International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206521

Opinion Mining Analysis of Twitter Users of Particular Topic

Sneha Naik[1], Mona Mulchandani[2]
[1]M. Tech Scholar, JIT College Govardhan, Maharashtra, India
[2]Professor, JIT College, Maharashtra, India

Article Info
Volume 6, Issue 5
Page Number: 102-108
Publication Issue :
September-October-2020

Article History
Accepted : 01 Sep 2020
Published : 12 Sep 2020

ABSTRACT

Opinion mining consists of many different fields like natural language processing, text mining, decision making and linguistics. Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and categorize opinions about a product. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to mine text for sentiment. This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange.

Keywords : Machine Learning, Sentiment Analysis, Feature Extraction, Opinion Mining, Natural Language Processing (NLP)

I. INTRODUCTION

Opinion mining is extract subjective information from text data using tools such as NLP, text analysis etc. Senitment analysis is also called as opinion mining and is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. Analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering "useful" patterns in large set of data, either automatically (unsupervised) or semiautomatically (supervised). The project would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of tweets and on "Machine Learning" techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them. This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream. Millions of messages are appearing daily in popular web-sites that provide services for microblogging. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to microblogging services. As more and more users post about products and services they use, or express their political and religious views, microblogging web- sites become valuable sources of people‟s opinions and sentiments. Such data can be efficiently used for adaptive user interface‟s. Data from these sources can be used in opinion mining and sentiment analysis tasks. Twitter social network is a service developed in order to facilitate communication between people by distributing short messages. [5][6]

Data Mining Hierarchy

Data mining is the process of finding anomalies, patterns and correlations within large **data** sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.

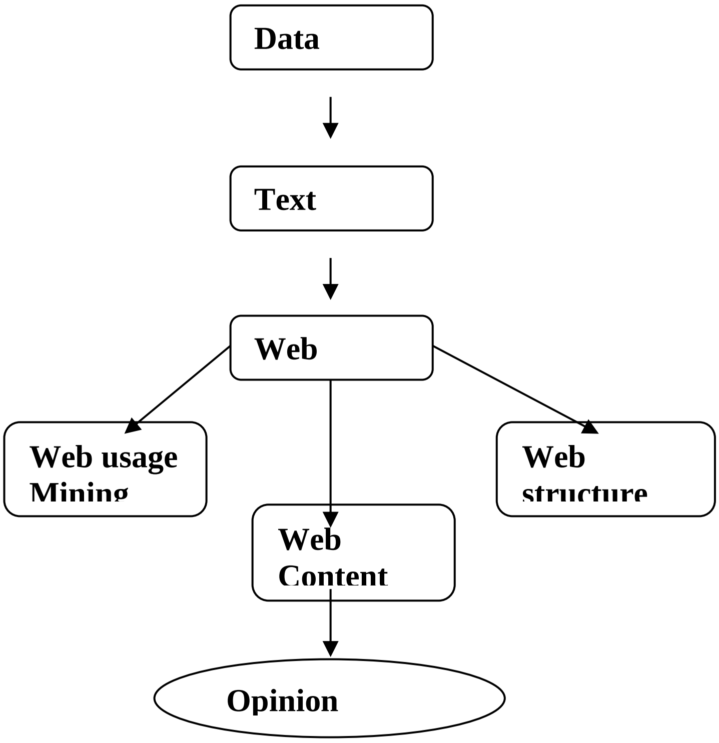

Figure 1. Data Mining Hierarchy

2 Opinion mining and sentiment analysis

Sentiment analysis systems help organizations gather insights from unorganized and unstructured text that comes from online sources such as emails, blog posts, support tickets, web chats, social media channels, forums and comments. In addition to identifying sentiment, opinion mining can extract the polarity (or the amount of positivity and negativity), subject and opinion holder within the text. Furthermore, sentiment analysis can be applied to varying scopes such as document, paragraph, sentence and sub-sentence levels. The sentiment may be his or her judgment, mood or evaluation. A key problem in this area is sentiment classification, where a document is labeled as a positive or negative evaluation of a target object (twitter comments, review on product etc.)

