Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python

Conference Paper · February 2018

CITATIONS
7

READS
16,112

SENTIMENT ANALYSIS ON TWITTER DATA USING MACHINE LEARNING ALGORITHMS IN PYTHON

S.SIDDHARTH 1, R.DARSINI 2

1 & 2 M.Sc. Software Systems,

Department of Computing

Coimbatore Institute of Technology
Coimbatore

ssk4n4ever@gmail.com,

darsinirajkumar1998@gmail.com

Dr. M. SUJITHRA

Assistant Professor,

Department of Computing

Coimbatore Institute of Technology
Coimbatore

sujisrinithi@gmail.com

ABSTRACT

With the rise of social networking epoch and its growth, Internet has become a promising platform for online learning, exchanging ideas and sharing opinions. Social media contain huge amount of the sentiment data in the form of tweets, blogs, and updates on the status, posts, etc. In this paper, the most popular micro blogging platform twitter is used. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), to extract user's opinions and sentiments. The main goal is to explore how text analysis techniques can be used to dig into some of the data in a series of posts focusing on different trends of tweets languages, tweets volumes on twitter. Experimental evaluations show that the proposed machine learning classifiers are efficient and performs better in terms of accuracy. The proposed algorithm is implemented in python.

Keywords – Machine Learning, Natural Language Processing, Python, Sentimental Analysis

1. INTRODUCTION

Micro blogging websites are one of the most important sources of varied kind of information. This is due to the fact that every people post their opinions on a variety of topics, discusses current issues, complains and expresses positive sentiment for products they use in daily life. Sentimental analysis is the process of deriving the quality information from the text. In other words, it is the process of deriving the structured data from unstructured data. This is used to measure opinions of the customer, feedback, product reviews Unstructured data not only

refers to the tables, figures from the organization but also consists of information from the internet i.e. chats, E-mail, pdfs, word files, E-Commerce websites and social networking sites.

On structured data analytics operation can be easily performed and the result can be obtained easily. But in case of unstructured data from E-mail, Twitter etc., it is quite difficult to conclude the output because of various problems such as virtual noise effect and unspecific data. In this paper, we look at one such popular micro blog called Twitter.

The paper consists of 5 different sections. 1] The first section explains what sentimental analysis is and what is its importance? 2] The second section clearly tells us about the proposed methodology that starts from the twitter data extractions till the feature extraction.3] The next section tells about the machine learning algorithms, here we have considered 2 main algorithms and a clear difference between them.4] The 4th section involves the applications and challenges of sentimental analysis.5] The last section carries with the conclusion and the future scope.

2. SENTIMENTAL ANALYSIS

Sentimental analysis is the process of computationally determining the opinion or attitude of the writers as positive, negative or neutral. Data mining is another name for sentimental analysis. In many fields like business, politics and public actions, determining the sentimental analysis is very important. Considering business, it is very useful to understand the customer's feelings in order to develop their company. Next in politics: It can be even be used to predict the election results. There are two ways of classifications and they are (1) machine learning (2) lexicon based approach. In this paper machine learning classifiers are implemented in sentimental analysis and is done in twitter because most of the politicians, famous personalities (even the president of various states) and even general people regularly update their moods in the form of tweets.

3. PROPOSED METHODOLOGY

Methodology of sentimental analysis in twitter mainly involves 5 steps.

3.1 DATA COLLECTION/ TWEET EXTRACTION

Twitter API — A Python wrapper for performing API requests.For fetching the twitter data from the twitter API includes the following steps 1] Installation of the needed software 2] authentication of twitters data. The main installation software's include tweepy, text blob, nltk etc, Authentication involves different steps

