International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT2063100

Opinion Mining and Sentiment Analysis on Big Data

Shruti Choudhari[1], Shrikant Zade[2]
[1]M.Tech Scholar, PIET College, Haryana, India
[2]Professor, PIET College Haryana, India

ABSTRACT

Opinion mining is extract subjective information from text data using tools such as NLP, text analysis etc. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to mine text for sentiment. Opinion mining, which is also called sentiment analysis, involves building a system to collect and categorize opinions about a product.In this project the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in terms of 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 : Artificaial Intelligence(AI), Machie Learning, Sentiment Analysis, Feature Extraction, Opinion Mining, NLP

I. INTRODUCTION

1.1 Introduction to sentiment analsysis

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]

1.2 Opinion Mining Methodology

As aforementioned, opinion mining is deployed to extract, classify, understand, and assess the opinions implicit in text content. Further, sentiment analysis is often used in opinion mining to identify sentiment, affect, subjectivity, and emotional states toward entities, events, and their attributes in such content Therefore, a social media opinion-mining methodology should have processes that involve computational techniques to aggregate, extract, analyze, andpresent the sentiment and attitude of authors in social media content. A large volume of information in current online systems is stored in text form. This is the way information is transmitted on the internet, being the most natural representation form and easier to read by the people [11]. In this context, applying of data mining techniques to the content of web pages, (text mining on web pages) or content mining become important [12]. Between web mining and data mining are important differences in terms of data collection. In data mining is assumed that the data is already collected and stored in databases, while in web mining, are used special mechanisms, taken from areas such as information retrieval - IR (Information Retrieval) and information extraction - IE (Information Extraction), to obtain data and to pre-process them to apply data mining techniques. From the terms of proposed objectives, Web mining is divided into three categories:

1. Web structure mining - knowledge discovery from hyperlinks to maximize information about the relations between web pages;
2. Web usage mining - extracting models and patterns of users, from web logs (logs), that stores data access and activities of each visitor to a website, detecting website users requirements;
3. Web content mining - extracting knowledge from web page content [13].

II. 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.

1.1 Architecture of Opinion Mining

Opinion mining and summarization process contain three main steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization.

1.1.1 Opinion Retrieval

It is the process of gathering comments from different websites. Comments involve reviews about products, movies, hotels and news. Information retrieval techniques like web crawler can be applied to accumulate the review text data from many sources and store them in database. This step includes retrieval of reviews, micro blogs, and comments of user.

1.1.2 Opinion Classification

Basic step in opinion mining is classification of review text. Given a review document D = {d1…..d1} and a categories set C = {positive, negative}, sentiment classification is to classify each di in D, with a tag expressed in C. The method involves classifying review text into two forms namely positive and negative.

1.1.3 Opinion Summarization

Summarization of opinion is a main part in opinion mining process.

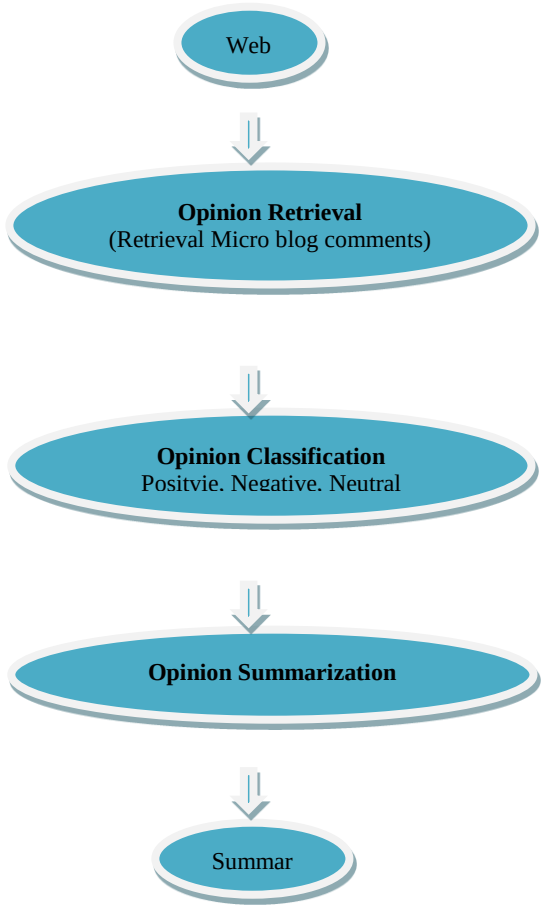

Figure 1: Architecture of Opinion Mining.

