Sentimental Data Analysis for prediction of public reaction using Hadoop Framework

Abhinav Agarwal, Aadiitya Chaturvedi, Priyanshi Singh, S. Aarthi

Abstract: Sentiment analysis is taken into account to be a sub-class of machine learning and natural language processing. It's accustomed disencumber, identify, or depict opinions from completely diverse content structures, as well as news, reviews and editorials and sorts them as positive, neutral and negative. In this paper, we have an inclination towards investigating the effectiveness of linguistic possibilities for sensing the sentiment of Twitter messages. We have an inclination towards evaluating the usefulness of present lexical sources in addition to qualities that seize information regarding the natural and artistic language employed in microblogging. We take a administered attitude to the issue, however control current hashtags within the Twitter data for making training data. We are making use of Pig Latin in our system. We record the stream data and store it in .csv format file. Then we compare the words stored in file with AFINN dictionary and based upon the keywords provided, it will rate each keyword ranging from -5 to +5 depicting most negative to most positive comments. Those ratings are combined to obtain a numerical value and that is what gives us our prediction of public opinion.

Index Terms: Sentimental Data Analysis, Public Reaction, Hadoop Framework, PigLatin, Natural Language Processing(NLP).

I. INTRODUCTION

Sentiment Analysis commonly known as Opinion Mining is discipline of Natural Language Processing (NLP) that creates systems that try to detect and unearth opinions within text. Typically, in addition to identifying the estimate, these systems unearth traits of the expression. Presently, sentiment analysis is an area of great curiosity and progress since it has abundant realistic applications. As publicly and privately available information over Internet is unceasingly increasing, a great number of texts articulating opinions are offered in study sites, forums, blogs, and social media. Using sentiment analysis systems, the shapeless information could be involuntarily converted into arranged facts of public views regarding products, services, brands, politics, or any matter that people can voice their views on. There are numerous methodologies and algorithms to implement sentiment analysis systems. Rule based systems which do sentiment analysis of the lexicon methodology it had been the Indian National Congress. SVM predicted a 78.4% likelihood that the BJP would triumph a lot of elections because of the positive response they got in tweets. because it clad, BJP won sixty out of 126 constituencies within the 2016 general election, much more than the ruling party because the ruling party (the Indian National Congress) alone won twenty six out of 126 constituencies. As it is incredibly troublesome to guess the outcome of elections in other ways, together with popular opinion surveys, and with the rising frequency of social media, like Facebook and Twitter, the researchers set to apply sentiment analysis of...
Twitter tweets to guess the outcome of the Indian general election. We conjointly computed the exactness and recollect. The results of the Simple Baye’s formula was .71 and .61. For Maintenance Vector Machine we have a tendency to gain .75 as precision and .78 as recollect. because it clad, the BJP did win sixty out of 126 of the constituencies in Bharat in 2016. The drawback of analysis is that we have a tendency to failed to thought-about the emoticons that also are a relevant facet once shaping the schism of a tweet. Since the information was tagged physically the amount i.e. 36,465 wasn’t massive enough to produce additional precise results, therefore we will gather additional tweets so label them. Within the future, we will conjointly increase the scale of the Hindi SentimentNet. In [2] In this paper, they have an inclination towards investigating the effectiveness of linguistic possibilities for detection the emotion of Twitter tweets. They have a affinity to estimate the utility of present lexical reserves likewise as choices that seize data related to informal and artistic language used in microblogging. We have a tendency to take a administered attitude to the matter, however leverage present hashtags within the Twitter information for building coaching information. In the past few years, there has been a large progression within the usage of microblogging proposals like Twitter. Encouraged by that advancement, firms and media societies are progressively finding ways that to mine Twitter for data regarding what folks assume and feel regarding their product and services. Firms like Twitratr (twittertratr.com), tweetfeel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are simply many United Nations agency advertise Twitter sentiment analysis jointly of their services, whereas there has been a good quantity of analysis on however emotions are voiced in genres like on-line audits and news columns, however emotions are voiced given the casual tongue and text-length limits of microblogging has been a lot of of under researched.