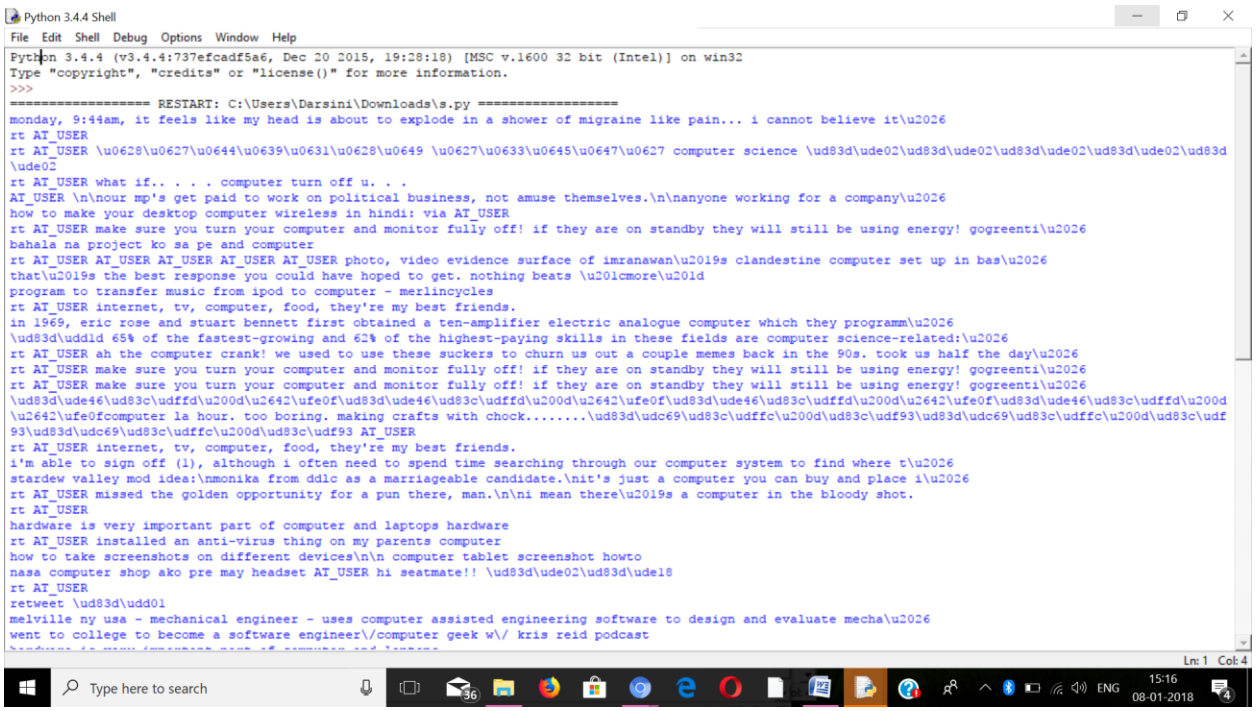
step1: visit the twitter website and click the button 'create new app'.
Step2:fill the details in the form provided and submit.
Step3:It will be redirected to the app page where the "'consumer keys', 'consumer access', 'access token' and 'access token secret' "that is needed to access the twitter data will be present.
Step4:implement in python.

There are different sources for storing the data taken from the twitter. They are like

MongoDB , open source document storage database and is the go-to "No SQL" database. It makes working with a database feel like working with JavaScript.

PyMongo, a Python wrapper for interfacing with a MongoDB instance. This library lets you connect your Python scripts with your database and read/insert records.

This is an example of the data that is been extracted from the twitter on the topic computer using python code.



3.2 PRE-PROCESSING

Once the data is collected from the twitter the next step is preprocessing that is implemented in python. There are several steps involved in the preprocessing stage. They are,

1. Converting all uppercase letters to lowercase.
2. Tokenization

Tokenization generally done by installing the NLP package. It generally means removal of hash tags, numbers (1, 2, 3 etc.,), URL's and targets (@). Once tokenization is over we move to the next step of preprocessing.

3. Removal of non-English words

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

4. Emoticon replacements

Emoticons are very important in determining the sentiment. Sothe emoticons are replaced by their polarity by seeing the emoticon dictionary.

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

5. Removal of stop words

Stop words play a negative role in sentimental analysis, so it is important to be removed. They occur both I negative and positive tweets. A list of stop words like he, she, at, on, a, the, etc. are created and ignored. Once the above four steps are over we move to the next main method called feature extraction.

3.3 FEATURE EXTRACTION

Selection of useful words from the tweet is called as feature extraction. In the feature extraction method, we extract the aspects from the pre-processed twitter dataset.

1. There are three different types of features namely unigram, bigram, n-gram features.

2. Parts Of Speech Tags such as like adjectives, adverbs, verbs and nouns are good indicators of subjectivity and sentiment.

3. Negation is another important but difficult feature to interpret. The presence of a negation usually changes the polarity of the sentiment.

3.4 FEATURE SELECTION

Correct feature selection techniques are used in sentiment analysis that has got a significant role for identifying relevant attributes and increasing classification (machine learning) accuracy. They are categorized into 4 main types namely,

1. Natural language processing
2. Statistical
3. Clustering based
4. Hybrid

Natural language processing mainly works on (1) Noun, noun phrases, adjectives, adverbs (2) Terms occurring near subjective expressions can act as features.

