Sentimental Analysis of Twitter Data

Bibin N
Department of Computer Science Eng
Lovely Professional University
Phagwara, Punjab, India.

ABSTRACT
With the growth of technology there are lot of data present in internet today. People started sharing their views, ideas, thoughts and even reviews on internet. There are lot of negative and positive reviews present. Sentimental Analysis technique can be used on those comments and reviews to understand the state of emotion of the reviewer. Social media has become the main tool where people share their thoughts. There are many social media networks but we are concentrating only on Twitter. The main idea of this project is to apply sentimental analysis on twitter data and find the emotion. In this project we use k-Nearest Neighbour algorithm.

KEY WORDS
Twitter, Sentimental, KNN, Text mining, NLP, Machine Learning, NB, K-Means

INTRODUCTION
Sentimental analysis is a type Natural Language Pre-processing technique which is used to determine the mood of a person about certain topic through his comments. Sentimental analysis is also called as opinion mining because it gives the opinion of a person.

NLP is Natural Language Pre-processing is a field of computer science that deals which interaction between computer language and human language. The text comment is transferred to a computer language by using Natural Language Pre-processing. The information provided by the NLP is used by our machine learning model to determine whether the statement is positive or negative.

Some challenges in NLP involve:

1. Natural language understanding.
2. Enabling computers to derive meaning from human or natural language inputs
3. Natural language generation.

Modern NLP is based on machine learning which makes solves the previous challenges faced by prior NLP. The prior NLP involves direct hand written coding of large sets and rules. But in modern NLP, instead of using general algorithms, it learns through already existing real-world examples.

LITERATURE SURVEY
Sentimental analysis has a huge influence in different industries today. Identifying the variation and context of the speech is the main challenges faced by most of the authors. Some of the research works are:

[1]. JALAJ S. MODHA and PANDI SANDIP J. MODHA have done their research on sentimental analysis on unstructured data.

Their approach:

1. Firstly, they separated sentences of documents into two types such as Opinionated and Non- Opinionated, without considering whether it is subjective or objective.

2. Secondly, they take the opinionated sentences and classify them as subjective sentences and Objective sentences.

3. Thirdly, they classify subjective sentences into positive, negative or neutral category. They use semantic orientation on complex sentence.

4. Fourthly, they classify objective sentences into positive, negative or neutral category. They use semantic orientation on complex sentence.

[2]. Aliza Sarlan, Shuib and Chayanit have done their research on twitter data analysis

In this experiment, the tweets were collected in Jason format and then they used python lexicon dictionary and assign the desired polarity to the extracted tweets. They used Support Vector Machine algorithm in their model (SVM).

[3]. Agarwal, Vovshaa, Rambow, conducted an experiment in which they firstly created a unigram model and used it as a baseline model and compared with other two models such as 1. model based on features. 2. model based on kernel tree. From these experiments they came to know that feature-based model was the best compared to unigram model and kernel tree model. In this experiment the kernel tree model very less accurate when compared with other two models.

[4]. Akshi Kumar and Teeja Mary Sebastian used the combination of lexicon based and corpus-based approach. This is a rarely used combination technique in machine learning. They have used adjectives and verbs as their features and for finding the semantic orientation of various adjectives present in the tweets, they have used corpus-based techniques. They have used lexicon dictionary for verbs. At last, they used an equation to find the polarity.

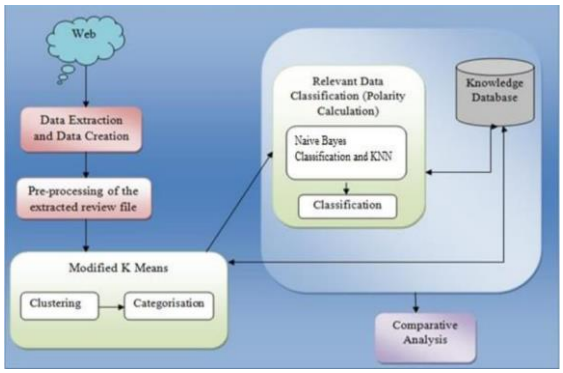
[5]. Harb conducted an experiment in which he used a document-level sentimental extraction approach. He completed the experiment in three steps.

1. Firstly, extracted a dataset from the Internet consists of documents containing opinions.

2. Secondly, learning dataset is provided to the model which contains positive and negative adjective sets.

3. Thirdly, test set documents are provided to the model and the accuracy is tested. Experiments were conducted numerous times on real data and F1 score of 0.717 for identifying positive documents and an F1 score of 0.622 for recognizing negative records.

PROPOSED METHODLOLGY



1. DATA COLLECTION

First way is data collection that is we must collect some organised data from Kaggle or any other websites. These websites provide with some organised datasets that can be used by developers to train their model. Or else one can also extract data directly from twitter using API.

2. DATA PREPROCESSING

The data that we get from these sites are not arranged properly or contain some unwanted data that are not needed for our model. Such unwanted data can affect the accuracy of our model. So, some manual data manipulations

are to be done to the datasets. And these types of manipulations are called data pre-processing.

3. FEATURE SELECTION

Modified K-Means

The popular clustering algorithm in machine learning is K-Means algorithm. This algorithm is very precise when it comes to handle small data. But the algorithm we used in this experiment is capable and good at handling large datasets. And the algorithm is called modified K-Means algorithm.

Algorithm

Let's consider,

The dataset to be D= {d1, d2, d3…. dn}, where n is the given num of data.

k be the number of clusters.

