International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

SENTIMENT ANALYSIS ON TWITTER DATA USING LEXICON-BASED AND NAÏVE BAYES APPROACH

Akshata Bahulekar[1] and Prof. Meenakshi Garg[2]

[1]Student, [2]Assistant Professor
[1]Department of MCA, [2]Department of MCA
[1]Vivekanand Education Society's Institute of Technology (VESIT), University of Mumbai, India.
[2]Vivekanand Education Society's Institute of Technology (VESIT), University of Mumbai, India.

ABSTRACT: Keeping current status of rapidly growing technology in mind - Internet has become a promising platform for online learning, exchanging ideas and sharing opinions. Social media contain tremendous measure of the assessment information as tweets, websites, and updates on the status, posts, and so on. In this paper, the most famous micro blogging platform twitter is utilized. Twitter sentiment analysis is used for analyzing the information from Twitter (tweets), to fetch user's opinions, thoughts and sentiments. The primary objective is to investigate how text analysis strategies can be utilized to discover a part of the information in a series of posts focusing on various patterns of tweets, languages, tweets volumes on twitter.

Keywords –Sentimental Analysis, Machine Learning, Natural Language Processing, Lexicon-based

INTRODUCTION

In today's world social networking sites plays a vital role in human life in that applications like Facebook, Instagram and Twitter are more famous ones. Amongst all these apps Twitter is widely used by people to express their thoughts, opinions in day to day life. People post their thoughts on a variety of topics, discusses current affairs, complains and expresses their sentiments for products and services they use.

Sentimental analysis is the process of deriving the quality data from the content. In other words, it is the way toward getting the organized information from unstructured information. This is used to measure opinion of the customers, criticism, reviews, feedback. Unstructured information not just refers to the tables, figures from the organizations but also additionally comprises of data from the web for example talks, E-mail, pdfs, word records, E-Commerce sites and social networking sites.

Monitoring Twitter permits organizations to understand their audience, keep on head of what's being said about their image and their competitors, and find new patterns in the business.

On structured information analytical activity can be easily performed and the result can be acquired without any problem. Contrary if the data is unstructured (eg. E-mail, Twitter), it is very difficult to analyze it and to conclude or to take any decision. To convert this unstructured data into structured form Natural Language Processing (NLP) and Data Mining procedures are mostly used. In this paper, we will use Twitter data for Sentiment Analysis.

The paper consists of 2 approaches:

1. Lexicon Approach
2. Machine-learning Approach(Naïve-Bayes Algorithm)

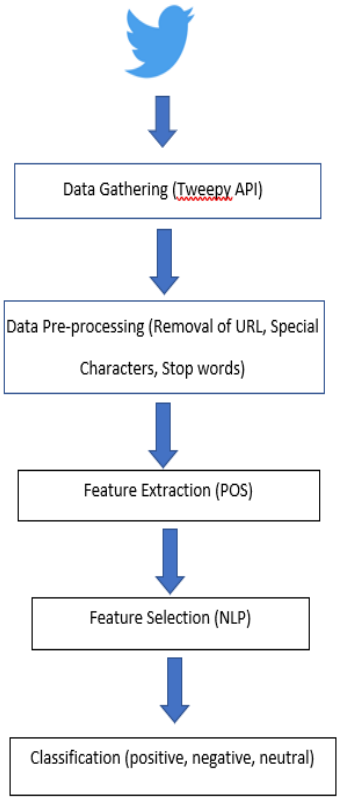
SENTIMENT ANALYSIS

Sentiment Analysis is the automated process of analyzing text data and classifying it into sentiments positive, negative or neutral.

Sentiment analysis is done by identifying and fetching the subjective data that underlies a text. This can be either an opinion, a judgment, or a feeling about a specific topic or subject. The most well-known type of sentiment analysis is called 'polarity detection' and consists in classifying a statement as 'positive', 'negative' or 'neutral'.

In many fields like business, Governmental Issues and public actions, determining the sentiment analysis is very important.

STEPS FOR SENTIMENT ANALYSIS



DATA GATHERING:

To gather data, Twitter API is used which is linked to the twitter account.