Clustering based feature extraction techniques are implemented by requiring few parameters. The major weakness of clustering is that only major features can be extracted and it is difficult to extract minor.

Statistical techniques are further divided into three sub types; they are univariate, multivariate and hybrid. Univariate methods, they are also called feature filtering methods, that take attributes separately, examples of this type include information gain (IG), chi–square, occurrence frequency, log likely-hood and minimum frequency thresholds. Univariate techniques have computational efficiency. Decision tree models, recursive feature elimination and genetic algorithms are the examples of multivariate methods .When compared to univariate; multivariate methods are expensive in terms of computational efficiency. Hybrid techniques are the one which combine the univariate and multivariate to achieve an efficient and accurate answer.

Hybrid techniques include POS Tagging with Word Net dictionary. Compactness and redundancy pruning methods were used for eliminating irrelevant features

3.5CLASSIFICATION

MACHINE LEARNING ALGORITHMS

Machine learning is the study of algorithms that can learn from and make predictions on data. It is also called as related to prediction-making on some data. There are many machine learning algorithms. But this paper explains about two of them. They are

- Naïve baye's
- Neural networks

3.5.1. NAÏVE BAYE'S

Naïve bayes is one of the most improved classification (classifier) methods. First in order to perform classification, we must select the features from the data

set. All the tweets in the data sets will be processed by the classifiers. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. A naïve bayes algorithm is very easy to build up and mainly used for a large set of data. It provides a way of calculating p(c|x) from p(c), p(x) and p(x|c). Here p(c|x) is called the posterior probability and it is given by the formula,

$$p(c|x) = \frac{p(x|c)\,p(c)}{p(x)} \qquad \text{where,}$$

P(*c*|*x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).

P(*c*) is the prior probability of *class*.

P(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.

P(*x*) is the prior probability of *predictor*.

Let us consider a simple example of naïve bayes. That is whether the players will play the game or not depending on the weather condition.

Step 1: collect the data set and store in frequency table

Step 2: create a table and find the probability of playing=0.64 and the overcast probability=0.29.

Step 3: use naïve bayes to calculate the posterior probability.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a method to predict the probability of different situations based on the various attributes.

ADVANTAGES OF USING NAÏVE BAYES

It is very easy and fast to predict the class of data set. It is also mainly used in multi class prediction.

When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression.

DISADVANTAGES OF NAÏVE BAYES

If categorical variable has a category in a test data set, which was not observed in the data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often called as “Zero Frequency”. To overcome this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

Many researchers have done sentiment classification by using Naive Bayes classifier. But Naive Bayes classifier has major limitation that the real world data may not always satisfy. Hence, it affects the accuracy of Naive Bayes classifier

APPLICATIONS OF NAÏVE BAYES

The applications of naïve bayes,

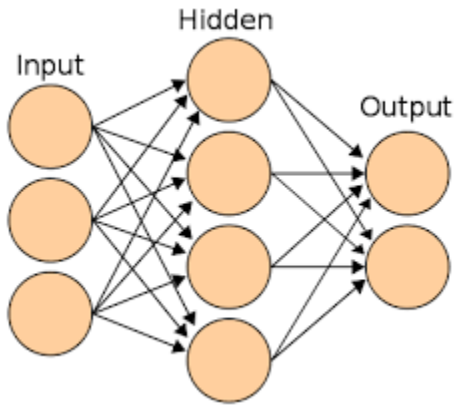
Real time Prediction: Naive Bayes algorithm is a also a fast learning algorithm. Thus, it is used for making predictions in real time.

Multi class Prediction: This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes also.

