International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

Twitter Sentiment Analysis using Machine Learning

Sourav Chavan[1], Vishal Navale[2], Sharad Waghmare[3], Uday Chavan[4], Prof. Nilesh Ghode[5]

[1,2,3,4]Department of Electronics & Telecommunication, Atharva College of Engineering, Mumbai
[5]Prof. Nilesh Ghode, Department of Electronics & Telecommunication, Atharva College of Engineering, Mumbai

---

Abstract - *The advent of social media has seen a drastic change in how content is made and shared on the web. This has shifted the main target of selling and advertising agencies from traditional methods to digital marketing. One thing that most of those agencies require is that the analysis of the content on such social media websites like Twitter. Twitter acts as a platform, for various sorts of users to share their views and sentiment on various sorts of topics in 140 characters or less. We have developed a tool that is able to extract tweets pertaining to a topic and analyse them to calculate their polarity i.e. positive, negative or neutral.*

***Key Words*: Sentiment, Machine Learning, Logistic Regression, Sci-Kit Learn.**

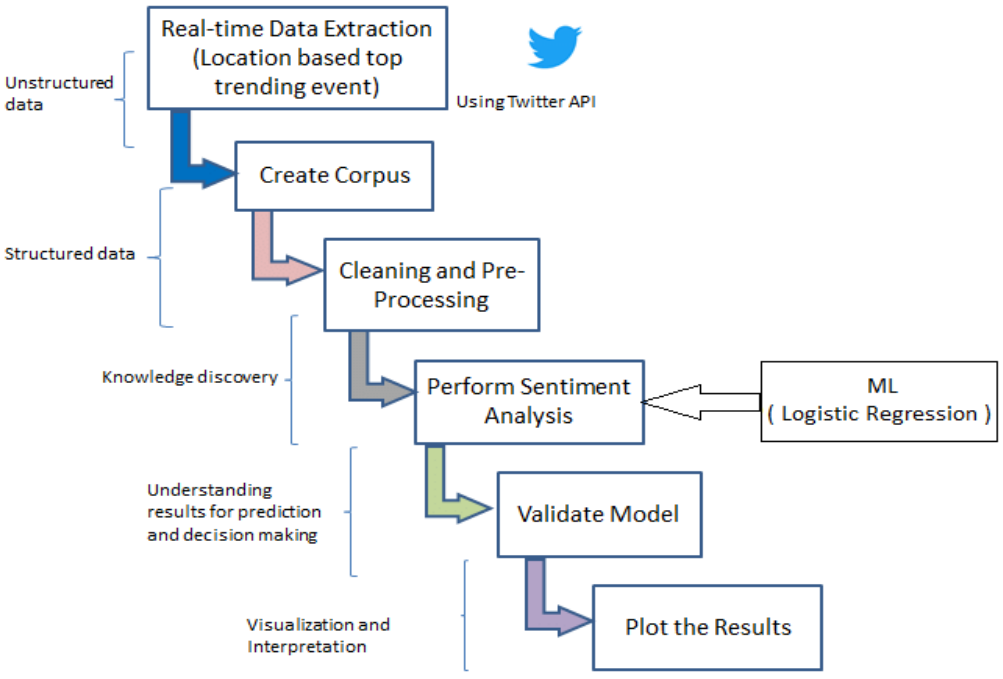
1. INTRODUCTION

The Emergence of social media has given web users a venue for expressing and sharing their thoughts and opinion on all quite topic and events. Twitter, with nearly 600 million users and over 250 million messages per day has become a gold mine for organization to watch their reputation and makes by extracting and analyzing the sentiments of the tweets posted by the people about them. To analyze the sentiments of the tweet the Sentiment Analysis came into the picture. Sentimental Analysis is that the method of computationally identifying and categorizing opinions from piece of text and determine whether the writer's attitude towards a selected topic or the merchandise, is positive negative or neutral. Here we are extracting the tweets based on the real time. Sentiment Analysis refers to the utilization tongue processing, text analysis and linguistics to systematically identify, extract, quantify, and study affective states and subjective information

