A Study of News Recommender System using Natural Language Cloud Computing Services

A.A. JALBANI, Mukhtiar.MEMON, Muniba.MEMON*, S. DEPAR, M. Y. KOONDHAR

Information Technology Centre, Sindh Agriculture University, Tandojam, Pakistan

Abstract: The Natural language processing is widely used for opinion mining, in which text are processed based on the sentiments. In world of digital society, news is ranked on the basis of their liking, disliking and number of times the news have been shared on the cowed-source networks. To compete with the other competitor, it is now necessary for the newspaper organizations to share their latest news on the crowd-sourcing platform that user can circulate news in a fast and impact full way to other users. Hence, It is difficult for the company to read all posts and recommend news to other users. For this, an automatic mechanism or framework is required to analyze the comments placed on the crowed source networks. The natural language toolkit can play its role for analyzing the comments and recommend the news. In this research, we used Google cloud natural language API for analyzing comments placed on the crowed source platforms and recommend the news based on the user feedback.

Keywords: Sentiment analysis, opinion mining, cloud computing, Natural language processing toolkit

1. INTRODUCTION

The recommendation is a process to get feedback or opinion from the users. If user recommends anything positively then the things gets an added value. The number of positive responses help in attracting more users. Currently organizations or communities get feedback of the users manually through questions or conducting surveys and the final conclusions are made based on the user feedback. For example, whenever we want to buy house or any other item, we always want to know what people think about it. Sometime even we ask our friends, family members about the item before making decisions to buy an item (Baccianella, 2010)

Now the communication is transformed into digital format with the advent of online crowed sourcing networks on the world wide web. Most of the networks conducts online surveys to recommend about their stories based on the feedback provided by their online friends. Therefore, it is easier to get recommendations online. There are many ways to get the feedback online. Some are paid surveys in which companies pay to the participants for providing feedback. With the advent of online platforms, now it become free to get feedback and people are very comfortable to critics easily that makes companies to put more efforts to launch anything with the quality. Hence, the Internet has become the one of the stronger platform to provide opinions freely. One of the survey results shows that among 81% of users have provided feedback online at least once The surveys include online reviews about hotels, travel agents, doctors and so on.

There are many applications of opinions, for example some are review related website, these automatically keep the opinions and aggregation of websites. Recommendation engines, the websites automatically recommend the feedback or the users search. Detection and classification technologies work here to recommend the best possible recommendations. Questioning and answering is another way to get recommendations on the websites. There are so many application, some are listed below:

- Government Intelligence
- Review related System
- Recommender Systems
- Movie ratings

It is easy to get cloud of words or big data of text data from any sources such as books, web sites, chat apps and so on. Sentiments can easily be extracted from such large text data by using natural language toolkit. In this research, the NLT with python programming is used to extract the sentiments from the large texts. The NLT is used to manipulate and analyze the text in various ways. Hence NLT is a toolkit and not a system. Hence the problems are resolved or solved through the combination of other languages such as python and its libraries, interfaces to external natural language toolkit.

Due to the importance of customer feedback, the interest of companies in opinion mining and machine learning algorithms has increased due to their reliable results. Sentiments or opinions are extracted from the text, for example simple sentence shows the positive,

++Correction author: E-mail: akjalbani@sau.edu.pk
*Department of computing Indus University Karachi

negative or neutral attitude in keywords of the text. For example:

- Awesome weather today → positive sentiment
- Hmm, Bad weather today→ negative sentiment
- Weather is normal → neutral sentiments

From above sentences, it is easy to extract the sentiments inside the text based on the keywords, but it becomes hard when we talk about reviews (Vyrva, 2016) The recommender systems use natural language processing toolkit to process the textual information, if the textual information is on the web than we require a web scraping tool to scrap the data and send it to the natural language processing toolkit to process it for sentiments. It is important to note that web scrapping has some limitations, in which hosting servers may block the IP from which scrapper actions have been initiated.

In this research, the textual posts of Facebook are scrapped, and scrapped data is provided to the natural language processing toolkit. There are many ways to use the natural language toolkit. However, this research is focused on the Google cloud Natural Language Processing API that helps to process the data to provide sentiments in the Facebook posts.

2.   **RELATEDWORK**

An article published by (Junaid *et. al.,* 2017) on sentiment analysis of Facebook posts. The research focuses on the grammatical aspects of words, the words are categorized adjective, noun, verb and anything else. This is one way of analyzing comments using natural language processing toolkit. The predications are calculated based on three metrics, these are; accuracy, error rate, precession and recall. The proposed framework uses rule based computational linguistics, which are applied on different types of sentiments. However, approach does not use any cloud-based computation for the sentiment analysis.