An important part of our information-gathering behavior is always to find out "what other people think ?". With the emergence of Web 2.0, a lot of opinion resources are available such as online review sites and personal blogs. They are a new opportunities and challenges to use information technologies to seek out and understand the opinions of costumers.

3. Text mining and Natural Langugage Processing

A. Text Mining

Text Mining is the process of deriving high-quality information from **text**. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text Mining is an Artificial Intelligence (AI) technology that uses Natural Language Processing (NLP) to transform the unstructured text in documents and databases into normalized, structured data suitable for analysis or to drive Machine Learning (ML) algorithms. The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive or predictive analytics.

Text mining identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data. Once extracted, this information is converted into a structured form that can be further analyzed, or presented directly using clustered HTML tables, mind maps, charts, etc. Text mining employs a variety of methodologies to process the text, one of the most important of these being **Natural Language Processing (NLP)**.

B. Natural Language Processing

Natural Language Understanding helps machines "read" text (or another input such as speech) by simulating the human ability to understand a *natural* language such as English, Spanish or Chinese. Natural Language Processing includes both Natural Language Understanding and Natural Language Generation, which simulates the human ability to create natural language text e.g. to summarize information or take part in a dialogue. Today's natural language processing systems can analyze unlimited amounts of text-based data without fatigue and in a consistent, unbiased manner. They can understand concepts within complex contexts, and decipher ambiguities of language to extract key facts and relationships, or provide summaries.

4 Data Preprocessing Task

The process of deriving information from the text. It usually requires a pre-processing of the input data. Some popular preprocessing steps are: tokenization, stop word removal, stemming, parts of speech (POS) tagging, and feature extraction and representation. Followings are the Pre-Processing Steps.

1. **Tokenization** is the process of **tokenizing** or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph. In lexical analysis, tokenization is the process of breaking up a stream of text into words, phrases, symbols or other meaningful elements called tokens [9].
2. **Stop words** are a set of commonly used **words** in any language. For example, in English, "the", "is" and "and", would easily qualify as **stop words**. Stop words is a list of words that doesn't have potential to contribute to characterize the content in the text. They can reduce the size of texts by 30% to 50%. In NLP and text mining applications, **stop words** are used to eliminate unimportant **words**, allowing applications to focus on the important **words** instead.
3. **Stemming** is basically removing the suffix from a word and reduce it to its root word.

For example: "**Flying**" is a word and its suffix is "**ing**", if we remove "**ing**" from "**Flying**" then we will get base word or root word which is "**Fly**" [12].

4. Part of Speech (POS) tagging also called grammatical **tagging** or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular **part of speech**, based on both its definition and its context—i.e., its relationship.

Different steps are involved for Data Preprocessing. These steps are described below –

1) Data Cleaning

This is the first step which is implemented in Data Preprocessing. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers, minimizing duplication and computed biases within the data.

2) Data Integration

This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for analysis.

3) Data Transformation

This step is used to convert the raw data into a specified format according to the need of the model. The options used for transformation of data are given below –

Normalization – In this method, numerical data is converted into the specified range, i.e., between 0 and one so that scaling of data can be performed.

Aggregation – The concept can be derived from the word itself, this method is used to combine the features into one. For example, combining two categories can be used to form a new group.

Generalization – In this case, lower level attributes are converted to a higher standard.

4) Data Reduction

After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data.

5 Data Pre-processing

Data Preprocessing is necessary because of the presence of unformatted real-world data. Mostly real-world data is composed of missing data, errorneous data and outliers, inconsisten data.

Inaccurate data (missing data) – There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.

The presence of noisy data (erroneous data and outliers) – The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.

Inconsistent data – The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more. [7]

6 Sentiment analysis techniques and approach

A. Approach

6.1 Knowledge-based approach

The main task in this approach is the construction of word lexicons that indicate positive class or negative class. The sentiment values of the words in the lexicon are determined prior to the sentiment analysis work. Lexicons can be created in different ways. It can be created by starting with some seed words and then using some linguistic heuristics to add more words to them, or starting with some seed words and adding to these seed words other words based on frequency in a text. SENTIWORDNET 3.0 is a publicly available lexical resource explicitly devised for supporting sentiment classification and opinion mining applications [8].