REAL-TIME DATA EXTRACTION

Data extraction may be a process that involves retrieval of knowledge from various sources. Frequently, companies extract data so as to process it further, migrate the info to a knowledge repository (such as a knowledge warehouse or a knowledge lake) or to further analyze it. It's common to rework the info as a neighborhood of this process

BLOCK DIAGRAM



STORY GENERATION & VISUALIZATION

Exploring and visualizing data, regardless of whether its text or the other data, is an important step in gaining insights. Before we start exploration, we must think and ask questions associated with the info in hand. A few probable questions are as follows:

1. What are the most common words in the entire dataset?
2. What are the most common words in the dataset for negative and positive tweets, respectively?
3. How many hashtags are there in a tweet?
4. Which trends are associated with my dataset?
5. Which trends are associated with either of the sentiments? Are they compatible with the sentiments?

CREATE CORPUS:

One of the primary things required for tongue processing (NLP) tasks may be a corpus. In linguistics and NLP, corpus (literally Latin for body) refers to a set of texts. Such collections could also be formed of one language of texts or can span multiple languages -- there are numerous reasons that multilingual corpora (the plural of corpus) may be useful. Corpora can also contain themed texts (historical, Biblical, etc.). Corpora are generally solely used for statistical linguistic analysis and hypothesis testing

CLEANING AND PRE-PROCESSING

The created dataset could have any redundant information or any unwanted garbage value which could cause Machine Learning algorithm to behave abruptly. So, this dataset is pre-processed to remove such unnecessary data.

EXTRACTING FEATURES FROM CLEANED TWEETS

To analyze a preprocessed data, it must be converted into features. Depending upon the usage, text features are often constructed using assorted techniques – Bag-of-Words, TF-IDF, and word Embeddings.

Fig.2 Word cloud indicating frequency and importance of words

MODEL BUILDING

We are now through with all the pre-modeling stages required to urge the info within the proper form and shape. Now we will be building predictive models on the dataset using the two feature sets — Bag-of-Words and TF-IDF.

We will use logistic regression to create the models. It predicts the probability of occurrence of an occasion by fitting data to a logit function

The following equation is used in Logistic Regression:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(\text{Age})$$

PERFORM SENTIMENT ANALYSIS

This dataset is then used to train the Machine Learning algorithm. Algorithm used for this project is Logistic Regression. In statistics, Logistic Regression model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick in our case it is whether text is racist/sexist or not.

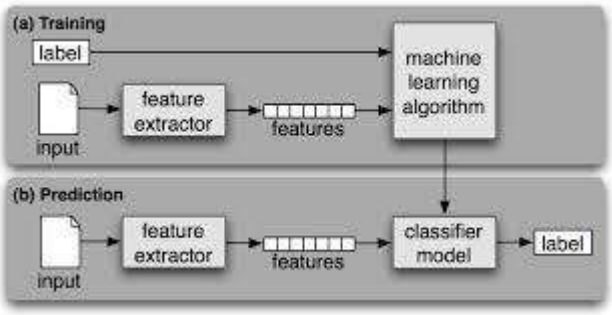


Fig.3 Training an algorithm

VALIDATE MODEL

Before plotting the results, the Machine Learning model should be validated using validation dataset. In machine learning, model validation is mentioned because the process where a trained model is evaluated with a testing data set. The testing model is evaluated with a testing data set. The testing data set may be a separate portion of an equivalent data set from which the training set springs. The main purpose of using the testing data set is to check the generalization ability of a trained model. Model validation is administered after model training. Together with model training, model validation aims to seek out an optimal model with the simplest performance

PLOT THE RESULT

The idea of building machine learning models works on a constructive feedback principle. You build a model, get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results

LITERATURE SURVEY

Sentiment analysis is currently one among the favored topic in research field. There are various works happening during this area for various languages not studied so far like Arabic, Hindi, Thai etc. There are various open source libraries available for various languages like python, R etc. which makes the work easy to research the text and process it. It are often used for various purposes like in reviewing movies, products of a companies, about companies, feeling or emotions of citizens for a rustic. The most popular thanks to get this information on social media and analyze it. To make it into something meaningful sense, the classifier techniques must be used.