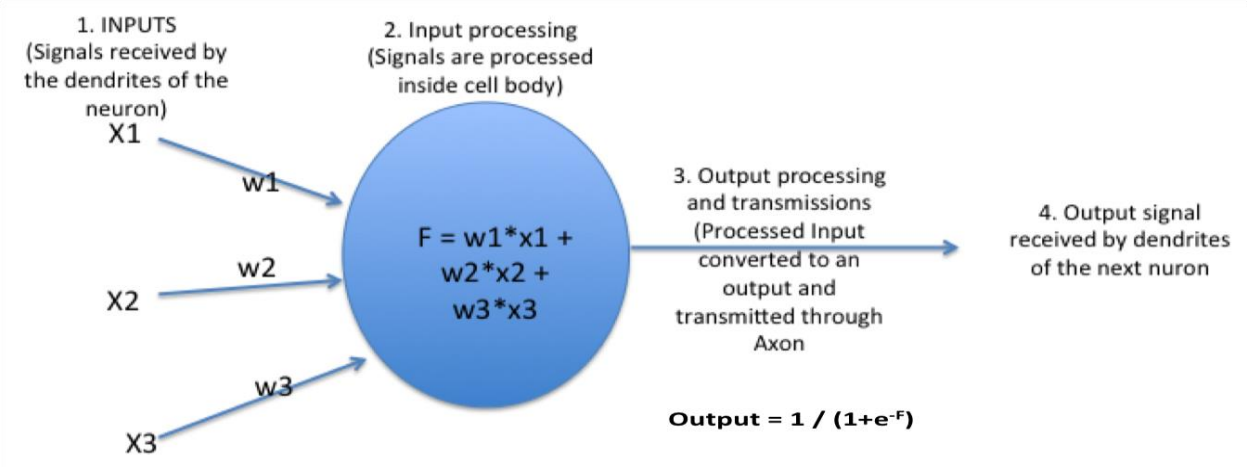
Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification as it has a better result in multi class problems and have higher success rate as compared to other algorithms. This is also used to identify spam e-mail. Main application is sentimental analysis it is used to predict whether a user would like a given resource or not.

3.5.2. NEURAL NETWORKS

A neural network is another important tool for classification. Is has also been a promising alternative to various classification methods. This classifier with the appropriate network structure can handle the correlation or dependence between the input variables. Artificial neural networks perform back propagation by activating the neurons in the hidden layer.



There are 2 phases (1) training (2) testing. In training phase the positive and negative comments are trained and assigned weights. The main purpose of training phase is to create the dictionary of positive comments. In the next phase testing is done based on the weighted dictionary. The artificial neural network is trained with labeled data to produce meaningful output. This process by which neural networks learn from labeled data is called as back propagation. A layer named as feed forward is made up of nodes and edges. Each node is a part of a layer and each node in a layer points to every node in the next layer. Input is fed into the first layer called the input layer, in which the input follows the edges to the nodes in the next layer until it reaches the output layer. Each edge has a weight and when the input travels it is multiplied by the weight associated with the edge. Evaluate the network performance. Then calculate whether the review is positive or negative.



ADVANTAGES OF NEURAL NETWORKS

The main advantage is that they are data driven self-adaptive methods where, they can adjust themselves to the data without any explicit specifications of functional or distributional form.

They can approximate any function with arbitrary accuracy since they are universal functional approximates.

DISADVANTAGES OF NEURAL NETWORKS

One main disadvantage is large complexity of the network structure.

APPLICATIONS OF NEURAL NETWORKS

Neural networks have been successfully applied to a variety of real world classification tasks in industry, business and science.

It is used in hand writing recognition, stock exchange prediction, image compression

Combining Naive Bayes Classifier with Neural Network will improve the accuracy and performance of sentiment classification in real world.

The computing World has a lot to gain from Neural Network. Thus, their ability to learn by example makes them very flexible and powerful.

CHALLENGES AND APPLICATIONS IN SENTIMENTAL ANALYSIS

Since the Opinion based or feedback based application are more fashionable, now a days, the natural language processing community shows much interest in Sentiment Analysis and Opinion Mining system. The explosion of internet has changed the people's life style, now they are more expressive on their views and opinions. And this tendency helped the researchers in getting user-generated content easily. The major applications are

- Purchasing Product or Service: While purchasing a product or service, taking right decision is no longer a difficult task. By this technique, people can easily evaluate other's opinion and experience about any product or service and also he can easily compare the competing brands. Now people don't want to rely on external consultant. The Opinion mining and sentiment analysis extract people opinion form the huge collection of unstructured content, the internet, and analyze it and then present to them in highly structured and understandable manner.
- Quality Improvement in Product or service: By Opinion mining and sentiment analysis the manufactures can collect the critic's opinion as well as the favorable opinion about their product or service and thereby they can improve the quality of their product or service. They can make use of online product reviews from websites such as Amazon and C|Net , RottenTomatoes.com and IMDb .
- Marketing research: The result of sentiment analysis techniques can be utilized in marketing research. By sentiment analysis techniques, the recent trend of consumers about some product or services can be analyzed. Similarly the recent attitude of general public towards some new government policy can also be easily analyzed. These all result can be contributed to collective intelligent research.
- Recommendation Systems: By classifying the people's opinion into positive and negative, the system can say which one should get recommended and which one should not get recommended.
- Detection of "flame": The monitoring of newsgroup and forums, blogs and social media is easily possible by sentiment analysis. Opinion mining and sentiment analysis can automatically detect arrogant words, over heated

words or hatred language used in emails or forum entries or tweets on various internet sources.