Authentication:

In order to extract tweets through Twitter API, one needs to register an App through their twitter account. After creation of an app, we will be redirected to the app page. The next step is copying 'Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret' from 'Keys and Access Tokens' tab.

PRE-PROCESSING:

The tweets retrieved can be unstructured or in any format. So, to get the structured data from it, we have to pre-process the tweets.

Pre-processed data can be achieved by:

- removing URLs, Special Characters, Stop Words
- Performing Word Stemming, tokenization and lemmatization.

Data Cleaning and Noise Reduction:

In order to properly analyze the data from tweets, it is necessary that this irregular data is removed, so the actual meaning and sentiments can be accounted from the sentence. Data Cleaning can be achieved by:

1. Conversion into lowercase:

Conversion of all uppercase letters to lowercase.

2. Removal of URLs:

Elimination of URLs with either regular expressions or generic word "URL".

3. Removal of @username:

Removal of @username with generic word "AT_USER".

4. Removal of # and other signs:

Elimination of # and other signs;

ex: #beautiful replaced with beautiful

5. Removal of white spaces:

Elimination of multiple white spaces with single white space.

6. Removal of non-English words:

Twitter generally supports more than 60 languages. But our project mainly includes English tweets; hence we remove the non-English words.

7. Emoticon replacements:

Emoticons are very important in determining the sentiment. So, the emoticons are replaced by their polarity by observing the emoticon dictionary.

Emoticon	Polarity
:-) :) :o) :] :3 :c)	Positive
:D C:	Extremely-Positive
:-(:(:c :[	Negative
D8 D; D= DX v.v	Extremely-Negative
: |	Neutral

8. Removal of stop words

Stop words play a negative role in sentimental analysis, so it is important to remove them. They occur in both negative and positive tweets. The examples of stop words are he, she, at, on, a, the, etc. While processing natural language, the words which frequently occur in the document are stop words (e.g. 'and', 'the', 'am', 'is'), which have little or no emotional meaning and it do not change the sentiment score when applied to lexical resources.

Ex: Removal of stop words

	Text Data
Tweet	I am shiny and fast
Stop word removal	['shiny', 'fast']

9. Removal of duplicate letters:

Elimination of duplicate letters in the word (not duplicate words).

Ex: delllicious will be replaced with delicious.

LEXICON BASED APPROACH

Data Processing using NLP:

Using Natural Language Processing (NLP) steps, one can process huge amount of unstructured data by analyzing sentence structure, and can calculate sentence or document level sentiment polarity using linguistic databases or lexical resources like WordNet, SentiWordNet, and treebanks. The techniques used in processing natural language are POS (Part of Speech) labeling, parsing, data extraction, Word Stemming and lemmatization, stop word removal, word tokenization, etc. This approach is known as Lexicon or Dictionary-based approach.

NLTK:

The Natural Language Toolkit (NLTK) is a program used for building Python programs that work with human language data for applying in statistical natural language processing (NLP).

NLTK includes text preparing libraries which are used for tokenization, classification, stemming, tagging and semantic reasoning. It also incorporates graphical demonstrations and sample data sets.

Word Tokenization:

Word tokenization is the process of dividing a large sample of text into words. This process is necessary in NLP where each word needs to be captured and subjected to further analysis like classifying and counting them for a particular sentiment. The Natural Language Tool kit (NLTK) is a library used to accomplish this.

Ex:

	Text Data
Tweet	Collusion is not a crime, but that does not matter because there was no collusion.
Word Tokenization	['collusion',' is', 'not', 'a', 'crime', 'but', 'that', 'does', 'not', 'matter', 'because', 'there', 'was', 'no', 'collusion']

Word Stemming:

Stemming and lemmatizing technics are used to get the word in normalized form. Stemming is used to get the base (root) word from the word in the text. It is done by removing the suffix from the word.

Ex: Identifying base word

	Text Data
Tweet	Technologies are rapidly changing.
Word Stemming	['Technologi', 'are, 'rapid, 'chang']

Word Lemmatization:

Lemmatization means assembling the inflected forms of word so they can be analyzed as single entity, meaning it will group words 'changing', 'changed', 'changes'.

