**AI ML PROJECT**

**TWITTER SENTIMENT ANALYSIS**

**Submitted by:**

**PRATHAM DEWAN (19CSU223)**

**SANYA PANDEY (19CSU275)**

**Submitted to:**

**DR. VIDHI KHANDUJA**

Contents

**ABSTRACT:**

**INTRODUCTION:**

**LITERATURE SURVEY**

**TECHNIQUES USED**

**DATA COLLECTION**

**PRE-PROCESSING**

**DATA VISUALISATION**

**FEATURE EXTRACTION**

**DATA SPLITTING**

**TRAINING MODLES**

**RESULTS**

**CONCLUSIONS**

**REFERENCES**

**ABSTRACT:**

Time keeps on changing and does not stop for anyone, with that being said the world as well doesn’t stop and moves along, Internet being one of the biggest achievements of mankind helps us go around our ways.

Internet being easily accessible to all is also one of the reasons why people tend to talk about their individual approaches and opinions over there more freely than in person.

In this project we have tried our best to use machine learning algorithms to help us decode the sentiments behind the tweets which are available on twitter. Twitter being one of the most open platforms (micro-blogging and social networking platform) that welcomes each and every opinion. Hence this paper explores the various sentiments that the Tweets in Twitter carry.

**INTRODUCTION:**

Sentiment analysis also can be known as “opinion mining”, is the automated process of identifying and classifying subjective information in text data. This might be an opinion, a judgment, or a feeling about a particular topic or product feature.it refers to identifying as well as classifying the sentiments which are expressed in our source.

Sentiment analysis uses NATURAL LANGUAGE PROCESSING(NLP) that helps to make sense of human language and machine learning to deliver accurate results automatically. Here text analysis and computational linguistics are also used.

The most common type of sentiment analysis is ‘polarity detection’ and involves classifying statements as positive, negative or neutral.

The reason why we chose to extract our data from twitter is because in the recent year the growth of websites like these, where public sentiments are rawer and more original, it’s not conceited into anything else than what it really is.

Example:

CASE 1=> Titanic is a good movie

CASE 2=>Titanic is a movie

CASE 3=>Titanic is a bad movie

In the above given examples, we can easily classify the fact in the case 1 we have a positive stance towards the movie titanic where as in the case 2 we can clearly see that the individual stance on titanic is very neutral in nature whereas in case 3 we can notice that the stance towards the movie is not very positive.

We can use this to help us analyse the public sentiment about various topics like politics, business, public actions, product reviews etc.

**LITERATURE SURVEY**

Through our research of various papers, the basic idea by which we could perform sentiment analysis is by extracting out data, pre-processing it and then dividing It in our testing and training set and then running it through the various machine learning algorithms. Some sentiment analyses are performed by analysing the twitter posts about electronic products like cell phones, computers etc. using Machine Learning approach. By performing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. Khan et.al. has presented a new feature vector for classifying the tweets as positive, negative or neutral and extract people’s opinion about products []. Another research tried to pre-processed the dataset, after that extracted the adjective from the dataset that have substantial meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning algorithms such as Naïve-Bayes, Maximum Entropy and SVM along with the Semantic Orientation based Word-Net which extracts synonyms and relation for the content feature [ ]. At the end, they measured the performance of classifier in terms of recall, precision and accuracy. Some researchers had an approach where posted tweets from the Twitter micro-blogging site are subjected to pre-processing and classified based on their emotional content as positive, negative and neutral or irrelevant; and compares the performance of various classifying algorithms based on their precision and recall in such cases. Further, the paper also discusses the applications of this research and its limitations. A number of machine learning like Naïve Bayes and Random Forest models performed sentiment analysis on product review data. Some work in this field included experiments with mood classification on blog posts. One of the researches also deals with review of aspect-based opinion polling from unlabelled free-form textual customer reviews without requiring customers to answer any questions.

**TECHNIQUES USED**

* In this project we have used various libraries like

NumPy

Pandas

Nltk

Porter stemmer

Sklearn

Tweepy

Text blob

Word cloud



* NLP – NATURAL LANGUAGE PROCESSING

Refers to the branch of computer science and artificial intelligence that Is concerned with giving the computers the ability to understand text and spoken works in pretty much the same way the humans can understand.

