Summary on Sentiments Analysis

Abstract

Currently social media has become a way to express the views. It is sometimes difficult to identify the sentiment of a person through written text. Thus, sentiment analysis is the method to identify the sentiment from the written text. Social Media like Facebook, Twitter, etc has data which we can use to identify the sentiments.

(I) Introduction

Sentiment analysis which is also known as opinion mining, is a method to automatic finding of opinions incarnated in text, is becoming a challenge in many research areas, particularly in data mining field for social media with several applications including product ratings and feedback analysis and customer decision making etc.

In general, sentiment analysis deals with detecting the polarity (e.g., positive, neutral, or negative) of the sentiment associated with a text, but the analysis can also consider the identification of the specific emotion

In sentiment analysis text is classified according to the following different criteria:

- the polarity of the sentiment expressed (into positive, negative, and neutral);
- the polarity of the outcome (e.g. improvement versus death in medical texts)
- agree or disagree with a topic (e.g. political debates)
- good or bad news
- support or opposition
- pros and cons

(II) Approaches for Sentiments Analysis

- Researchers investigated the content-based correlations among the topics, and calculated TCS to measure them. The assumptions about social context and topical context were both corroborated by the hypothesis testing over the Twitter data set they created.
- The performance of Bag of words sometimes remains limited due to some fundamental deficiencies in handling the polarity shift problem.So,to address this problem for sentiment classification they proposed a model called dual sentiment analysis (DSA). They first proposed a novel data expansion technique by creating a sentiment-reversed review for each training and test review. On this basis, they proposed a dual training algorithm to make use of original and reversed training reviews in pairs for learning a sentiment classifier.
- For textual mining of the user data, we need a lexicon dataset that reflects the keywords one may interpret as negative, positive or neutral. Negativity and Positivity are both incredibly large domains. The second edition of the 20 volume Oxford Dictionary contains 171,476 words. All words contained in the lexicon are then mined for, in the user's data set. Depending on the score/rank assigned to each word in the lexicon, the respective user tweet is assigned a value which indicates the polarity of the tweet content Approximately 200 new words synonymous with anger, hatred, jealousy, depression, alcoholism, suicide, apathy and other extreme emotions were added to the negative lexicon.

- Machine learning technique uses a labelled training set and a real test set for classification. Labelled Training set contains input feature vectors and their corresponding class labels. Using this training data set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. Several machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews.

- NLP techniques are based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules. Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level. NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.

- Case-Based Reasoning (CBR) is one of the techniques available to implement sentiment analysis. CBR is known by recalling the past successfully solved problems and use the same solutions to solve the current closely related problems. identified some of the advantages of using CBR that CBR does not require an explicit domain model and so elicitation becomes a task of gathering care histories and CBR system can learn by acquiring new knowledge as cases. This and the application of database techniques make the maintenance of large columns of information easier.

- Artificial Neural Network (ANN) or known as neural network is a mathematical technique that interconnects group of artificial neurons. It will process information using the connections approach to computation. ANN is used in finding the relationship between input and output or to find patterns in data.

- Support Vector Machine is to detect the sentiments of tweets together with stated SVM is able to extract and analyse to obtain upto70%-81.3% of accuracy on the test set. Collected training data from three different Twitter sentiment detection websites which mainly use some pre-built sentiment lexicons to label each tweet as positive or negative. Using SVM trained from these noisy labelled data, they obtained 81.3% in sentiment classification accuracy.

(III) Extraction of the Sentiment

The availability of free semantic networks is very rare. The semantic network which can be advocated and used is WordNet. Among the available semantic networks, WordNet covers relatively larger number of semantic concepts and hence it is preferred to be used. The selected semantic network can be used to calculate the sentiment polarity values for adjectives and adverbs. Every semantic concept in the sets can be assessed by a lexical unit and the index for that lexical unit conveys the desired concept. The Domain knowledge of the semantic graph and the distance between the entities plays a vital role in judging the polarity of the sentiment.

(IV) Sentiments analysis levels

- Level 1 is the sentence level, which detects positive, negative and neutral sentiment for each sentence.

- Level 2 is the document level, which detects the whole document sentiment as one unit or one entity positive or negative or neutral.

- Level 3 is the aspect level and it is used in case of the availability of attributes inside entity, post or input text. Each attribute can hold a sentiment in its own.