Ex:

	Text Data
Tweet	Technologies are rapidly changing.
Word Tokenization	['Technology', 'are, 'rapid, 'change']

The only difference between lemmatization and stemming is in lemmatization the retrieved word will be the actual meaningful word but in stemming the word retrieved will not necessarily actual meaningful word.

Part-Of-Speech (POS) Tagging:

POS tagging is important to discover the meaning of word used in the sentence by identifying how it is used in the sentence.

Ex. "Select a seat" - in this sentence seat is used as Noun, but in "Seat here" seat is used as verb.

For Words with Negation:

For words with negation we can use negation module available in sentiment utility under NTLK package. This technique is used to assign the '_NEG' tag for the words which are followed by the negation words. In this approach, the words which comes after the negative word (the words with negative meaning) will be tagged as '_NEG'. During the sentence analysis, if the word like 'not', didn't, do not appeared in the sentence, all the words till the last word of sentence are classified with '_NEG' tag. The result of first step is stored in the list for further analysis, that will further provide negation score (1 or 0) to the word. In the second step, by iterating the results obtained from a first step and by parsing the word with the tag '_NEG' it will assign the score 1 to the words with negation tag and 0 to the words without tag. The list of lemmas will be returned by word with score '1' which include antonym to the original word and will reverse the meaning as well the polarity of the sentence when applied to the lexical resources to achieve accuracy in sentiment analysis from the Twitter data.

Ex: Negation Words	Example-1	Example-2
Input	['Happy', 'to', 'be', 'the', 'Pradhan', 'Sevak', 'for', 'each', 'and', 'every', 'Indian']	['I', 'am', 'not', 'looking', 'forward', 'to', 'the', 'weekend']
Mark Negation	['Happy', 'to', 'be', 'the', 'Pradhan', 'Sevak', 'for', 'each', 'and', 'every', 'Indian']	['I', 'am', 'not', 'looking_NEG', 'forward_NEG', 'to_NEG', 'the_NEG', 'weekend_NEG']
Step List of Score	['0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0']	['0', '0', '0', '1', '1', '1', '1', '1']

WordNet:

Wordnet is a large lexical database of English used in NLP. Adverbs, Adjectives, Nouns and verbs are assembled into sets of similar words (synsets), each expressing a distinct concept.

Wordnet's formation makes it very useful tool for computational semantics and natural language processing.

WordNet tag	Treebank tag
n	NN
a	JJ
s	JJ
r	RB
v	VB

Ex. For word "Program"

Plan.n.01

(In above example 'Plan' is synonym, 'n' is wordNet tag as noun and '01' represents 1st synset.)

The words are tagged with the generic POS tags and the words which are not in wordnet are set to 'NONE'. The words which are labeled as 'NONE' will not return any opinion or emotional attribute and hence, during further analysis it will be ignored if it does not contain any sentiment score. Further in the subsequent step, the possible synset terms are obtained for the given word. It is done by iterating through the loop to find a correct lemma for the given word in the synsets.(Ravi_Patel_Thesis)

The extracted term from the WordNet database will be assigned with the numerical score using SentiWordNet dictionary (Ohana et al. 2009). Each set of terms in SentiWordNet (synsets) is linked with two numerical scores ranging from 0 to 1, each value indicates the positive and negative bias of synset(Ravi_Patel_Thesis).

The next step is to calculate the scoring of the polarity. We can calculate the totalPosScore (total positive score) by adding all the positive scores and totalNegScore by adding all the negative scores. If totalPosScore is greater than totalNegScore, it will be categorized as 'POSITIVE'; If totalNegScore is greater than totalPosScore, it will be categorized as 'NEGATIVE'; otherwise it will be categorized as 'NEUTRAL'.

MACHINE LEARNING ALGORITHM

Machine learning is the study of algorithms in which training data is used to train the model and test data is used to check accuracy of the model. ML approaches are used to make predictions on given data.

