JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Linear SVM Classification of Sentiment in Tweets about Airlines on Twitter in Machine Learning

Sashi Jangra

shashijangra101@gmail.com

M. Tech Scholar, Department of Computer Science & Engineering, BRCM CET, Bahal, (Haryana), India

Mrs. Neha

neha@brcm.edu.in

Assistant Professor, Department of Computer Science & Engineering, BRCM CET, Bahal, (Haryana), India

ABSTRACT

In this research SVM classifier is used to analysis of sentiment in data mining for predicting the mined information on particular topic like statements comments etc. Is either in the form of positive, negative or natural? There are many modals have been created over sentiment analysis using SVM to predict the correct result but till now created modals could not predict the correct result. In this we have used SVM Classifier that is a supervised machine learning modal but more advanced. It uses algorithms to train and classify text within our sentiment polarity model by taking a step beyond X/Y prediction. It will divide data in two segment with a line one side of the line the data will be positive and other side the data will be negative by using Linear Classifier modal that Is type of SVM classifier. To implement this research modal I have used Python programming language. In this, there is taken a dataset on which SVM classifier is applied to predict the information in the form of positive, negative and natural.

Keywords:-Dataset, SVM classifier, Data Mining, Python programming Language.

Introduction

In this research SVM classifier is used to analysis of sentiment in data mining for predicting the mined information on particular topic like statements, comments etc. is either in the form of positive, negative or natural.

There are many modals have been created over sentiment analysis using SVM to predict the correct result but till now created modals could not predict the correct result.

In this we have used SVM Classifier that is a supervised machine learning modal but more advanced. It uses algorithms to train and classify text within our sentiment polarity model by taking a step beyond X/Y prediction.

It will divide data in two segment with a line one side of the line the data will be positive and other side the data will be negative by using Linear Classifier modal that Is type of SVM classifier.

To implement this research modal I have used Python programming language.

Sentiment Analysis using SVM in Data Mining

Sentiment Analysis is a process of computationally and categorizing opinions from piece of text and determine whether the writer's attitude towards a particular topic or the product , is positive, negative or neutral.

SVM Classifier

- It is machine learning algorithm which is used for classification and regression both .
- In this research SVM is used for classify positive and negative sentiment different –different segment by drawing a line between them.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

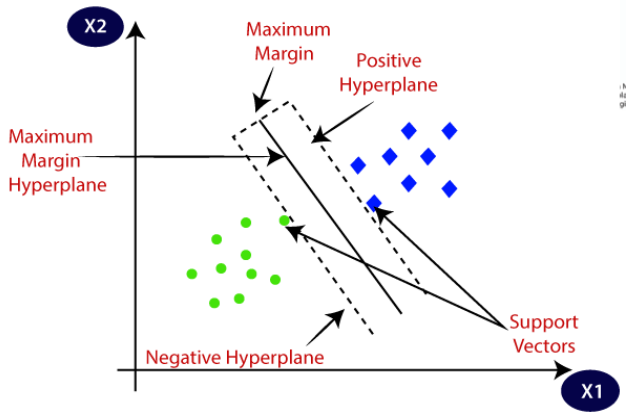

Fig. (1)

Data Mining

- It is basically the process carried out for the extraction of useful information from a bulk of data.
- Here Required data is mined by using data mining on which Sentiment Analysis is performed.

In general terms, "**Mining**" is the process of extraction of some valuable material from the earth e.g. coal mining, diamond mining, etc. In the context of computer science, "**Data Mining"** can be referred to as **knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging**. It is basically the process carried out for the extraction of useful information from a bulk of data or data warehouses. One can see that the term itself is a little confusing. In the case of coal or diamond mining, the result of the extraction process is coal or diamond. But in the case of Data Mining, the result of the extraction process is not data!! Instead, data mining results are the patterns and knowledge that we gain at the end of the extraction process. In that sense, we can think of Data Mining as a step in the process of Knowledge Discovery or Knowledge Extraction.

Classification Algorithms

The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks:

- Naïve Bayes: a family of probabilistic algorithms that uses Bayes's Theorem to predict the category of a text.

- Linear Regression: a very well-known algorithm in statistics used to predict some value (Y) given a set of features (X).
- Support Vector Machines: a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they're mapped to.
- Deep Learning: a diverse set of algorithms that attempt to mimic the human brain, by employing artificial neural networks to process data.

SVM (Support Vector Machine)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Research Methodology

There are 5 steps to analyze sentiment data and here's the graphical representation of the methodology to do the same.