- Opinion spam detection: Since internet is available to all, anyone can put anything on internet, this increased the possibility of spam content on the web. People may write spam content to mislead the people. Opinion mining and sentiment analysis can classify the internet content into' spam' content and 'not spam' content.
- Policy Making: Through Sentiment analysis, policy makers can take citizen's point of view towards some policy and they can utilize this information in creating new citizen friendly policy.
- Decision Making: People's opinion and experience are very useful element in decision making process. Opinion mining and Sentiment analysis gives analyzed people's opinion that can be effectively used for decision making.

The main challenges that are faced by and sentiment analysis include,

- Detection of spam and fake reviews: The web contains both authentic and spam contents. For effective Sentiment classification, this spam content should be eliminated before processing. This can be done by identifying duplicates, by detecting outliers and by considering reputation of reviewer.
- Limitation of classification filtering: There is a limitation in classification filtering while determining most popular thought or concept. For better sentiment classification result this limitation should be reduced. The risk of filter bubble gives irrelevant opinion sets and it results false summarization of sentiment.
- Asymmetry in availability of opinion mining software: The opinion mining software is very expensive and currently affordable only to big organizations and government. It is beyond the common citizen's expectation. This should be available to all people, so that everyone gets benefit from it.
- Incorporation of opinion with implicit and behavior data: For successful analysis of sentiment, the opinion words should integrate with implicit data. The implicit data determine the actual behavior of sentiment words.
- Domain-independence: The biggest challenge faced by opinion mining and sentiment analysis is the domain dependent nature of sentiment words. One

features set may give very good performance in one domain, at the same time it perform very poor in some other domain.

- Natural language processing overheads: The natural language overhead like ambiguity, co-reference, Implicitness, inference etc.

CONCLUSION AND FUTURE SCOPE

Applying sentimental analysis to extract the sentiment became an important work for many organizations and even individuals. Sentiment analysis is an emerging field in decision making process and is developing fast. Our project goal is to analyze the sentiments on a topic which are extracted from the Twitter and determine its nature (positive/negative/neutral) of the defined topics. The development of techniques for the document-level sentiment analysis is one of the significant components of this area. Recently, people have started expressing their opinions on the Web that increased the need of analyzing the opinionated online content for various real-world applications. A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and commonsense knowledge.

REFERENCES

[1] Suchita V Wawre , Sachin N Deshmukh "Sentiment Classification using Machine Learning Techniques" Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India

[2] Riya Suchdev , Pallavi Kotkar , Rahul Ravindran , Sridhar Swamy "Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach" Computer Engineering VES Institute of Technology, University of Mumbai, Mumbai, India.

[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"

[5]I.Hemalatha , Dr. G.P.S.Varma ,Dr. A.Govardhan “Automated Sentiment Analysis System Using Machine Learning Algorithms”

[6] Pranali Borele , Dilipkumar A. Borikar “An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques”

[7] Lina L. Dhande , DR. Girish K. Patnaik “Review of Sentiment Analysis using Naive Bayes and Neural Network Classifier “.

[8] Rudy Prabowo , Mike Thelwall” Sentiment Analysis: A Combined Approach”.

[9]David Osimo, Francesco Mureddu” Research Challenge on Opinion Mining and Sentiment Analysis”.

[10] Bo Pang, Lillian Lee2 “ Opinion mining and sentiment analysis”.

[11] Ashish Katrekar “ An Introduction to Sentiment Analysis”.

[12] Prof. Ronen Feldman Hebrew University, “JERUSALEM Digital Trowel, Empire State Building SENTIMENT ANALYSIS TUTORIAL “.

[13] K S Kushwanth Ram , Sachin Araballi ,Shambhavi B R ,Shobha G” Sentiment Analysis Of Twitter Data”.

[14] Apoor v Agarwal, Jasnee t Singh Sabharwal “End-to-End Sentiment Analysis of Twitter Data” .

[15] Alexander Pak, Patrick Paroubek” Twitter as a Corpus for Sentiment Analysis and Opinion Mining”