Real Time Text Analysis

To cite this article: K. Senthilkumar and E. Ruchika Mehra Vijayan 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **263** 042005

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042005 doi:10.1088/1757-899X/263/4/042005

Real Time Text Analysis

Senthilkumar .K, Ruchika Mehra Vijayan .E

VIT Univeristy, Vellore-632014, Tamilnadu, India.
Email :ksenthilkumar@vit.ac.in

Abstract. This paper aims to illustrate real time analysis of large scale data. For practical implementation we are performing sentiment analysis on live Twitter feeds for each individual tweet. To analyze sentiments we will train our data model on sentiWordNet, a polarity assigned wordNet sample by Princeton University. Our main objective will be to efficiency analyze large scale data on the fly using distributed computation. Apache Spark and Apache Hadoop eco system is used as distributed computation platform with Java as development language

1. Introduction

The instant breakup of the use of the internet in any media is increasing the day by day activities especially taking into accounts the online activities like chatting, ticket booking, online studying, streaming, and downloading. There is an very huge quantity of structured and unstructured data [1]. That is being presented in the field. Within the last half century every single field of the internet services are evolved including the research, academics, public and private sector services that are heavy efforts their studies on sentiment analysis and research development. We here present an small section of sentiment analysis that is yet presented in any social media by recognizing any person's personal status updates or reviews by associating them with the database techniques and analyzing them through the server created. Sentiment Analysis as the definition suggest is the method of text classification that categorizes the texts based on the different emotions and conditions a person suggests or comments. it can be determined by computing them into positive negative and neutral [2]. We are performing an document level access based specification computing using Apache Spark in the Hadoop platform respectively. Instead of peeping into the documents and paragraphs our implementation stares directly into the opinion.

2. Sentiment Analysis Approaches used

The specific levels of Sentiment Analysis are subdivided into positive negative and neutral sentiments. The opinions are classified into the basis of classification that differentiates them into subjects of views and the opinions. The figure-1 shows abstract view of text analysis [3]

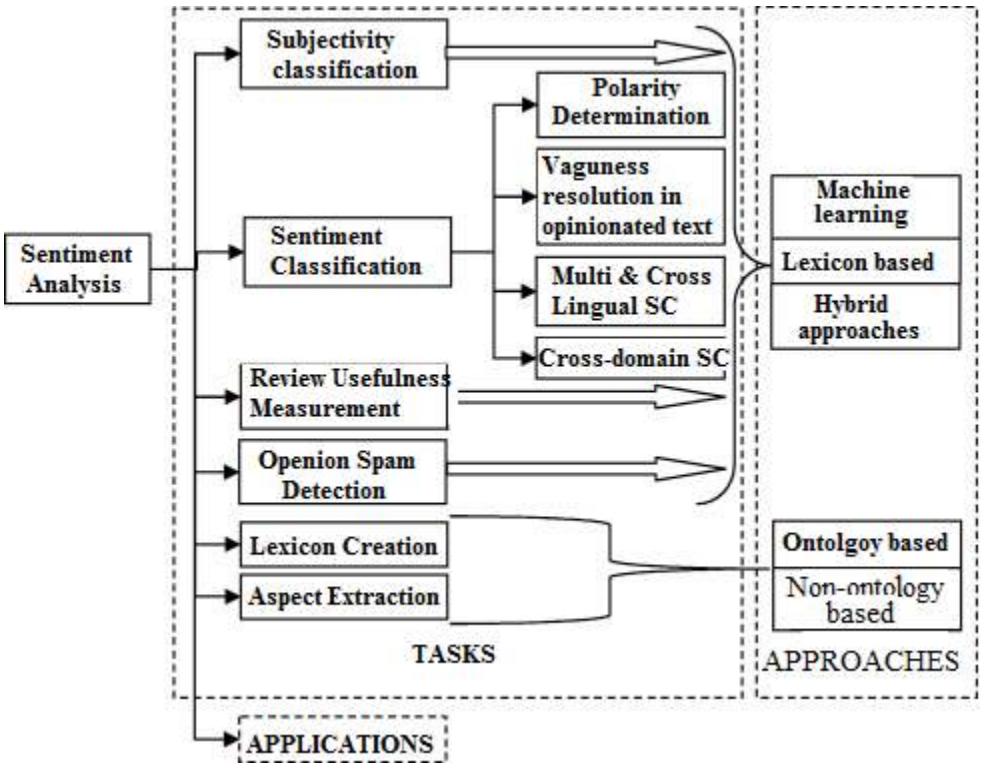

Fig. 1. Abstract view of text analysis

The segregation of the sentiment Analysis that is performed is been indicated in the following tables.