- Level 4 is the user level which handles the social relationships between different users using graph theory

(V) SENTIMENT ANALYSIS ENHANCEMENT METHODS

1. Sentiment analysis data cleaning
   - Data cleaning operations include tokenizing, stemming and filtering.
2. Dimension Reduction
   - Feature extraction is a transformative method which applies a transformation on the data to project it into a new feature space with lower dimensions
3. Sentiment Analysis and Data Integration.
   - Data is integrated, from different sentiment lexicons for sentiment analysis classification. This integration is performed by combining, filtering and deleting the duplicated data from individual dictionaries.
4. Sentiment Analysis and Crowdsourcing
   - Crowdsourcing can help in providing more accurate sentiment analysis results. Crowd can help in assigning labels to training data set or giving feedback about sentiment classification results, which can enhance the predication and the classification models.
5. Sentiment Analysis and Ontologies
   - Ontology based sentiment analysis systems produces more efficient classifiers and presents more detailed analysis about the results.

6. Sentiment Analysis and Spam Detection

7. Sentiment Analysis and User Profiling
   - Online user identity and user profiling helps in making the sentiment analysis results more accurate as it measures polarity based on the user profile in addition to the post polarity
8. Sentiment Analysis and Text Summarization
   - Sentiment summarization is the process of summarizing sentiment according to a specific domain or a topic, also called target based summarization.

(VI) Sentiment Analysis using Naïve Bayes classifier

The main reason behind designing and modeling sentiment analysis using Naive Bayes classifier is due to the following facts:

- It is very easy to build and comes very much handy while working with large data sets. Being the simplest among the analysing algorithms, it is known to outperform even with highly sophisticated methods of classification.
- It provides a way of classifying and calculating the subsequent probability. Naive Bayes classifier is one of the machine learning algorithms that uses Bayes algorithm with the assumption of strong independence among the features.

The sentiment analysis with Naive Bayes classifier is used to find the positive, neutral and negative sentiment of the tweets that are extracted from the data set.

Algorithm:

1. Consider a training data set D consists of documents which belongs to different classes say class A and B.
2. Prior probability of both classes A and B is calculated as shown Class A=number of objects of class A / total number of objects. Class B=number of objects of class B / total number of objects.
3. Now calculate the total number of word frequencies of both classes A and B i.e., ni na = the total number of word frequency of class A. nb =the total number of word frequency of class B.
4. Calculate the conditional probability of keyword occurrence for given class

   P(word1 / class A) = wordcount / ni(A)

   P(word1 / class B) = wordcount / ni(B)

   P(word2 / class A) = wordcount / ni(A)

   P(word2/classB)=wordcount/ni(B)

   …………………………………………

   P(wordn / class B) = wordcount / ni (B)

5. Uniform distributions are to be performed in order to avoid zero frequency problem.
6. Now a new document M is classified based on calculating the probability for both classes A and B P (M/W).
   a. Find P(A / W) = P(A) * P(word1/class A)* P(word2/ class A)…* P(wordn / class A).
   b. Find P(B / W) = P(B) * P(word1/class B)*P(word2/ class B)……* P(wordn / class B).
7. After calculating probability for both classes A and B the class with higher probability is the one the new document M assigned.

(VII) Approach using Naïve Bayes classifier

1. The dataset is collected which contains 65536 tweets these tweets are collected based on the situation on all topics.
2. the tweet containing sentiment is taken as 0 and tweet without any sentiment is declared as 1, and the third attribute sentiment source represents the source from the tweet is taken and of maximum length 140 characters, and the last attribute sentiment text represents the text or tweet based on all situations either containing sentiment or not.
3. So, in order to classify data first, we need to perform the following steps.
   - Tokenization: It is a method that divides the variety of document into small parts called tokens. These tokens may be in the form of words or numbers or punctuation marks. Ex: it is going to rain today After performing tokenization the sentence is divided into tokens as follows "It", "is", "going", "to", "rain", "today".
   - Stop words: These are the common words that are to be ignored which reduces the size of the dataset also the no of words (tokens). In our programming language python we use a tool called natural language tool kit (NLTK) in which there is list of stop words in 16 different languages. Ex: I like dancing, so I dance. After removing stop words the sentence will be as follows Like, dancing, dance.
   - Bag of words concept is applied to these tokens.
   - Finally, our classification technique Naïve Bayesian classifier is applied which calculates the probability of all words in the document and gives the result i.e., probability of each tweet in both positive and negative.
   - Results show the probability of each tweet saying whether the tweet is either positive or negative.

❖ *Bag-of-words*

A bag-of-words is a representation of text that describes the occurrence of words within a document. The occurrence of words is represented in a numerical feature. It is a way of extracting features from the text for use in modelling, such as with machine learning algorithms. The approach is very simple and flexible and can be used for extracting features from documents. But there is some complexity on two cases i.e., one is on designing the vocabulary of known words and the other is on scoring the presence of known words.

