

Crowdsourcing

Vishal¹ Nitin Mohan² Vipin³ Mrs. Shallu Juneja⁴

^{1,2,3}Student ⁴Assistant Professor

^{1,2,3,4}Department of Computer Science and Engineering

^{1,2,3,4}Maharaja Agrasen Institute of Technology

Abstract— The aim of this project is to analyze the crowd's opinions on Telecom Operators and the services they provide. We have used twitter to gather opinions, although some other platform or site could be used. Sentiment analysis and data mining techniques are applied to find out sentiments of people of India on different Telecom companies. In this project, we are tackling the problem of evaluation of data submitted by crowd. Sentiment analysis, a well-known task in Text analysis has been done on Tweets to evaluate and understand people's opinion. The evaluated result has been displayed on the webpage. This webpage has been made using HTML (Hyper Text Markup Language) and PHP (Hypertext Preprocessor). The main goal of such a sentiment analysis is to discover how the customers feel about various companies since the launch of Reliance Jio free data and voice services. Our focus is mainly on the tweets around 31st March, 2017. Millions of Indians became used to free Jio, its free calling and data. From 1st April, 2017, Jio stopped its free 4G data offer and started charging customers depending on Jio plan (with Prime or without). Due to Jio's offer, other telecom companies have also changed their data plans in order to protect their existing customers and revenue per user. The Twitter data that is collected will be classified into three categories: positive, negative or neutral. An analysis will then be performed on the classified data to investigate the count of the audience sample falls into each category.

Key words: Sentiment, Data Mining, HTML, PHP, Investigate

I. INTRODUCTION

Crowdsourcing is a word that is the combination of two words "crowd" and "outsourcing". It describes any activity that includes outsourcing which is not limited to companies only, but is addressed to the crowd as an "open call", via an Internet Web platform. It is an online model of innovation and collaboration that provides businesses with the opportunity to receive more inflow of solutions compared to traditional practices.

The major challenges faced by crowdsourcing systems include designing and dividing tasks among crowd workers, evaluating and assessing quality of the contributed work, motivating crowd and retaining them on platform.

Evaluation and quality assessment of contributed work is important for successful working of any crowdsourcing platform. Text mining could be used to evaluate the contribution. It is defined as a process of using data mining techniques to derive useful pattern from unorganized data. It can be used to analyze solutions in order to filter out similar submissions. It can also be used to analyze the feedback or reviews of public in order to know people's opinion about certain products or services.

As content is created and shared online through social channels, blogs, review sites etc. the requirement for

businesses to mine this information, in order to gain business insight from it, has also increased.

Sentiment Analysis is a well-known task in Text Analysis. Sentiment analysis is predominantly implemented in software, which can autonomously extract emotions and opinions in text. It has many real world applications such as it allows companies to analyze how their products or brand is being perceived by their consumers; this usage is particularly applicable to this project. It is difficult to classify sentiment analysis as one specific field of study as it incorporates many different areas such as linguistics, Natural Language Processing (NLP), and Machine Learning or Artificial Intelligence.

Crowdsourcing involves classification of tweets into three main sentiments: positive, negative and neutral. In our project, we have tried to record the sentiments of people regarding telecom companies. The consolidation in the crowded Indian telecom industry has been hastened by Jio's free voice call and data plans, forcing other telecom giants to slash tariff at the cost of profits and people to express their views online.

II. BACKGROUND

In previous semester, we have researched about various crowdsourcing platforms and found the various problems faced by them. One of the major problems was to analyze the solution submitted by crowd. Data mining is one of the preferable solutions to this problem. It is the process of finding patterns in large datasets. The objective of data mining is to extract information or knowledge from a dataset and transform it into a structure that can be understood. Data preparation is an important part of any data analysis. To properly prepare data it is necessary to understand the application domain, this is important as the researcher must be able to identify pertinent data and cleansing the dataset removing any data which is deemed as unimportant to the analysis.

In this project, we are tackling the problem of evaluation of data submitted by crowd. Sentiment analysis, a well-known task in Text analysis has been done on Tweets to evaluate and understand people's opinion. The evaluated result has been displayed on the webpage. This webpage has been made using HTML (Hyper Text Markup Language) and PHP (Hypertext Preprocessor).