In this project we have used NLTK (natural language toolkit) which is one of the libraries of NLP that is a platform used in python programs that work with human language data for applying in statistical NLP.

NLTK contains text pre-processing libraries for tokenization, parsing, classification, stemming, tagging, semantic reasoning etc.

* LOGISTIC REGRESSION

It is a machine learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and is based on the concept of probability.

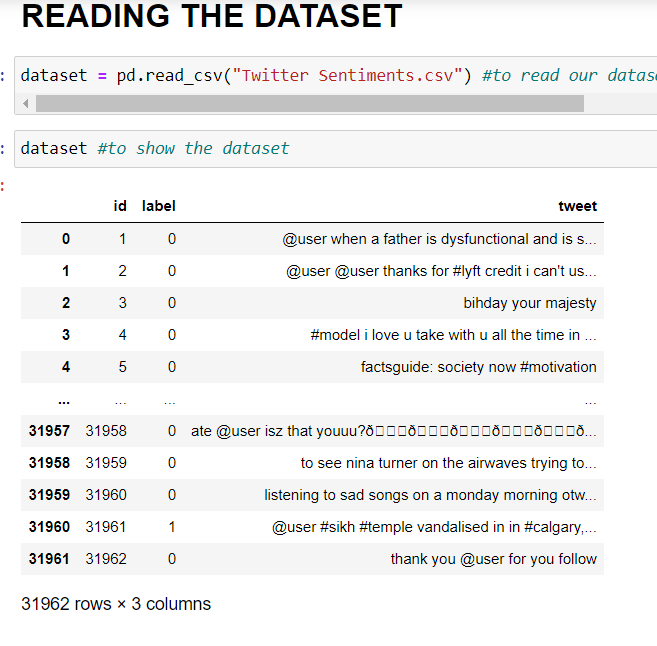
It is a statistical analysis method used to predict a data value based on prior observation of data set

It also predicts a dependent data variable by analysing the relationship between

one or more existing independent variables.

**DATA COLLECTION**

We have collected our dataset using the tweepy extension of twitter, where we first had to apply for the developer id and got our API key, then we stored our dataset into a csv file, by the name of Twitter Sentiments.csv.



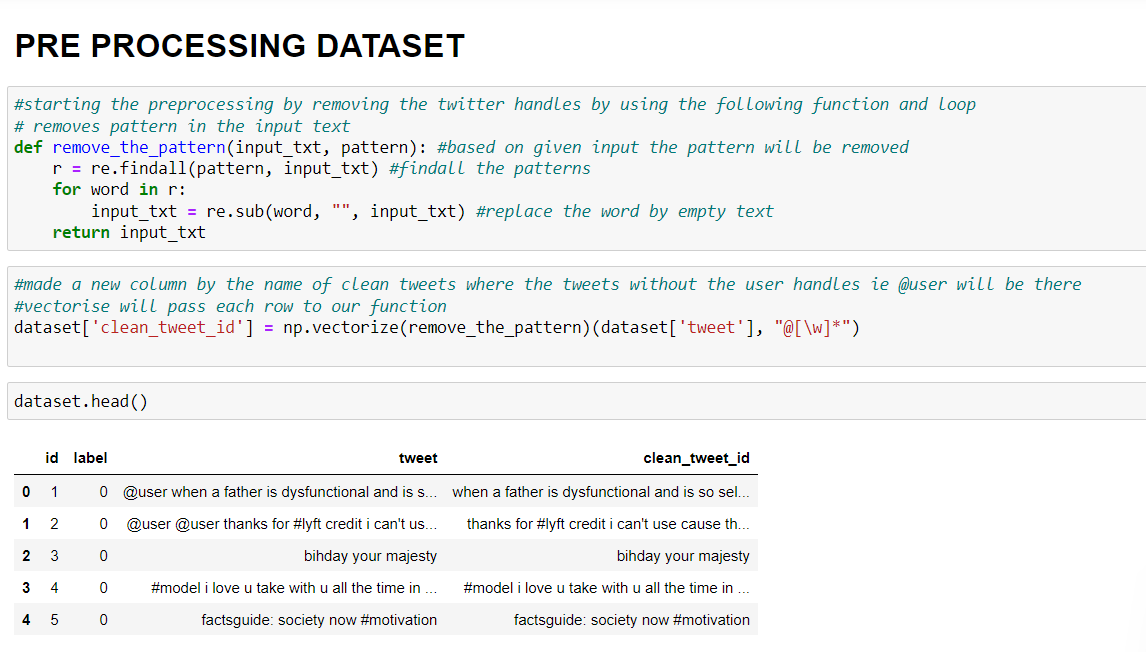
Following that we have also explored out dataset and got to know about the columns, length, shape, type of dataset, null values, unique id values etc.