6.2 Relationship-based approach

In this approach the different relationships between features and components is analyzed for sentiment classification task. Such relationships may be relationships between different participants, relationships between product features. For example, if one wants to know the sentiment of customers about a product brand, one may compute it as a function of sentiments on different features or components of it.

B. Techniques

Sentiment analysis can be implemented using both supervised and unsupervised methods of classification. Supervised methods have shown better performance than the unsupervised methods. However, unsupervised methods are also important because supervised methods demand large amounts of labeled training data that very expensive whereas acquisition of unlabeled data is easy.

6.3 Supervised Techniques Supervised techniques can be implemented by constructing a classifier. This classifier is trained by examples which can be manually labeled. The popular supervised algorithms are Support Vector Machines (SVM), Naive Bayes classifier and Maximum Entropy. Supervised Techniques proved that they provide efficient performance.

6.4 Unsupervised Techniques In unsupervised technique, classification is done by comparing the features of a given text against word lexicons whose sentiment values are determined prior to their use. For example, start with positive, negative word lexicons, analyze the document for which sentiment Name of the Tool Purpose TweetMotif Tokenization of tweets POS tagger Twitter POS tagger TweetNLP6 Twitter natural language processing Lancaster stemming algorithm Stemmer GNU Aspell Spell Checker Snowball English stemmer Stanford Log-linear Part-Of-Speech Tagger POS tagger TweeboParser Tweet Dependency parser IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 1, February 2018 ISSN (Online) : 2277-5420 www.IJCSN.org Impact Factor: 1.5 32 Copyright (c) 2018 International Journal of Computer Science and Network. All Rights Reserved. need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative.

II. Implementation

If you're using RStudio, you can quickly connect and pull data that is publicly available. In the case of Twitter, one can pull lists of users, trending topics in different regions, as well as lists of followers. This type of assessment would prove helpful for corporate analysts, as well as people involved in the political realm, or those of us who are simply curious about the reach of certain data science tools. There are countless ways to analyze this type of data (thinking of boxplots, histograms and text mining to name a few!). After authenticating to the Twitter API, I wanted to know how to assess and visualize what the social media site's users are not only **saying**, but also may be **feeling**, about a given topic [14].

Step 1: Load the required packages in RStudio using command

```
install.packages(“rtweet”)
install.packages("tm")
install.packages("TwitteR")
```

Step 2: Authenticate using your credentials to Twitter API's by creating an access token.

a. developer.twitter.com
b. click on Sign in
c. enter user id and password
d. click on apps
e. create a new apps
f. click on details
g. go to keys and tokens tab

Step 3: Search tweets of your choice for example- Covid-19, Lockdown, IPL2020, CBI Enquiry etc.

Step 4: Process each set of tweets into tidy text or corpus objects. For example- Convert the tweets to a text format.

Step 5: Use pre-processing text transformations to clean up the tweets; this includes stemming words.

Step 6: Perform sentiment analysis using the get_sentiments function from the tidytext package.

Step 7: Get the sentiment score for each tweet like positive, negative and neutral sentiments.

III. REFERENCES

[1]. Golnar Assadat Afzali, Shahriar Mohammadi, "Privacy preserving big data mining: association rule hiding using fuzzy logic approach",published in Information Security IET in the year 2018.

[2]. Rajeshwari Dembala, S. Vagdevi, "Conceptual notion for opinion mining from upcoming big data", Electrical Electronics Communication Computer and Optimization TechniquesICEECCOT) 2017 International Conference.

[3]. M. Trupthi, Suresh Pabboju, G. Narasimha, "Sentiment Analysis on Twitter Using Streaming API", Advance Computing Conference (IACC) 2017 IEEE 7th International.