❖ *Application of sentiment analysis*

1. Naïve Bayes classifier is one of the supervised classification techniques which classifies the text/sentence that belongs to particular class. It is the probabilistic algorithm which calculates the probability of each word in the text/sentence and the word with highest probability is considered as output.
   a. Let us consider a document a
   b. A document a with a set of classes B = { b1, b2, … , bn }
   c. Consider a training set having m documents which is pre-determined that belongs to a particular class.

2. Now we train our classification algorithm using this training set and we get trained classifier. By using this trained classifier, we can classify the new document

(VIII) Advantages of Sentiment analysis

Twitter dataset contains users posts, their views and opinions based on the situation. The main objective of our proposed system is to perform analysis on tweets having sentiment which causes the great help to business intelligence on predicting the future. This paper addresses the sentiment analysis on twitter dataset; that is at first classification is performed on tweets using naïve bayes classifier. Each tweet is represented in the form of sentiment asserted in terms of positive, negative and neutral.

Performing sentiment analysis is vital which is used to find out the pros and cons of their products in the market by public that results in improving their business productivity. The aim of this project is to develop a classification technique using machine learning which gives accurate results and automatic sentiment classification of an unknown tweet by predicting the future

(IX) Implementation on Twitter Dataset.

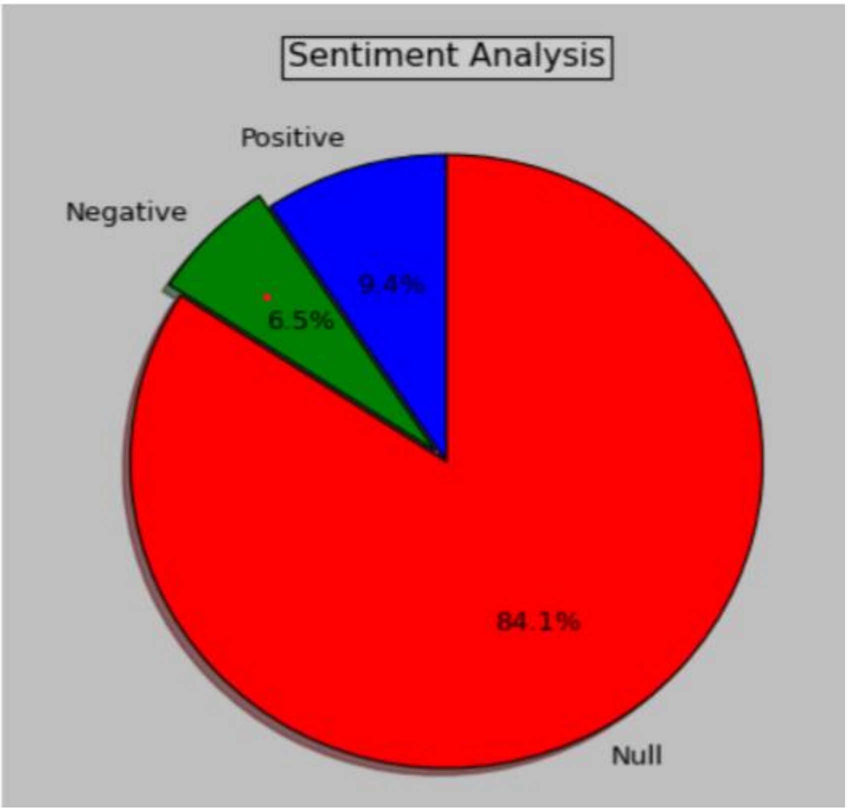
1. To associate with Twitter API, developer need to agree in terms and conditions of development Twitter platform which has been provided to get an authorization to access a data. The output from this process will be saved in JSON file.
2. The reason is, JSON (JavaScript Object Notation) is a lightweight data-interchange format which is easy for humans to write and read. Moreover, stated that, JSON is simple for machines to generate and parse. JSON is a text format that is totally language independent, but uses a convention that is known to programmers of the C-family of languages, including Python and many others
3. Nevertheless, the output will be categorized into 2 forms, which are encoded and un-encoded. According to security issue for accessing a data, some of the output will be shown in an ID form such as string ID. Sentiment Analysis. The tweets will be assigned the value of each word, together with categorize into positive and negative word, according to lexicon dictionary. The result will be shown in .txt, .csv and html.