III. Text mining and Natural Langugage Processing

3.1 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)**.

3.2 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.

IV. 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**" [15].

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.

V. Why is Data Preparation is Important?

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]

VI. Application of sentiment analysis

1. Voice of Customers

At any time a product is to be commenced by a definite company, the customers would to recognize about the product ratings, reviews and comprehensive metaphors about it. Sentiment Analysis can assist in analyzing marketing, advertising and for making new tactics for endorse the product. It offers the customer a chance to prefer the best among the all.

2. Market Research, Competitor Analysis

Market research is probably the most prominent field of sentiment analysis application. It is important to note that sentiment analysis is not the primary tool for market research. However, it can bring an additional perspective on the market and give a couple of handy insights about how the state of things is seen from the ground level i.e. consumers.

Here's what you can do with sentiment analysis:

1. Gather information across different platforms
2. User-generated content (comments, reviews, etc)
3. News articles
4. Extract numerous insights on different criteria
5. Provide results in real-time

3. Customer Support- feedback analysis

3.1 Insight into customer's opinions regarding the product:

1) The general perception of the product - whether it is positive or negative;
2) Aspect-based - regarding specific elements of the product;
3) Reaction to the Service - whether it is effective or not. May also include more detailed analysis regarding particular aspects such as response time or quality of interaction;

3.2 Intent Analysis for process automation - so that routine queries will be handled automatic scenarios, such as frequently asked questions and basic product use information.

3.3 Workflow management and customer prioritization. For example, you have a disgruntled customer - his ticket is prioritized to be processed as soon as possible.

VII. REFERENCES

[1]. Ahmed Yousuf Saber, "IoT based Online Load Forecasting", 2017 Ninth Annual IEEE Green Technologies Conference, 2017.

[2]. H. S. Hippert, C. E. Pedreira, and R. C. Souza. "Neural networks for shortterm load forecasting: A review and evaluation," IEEE Trans. Power Syst., vol. 16, no. 1, pp. 44-55, Feb 2001.

[3]. J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in Proc.

[4]. IEEE Int. Conf. Neural Networks, 1995, vol. 4, pp. 1942–1948.

[5]. Malhar Anjaria and Ram Mohana Reddy Guddeti, " Influence Factor based opinion mining of Twitter data using supervised learning" sixth IEEE conference on COMSNETS, 6-10 Jan 2014, Bangalore, India.

[6]. Ana c E S Lima and Leandro N de Castro, " Automatic sentiment Analysis of Twitter messages", fourth IEEE conference on CASoN, 21-23 Nov 2012, Sao Carlos.'

[7]. https://www.xenonstack.com/blog/data-preparation/

[8]. Tuan Anh Hoang, William w Cohen, Ee-Peng Lim, Dovy Pierce, David R Redlawsk, "Politics, Sharing and emotion in Microblogs", IEEE/ACM conference on ASONAM, 2013, New York, USA

[9]. N. Madnani, "Getting Started on Natural Language with python", pp. 1-16, 2013

[10]. 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.

[11]. 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.

[12]. 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.

[13]. B. Liu, Web Data Mining - Exploring Hyperlinks, Contents and Usage Data, Secound ed.: Springer, 2011.

[14]. Mahnaz Roshanaei 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.

[15]. https://medium.com/@tusharsri/nlp-a-quick-guide-to-stemming-60f1ca5db49e

[16]. 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.

[17]. Processing with Python," pp. 1–16, 2013.

Cite this article as :

Shruti Rajkumar Choudhary, "Opinion Mining and Sentiment Analysis on Big Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.410-415, May-June-2020. Available at doi : https://doi.org/10.32628/CSEIT2063100

Journal URL : http://ijsrcseit.com/CSEIT2063100