Fig. (2)

Methods of Sentiment Analysis

- **Data Collection**

Consumers usually express their sentiments on public forums like the blogs, discussion boards, product reviews as well as on their private logs – Social network sites like Facebook and Twitter. Opinions and feelings are expressed in different way, with different vocabulary, context of writing, usage of short forms and slang, making the data huge and disorganized. Manual analysis of sentiment data is virtually impossible. Therefore, special programming languages like 'R' are used to process and analyse the data.

- **Text Preparation**

Text preparation is nothing but filtering the extracted data before analysis. It includes identifying and eliminating non-textual content and content that is irrelevant to the area of study from the data.

- **Sentiment Detection**

At this stage, each sentence of the review and opinion is examined for subjectivity. Sentences with subjective expressions are retained and that which conveys objective expressions are discarded. Sentiment analysis is done at different levels using common computational techniques like Unigrams, lemmas, negation and so on.

- **Sentiment Classification**

Sentiments can be broadly classified into two groups, positive and negative. At this stage of sentiment analysis methodology, each subjective sentence detected is classified into groups- positive, negative, good, bad, like, dislike.

- **Presentation of Output**

The main idea of sentiment analysis is to convert unstructured text into meaningful information. After the completion of analysis, the text results are displayed on graphs like pie chart, bar chart and line graphs.

Related Work

Linear SVM classification of sentiment in tweets about airlines

This notebook describes an attempt to classify tweets by sentiment. It describes the initial data exploration, as well as implementation of a linear one-vs-rest Support-Vector-Machine (SVM) classifier.

Dataset

It's always good to start by exploring the data that we have available. To do this we load the raw csv file using Pandas.

Characterizes text of different sentiments

While we still haven't decided what classification method to use, it's useful to get an idea of how the different texts look. This might be an "old school" approach in the age of deep learning, but lets indulge ourselves nevertheless.

To explore the data we apply some crude preprocessing. We will tokenize and lemmatize using Python NLTK, and transform to lower case. As words mostly matter in context we'll look at bi-grams instead of just individual tokens.

As a way to simplify later inspection of results we will store all processing of data together with it's original form. This means we will extend the Pandas dataframe into which we imported the raw data with new columns as we go along.

Linear SVM classifier

We will build a simple, linear Support-Vector-Machine (SVM) classifier. The classifier will take into account each unique word present in the sentence, as well as all consecutive words. To make this representation useful for our SVM classifier we transform each sentence into a vector. The vector is of the same length as our vocabulary, i.e. the list of all words observed in our training data, with each word representing an entry in the vector. If a particular word is present, that entry in the vector is 1, otherwise 0.

To create these vectors we use the Count Vectorizer from sklearn.

Fitting the classifier

We're now ready to fit the classifier to our data. We'll spend more time on hyper parameter tuning later, so for now we just pick some reasonable

guesses. Note here that we use the OneVsRestClassifier. This allows us to get the probability distribution over all three classes. Behind the scenes we actually create three classifiers. Each of these classifiers determines the probability that the datapoint belongs to it's corresponding class, or any of the other classes. Hence the name OneVsRest.

Evaluation of results It's most likely possible to achieve a higher score with more tuning, or a more advanced approach. Lets check on how it does on a couple of sentences.

Result and Discussion

SVM classifier is a modern classifier to classify data of any dataset to predict that given information is in the form of positive, negative and natural. Clearly the positive data is much harder for the classifier. This makes sense since there's a lot less of it. An important challenge in building a better classifier will then be how to handle positive data.

Conclusions

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analysing the sentiments of the tweets and feeding the data to a machine learning model to train it and then check its accuracy, so that we can use this model for future use according to the results. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 85%-90%.

FUTURE OUTLOOK

But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analyzers don't perform well when the number of classes are increased. Also, it's still not tested that how accurate the model will be for topics other than the one in consideration. Hence sentiment analysis has a very bright scope of development in future.

References

- 1) What is sentiment analysis from https://monkeylearn.com
- 2) More about sentiment analysis from https://greeksforgreeks.com
- 3) Literature Review on sentiment analysis using SVM from https://medium.com
- 4) Literature Review on sentiment analysis using SVM from https://thesai.org.com
- 5) What is Support Vector Machine from https://greeksforgreeks.com , https://analyticsvidya.com
- 6) Learn about Implantation of Sentiment Analysis from www.kaggle.com and monkey learn.com
- 7) Sentiment Analysis A Complete Guide by Gerardus Blokdyk.
- 8) Feature selection and model Selection in Supervised Learning by Jianbo Yang.
- 9) Introduction of Machine Learning.
- 10) Pervious Journal paper on sentiment analysis.
- 11) Machine Learning from Tutorial Point.
- 12) SVM from Java t point.