NAÏVE-BAYES APPROACH

In Naive Bayes classifiers, each attribute impacts which label should be assigned to a given input value. For selection of appropriate label for an input value, the naive Bayes classifier begins by calculating the prior probability of each label, which is determined by checking frequency of each label in the training set. At this point, we have a training set, so we will use classifier to classify the tweets.

Naive Bayes Classifier is a classification algorithm which is based on Bayes' Theorem. In this approach, the posterior probability is calculated, in which the probability of an event A occurring is reliant on probabilistic known background (e.g. event B evidence).

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

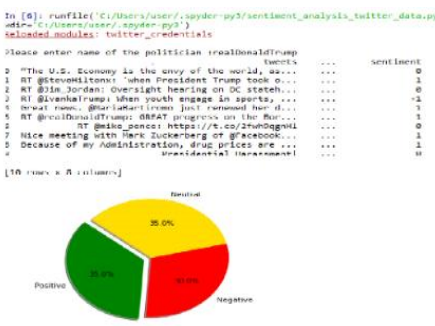
using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

STEPS FOR SENTIMENT ANALYSIS USING NAIVE-BAYES CLASSIFIER:

Train the classifier:

1- Determine Training set and Test set:
Training set contains the data and associated labels in our case whether the tweet is positive or negative.
Test set contains the data i.e. tweets which needs to be classified.
2- Bag-Of-Words:
In this step, the frequency of each individual word is calculated i.e. number of times the word occurred in the positive and negative tweet.
3- Calculating Prior probability:
This step includes calculating prior probability for labels positive and negative. For positive class it is calculated using total number of positive tweets divided by total number of tweets and for negative class, total number of negative tweets divided by total number of tweets.
4- Computing conditional probability:
It is the likelihood of each word or attribute.
5- Computing posterior probability:
In this step posterior probability is calculated. Posterior probability is the probability of event A happening given that event B has happened.
6- Assigning the class or label:
This is the last step in which whichever is the greater probability, that class will be assigned to the data.



LIMITATIONS OF SENTIMENT ANALYSIS:

1. Possible to have errors while analyzing large volume of data.

2. Can predict the sentiment wrongly by not considering sarcasm and irony, negations, jokes, exaggeration.

3. When using ML approaches like Naïve-Bayes Algorithm, we have to first train the classifier.

4. In ML approaches, if the word in test data set is not present in Vocabulary, the model can predict it wrongly.

CONCLUSION AND FUTURE SCOPE

Applying sentiment analysis to fetch the sentiment became an important work for many organizations and even individuals. Sentiment analysis is an emerging field in decision making process and is growing fast. Recently, people have started expressing their thoughts on the Social media that increased the need of analyzing the opinionated online content for various real-world applications. Existing sentiment analysis models can be improved with more semantic and realistic information.

Sentiment analysis can be used for estimating feedback from people regarding specific service, movie or product, to evaluate positive and negative leaders, etc. But both naïve-bayes and lexicon approach have some drawbacks. So none of the approach can provide the accurate results. But we can use combination of both the approaches to achieve more accurate results.

REFERENCES

[1] Suchita V Wawre, Sachin N Deshmukh "Sentiment Classification using

Machine Learning Techniques" Department of Computer Science & Information

[2] Riya Suchdev, Pallavi Kotkar, Rahul Ravindran , Sridhar Swamy "Twitter

[3] Dipak R. Kawade, Dr.Kavita S. Oza "Sentiment Analysis: Machine Learning Approach".

[4] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan "Thumbs up? Sentiment

Classification using Machine Learning Techniques"

- https://zone.biblio.laurentian.ca/bitstream/10219/2963/1/Ravi%20Patel_Thesis_Final.pdf
- https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk
- https://www.tutorialspoint.com/python_data_science/python_word_tokenization.htm
- https://www.knowledgehut.com/tutorials/machine-learning/tokenize-text-nltk-python-machine-learning
- Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." 9th. IT & T Conference. 2009.
- https://www.researchgate.net/publication/317058859
- https://www.researchgate.net/publication/327160507
- https://www.researchgate.net/publication/268692721
- https://www.researchgate.net/publication/307382684