The research on Twitter is focused by (Rosenthal *et. al.,* 2017).This research is continuity of their previous research on sentiment analysis of twitter data. The data is collected from the Twitter participants, various techniques have been applied by using natural language processing toolkit. A large data set have been maintained for this research. This new linguistics technique has seen applied on Arabic text. In parallel, (Poria, *et. al.,* 2017) and (Zehe, *et.al*., 2017) have proposed multiple ways to analyze sentiments from videos, the proposed model captures information from surroundings.

(Hussein *et. al.,* 2016) have investigated the sentiment analysis challenges through an extensive survey and identified various domain where the sentiments can play an important role. The two most

important keywords categorized by this research are nature and review that relies on challenging for sentiment study. The comparative study intimates the challenges for the specific domain and did not give any generic results for the study. This study surveyed only few specific research papers and did not use cloud environment for the sentiment analysis.

The research published by (Sidney*et.al.,*2016) explores the limitations of using natural language toolkit. The authors identified three limitations of Facebook as compared to Twitter, i.e., variety of newsgroups, time series and analysis of association or grouping words that have been tweeted on twitter or posts or messages added on the Facebook pages. To identify these limitations, one must have prior knowledge and strategies for combination of words.

(Raj *et, al.,*2016)presented an overview on present, past and identified some suitable directions for the future work on analytics. The research(Chopra *et, al. ,*2016) proposed a framework for sentiment analysis of words instead of complete paragraph. The complete graphs may increase the complexity and conjunction of words. However, the researchers did not use any cloud-based environment or approach for sentiments storage or analysis.

A framework to analyze comments and predict gender, age, entity and sentiments proposed by (Kowcika *et. al.,* 2013). The approach is based on the online data and for each attribute of predication, the statistical calculation has been made to prove the proposed framework for real case study scenario. Natural Language Processing methods vary from different point of view, researchers have established various models for probabilistic tagging. The (Schmid 2013)proposed speech tagger for parts of speech tagging, that is based on decision tree algorithm. However, this research only focusses on sentiment analysis.

The blogs, review sites, data sets, Micro blogging are the classification of datasets for the sentiment analysis. The most favorable data set focused by (Vinodhini 2012) for this research, in which blogs and reviews of MovieData sets are discussed, analyzed, predicted based on the reviews. The research provides useful predication-based algorithm for the recommender system for the blogs and micro blogging system. Another work related to sentiments analysis on twitter have been published by (Bakliwal 2012) The author focused on tweets and these tweets are mined, and related information has been extracted. Realtime micro blogging have been focused by (Bermingham 2012) in his PhD research, the author has proposed posts on blogs and sentiments are extracted for other formal verifications.

The huge volume of data on the Internet have been uploaded almost every second, the portion of online data is much bigger than any type of the data (Thelwall *et.al.*2010). This huge data (i.e.; big data) has more variety as compared to any traditional data, which is stored in the traditional relational database software. Hence, this variety makes the data to be handled in a more complicated way. The strength of the text data has been calculated from the informal and formal behaviors of the users. It is obvious that the strength of the data provides the attitude or the impression of the user for the discussion. In this research, authors have proposed the SentiStrength algorithm to identify the strength of the data from the informal text. The application of the study was tested on MySpace. The proposed algorithm was able to identify the strength in the sentiments to predict the emotions with sufficient accuracy for negative emotions, which has almost high strength in the sentences. The research focused on the baseline data and did not use virtual or cloud engine to predict the behaviors.

The sentiment classifier training based on the automatic collection of corpus proposed by (Pak*et.al;*2010). The idea was further analyzed based on three factors, these are: positive, negative and neutral. The difference was computed from the POS-tagging and Tree Taggers to achieve their goals. However, this research focuses on the sentiments scrapped from the Facebook posts.

A new concept of extracting information from website posts proposed by Prabowo(2009). The concept was implemented based on their application developed for the detection of significant emotion changes in various posts of Facebook in particular. The authors also classified messages and statistics among the users, which gives us good indication of their scales for the polarities. The purpose of their application was to check whether the sentiments are helpful in e-learning environment or not. Their result shows the positiveness of the application for the e-learning platforms. The research gives us good indication and makes a roadmap for our research. However, this research uses cloud computing technology to analyze sentiments. Another research conducted on the students' networking on Facebook by(Pempek*et.al.,* 2009*)*. Their research goal is achieved by conducting the survey and the results show the interest of students using online websites.

(Pang,2008) discussed on the sentiment analysis and opinion mining in general and investigated various traditional ways of analysis sentiments or opinions. The researcher main focused on the privacy attributes of the information, the manipulation of the text to identify the effects of the sentiments, opinions and finally economic impact of analyzing sentiments. The researchers also provide directions for future benchmarks for the sentiment analysis, reviews and assessments of further campaigns that could be possible for the opinions to be extracted from the rich data set. Another similar work is proposed by (Koppel *et.al;*2006) and (Kim *et.al.*2004*)*.

