IOP Conference Series: Materials Science and Engineering

Sentimental Analysis for Airline Twitter data

To cite this article: Deb Dutta Das *et al* 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **263** 042067

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042067 doi:10.1088/1757-899X/263/4/042067

Sentimental Analysis for Airline Twitter data

Deb Dutta Das, Sharan Sharma, Shubham Natani, Neelu Khare and Brijendra Singh

School of Information Technology and Engineering, VIT University, Vellore-632014, Tamil Nadu, India.

E-mail : neelu.khare@vit.ac.in

Abstract. Social Media has taken the world by surprise at a swift and commendable pace. With the advent of any kind of circumstances may it be related to social, political or current affairs the sentiments of people throughout the world are expressed through their help, making them suitable candidates for sentiment mining. Sentimental analysis becomes highly resourceful for any organization who wants to analyse and enhance their products and services. In the airline industries it is much easier to get feedback from astute data source such as *Twitter*, for conducting a sentiment analysis on their respective customers. The beneficial factors relating to *twitter* sentiment analysis cannot be impeded by the consumers who want to know the who's who and what's what in everyday life. In this paper we are classifying sentiment of *Twitter* messages by exhibiting results of a machine learning algorithm using R and Rapid Miner. The tweets are extracted and pre-processed and then categorizing them in neutral, negative and positive sentiments finally summarising the results as a whole. The Naive Bayes algorithm has been used for classifying the sentiments of recent tweets done on the different airlines.

1. Introduction

The need for sentiment analysis through *Twitter* has been found a more captivating area for research it provides more effective sentiment's of the general public, especially in the case of airline industries where people are pleased and displeased very easily and more often take their views to *Twitter*. As compared to other sources of data such as review websites or blogs where the data found is not only fully informative but also lacks number, the amount of data present in *Twitter* is an unimaginable count. With 1/60th of the world's population in *Twitter*, although the number may seem small but it actually amounts to 100 million people, more than half a billion tweets are tweeted on a daily basis and the numbers keep growing each passing day[4]. Since the discovery and implementation of Big Data the collection of such huge amount of data and conducting research on them has now become almost too easy. The tweets sent by the mass public can be treated as their opinion and their perspective on situations making it a valuable source of information for the any organization trying to attract more number of people by improving and attending customer needs[18]. *Twitter* Sentiment Analysis can be used as a great instrument for customer review and feedback especially for new product released in the market. For example, with the Samsung Galaxy 8 market release the company can quickly do a customer feedback and review but carefully analyzing tweets related to the product itself then later processing for a more definitive result.

The main motive of this paper is to provide the airline industry a more comprehensive views about the sentiments of their customer and provide to their needs in all good ways possible[18]. We would be taking in *200 tweets* directed at Emirates and Jet Airways and conducting the analysis on them and more than *1000* tweets to classify the recent traumatic incident involving the United Airlines.

Sentiment here has been classified as a feeling of being positive or negative. These are personal to each and every person. While neutral can be treated as a tweet which shows neither a positive or negative, in other words no conclusive sentiment can be drawn out of it [not headline defining[16].

The characteristic of the tweets will that be that they will be having certain basic characteristics such as a *language model*, such as frequency of words or slang used. *Length* of the tweets should be in 140 characters [9], which is the basic length of tweets.

2. Literature Survey

The authors Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore [11] have surveyed of a linguistic feature of checking the sentiment of *Twitter* tweets. They have evaluated the usefulness of lexical resources as well as features collect information in informal and creative language used in blogging. A supervised approach has solved this problem, but existing hash tags in *Twitter* data for a dataset. In past years there was a huge growth in blogging platforms like *Twitter* which is better in media, company, and growth. In *Twitter*, people provide their information in the form of feeling and opinion. There are some companies which provide *Twitter* sentiment analysis as a service. The challenge of blogging is the incredible size of the topic. To identify the *Twitter* sentiment about any topic which is given, we need a method for identifying data that is used as a dataset. I this they explore a method for building data like hash tag to identify positive, negative and neutral tweets to use for sentiment classifiers. Gao, Wei, and FabrizioSebastiani [12] paper states that estimation order has turned into a pervasive empowering innovation in the *Twitter* sphere, since grouping tweets as indicated by the notion they pass on towards a given substance (be it an item, a man, a political gathering, or a strategy) has numerous applications in political science, sociology, showcase investigate, and numerous others. In this paper, we battle that most past reviews managing tweet notion arrangement (TSC) utilize an imperfect approach. The reason is that the last objective of most such reviews is not evaluating the class mark (e.g., Positive, Negative, or, on the other hand Neutral) of individual tweets, however evaluating the relative recurrence (a.k.a. "pervasiveness") of the diverse classes in the dataset. The last undertaking is called evaluation, and late research has convincingly appeared that it ought to be handled as its very own assignment, utilizing learning calculations and assessment measures distinctive from those utilized for the arrangement. In this paper, we appear, on an assortment of TSC datasets that utilizing an evaluation particular calculation delivers generously preferred class recurrence appraises over a state of-the-craftsmanship characterization situated calculation routinely utilized as a part of TSC. We, therefore, contend that analysts intrigued by tweet feeling commonness ought to switch to measurement particular (rather than classification specific) learning calculations and assessment measures also learning a lot in the process of this.
In the paper of Torunoğlu, Dilara [13] it depicts that assumption characterization is one of the critical and well-known application ranges for content order in which writings are marked as positive and negative. Also, Naïve Bayes (NB) is one of the, for the most part, utilized calculations around there. NB having a few points of interest on lower multifaceted nature and more straightforward preparing system, it experiences sparsity. Smoothing can be an answer for this. The issue, for the most part,

