllable-to-syllable relationships and investigate its effect on word polarity detection over a limited word corpus for Manipuri language (a resource-poor language spoken in Manipur, a state in North Eastern part of India). The extracted features are subjected to various classification frameworks. From various experimental observations, it was evident that syllable-to-syllable relationship based features outperform its word-to-word relationship based counterparts. Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F. Smeaton [10] proposed text extraction techniques to create topic-specific sub-documents, which were used to train a sentiment classifier. They showed that such approaches provide a substantial improvement over full document classification and that word-based approaches perform better than sentence-based or paragraph-based approaches.

Machine Learning models in Analyzing Sentiments

Machine learning models are widely used in sentiment analysis, a natural language processing (NLP) task that involves determining the sentiment or opinion expressed in a piece of text. There are several machine learning algorithms and models that can be used for sentiment analysis, depending on the specific requirements and characteristics of the problem. The various models which are used for the sentiment analysis are as follows:

a) Naive Bayes: Naive Bayes is a probabilistic classifier commonly used in sentiment analysis. It calculates the probability of a particular sentiment label given the features (words or n-grams) present in the text and it can be used when the data is not extremely complex and the features are adequately informative. Naive Bayes assumes independence between the features, which simplifies the computation and is a popular algorithm used in sentiment analysis due to its simplicity, efficiency, and effectiveness.

b) Support Vector Machines (SVM): SVM is a supervised learning algorithm that can be used for sentiment analysis. It finds an optimal hyperplane to separate the positive and negative sentiment classes based on the input features. SVM can handle high-dimensional feature spaces and is effective when the data is not linearly separable.

c) Logistic Regression: Logistic regression is a popular classification algorithm used in sentiment analysis. It models the relationship between the input features and the probability of a particular sentiment class using a logistic function. Logistic regression can handle both binary and multi-class sentiment classification problems.

Data Preparation:

The data set used in the paper is composed of the 1998 sentences of Dogri. The data set has been taken from online resources of other languages which was not available in Dogri. The data set was available in English language which was converted manually to Dogri. Before the data can be put to training the following steps are also performed on the data:

Text Preprocessing: Before applying any machine learning algorithm the text data needs preprocessing. This involves steps like tokenization, removing stop words, stemming or lemmatization, and sometimes handling negations or emoticons to extract features that represent the sentiment.

Vectorization and Feature Extraction: Once the text is preprocessed, it needs to be converted into a format that machine learning algorithms can understand. This usually involves creating a numerical representation of the text data and bag-of-words model is used for the purpose where each sentence is represented as a vector of word frequencies or presence/absence indicators.

Sentiments Analysis in Dogri Language

The experiment is conducted for comparing the performance of Naïve Bayes, LR and SVM (Support Vector Machine) classifiers in analyzing the sentiment in the text of the Dogri language. The text is classified into three categories that is Positive, Negative, and Unknown State. The training data is composed of 1998 samples and test data is composed of 395 samples. The precession, recall and F1 score of the three classifiers is tabulated as under.

Machine Learning Model	Precision	Recall	F1 Score
Naive Bayes	0.65	0.61	0.63
LR	0.71	0.70	0.69
SVM	0.72	0.70	0.68

Table 1: Experimental Data

Results and Observations

In the experiment, it is observed that training an SVM is generally faster as compared to Naïve Bayes and LR because of its ability to handle high-dimensional feature spaces and their flexibility in capturing complex relationships between features and labels.

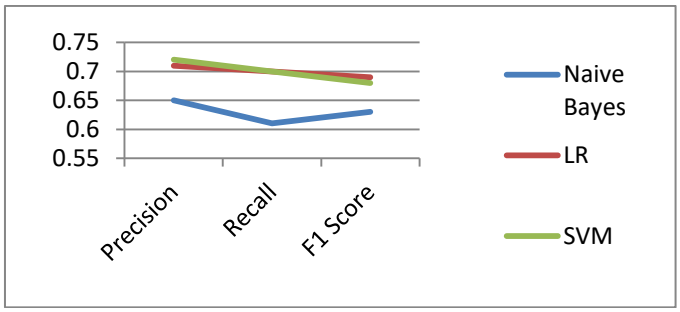

Fig1: Pictorial Representation of the Results

The training of LR and Naïve Bayes is slower than SVM. Naive Bayes is simple, efficient, and good for quick sentiment analysis tasks but may lack sophistication in capturing complex relationships. Logistic Regression provides probabilistic interpretation, feature importance, and good efficiency but assumes linear relationships and SVM can capture complex relationships using non-linear kernels.

References

[1] N. Hadiya and N. Nanavati, "Indic SentiReview: Natural Language Processing based Sentiment Analysis on major Indian Languages," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 322-327, doi: 10.1109/ICCMC.2019.8819786.

[2] G. I. Ahmad, J. Singla and N. Nikita, "Review on Sentiment Analysis of Indian Languages with a Special Focus on Code Mixed Indian Languages," 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 2019, pp. 352-356, doi: 10.1109/ICACTM.2019.8776796.

[3] P. Impana and J. S. Kallimani, "Cross-lingual sentiment analysis for Indian regional languages," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, India, 2017, pp. 1-6, doi: 10.1109/ICEECCOT.2017.8284625.

[4] R. Naidu, S. K. Bharti and K. Sathya Babu, "Building SentiPhraseNet for Sentiment Analysis in Telugu," 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 2018, pp. 1-6, doi: 10.1109/INDICON45594.2018.8987162.

[5] K. Chakraborty, R. Bag and S. Bhattacharyya, "Relook into Sentiment Analysis performed on Indian Languages using Deep Learning," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2018, pp. 208-213, doi: 10.1109/ICRCICN.2018.8718709.

[6] P. Sharma and T. -S. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016, pp. 1966-1971, doi: 10.1109/BigData.2016.7840818.

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

[8] K. Sarkar, "Using Character N-gram Features and Multinomial Naïve Bayes for Sentiment Polarity Detection in Bengali Tweets," 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, India, 2018, pp. 1-4, doi: 10.1109/EAIT.2018.8470415.

[9] L. G. Singh and S. R. Singh, "Word polarity detection using syllable features for manipuri language," 2017 International Conference on Asian Language Processing (IALP), Singapore, 2017, pp. 206-209, doi: 10.1109/IALP.2017.8300580.

[10] Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. 2009. Topic-dependent sentiment analysis of financial blogs. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (TSA '09). Association for Computing Machinery, New York, NY, USA, 9–16. https://doi.org/10.1145/1651461.1651464