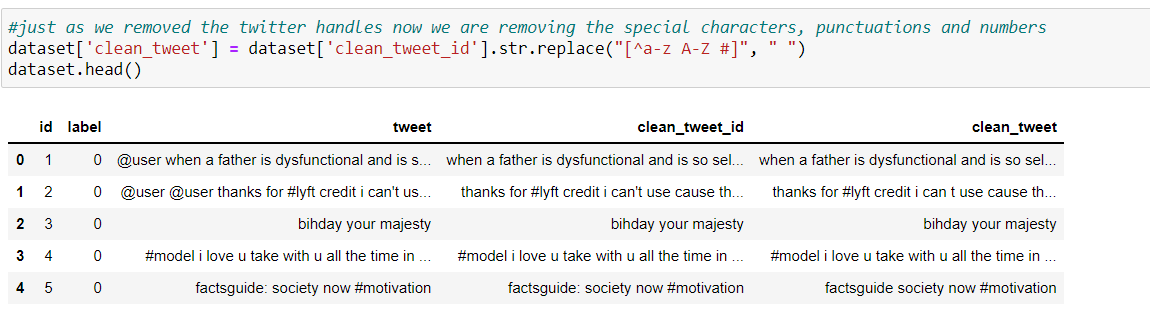
**PRE-PROCESSING**

1. We started the pre-processing by removing the:

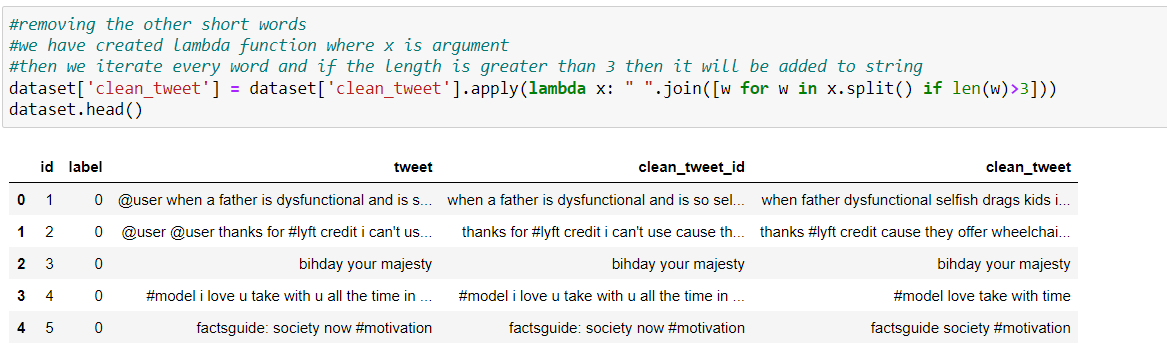
* Twitter handles



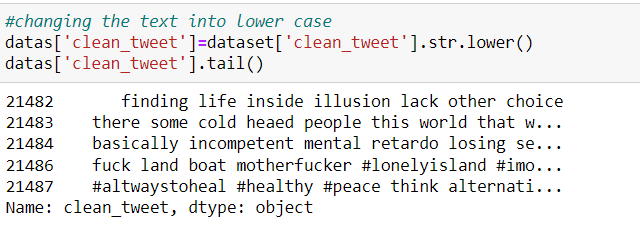
* Patterns in input text
* Special characters