1. Calculate the distance between each data point and all other data points from the given dataset D.

2. Find the closest pair of data points from the set D, and add it the new set called A. Now, delete those two data points from the set D. A(1<=p<=n+1), p is number of data points in set A.

3. Find the data point in D that is closest to set A and add that to A and delete that data point from set D.

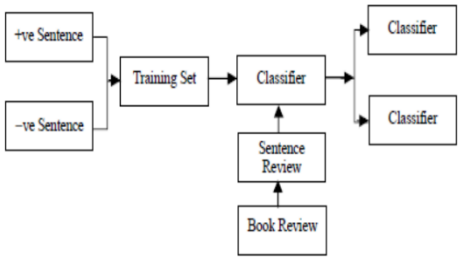
4. Repeat step 3 till the number of data points in the set A reaches n. (p=n).

5. If p<n+1, then p = p+1, find another pair of data points from D that is closest to set A and repeat the step 3.

4. MODEL SELECTION

Navies Bayes

We have used Navies Bayes (NB) classifier algorithm in this experiment. This algorithm finds the probability of the event if the probability of an already occurred similar event is given. This algorithm works perfectly well on linear separable problems. And performs decently well on non-linearly separable problems. The biggest advantage of using this algorithm is that it is very easy to interpret.



K-Nearest Neighbour

We have used K-Nearest Neighbour algorithm to train our model. It is the easiest Machine Learning algorithm based on Supervised Learning.

Algorithm

1. K-NN algorithm finds the similarity between the new data and already existing data and add the new data into the group that is most similar than the other groups.

2.K-NN algorithm is most commonly used for classification problems.

3. K-NN is a non-parametric algorithm because this algorithm does not make assumption using the given data.

4. K-NN is called a lazy learner algorithm because this algorithm does not learn when the training data is provided instead it stores the data and learns only when the classification is done.

5. MODEL EVALUATION

The best way to do model evaluation is by using confusion matrix. The following confusion matrix is used in our experiment for evaluation.

CONFUSION MATRIX

	PREDICTED CLASS 1 (p)	**PREDICTED CLASS 2 (x)**	**PREDICTED CLASS 3 (n)**
ACTUAL CLASS 1 (p)	True positive **(tp)**	False neutral **(fx1)**	False negative **(fn1)**
ACTUAL CLASS 2 (x)	False positive **(fp1)**	True neutral **(tx)**	False negative **(fn2)**
ACTUAL CLASS 3 (n)	False positive **(fp2)**	False neutral **(fx2)**	True negative **(tn)**

Accuracy

Accuracy is used to determine how accurately how model has predicted the result.

$$\text{Accuracy} = \frac{tp + tx + tn}{tp+tx+tn+fp1+fp2+fx1+fx2+fn1+fn2}$$

RESULTS ANALYSIS

By using K-Nearest Neighbour algorithm and Navies Bayes algorithm we have done our experiments and have turned out to be working well. By using these algorithms, we got higher accuracy compared to other approaches.

	PREDICTED POSITIVE	**PREDICTED NEUTRAL**	**PREDICTED NEGATIVE**
ACTUAL POSITIVE	1823	36	112
ACTUAL NEUTRAL	40	215	19
ACTUAL NEGATIVE	87	44	1284

Accuracy = 0.908

Percentage Accuracy = 90.8 %

CONCLUSION AND FUTURE WORKS

In this paper, I have done sentimental analysis on twitter data. I have used modified K-Means algorithm for clustering and Navies Bayes along with KNN algorithm for classification. The model that I have trained is 90 percentage accurate and can be used to predict twitter tweets emotion. For sentimental analysis model 90 percentage is a huge success and can be preferred over the other models.

In future, I would like to work more on this project and further improve the accuracy of the model by using some advanced concepts of deep learning such as neural networks.

REFERENCES

[1] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, "Opinion Mining on Clustering Product Features", WSDM'11, February 9–12, 2011, Hong Kong, China.

[2] Singh and Vivek Kumar, "socio-political analysis using opinion mining", Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[3] Alexander Pak and Patrick Paroubek, "Sentiment Analysis on twitter data using Opinion Mining technique"

[4] Bing Liu. "Sentiment Analysis ", Morgan & Claypool Publishers, May 2012.

[5] V. S. Jagtap and Karishma Pawar, "Sentence-Level Sentiment Classification by different approaches and comparing them", International Journal of Scientific Engineering and Technology (ISSN: 2277-1581) Volume 2 ,1 April 2013

[6]. K. Bun and M. Ishizuka, "TF*PDF algorithm is done on News Topic extraction to predict the emotion", In Proceedings of Third International Conference on Web Information System Engineering.

[7]. Jacques Savoy, Olena Zubaryeva, "Specific Vocabulary by using Classification algorithm" published in 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 4/11 2011 IEEE

[8] Jalaj S. Modha, Pandi Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data using semantic orientation", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.

[9] R M. Chandrasekaran, G.Vinodhini, "Sentiment Analysis and Opinion Mining: A Survey with people", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012

[10] Bing Liu., "Sentiment Analysis", Morgan & Claypool Publishers, published in May 2012.

[11] Aliza Sarlan, Shuib and Chayanit have done their research on twitter data analysis.

[12] Arun. K, Sinagesh. A & Ramesh. M, "sentiment analysis on twitter data about Tweets during demonetization", international journal of computer engineering in research trends, vol.4, no.6, (2017), pp.252-258".