3. METHODOLOGY

In this research, google cloud natural language REST API has a main role to process the Facebook network comments. To use the Google natural Language API, the Figure 1 shows the complete process flow to access the Google Cloud Natural Language API and apply the sentiment analysis library using the python language to get knowledge about the sentiments in the required Facebook posts.

The process starts with creating an account in the Google Cloud and using the credentials, the account is for testing and learning purpose, hence the Google gives free credits access to their cloud, as our research is based on the sentiment analysis of Facebook posts for this it is essential to have a Facebook developer account. The Facebook graphic API will be used to get access token that will be used to download the Facebook posts from the Facebook groups for this research case study.

Once, the access token is created into the Facebook developer page, then it will be easy to scrape the data from Facebook posts into CSV or text files. This text file contains lots of data related to the posts posted on the Facebook group for our case study. It is important to note that this research will use the same data in text or CSV file to upload into the cloud computing facility of the Google Cloud for Natural language toolkit API. Finally, the python code is used to access the Google Cloud Natural Language API to get the sentiment from the posts which have been downloaded from the Facebook graph API. The entire process is sketched in the figure below:
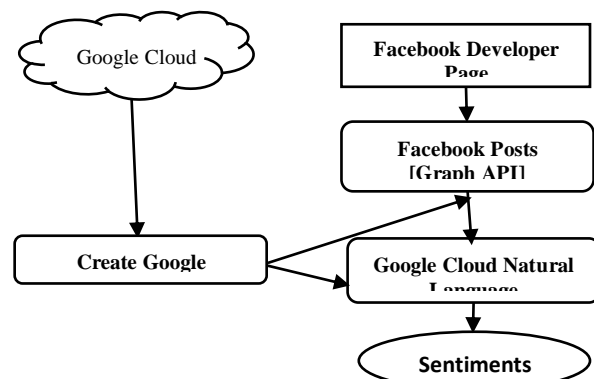

Fig. 1: Process Flow Diagram for Extracting Sentiments

4. RESULTS AND DISCUSSIONS

This section provides the results and discussion of the experimental study on the Facebook page, in which the data set is based on the visitor's comments, reviews

or feedback on the individual post. The data set is collected from the well-known Pakistan's newspaper "Daily Dawn". In this experimental study, we used Google Cloud Natural Language API to classify the comments on the Dawn Facebook page as either positive, neutral or negative. Finally, we calculate the proportion of positive, neutral or negative comments on these posts

The comments of the posts show the sentiments as illustrated in Figure 2, there are four columns should be noted here that first column shows the count of the comments and the second attribute is the sentiment is Positive, it gives the proportion of Positive, Neutral and negative sentiments, respectively

It is possible to calculate the proportion of each sentiment. the proportion is calculated for positive, neutral and negative sentiments accordingly. The mathematical foundation is necessary for the proof-of-concept, hence first all the comments were counted, i.e.; the number of posts on the comments were counted and then sentiments were evaluated using google could natural API to get count of positive, neutral and negative sentiments.
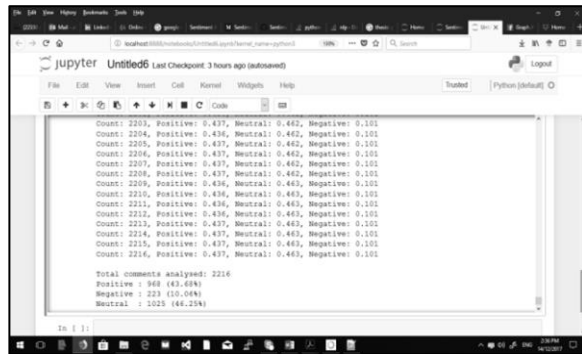

Fig. 2: Analysis of Comments using Google NLTK API

The following mathematical model have been used for the calculation:

$$\%age\,of\,sentiment\,Type = \frac{Total\ Number\ of\ Sentiment\ Types}{Total\ Number\ of\ Comments}$$

Total number of comments extracted on 14th August 2017 = 2216.