Options like self-part-of-speech tags and reserves like emotion lexicons have tested helpful for sentiment analysis in different domains, however can they conjointly demonstrate helpful for sentiment analysis in Twitter? They start to analyze this query. One big encounter of microblogging is that the unimaginable width of matter that’s coated. it’s not associate overemphasis to mention that individuals tweet regarding something and everything. Therefore, to be ready to create systems to devour Twitter sentiment regarding any given topic, we want a way for swiftly characteristic information which will be used for coaching. During this paper, we have a tendency to explore one technique for creating such data: mistreatment Twitter hashtags (e.g., #bestfeeling, #epicfail, #news) to spot positive, negative, a006Ed neutral talks to use for coaching three-party sentiment classifiers. In [3] We performed sentiment analysis on article citation sentences corpora bearing 3 polarities viz. positive, negative, and neutral. thanks to insufficiency of negative citation sentences, the dataset suffers from a large categoryimbalance issue. To tackle this, we tend to planned associate ensemble feature engineering technique for deep learning, that uses embedding of text and its dependency relationships. The performance of deep learning models was compared with a support vector machine and provision regression approach victimization bag of words. Experimental results show that deep learning may be used effectively for associate unbalanced dataset by applying the planned ensemble options. applied math significance check indicates that one-hot supervised LSTM is statistically not totally different from the baseline strategies for 2 datasets, one developed by North American country and also the alternative taken from literature. Citation Sentiment Analysis (CSA) is that the method of deciding the sentiment expressed by the author vis-à-vis the cited literature victimization opinion words... the quantity of printed analysis articles is growing at an amazing rate, which, in turn, will increase the citation counts for existing literature, albeit, unevenly. However, a straightforward citation count doesn't mirror actuality quality of a look article, so as to work out the standard and subjective assessment of a broadcast article, CSA has been planned within the literature. Citation analysis has been studied for a protracted time to investigate the impact of a scientific article on the event of techniques and technologies. Citation analysis may be classified into quantitative and qualitative varieties. The previous usually depends on the quantity of citations found for a publication, whereas the latter utilizes the characteristics of a citation like appreciation, criticism, individualism elaboration, seminal study, etc. so as to capture one dimension of the standard of citation, citation sentiment analysis (CSA) has been planned in. CSA usually is employed to work out whether or not the citation text is appreciative or crucial of the cited publication. ottar was the primary (to associateanalyzego research after) to examine the sentiment polarity of a sentence (of an article), wherein a citation happens. He devised 3 rules to work out whether or not the sentence indicated appreciation or criticism of associate existing study. In [4] This project has been divided into a pair of phases. Paper review includes performing studies on numerous sentiment analysis procedures and technique that presently in practice. In stage 2, application necessities and functionalities are outlined before its advancement. Also, design and interface style of the platform and the way it'll act are known. In progressing the Twitter Sentiment Analysis application, many tools are used, like Python Shell a pair of.7.2 and tablet. The goals of the study are one, to review the emotion Associate in Nursinganalysis in microblogging that visible to research response from a client of an establishment’s product; and other, is to create a program for users’ feedback on a product that permits a corporation or person to sentiment and examines a huge quantity of tweets into a helpful way. As a result, program are going to be categorised emotion into positive and negative, that is described in a very chart and markup language page though, the program has been prearranged to be created as an online application, because of drawback of Django which might solely work on UNIX system server or LAMP. Thus, it can't be realised. Therefore, more improvement of this part is suggested in future analysis.