❖ *Sentiment Analysis*

Tweets from JSON file will be assigned the value of each word by matching with the lexicon dictionary. As a limitation of words in the lexicon dictionary which is not able to assign a value to every single word from tweets. However, as a scientific language of python, which is able to analyse a sense of each tweet into positive or negative for getting a result.

❖ *Information Presented*

The result will be shown in a pie chart which is representing a percentage of positive, negative and null sentiment hash tags. For null hash tag is representing the hash tags that were assigned zero value. However, this program is able to list a top ten positive and negative hash tags.



(X) Why to use python for Analysis

Python is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best languages used by data scientist for various data science projects/application. Python provide great functionality to deal with mathematics, statistics and scientific function. It provides great libraries to deals with data science application.

One of the main reasons why Python is widely used in the scientific and research communities is because of its ease of use and simple syntax which makes it easy to adapt for people who do not have an engineering background. It is also more suited for quick prototyping.

In terms of application areas, ML scientists prefer Python as well. When it comes to areas like building fraud detection algorithms and network security, developers leaned towards Java, while for applications like natural language processing (NLP) and sentiment analysis, developers opted for Python, because it provides large collection of libraries that help to solve complex business problem easily, build strong system and data application.

Following are some useful features of Python language:

- It uses the elegant syntax, hence the programs are easier to read.
- It is a simple to access language, which makes it easy to achieve the program working.
- The large standard library and community support.
- The interactive mode of Python makes its simple to test codes.
- In Python, it is also simple to extend the code by appending new modules that are implemented in other compiled language like C++ or C.
- Python is an expressive language which is possible to embed into applications to offer a programmable interface.
- Allows developer to run the code anywhere, including Windows, Mac OS X, UNIX, and Linux.
- It is free software in a couple of categories. It does not cost anything to use or download Pythons or to add it to the application.

Most Commonly used libraries for data science:

1. Numpy: Numpy is Python library that provides mathematical function to handle large dimension array. It provides various method/function for Array, Metrics, and linear algebra. NumPy stands for Numerical Python. It provides lots of useful features for operations on n-arrays and matrices in Python. The library provides vectorization of mathematical operations on the NumPy array type, which enhance performance and speeds up the execution. It’s very easy to work with large multidimensional arrays and matrices using NumPy.

2. Pandas: Pandas is one of the most popular Python library for data manipulation and analysis. Pandas provide useful functions to manipulate large amount of structured data. Pandas provide

easiest method to perform analysis. It provide large data structures and manipulating numerical tables and time series data. Pandas is a perfect tool for data wrangling. Pandas is designed for quick and easy data manipulation, aggregation, and visualization. There two data structures in Pandas –

3. Matplotlib: Matplolib is another useful Python library for Data Visualization. Descriptive analysis and visualizing data is very important for any organization. Matplotlib provides various method to Visualize data in more effective way. Matplotlib allows to quickly make line graphs, pie charts, histograms, and other professional grade figures. Using Matplotlib, one can customize every aspect of a figure. Matplotlib has interactive features like zooming and planning and saving the Graph in graphics format.

4. Scipy: Scipy is another popular Python library for data science and scientific computing. Scipy provides great functionality to scientific mathematics and computing programming. SciPy contains sub-modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, Statmodel and other tasks common in science and engineering.

5. Scikit – learn: Sklearn is Python library for machine learning. Sklearn provides various algorithms and functions that are used in machine learning. Sklearn is built on NumPy, SciPy, and matplotlib. Sklearn provides easy and simple tools for data mining and data analysis. It provides a set of common machine learning algorithms to users through a consistent interface. Scikit-Learn helps to quickly implement popular algorithms on datasets and solve real-world problems.

References

https://www.ijcaonline.org/research/volume125/number3/dandrea-2015-ijca-905866.pdf

https://www.ijcaonline.org/research/volume125/number3/dandrea-2015-ijca-905866.pdf

https://www.ijitee.org/wp-content/uploads/papers/v8i8/H6330068819.pdf

DOI: 10.1109/ICIMU.2014.7066632

DOI: 10.1109/ICDMW.2015.142

DOI : 10.1109/IRI.2015.37

https://www.geeksforgeeks.org/python-for-data-science/

https://sproutsocial.com/insights/social-media-sentiment-analysis/

https://www.researchgate.net/publication/268817500_Sentiment_Analysis_for_Social_Media

https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed

Prepared By

Akrit Khanna (D18CE159)

akritkhanna12@gmailc.com

Esha Shah(D18Ce148)

d18ce148@charusat.edu.in