Approaches	Classification	Advantages	Disadvantages
Machine Learning	Supervised and Unsupervised	• Dictionary is not necessary. • Demonstrate the high accuracy of classification.	• Classifier trained on the tests in one domain in most cases does not work with other domains
Rule Based Approach	Supervised and Unsupervised	• Performance accuracy of 91% at the review level and 85% at the sentence level. • Sentence level sentiment classification performs better than the word level	• Efficiency and accuracy depend the defining rules
Lexicon based Approach	Unsupervised Learning	• Labelled data and the procedure of learning is not required	• Require powerful linguistic resources' which is not always available

Fig 2. View of Classification

3. Data Acquisition

With increasing availability of social media websites that are used worldwide, it is supported by the data format that is needed to perform the analysis. We have taken into account the twitter media for our implementation of sentiment analysis using the Application Programming Interface that reduces the streaming information from the other websites [4]. The API helps to review a large data that is cloud stored into the account that are taking 6000 members of 500000 friends of a specific person using a pre-existing blog search engine that is available. We here follow the technique of SENTIWORDNET 3.0,which notates the small and manual usage of the automatic signals that are from the method so prescribed. The developed automatic condition prescribes the synsets that are identified and categorized into the positives, negatives and objectives. Following table list the ranked positive and negative syn-sets in the SENTIWORDNET.

Rank	Positive	Negative
1	dood#n#2goodness#n#2	abject#a#2
2	better.off#a#1	deplorable#a#1 Distressing#a#2 lamentable#a#1 pitiful#a#2 sad#a#3 sorry#a#2
3	divine#a#6elsiann#a#2inspired#a#1	bad#a#10 unfit#a#3 unsound#a#5
4	good.enough#a#1	scrimy#a#1
5	solid#a#1	cheapjack#a#1 shoddy#a#1 tawdry#a#2
6	superb#a#2	unfortunate#a#3
7	good#a#3	inauspicious#a#1 unfortunate#a#2
8	goody-goody#a#1	unfortunate#a#1
9	amiable#a#1 good-humored#a#1 good- humoured#a#1	dispossessed#a#1 homeless#a#2 roofless#a#2
10	gainly#a#1	hopless#a#1 miserable#a#2 misfortunate#a#1 pathertic#a#1 piteous#a#1 pitiable#a#2 pitiful#a#3 poor#a#1 wretched#a#5

Fig 3 Data Acquisition method

4. Methodology

Analyzing the sentiment of a particular sentence can initiate the access of the present scenario in the enormous data that are collected. We use the particular machine language [5] that changes the data analytics in collection for practical use. The base implementation depends on the tweets that may can determine the political social and many other individual social situations so determined. The popularity of a particular issue is customized to positive negative and neutral issues in the existing scenario. We have come with the usage of Hadoop which is the traditional name for data analytics that can determine the sentiment of a particular item. This does not help to levelize the results immediately. We have used the Apache Spark that is used in the Hadoop platform that issues the
Present implementation in easy access of determination. Apache Spark initiates the use of Spark Streaming API, they help them in twitter stream instead of the filter streams used in the machine learning.

Fig 4 Machine learning stream

We are using Hadoop platform for cluster programming and large data set because it's very difficult analyst sentiment on real time data like twitter tweets. In Hadoop include three main module HDFS which is use for storing and accessing application data. YARN is framework which is use for scheduling and clustering resources. Map-reduce is a highly efficient distributed computation system, is only good for batch computation. Map-reduce has one major disadvantage The Map-reduce execution engine has to write the results to disk between intermediate mappers and reducers then next step of mapper or reducer picks the same results back again. This high disk IO significantly reduces the speed of complete cycle. We can't analyze the real time data. That's why we are using Apache Spark on replace of map reducing module on Hadoop framework. Apache Spark is a in memory distributed computing system which does not have to spill intermediate results to disk before passing to the next step. So we don't need high disk IO and it provides high speed.