III. PROPOSED ALGORITHM

In our experiment, we implement Twitter Sentimental data analysis, with the help of the Hadoop framework. The Apache Hadoop software library is a framework that gives us the freedom of storing and processing large data sets in a parallel or distributed fashion. Hadoop provides us with a distributed framework to store data called HDFS (Hadoop Distributed File System), it allows us to store any kind of data across the cluster, it can be structured, semi-structured or unstructured data. It makes use...
of NameNode and DataNode architecture (also known as Master/Slave architecture) to implement the distributed file system and gives a high performance during data access, parallel processing and is highly tolerant. For processing, Hadoop provides us with YARN, which stands for Yet Another Resource Negotiator. It provides a paradigm for parallel processing of data that is stored in HDFS. It supports a variety of processing engines and applications, separates its duties across multiple components and has the ability to dynamically allocate resources to applications. Hadoop has the ability to integrate different kinds of tools that together comprise the entire Hadoop ecosystem because of YARN.

We are going to use a tool from the Hadoop ecosystem to perform data analysis: FLUME and a dictionary called the AFINN dictionary. The first step of the implementation is to fetch all the eligible tweets using FLUME and store them into the HDFS. After the storage is complete, the sentimental analysis of the tweets shall be performed using AFINN.

We calculate the rating of every word and then group them together to get an average. This average score can be used to classify the tweets as negative or positive. If the average rating of the whole tweet is greater than zero, it is classified as positive else negative. We can filter the positive and negative tweets by using the commands:

\[
\text{positive_tweets} = \text{filter avg_rate by tweet_rating} \geq 0; \\
\text{negative_tweets} = \text{filter avg_rate by tweet_rating} < 0;
\]

That will give us the cumulative result and exact percentage of positive and negative opinions or tweets.

IV. EXPECTED RESULT

For this experiment, we have chosen the keywords to be Abhinandan, India. So first we will create a flume file which will hold our streamed data.

```plaintext
TwitterAgent.sources.Twitter.keywords= abhinandan, India
TwitterAgent.sinks.HDFS.channel=MesosChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:8020/user/flume/tweets
```

Figure 2. Assigning keyword to flume agent.

After we have successfully created our flume.conf file, we run it on the terminal, the connection is established and streaming of data begins.

Figure 3. Streaming Data from twitter using flume.

The agent will keep dumping the streamed data in HDFS file until it is done or manually interrupted. The tweets are stored in semi-structured data in the HDFS. After the data is stored, we enter the grunt shell in the terminal and load that into Pig using Pig Storage. From the loaded data we extract the id and tweet text by using the following command:

\[
\text{extract_details} = \text{FOREACH load_tweets GENERATE } \$0 \text{ as id, } \$1 \text{ as text;}
\]

After extraction is finished, we divide the tweet text into separate words to calculate the whole sentiment of the tweet by using the AFINN dictionary. The division is done by using the TOKENIZE command.

\[
\text{tokens} = \text{foreach extract_details generate id, text, FLATTEN(TOKENIZE(text)) As word;}
\]

The AFINN dictionary comprises of a list of words that are rated for valence with an integer on a scale of minus five (negative) to plus five (positive). The dictionary is loaded into pig by using the following command:

\[
\text{dictionary} = \text{load ‘AFINN.txt’ using PigStorage(‘ut’) AS(word:chararray, rating:int);}
\]
After all data is stored we give the command to compare with AFINN dictionary and get the tweets some numeric value.

Figure 5. Numeric value assigned to tweets.

V. CONCLUSION

Our research depicts that the area of sentiment analysis is being nicely studied by researchers currently and also in the past couple years. Various varying ways have been created and verified. However, still, a huge amount of work is to be accomplished. One of the very common tactic is machine learning, a technique that requires a substantial data set for training and learning the features and emotions linked. Also, models tend to aim a naive worldwide grouping of audits, instead of rating individual features of the reviewed product. Only a limited number of the tactics were able to attain little higher level of exactness. Thus, the answer to sentiment analysis still has a long way to go before touching the confidence level needed by practical applications.

REFERENCES

2. Efthymios Kouloumpis, Theresa Wilson and Johanna Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!” in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media

AUTHORS PROFILE

Abhinav Agarwal a pre final year student graduating from SRMIST, keen interest in development in field of IoT