Twitter Sentiment Analysis

Swetabh Suman , Anmol Kumar , Shyam Sundar Mishra , Raj Mishra

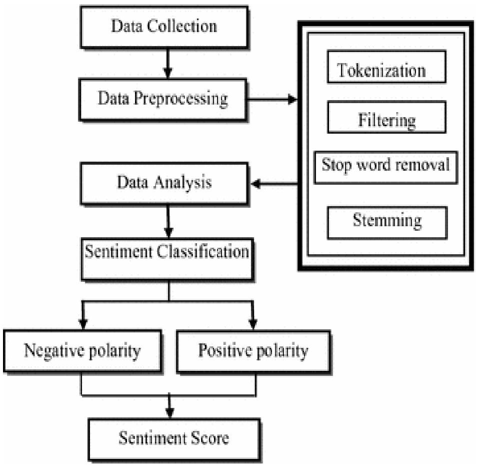
Abstract **Sentiment analysis is one of the Natural Language Processing fields, dedicated to the exploration of subjective opinions or feelings collected from various sources about a particular subject. Sentiment Analysis is a set of tools to identify and extract opinions and use them for the benefit of the business operation.**

I. INTRODUCTION

Sentiment analysis is the process of classifying whether a block of text is positive, negative, or neutral. Sentiment analysis is contextual mining of words which indicates the social sentiment of a brand and helps the business to determine whether the product which they are manufacturing is going to make a demand in the market or not. The goal which Sentiment analysis tries to gain is to analysed people's opinion in a way that it can help the businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It is important because huge data are produced by humans every day and most of the data are unstructured. So, it is difficult as well as time-consuming to understand. That's why Sentiment analysis makes it easy for us to shed light on the unstructured data using automated methods and algorithms.

Approaches for performing sentiment analysis:

- **Lexicon based approach:**Count number of positive and negative words in each text and the larger count will be the sentiment of the text.



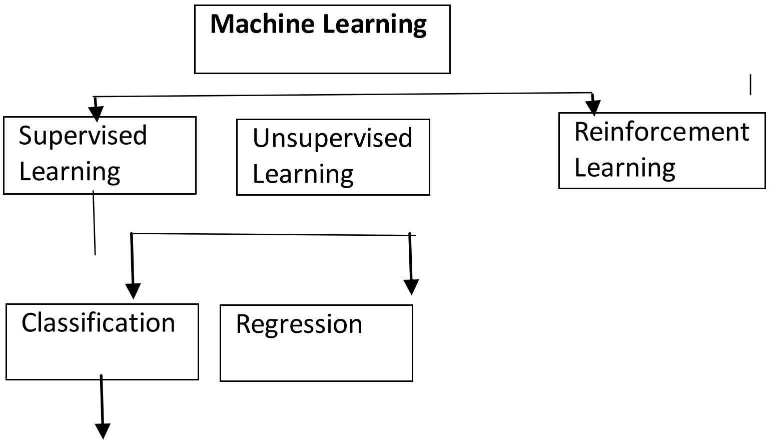
– Pre-processing. The text undergoes pre-processing steps that is POS tagging, stemming, stop-words removal, tokenisation into N-grams. The outcome of the pre-processing is a set of tokens ora bag-of-words.

– Checking each token for its polarity in the lexicon. Each word from the bag of-words is compared against the lexicon. If the word is found in the lexicon, thepolarity W_i of that word is added to the sentiment score of the text. If the word isnot found in the lexicon its polarity is considered to be equal to zero.

– Calculating the sentiment score of the text. After assigning polarity scores toall words comprising the text, the final sentiment score of the text is calculated bydividing

$$\text{Score}_{\text{AVG}} = \frac{1}{m}\sum_{i=1}^{m} W_{\text{i}}$$

- **Machine learning-based approach:** Develop a classification model, which is trained using the pre-labelled dataset of positive, negative, and neutral.



