Minor Project- Report

Aug-2019-2020

Course Faculty: Prof. Ravichandran H

Course Name & code: Big Data Analytics[17CS7DCBGD]

Semester: VII

Date: 01/01/2021

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TITLE OF THE PROJECT | Twitter Sentiment Analysis | | | |
|  |  | | | |
| STUDENT NAME | K CHANDRAHAS | HEMANTH KUMAR M | MANOJ V KHATOKAR | PREETHAM A |
| USN | 1DS17CS050 | 1DS17CS059 | 1DS17CS063 | 1DS17CS079 |
| INDIVIDUAL  CONTRIBUTION | Twitter Developer Account Setup and Python Scripting. | Apache Nifi installation and setup. | Hadoop installation and configurations. | Tweets analysis and formatting. |
| GUIDE | Prof. Poornima K S | | | |
|  |  | | | |
| PROJECT ABSTRACT : | Blogging and networking platforms like Facebook, Reddit, Twitter and LinkedIn are social media channels where users can share their thoughts and opinions. Since online chatter is a vital and exhaustive source of information, these thoughts and opinions hold the key to the success of any endeavor. Tweets which are posted by millions all over the world can be used to analyze consumers‟ opinions about individual products, services and campaigns. These tweets have proven to be a valuable source of information in the recent years, playing key roles in success of brands, businesses and politicians. We collect dataset, i.e. the tweets are fetched from Twitter and text mining techniques like tokenization are executed to use it for building classifier that is able to predict sentiments for each tweet. | | | |
| PLATFORM USED  (H/W & S/W TOOLS TO BE USED | Apache Nifi, Hadoop, Python, VS Code. | | | |
|  |  | | | |
| INTRODUCTION | With increase in penetration of social media in everyone’s daily lives, people can share their opinions on social networking websites like Facebook, Twitter, etc. People freely share their unbiased opinions on these platforms and thus, an opportunity has arisen where these free thoughts can be utilized to cater to the needs of the general public. Sentimental Analysis is the process of determining the emotion of any form of expression. Data is collected from various sources of information and then cleaned and structured for Sentiment Analysis to take place. Sentiment Analysis can help in improving user experience by analyzing negative emotions. It can generate additional revenue streams for a business by determining market for a new product/experience which will improve the emotion of the tweets by the target audience. It can improve customer service and thus, determine the success of the product or experience in question. This feedback loop results in corrective production and design, leading to continuous improvement of product or experience. Hidden Sentiment Identification is defined as the science behind realizing the actual intention behind a tweet as a tweet may not necessarily be restricted to a singular emotion.  Informal Language is widespread on social networks. It includes Elision and Assimilation. They are common when encountering informal ways of communication online. Elision of sounds is common in formal and informal language in abbreviated forms like isn't, I'll, who's, they'd, haven't and so on. It is very necessary to remember that sounds do not simply vanish. Meanwhile, Assimilation is commonly found in languages all over the world by which one sound becomes similar to a nearby sound. This can occur either inside the structure of a word or between more than one words. Complete assimilation takes place when the assimilated words‟ sound is a replica of the sound that it assimilates to. Partial assimilation is very common and very difficult to differentiate from others. When it occurs, the assimilated word creates the same sound, but disguised to be more similar, phonetically, to another word. Intermediate Assimilation is common in many languages, including English.  Our approach to tackle the problem of analyzing each tweet is to perform Sentiment Analysis by tokenization of each tweet (user opinion on twitter). We will split each tweet into its building blocks called tokens and then compare these tokens with dictionaries of positive and negative words. By comparing each token with these dictionaries, we will allot a positive, negative or neutral score for each token to the tweet. We will then calculate the overall weight of each tweet by addition of all the score of a tweet, thus obtaining the overall emotion of the tweet. By using this approach, we can have a fairly accurate idea of the overall emotion of the user and then analysis can be performed to take the input of the user for the next iteration of the product/experience. Using this approach can deliver fairly accurate results as dictionaries of positive and negative tokens are fairly exhaustive in nature. This will allow us to take all the tokens of a tweet into account while performing Sentiment Analysis. | | | |
|  |  | | | |
| DESIGN |  | | | |
|  |  | | | |
| PROJECT SOURCE CODE LINK (GITHUB/ GOOGLE DRIVE) | <https://github.com/manojvkhatokar/BGD_HADOOP_TwitterSentimentAnalysis/tree/master> | | | |
|  |  | | | |
| CONCLUSION /FUTURE ENHANCEMENT | In this work, we proposed a distributed solution to process sentiment analysis for twitter. In the future, we aim to test the code with larger datasets and look for some of the complex issues. After downloading the results of the Sentiment Analysis from HDFS, we can see the holistic polarity of each tweet, with the weight deciding the true sentiment behind the tweets. We can then observe for each tweet whether the result is positive, negative or neutral. We plan to include emoji as well during classification in future. This method of Sentiment Analysis can be extrapolated to other Indian languages like Hindi, Marathi, Tamil and their regional dialects as well. | | | |
|  |  | | | |
| UI SCREENSHOTS | C:\Users\manoj.LAPTOP-AKB0DTBI\Downloads\Screenshot from 2021-01-08 13-39-58.pngC:\Users\MANOJ~1.LAP\AppData\Local\Temp\Rar$DRa18484.3347\Screenshot from 2021-01-08 13-40-00.pngC:\Users\MANOJ~1.LAP\AppData\Local\Temp\Rar$DRa18484.3347\Screenshot from 2021-01-08 13-40-12.pngC:\Users\MANOJ~1.LAP\AppData\Local\Temp\Rar$DRa18484.3347\Screenshot from 2021-01-08 13-48-36.png  C:\Users\MANOJ~1.LAP\AppData\Local\Temp\Rar$DRa18484.3347\Screenshot from 2021-01-08 13-42-03.png | | | |