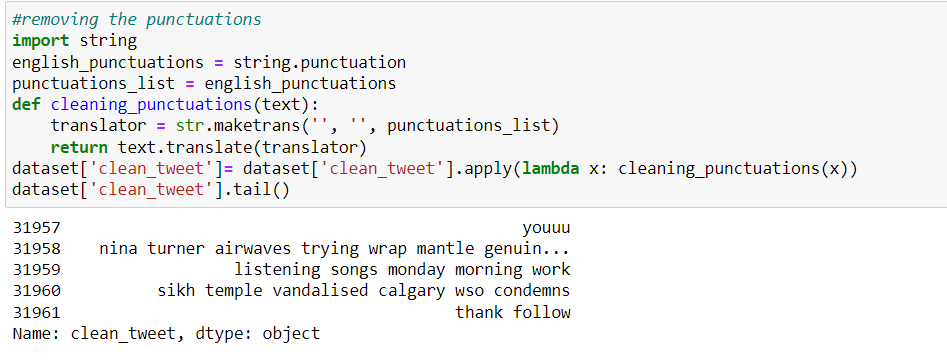
* Emoticons
* Words shorter than 3 letters



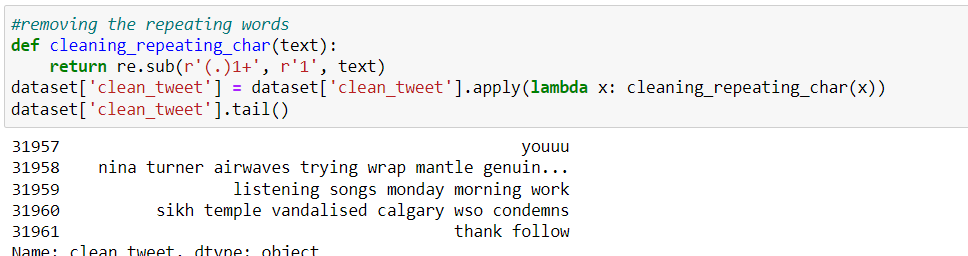
* Changing them to lower case



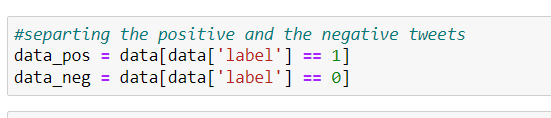
* Punctuations



* Repeating words



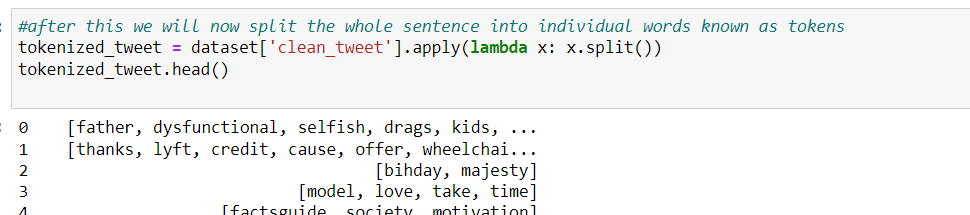
1. Then we had to assign values to our positive and negative sentiment. We labelled positive as 1 and negative as 0.