5. Implementation

We are using twitterAPI to fetch the live data onto Spark Stream from twitter. API key, API secret, access token and access token secret these are four important keys is required for fetching live data onto twitter. Twitter API is you have an account on twitter for obtaining credential from twitter. Need to install a library. Streaming API is being able to provide all access which is publishing on a twitter for connecting streaming API.We require API secret, access token and access token secret. After creating this key we can join the hadoop platform with streaming API and able access all live tweets on twitter. We have used nearly 4 important steps in the stream API for analysis sentiment.

- Stop word removal-The stop word removal. Stop word removal in short refers to the removal of the repeated words so we can focus on the important word which is in tweets. Stop word is a set which is basically used in language, not only English it's able to support every languages example if someone tweet on a twitter "The galaxy Smartphone is very good phone and it's waterproof" in this tweet stop removal can remove the 'the','is','and','its' etc words. These frequent word don't have any real meaning and only hijack scores.

- Lemmatization-After removing frequent word by stop word removal. We have to do lemmatization i.e limiting the grammatical words in the sentence in form of organizing them into a structural format of a single similar derivation of an word.

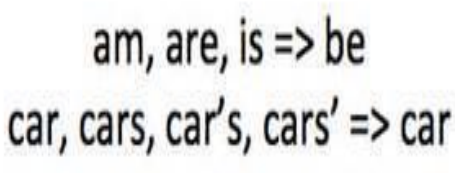

Fig 5 Lemmatization procedure

- Vectorization technique-In computing terms, the specialization of a computer program is converted into a scalar implementation, where the lists of words are ordered using their specific identification number. This is helpful in analyzing the computer programming technical streaming of certain words into algorithms. The text corpus maintains the unique identification number in

complementing of feature set. They are basically converted into array variable from string variable i.e each array is stored as array identification number.

- Score computing-this method generally explains that each and every word has an sentiment attached to it. Be it positive negative or neutral sentiment, the score associated with the words are computed with a specialized algorithm that identifies our output. We have corpses from the Princeton University database called the Sentiwordnet 3.0 that saves them into the real memory. The basic computation schema is by adding all the positive, negative and neutral comments from the sentence and analyzing the aggregate positive output of the former.

6. Conclusion

We have taken into consideration the difference in accuracy between the ratings that are obtained from the SENTIWORDNET in the outcome that are evaluated from the process with beneficiaries. It has been an overwhelming situation on the study of experiencing the differences that are obtained through the comparison on the random walk up statement obtained in the search engines. We are also been approached by some lexicon and rule based information on the analysis for these methods. This model of analyzing the people's behaviour can help the internet world to understand and carry on with decision making.

References

[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2014 An Enhanced Lexical Resource for Sentiment Analysis, *SENTIWORDNET 3.0:* **5(4)** pp5422-5425.
[2] Devika .M.D, Sunitha C, Amal Ganesha, 2014 Sentiment Analysis:A Comparative Study On Different Approaches *IJRCCE,* **3(3**) pp2-5
[3] Kumar Ravi, Vadlamani Ravi, October 24-25, 2013, A survey on opinion mining and sentime|nt analysis: Tasks, app|roaches and applications, *11th IEEE International Conference on Emerging eLearning Technologies and Applications*, pp3-5
[4] Subramaniyaswamy. V , Vijayakumar. V. , Logesh R. , Indragandhi.V. 2008, Unstructured Data Analysis on Big Data Using Map Reduce, *ICDE Workshop 2008*, pp2-6
[5] Hassan Saif,,YulanHe, Miriam Fernandez, Harith Alani 2016, Contextual semantics for sentiment analysis of Twitter, *Information Processing & Management* **52(1)**, pp5-19