[4]. Shiv Dhar, SuyogPednekar, KishanBorad, Ashwini Save, "Sentiment Analysis Using Neural Networks: A New Approach", Inventive Communication and Computational Technologies (ICICCT) 2018 Second International Conference.

[5]. Xiaoya Xu, Qingsong Hua, "Industrial Big Data Analysis in Smart Factory: Current Status and Research Strategies",published in Access IEEE in the year 2017.

[6]. Gongqing Wu, Ying He, Xuegang Hu, "Entity Linking: An Issue to Extract Corresponding Entity With Knowledge Base" published in Access IEEE in the year 2018.

[7]. AobakweSenosi, George Sibiya, "Classification and evaluation of Privacy Preserving Data Mining: A review" published in AFRICON 2017 IEEE in the year 2017.

[8]. Kai Peng, Victor C. M. Leung, Qingjia Huang, "Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System Over Big Data" published in Access IEEE in the year 2018.

[9]. AichaAggoune, AbdelkrimBouramoul, Mohamed-KhiereddineKholladi, "Big data integration: A semantic mediation architecture using summary", Advanced Technologies for Signal and Image Processing (ATSIP) published in 2016 2nd International Conference in the year 2016.

[10]. Todd Bodnar, Conrad Tucker, Kenneth Hopkinson, Sven G. Bilén, "Increasing the veracity of event detection on social media networks through user trust modeling", Big Data (Big Data) published in IEEE International Conference in the year 2014.

[11]. Dazhi Yang, Gary S. W. Goh, Siwei Jiang, Allan N. Zhang, "Spatial data dimension reduction using quadtree: A case study on satellite-derived solar radiation", Big Data (Big Data) published in IEEE International Conference in the year 2016.

[12]. Muhammed O. Sayin, N. DenizcanVanli, Ibrahim Delibalta, Suleyman S. Kozat, "Optimal and Efficient Distributed Online Learning for Big Data", Big Data (BigData Congress) published in IEEE International Conference in the year 2015.

[13]. Ch Srinivasa Rao, Dr. G. Satyanarayan Prasad, ―A Survey on Opinion Mining on Twitter Data: Tasks, Approaches, Applications and Challenges for Sentimental Analysis‖, IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 1, February 2018 ISSN (Online) : 2277-5420,pp. 27-35.

[14]. I. Smeureanu, M. Zurini, ―Spam Filtering for Optimization in Internet Promotions using Bayesian Analysis,‖ Journal of Applied Quantitative Informatica Economică vol. 16, no. 2/2012 91 Methods, Vol. 5, Issue.2, pp. 198-211, 2010.

[15]. C. Bucur, T. Bogdan ―Solutions for Working with Large Data Volumes in Web Applications―, Proceedings of the 10th International Conference on Informatics in Economy - IE 2011 „Education, Research & Business Technologies‖, 5-7 Mai 2011, Printing House ASE, Bucharest, 2011.

[16]. MahnazRoshanaei and Shivakant Mishra, ―An Analysis of positivity and Negativity attributes of users on Twitter ―, IEEE/ACM conference on ASONAM, 17-20 Aug 2014, Beijing, China.

[17]. https://medium.com/@tusharsri/nlp-a-quickguide-to-stemming-60f1ca5db49e

[18]. John P Dickerson, vadim Kagan, V S Subrahmanian, ―Using sentiment to detect Bots on Twitter: Are Humans more opinionated than Bots? ―, IEEE/ACM conference on ASONAM, 17-20 Aug 2014, Beijing, China.

[19]. Sumbaly, R., Kreps, J., Shah, S.: The big data ecosystem at linkedIn. In: Ross, K.A., Srivastava, D., Papadias, D., (eds.) SIGMOD Conference, pp. 1125–1134. ACM (2013)

Cite this article as :

Sneha Naik, Mona Mulchandani, "Opinion Mining Analysis of Twitter Users of Particular Topic", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 102-108, September-October 2020.
Available at
doi : https://doi.org/10.32628/CSEIT206521
Journal URL : http://ijsrcseit.com/CSEIT206521