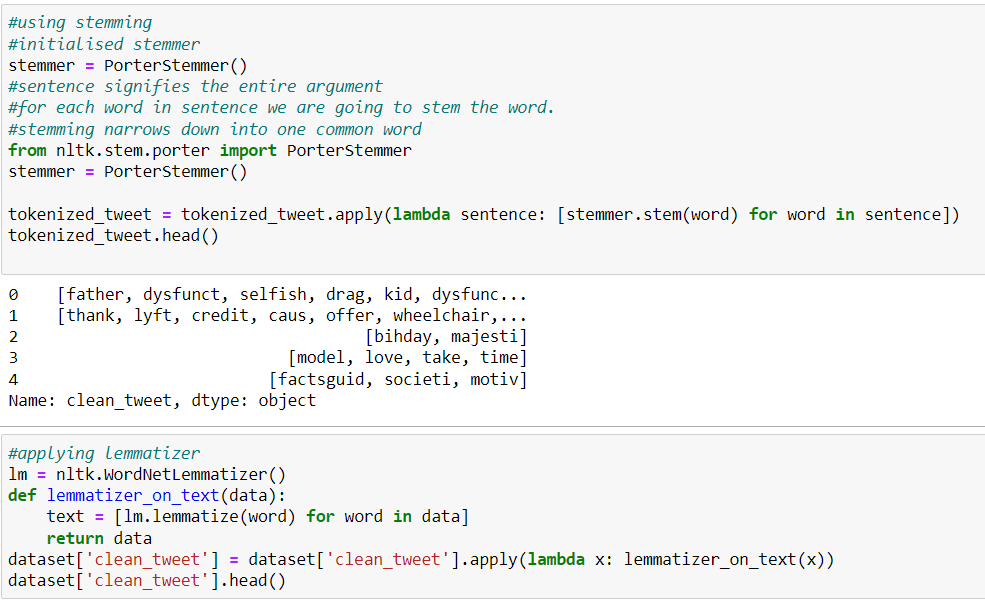


1. We also defined our stop words also made sure to have them cleaned out.

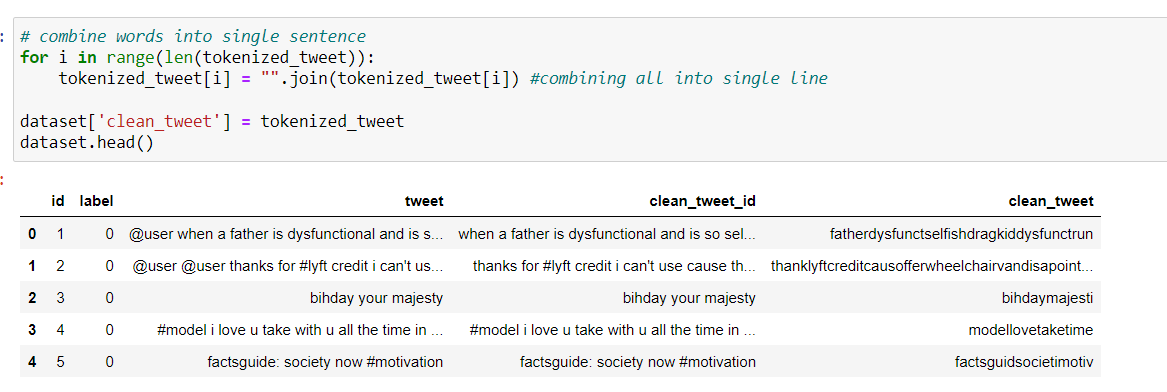


1. Later we also split the whole sentence into individual words known as tokens and after the process of tokenisation we did stemming by initialising out Porter Stemmer, with stemming it was also essential to make sure that we did lemmatization as well.



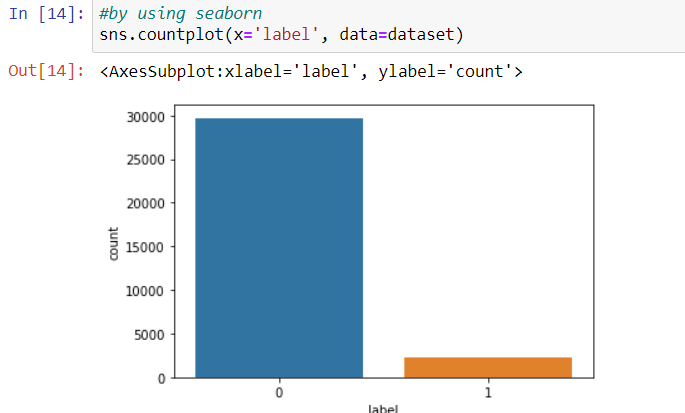


1. Later after all the pre processing the data was combined again into a single sentence.



**DATA VISUALISATION**

The data collected was visualised using seaborn.