Support Vector Machines
Discriminant Analysis
Naïve Bayes
Nearest Neighbour

Fundamentals of Supervised Sentiment Analysis

Sentiment analysis is used to identify the affect or emotion (positive, negative, or neutral) of the data. For a business, it is a simple way to determine customers' reactions towards the product or service and to quickly pick up on any change of emotion that may require immediate attention. The most basic approach to this problem is to use supervised learning. We can have actual humans to determine and label the sentiment of our data and treat it like a text classification problem.

Supervised learning is a machine learning approach that is defined by its use of labelled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labelled inputs and outputs, the model can measure its accuracy and learn over time.

Supervised learning can be separated into two types of problems when data mining: classification and regression:

- **Classification** problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Or, in the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox. Linear classifiers, support vector machines, decision trees and random forest are all common types of classification algorithms.

Naive Bayes classifier is the one which is most commonly used for classification of data in Sentiment Analysis.

Naive Bayes classifier is a supervised machine learning approach. This supervised classifier was given by Thomas Bayes and hence the name. According to this theorem, suppose there are two events say, p1 and p2 then the conditional probability of occurrence of event p1 when p2 has already occurred is given by the following mathematical formula:

$$\boxed{P(p1|p2) = P(p2|p1)P(p1)/P(p2)}$$

Where A = Sentiment, B=Sentence
And the conditional probability of a word is given by-

$$\boxed{P(word|A)=C+1/(D+E)}$$

C=no. of word occurrence in class
D= no of words belonging to a class
E= total no. of words

- **Regression** is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.

Applications

Sentiment Analysis has a wide range of applications as:

Social Media: If for instance the comments on social media side as Instagram, over here all the reviews are analysed and categorized as positive, negative, and neutral.
Customer Service: In the play store, all the comments in the form of 1 to 5 are done with the help of sentiment analysis approaches.
Industry Sector: In the industry area where a particular product needs to be reviewed as good or bad.
Reviewer side: All the reviewers will have a look at the comments and will check and give the overall review of the product.

Challenges of Sentiment Analysis

There are major challenges in sentiment analysis approach:

If the data is in the form of a tone, then it becomes difficult to detect whether the comment is pessimist or optimist. If the data is in the form of emoji, then you need to detect whether it is good or bad. Even the ironic, sarcastic, comparing comments detection is hard. Comparing a neutral statement is a big task.

Motivation

Tremendous amount of work has been done on Sentiment Analysis of tweets using Machine Learning techniques, such as Naïve Bayes used for sentiment classification, BFTREE algorithm used for sentiment prediction etc. Twitter is a social networking service on which users post and interact with messages known as "tweets" and it's creates huge data which helps us to analyse that tweets are positive or negative or neutral. These opinions are important in many business-related decisions and even political sentiments about a candidate.

Nowadays, many industries are developing more robust machine learning models capable of analysing bigger and more complex data while delivering faster, more accurate results on vast scales. Machine learning tools enable organizations to identify profitable opportunities and potential risks more quickly.

Types of Sentiment Analysis

Fine-grained sentiment analysis: This depends on the polarity based. This category can be designed as very positive, positive, neutral, negative, very negative. The rating is done on the scale 1 to 5. If the rating is 5 then it is very positive, 2 then negative and 3 then neutral.

Emotion detection: The sentiment happy, sad, anger, upset, jolly, pleasant, and so on come under emotion detection. It is also known as a lexicon method of sentiment analysis.

Aspect based sentiment analysis: It focuses on a particular aspect like for instance, if a person wants to check the feature of the cell phone then it checks the aspect such as battery, screen, camera quality then aspect based is used.