III. PROBLEM ANALYSIS

What should other people think has always been an important piece of information for most of us during the decision-making process. Before awareness of the World Wide Web became widespread, many of us asked our friends to recommend and auto mechanic or to explain who they were planning to vote for in local elections, requested reference letters regarding job applicants from their colleagues, or

consulted Consumer Reports to decide what dishwasher to buy. Twitter is one of the opinion-rich resource which includes opinions of people in the form of tweets.

As content is created and shared online through social channels, blogs, review sites etc. the requirement for businesses to mine this information, in order to gain business insight from it, has also increased. Businesses try to unlock the hidden value of text in order to understand their customer's opinions and needs and make better, more informed, business decisions.

IV. SOLUTION DESIGN

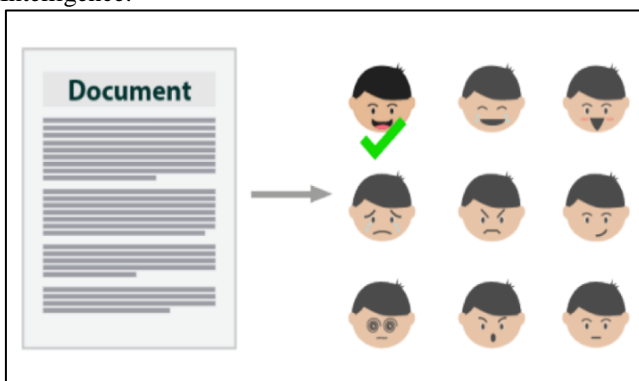
To deal with the problem that is to understand the large amount of unstructured data that is available on blogs, reviews or micro blogging sites like twitter, text mining and analysis of data could be done. The opinions of crowd through tweets are extracted and analyzed to understand and optimize what is being thought about any telecom company and the services they offer.

A. Text mining

Text mining could be used to evaluate the contribution of crowd. It is defined as a process of using data mining techniques to derive useful pattern from unorganized data.

B. Sentiment Analysis

Sentiment Analysis is a well-known task in Text Analysis. Sentiment analysis is predominantly implemented in software which can autonomously extract emotions and opinions in text. It has many real world applications such as it allows companies to analyze how their products or brand is being perceived by their consumers; this usage is particularly applicable to this project. It is difficult to classify sentiment analysis as one specific field of study as it incorporates many different areas such as linguistics, Natural Language Processing (NLP), and Machine Learning or Artificial Intelligence.



The main hurdle with understanding sentiment and opinions is the complexities involved in how we express our thoughts and opinions, and form a message. When people give feedback about something, it's a combination of things they liked and disliked about that thing. Example: "I like the battery life of this phone, but the screen sucks!". Similarly some messages are commonly expressed as a negation of another message, like "I don't like the food" instead of "I dislike the food". Above described complexities would impose challenges for Sentiment Analysis systems.

So required to take the general structure of the sentence into account and, their context in order to make better judgment rather than just rely on only the "polar" words.

V. METHODS AND APPROACHES

We are using RapidMiner Studio for text mining and to perform sentiment analysis. The tweets are extracted using Search Twitter operator and sentiments are analyzed using a RapidMiner extension named Text Analysis by AYLIEN which has an operator called 'Analyze Sentiment' for analyzing sentiments of tweets.

On RapidMiner, we are using its operators to perform the following tasks:

- Extracting tweets using the Search Twitter operator which is available under Data Access category. The term or phrase to be searched need to be mentioned in the query in parameter view. We have extracted tweets before and after 31st march, 2017.
- Sentiment Analysis on the extracted tweets is performed using Analyze Sentiment operator that is available in Aylien Text analysis extension of RapidMiner.
- Preprocessing of data i.e. tokenization, stop word removal and removal of duplicate tweets is done through respective operators.
- Wordlist is created from the data using polarity and filtered out depending on the names of telecom companies to compare the polarity of various companies.
- Clustering (k-means) is performed to find out similarity in opinions or thoughts in various tweets. It is concerned with grouping together objects that are similar to each other and dissimilar to the objects that are belonging to other clusters. It is a technique for extracting information from unlabelled data. After getting list of words occurring in tweets and their corresponding frequency, clusters have been formed in which all the words having number of occurrences relatively close to each other fall in same cluster.