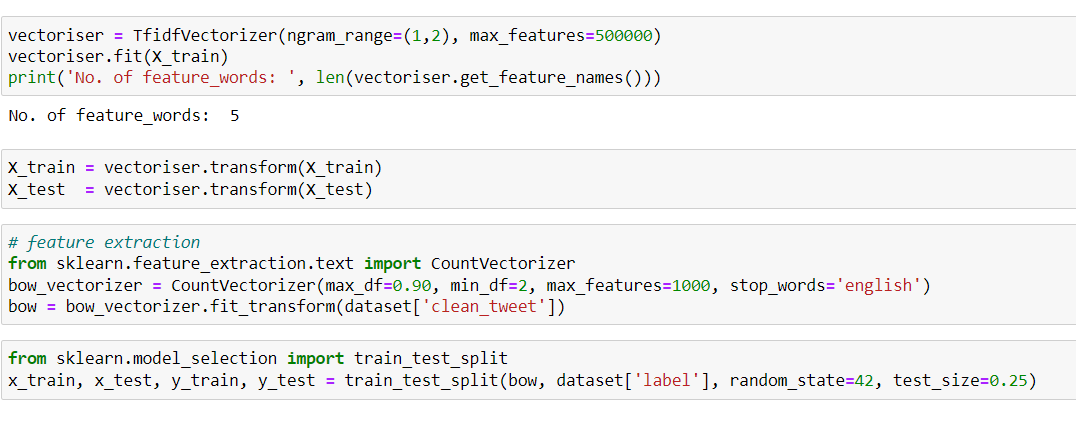


Word cloud was also used to visually represent the tweets, positive and negative as well.



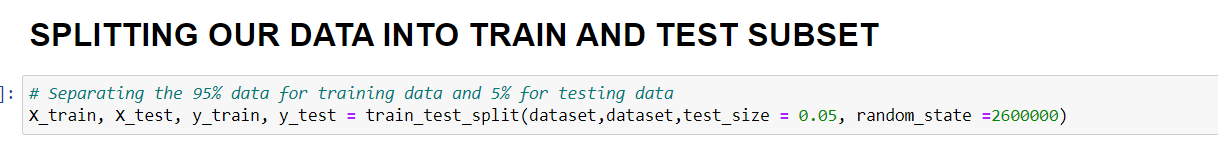
**FEATURE EXTRACTION**

We used various techniques like sklearn feature extraction and also TF-IDF (term frequency inverse document frequency) is used to transform meaningful text into representation of numbers which is used to fit machine learning algorithms.



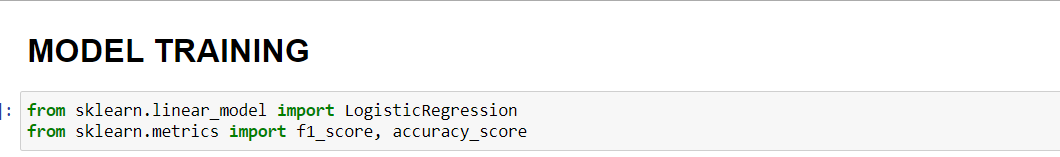
**DATA SPLITTING**

We Separated the 95% data for training data and 5% for testing data,



**TRAINING MODLES**

Logistic regressions are used as the training model in our project, the reason why we choose Logistic regression is a supervised learning algorithm which is mostly used to solve binary “**classification”**tasks although it contains the word “**regression**”. “Regression” contradicts with “classification” but the focus of logistic regression is on the word “logistic” referring to **logistic function** which actually does the classification task in the algorithm. Logistic regression is a simple yet very effective classification algorithm so it is commonly used for many binary classification tasks. Customer churn, spam email, website or ad click predictions are some examples of the areas where logistic regression offers a powerful solution. It is even used as an activation function for neural network layers.



**RESULTS**

****

**CONCLUSIONS**

To conclude we can say that twitter sentiment analysis using various machine learning models Is not just useful in today’s time but will be something that is useful in all our coming days, the more we take help of these models the better we will be able to understand how the stances are taken in the world.

**REFERENCES**

<https://www.upgrad.com/blog/sentiment-analysis-projects-topics-for-beginners/>

<https://www.upgrad.com/blog/sentiment-analysis-using-python/>

<https://www.cse.ust.hk/~rossiter/independent_studies_projects/twitter_emotion_analysis/twitter_emotion_analysis.pdf>

<https://www.rcciit.org/students_projects/projects/it/2018/GR33.pdf>

<https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python?utm_source=adwords_ppc&utm_medium=cpc&utm_campaignid=14989519638&utm_adgroupid=127836677279&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332602034364&utm_targetid=aud-392016246653:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9061702&gclid=Cj0KCQiA-eeMBhCpARIsAAZfxZDu2xOw2BJH4nxHY9DQ5ASGmIMDwXuXppImIb2hK3D01aE31CpvS68aAlRZEALw_wcB>

<https://towardsai.net/p/nlp/sentiment-analysis-with-logistic-regression>