Multilingual sentiment analysis: Multilingual consists of different languages where the classification needs to be done as positive, negative, and neutral. This is highly challenging and comparatively difficult

Paper Name/year	Techniques used	Corpus/Datasets	Features Used	Accuracy/Performance	Standard tools used for implementation(if any)	Observations
1.English						
Sentiment Analysis of twitter data/2016	SVM, Naïve Bayes, CoTraining SVM, Deep Learning	Twitter Data/WordNet	emotional dictionary or sentiment lexicon.	SVM with unigram-76.08 SVM with bigram-77.73	Opnion,View, Sentiment,Belief	SVM and Naïve Bayes have highest acuuracythus,more accurate results can be obtained.
Analyse the sentiment of different groups of people, namely politi- cians, doctors, comedians, motivational speakers and graduate students/2019	Twitter API is used to collect tweets to create a dataset	Twitter tweets from the group comprised of doctors, politicians and Comedians(corpus of 3200 tweets were created).	lexical features, crowd wisdom approach, coding analysis	doctors had the highest median sentiment score of 0.38 Graduate students displayed the lowestmedian sentiment score of 0.17	VADER library, Normalize and correct sp0elling	politicians have a similar propensity for positive and negative tweets. Comedians' tweets have the best positive median score, while graduate students' tweets have the least positive median score among the tested groups.
2.Hindi						
Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter/2016	Dictionary Based, Naive Bayes and SVM algorithm	tweets with the names of Indian political parties such as #BJP, #Congress, #NCP, #AAP. We collected a total of 23,998 tweets relevant to these hashtags.	classification algorithms Naive Bayes, Support Vector Machine and unsupervised approach as Dictionary based	Naive Baye's algorithm = 62.1% Support Vector Machine was =78.4%	a lexical resource of Hindi SentiWordNet (HSWN) was created, utilizing its English format	Final prediction is made by utilizing SVM, since the accuracy of the algorithm is higher. We predicted that the party thathad a better chanceof winning the 2016 general election is BJP.
Practical Approach to Sentiment Analysis of Hindi Tweets	Subjective Lexicon , N-gram modeling ,Machine Learning,Proposed Algorithm	hindi tweets with the help of twitter archiever	Subjective Lexicon • N-gram modeling • Machine Learning	Average accuracy of #jaihind and #worldcup2015 =75.39	Result of our approach is compared with Unigram count the words with positive and negative polarity and choose the one Presence Method in which we with dominating polarity.	we count the positive and negative words in the tweet and choose the dominating one. The results indicate that proposed algorithm gave better accuracy
3.Punjabi						
Sentiment Analysis of Twitter User Data on	Naive Bayes algorithm,Python language	Twitter API	Pattern Analyzer (based on the pattern library) and NaiveBayesAnalyzer.	Overall accuracy is 90.29%.	classification of tweets in the different class (positive and negative),library of Python called Textblob.	AAP has more negative tweets as compared to other two parties (i.e. INC

Punjab Legislative Assembly Election/2017						and BJP-Akali), LIWC result what we found is that all the parties show more positive emotions as compared to negative words in the tweets.
Analyzing research work done for mining sentiments written in Punjabi language/2017	Lexicon based, N-grams modelling, ML	Manually developed a seed list of Punjabi words, Punjabi websites, newspapers and Punjabi blogs	*Linguistics:* Positive/ Negative Polarity. *Linguistics:* Joy, Sadness, Fear, Surprise, Disgust and Anger	accuracy of 54.2%	determining the polarity of given text at the document level, sentence level or the feature/aspect level and it could be an emotional state also such as „angry", „happy" or „sad"	The availability of linguistic resources for Punjabi language is very scarce such as automatic tools for tokenization, feature selection and stemming etc. Problems like Word Sense Disambiguation, word order, co-reference of words, morphological variations etc. need to be worked out.
4.Bengali						
Analyze the sentiment from Bangla text	rule-based algorithm ,lexicon data dictionary (LDD)	active word list in restaurantdataset ,contradict word list in cricketdataset.	UniGram and BiGram features	accuracy is attained in both dataset 80.58% and 82.21%	1. specific domain-based categorical weighted LDD for analyzing sentiment classification from Bangla dataset. 2. To develop a novel and effective rule-based algorithm for detecting sentence polarity classification by extracting score from a chunk of Banglatext.	This analysis shows that cricket data has higher accuracy than restaurant dataset, because cricket dataset has trained more data than the restaurant data. Every dataset has its owned variabilities. If we use i.e., fifty (50) thousand dataset in our machine learning process, our result will predict more accuracy than the obtained accuracy with the current dataset.
Sentimental Analysis of Bengali Language/2016	*Multinomial Naive Bayes classifier*	*SAIL dataset*	*simple, robust, scalable, and language-independent*	*accuracy of 54.05%*	Word n-grams, Character n-grams ,Surface features, SentiWordNet features.	We obtained 51.25% for Bengali compared to 43.2%, 56.96% for Hindi compared to 55.67%, and 45.24% for Tamil compared to 39.28% – in the constrained version of the task. for Bengali we used the 2-class model, because Bengali development data did not have any samples from the "neutral" class.