The result of all the above process will be ExampleSet (like Excel sheet) and that will be used to make graph to get complete visualisation of the result. some important terms used in RapidMiner are:

- Process - A process can be defined as a chain of operators that can be subsequently applied to get the result. It has an input and result port.
- Example - In RapidMiner, "example" means "a row" in a table.
- ExampleSet - It is a table containing rows and columns.
- Attributes - A column in RapidMiner is known as attribute in ExampleSet.
- Subprocess - A subprocess can be considered as a small unit of process, like in process, all operators and combination of operators can be applied in a subprocess.

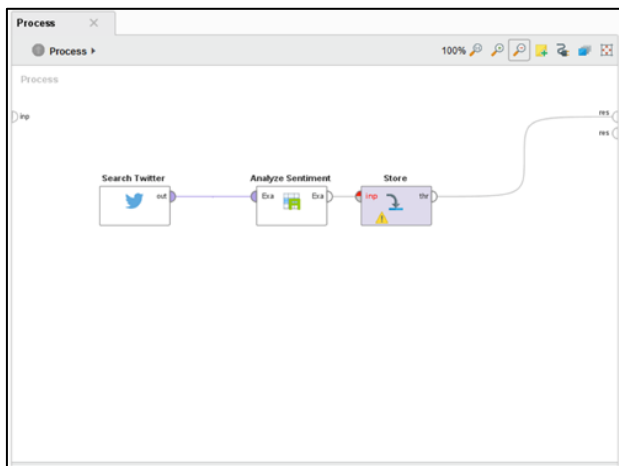


Fig. 1: Process of Extracting Tweets and Analyzing Sentiments

VI. BENEFITS OF CROWDSOURCING

- Crowdsourcing is cost-effective: the company only pays for bugs which are found instead of an hourly or salaried rate which professional testers would receive
- The vast range of users provide huge diversity in their experiences
- Crowdsourcing allows for testing with all different kinds of parameters, such as with different connection speeds, browsers, and devices to which the core testing team may not have access
- Larger group is more likely to find reproducible bugs than a handful of testers
- Lack of bias towards the company can be expected of testers

VII. CONCLUSION

Crowdsourcing is a project which aim towards analyzing the crowd's opinions on Telecom Operators and the services they provide. We have used twitter to gather opinions, although some other platform or site could be used. Sentiment analysis and data mining techniques are applied to find out sentiments of people of India on different Telecom companies.

The main goal of such a sentiment analysis is to discover how the customers feel about various companies since the launch of Reliance Jio free data and voice services. Our focus is mainly on the tweets around 31st March, 2017. Millions of Indians became used to free Jio, its free calling and data. From 1st April, 2017, Jio stopped its free 4G data offer and started charging customers depending on Jio plan (with Prime or without). Due to Jio's offer, other telecom companies have also changed their data plans in order to protect their existing customers and revenue per user. The Twitter data that is collected will be classified into three categories: positive, negative or neutral. An analysis will then be performed on the classified data to investigate the count of the audience sample falls into each category.

The proposed system generates the result with an average accuracy of 68 to 71 percent. It can be of great use for the industries to reach their prospective.

REFERENCES

- [1] Antonio Ghezzi ,Donata Gabelloni ,Antonella Martini ,Angelo Natalicchio, "Crowdsourcing: A Review and Suggestions for Future Research", 19 January 2017
- [2] Enrique Estellés Arolas," Protocol: A literature review about the use of crowdsourcing in educational environments", Vol 7, No 2 (2016)
- [3] Vimi Grewal-Carr, Carl Bates,"The three billion Enterprise crowdsourcing and the growing fragmentation of work"
- [4] Oskar Ahlgren,"Research On Sentiment Analysis:TheFirstDecade", Published in: Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on, Date Added to IEEE Xplore: 02 February 2017
- [5] Alireza Amrollahi," A Process Model for Crowdsourcing: Insights from the Literature on Implementation", November 2016
- [6] Motomichi Toyama," Automatic vs. Crowdsourced Sentiment Analysis", July 2015
- [7] Adam Bermingham and Alan F. Smeaton," Crowdsourced Real-world Sensing: Sentiment Analysis and the Real-Time Web", 07/CE/I1147