Table 1: Sentiments Proportions of Dawndotcom

S.No.	Sentiment Type	No.of Comments	Percentage
1.	Positive Sentiment	968	43.68%
2.	Negative Sentiments	223	10.06%
3.	Neutral Sentiments	1025	46.25%

The **(Table-1)** shows that the neutral sentiments are more posted on the dawn dotcom Facebook page, while the negative comments are almost 10%. The **(Fig.3)** shows the illustration of the results in the bar chart graph
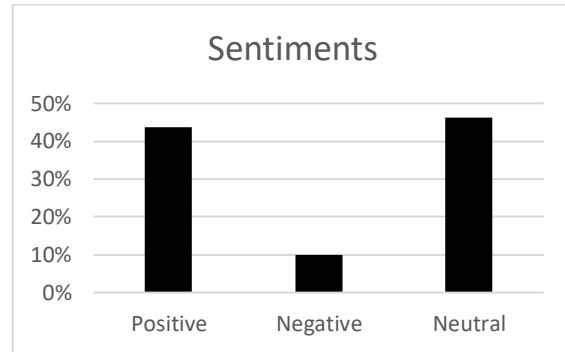

Fig.3: Sentiments Proportion

5. SUMMARY AND OUTLOOK

The research conducted on the Facebook platform to process their user's feedbacks on various posts. The Facebook platform is chosen, one of the most widely circulated newspaper as paper based as well as online newspaper named Dawn. Dawn newspaper have their Facebook page dawn dotcom. All the data of the dawndotcom is scrapped using Facebook Graph API and stored into the textual format. This text file contains the posts and their feedbacks from the visitors. To analyze the sentiments of the users, the textual file is further processed into the Google cloud natural language API. Google Cloud has provided the key to use the API. The whole process is integrated in python language.

The results show that the dawn dotcom Facebook page contains good number of positive and neutral sentiments and less number of negative sentiments. The process is again verified from the actual posts to check the actual textual information presented in the posts. It has been observed that the Google Natural Language API is powerful tool to process such type of textual information. Further, the Google Cloud provide various ways to access their API, this research can further be extended to analyze the quality of the results and it is necessary to compare Google Natural Language API to other sentiment analysis API's.

REFERENCES:

Baccianella, S., A. Esuli,and F. Sebastiani (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC Vol. 10, 2200-2204.

Bermingham. A., (2012) Sentiment analysis and real-time microblog search (PhD thesis) Dublin City University,

Bakliwal. A., P. Arora. S. Madhappan. N. Kapre. M. Singh, and V. Varma, (2012) Mining Sentiments from Tweets. In WASSA@ ACL 11-18,

Chopra.F. K., and R. A .Bhatia. (2016) Critical review of sentiment analysis. International Journal of Computer Applications, 149(10)

Hussein, D. M. E. (2016) Survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences.

Junaid,S. M., S. W. Jaffry. M. M .Yousaf. L. Aslam.and S. Sarwar. (2017) Sentiment analysis and opinion mining-A Facebook posts and comments analyzer. Technical journal, University of Engineering and Technology (UET) Taxila, Pakistan 22.

Kowcika. A., A.Gupta. K. Sondhi. N. Shivhre.and R. Kumar. (2013) Sentiment analysis S.M. International journal of advanced research in computer science and software engineering,

Koppel. M. and J. Schler (2006) The importance of neutral examples for learning sentiment. Computational Intelligence, 22(2), 100-109

Kim. S. M. and E. Hovy. (2004) Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics Association for Computational Linguistics.

Pak.A. and P. Paroubek.(2010) Twitter as a corpus for sentiment analysis and opinion mining. In LREc Vol. 10, 89Pp

Poria S., E. Cambria. D. Hazarika.N. Majumder. A. Zadeh. and L. P. Morency. (2017) Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Vol. 873-883.

Prabowo.R. and T. Mike. (2009) Sentiment analysis: A combined approach. Journal of Informetrics, Vol 3, Issue 2, 143-157,

Pempek. T., A. Yermolayeva. and S. L. Calvert. (2009) College students'networking experiences on Facebook. Journal of applied developmental psychology, 30(3), 227-238,

Raj. K., V. T. Ritu. and R. Nirmal. (2016) Sentiment analysis of web data: A Review. In proceeding of ACEIT

Rosenthal, S., N. Farra.and P. Nakov. SemEval- (2017) task 4: Sentimentanalysis Twitter In Proceedings of the 11th International Workshop on Semantic Evaluation

Schmid. H., (2013) Probabilistic part-of-speech tagging using decision trees. In New methods in language processing,

Sidney. B., O. Ochieng. M. Loki. N. Sambulis.(2016) Limitation of sentiment analysis of Facebook. International Journal of SS and Information Technology Vol. 2, 10-15

Thelwall. M., K. Buckley. G. Paltoglou. D. Cai. and A. Kappas. (2010) Sentiment strength detection in short informal text. Journal of the association for information science and technology, 61(12), 2544-2558,

Vinodhini. G. and R. M. Chandrasekaran. (2012) Sentiment analysis and opinion mining: a survey. International Journal, 2(6), 282-292

Vyrva, N. (2016) Sentiment analysis in SM (Master'.s thesis). Ostfold University College, Norway,

Zehe. A., M. Becker. F. Jannidis. and A. Hotho. (2017) Towards Sentiment Analysis onGerman Literature. In Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz) 387-394. Springer.
.