5.Tamil						
Sentiment classification of Tamil movie tweets using synaptic patterns	Corpus categorisation ,Creating Domain Specific Tags,Calculation of Accuracy using TF and IDF scores,length cutoff.	Own corpus developed .7418 tweets about 100 movies	TF-IDF values,newly created lexicon(DST) and sentence length.	TF-IDF-29.87% TF-IDF+DST-35.64% Tweet weight age model-40.07%		Word level feature representation using TF-IDF gives notable improvement compared to feature representation technique using character-level give best result when compared to other classifiers.
Sentiment analysis of Tamil movie reviews via feature frequency count	Google Translate ,Opinion Lexicon Extraction Module , Key Sentence Extraction Module, Supervised Learning Module Result Analysis Module	Own corpus developed using websites. 1160 positive and 1160 negative reviews	Features: TF of unigram and bigram. Classifiers: MNB, BNB, Logistic regression, RKS and SVM.	the machine learning algorithms such as Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), Logistic Regression (LR), Random Kitchen Sink (RKS) and SVM	SVM accuracy-64.69% (Using bi-grams	Neutral reviews in both classes, unstructured Corpus, implicit meaning, negations, stop words, short abbreviations and unlemmatised words. RKS yielded lowest when compared to MNB and SVM. MNB yielded highest time of all classifiers.

Future Scope

Sentiment analysis can be used on a wide variety of topics to figure out the sentiments of the general public about any matter. It has a wide variety of applications from getting reviews on a new movie, or a series to more serious applications like finding out the temperament of the public on any new bill or overall satisfaction with the government. Business Intelligence is a field which regularly uses sentiment analysis for figuring out people's emotions about the products to help corporate bodies take decisions to earn better profits and make their consumers happier.

All of it can be achieved by carefully selecting the keywords and extracting the related tweets.

Majorly the project can be divided into two parts:

- Efficiently collecting as much relevant data as possible.
- Analysing that data to predict the sentiments.

Right now, we are just collecting the data from twitter itself. We are not using any data from Facebook. We can include that. Along with the additional data from Facebook we can also explore various personal blogs and scrape relevant information from them. The better our data extraction techniques are, the better results we will be getting.

We can expand our field of search based on the type of keywords being searched for. For example, if someone is looking for sentiments of users about a laptop or a mobile phone, then the application should also scrape data from reviews written by various users on websites like amazon, flipkart etc. who have already bought the product.

We want to create a simple sentiment analysis web application with a very easy UI so that a general person with no technical background can use it and make sense of the results. We are inclined to go in the direction of using this tool for comparison of various products.

Suppose a person wants to buy a car and is confused between 3-4 models, then he should be able to use our tool in order to analyse the sentiments of other users actually using those models as an additional factor to make his decision.

Bibliography

[1] Amitava Das, Sivaji Bandopadaya, SentiWordnet for Bangla, KnowledgeSharing Event -4: Task, Volume 2,2010

[2] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326