The data must be in readable format, in English. The classifiers are wont to tokenize of classify the info. The Superwised learning technique is employed with machine learning approach to detect sentiments and analyze the emotions of the remainder of the text. Un-Superwised learning is linguistic approach during which text is first tokenized into tokens and added with tags to guage the emotions of the text.

How to get many data to evaluate:

1) Social sites

a) Facebook.com

b) Twitter.com

c) LinkedIn.com

d) LinkedIn.com

2) News websites and comments

3) Movie reviewing sites

4) Products selling sites

i) Flipkart

ii) Snapdeal

5) Blogs etc.

6) Techniques used presently are:

a) Machin Learning

i) Logistic Regression Text Structure:

1. An array of sents/sentences

2. Each sent is again tokenized called tokens

3. Each word or token is padded with 2 other tags in dictionary format. These added tags make each token to be recognized as verbs, nouns, adjectives, adverbs etc. to verify if that token is polar word or not.

4. Separate datasets are there so that each token can be matched with words present in the datasets. First, collection of data could also be a priority. Useful data is what's required before analyzing the data. Sentiment analysis is performed on the info which is a few product or review and user wants to understand about if it's good or not. Sentiments can have various sorts of polarity or emotions about something.

Summarizing the opinions is additionally one among the good concern for today's researchers. summarizing the emotions doesn't affect subset of text or its one a part of text to be printed. It is printing the info with a particular sense in fewer number of words and it also contains the topic of the text.

3. CONCLUSIONS

In conclusion the tool developed by us will be a simple showcase of a system which will have a number of applications in the near future.
With the shift of advertising and marketing from print to digital and social media, sentiment analysis will have a huge role in deciding how to push products to the consumers and how to interact with them and twitter are going to be one among the most platforms for users to take advantage of this untapped market

- **APPLICATIONS AND FUTURE SCOPE**

1. Feedback on Pilot Releases And Beta Versions:

   When a corporation releases a replacement product or service, it's released as a pilot or beta version. The monitoring of public feedback at this stage is extremely crucial. So, text mining from social media platforms and review sections greatly helps accelerate this process.

2. Employee Feedback:

   Sentimental analysis also can be wont to receive feedback from the workers of the corporate and analyze their emotions and attitude towards their job. And to work out whether or not they are satisfied with their job or not.

3. Better Services:

   Text mining can provide a filter about, which service of the corporate is getting more negative feedback. This will help the corporate to understand, what are the issues arising thereupon particular service. And supported this information the corporate can rectify these problems.

- **RESULT AND DISCUSSIONS**

In conclusion the tool developed by us will be a simple showcase of a system which will have a number of applications in the near future.

With the shift of advertising and marketing from print to digital and social media, sentiment analysis will have a huge role in deciding how to push products to the consumers and how to interact with them and twitter are going to be one among the most platforms for users to take advantage of this untapped market

REFERENCES

1. "Sentiment Analysis of Twitter Data" by El_Rahman, AlOtaibi and AlShehri (IEEE 2019)

2. "Sentiment Analysis of Polarity in Product Reviews In Social Media" by Marium Nafees, Hafsa Dar, Ikram Ullah Lali, Salman Tiwana (IEEE 2018)

3. Chen, Y., & Zhang, Z. (2018). Research on text sentiment analysis supported on CNNs and SVM. 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA).

4. "Investigating sentiment analysis using machine learning approach" , Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9

5. https://pandas.pydata.org/pandas-docs/stable/

https://www.nltk.org/

6. https://docs.python.org/

7. https://scikit-learn.org/

8. https://machinelearningmastery.com/logistic-regression-for-machine-learning/

9. https://www.coursera.org/learn/machine-learning/

10. https://elitedatascience.com/feature-engineering

11. https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/

12. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

13. https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

14. https://machinelearningmastery.com/statistics-for-machine-learning-mini-course/