Laplace smoothing is utilized; however, in this paper, we propose Wikipedia based semantic smoothing approach. In our review, we augment semantic approach by utilizing Wikipedia article titles that exist in preparing reports, classifications, and side-tracks of these articles as theme marks. After effects of the broad analyses demonstrate that our approach enhances the execution of NB and even can surpass the precision of SVM on *Twitter* Sentiment 140 dataset. Content grouping is one of the vital strategies to consequently, arrange a lot of printed information gathered in associations, online networking, and the Internet. Content arrangement picking up significance with quick increment in the use of the web and particularly online networking locales, for example, *Twitter* and Facebook. Accordingly, a gigantic measure of printed data is created by people and additionally the business substances, and associations. One of the vital furthermore, prominent application regions of the content grouping are the feeling characterization in which the remark writings are normally sorted as positive or negative. Ordinarily utilized machine learning calculations in content order are Naïve Bayes (NB) [1], the k-closest neighbour [2], Support Vector Machines (SVM) [3]. In spite of the fact that SVM is one of the best performing calculations in this space, NB can perform better in a few cases and also it has a few preferences, for example, bring down many-sided quality and easier preparing technique. In any case, NB extraordinarily experiences sparsity when connected to the especially high dimensional information as in content arrangement. This is particularly the situation when the preparation information comprises of short archives, for example, tweets and when the preparation set size is restricted due to the cost of manual marking forms. With a specific end goal to stay away from zero likelihood issue smoothing techniques are utilized. Most regularly utilized and default smoothing procedure is called Laplace Smoothing which adds one include to all terms the vocabulary. Torunoğlu, Dilara paper [14] states that feeling order over *Twitter* is typically influenced by the loud nature (shortened forms, unpredictable structures) of tweets information. A well, known system to lessen the clamour of printed information is to expel stop words by utilizing pre-gathered stop word records or more advanced techniques for element stop word distinguishing proof. Be that as it may, the adequacy of expelling stop words with regards to *Twitter* notion grouping has been wrangled over the most recent couple of years. In this paper, we explore whether expelling stop words aides or hampers the viability of *Twitter* supposition grouping techniques. To this end, we apply six distinctive stop word recognizable proof techniques to *Twitter* information from six distinctive datasets and watch how expelling stop words influences two surely understood managed notion characterization techniques. We survey the effect of evacuating stop words by watching vacillations on the level of information sparsity, the extent of the classifier's component space and its characterization execution. Our outcomes demonstrate that utilizing pre-ordered arrangements of stop words contrarily impacts the execution of *Twitter* assumption order approaches. Then again, the dynamic era of stop-word records, by evacuating those occasional terms seeming, just once in the corpus gives off an impression of being the ideal technique to keeping up a high characterization execution while diminishing the information sparsity and significantly contracting the component space. Vo, Duy-Tin, and Yue Zhang paper [15] provides with insight that, target-subordinate assumption investigation on *Twitter* has pulled in expanding research consideration. Most past work depends on the linguistic structure, for example, programmed parse trees, which are liable to commotion for casual content, for example, tweets. In this paper, we demonstrate that aggressive outcomes can be accomplished without the utilization of sentence structure, by extricating a rich arrangement of components. Specifically, we split a tweet into a left setting and a correct setting as indicated by a given target, utilizing dispersed word portrayals and neural pooling capacities to concentrate highlights. Both opinion driven furthermore, standard embeddings are utilized, and rich arrangements of neural pooling capacities are investigated. Feeling vocabularies are utilized as an extra wellspring of data for highlight extraction. In the standard assessment, the thoughtfully basic technique gives a 4.8% outright change over the cutting edge on three-way focused on the feeling arrangement, accomplishing the best outcomes for this undertaking.

3. Existing System

Sentimental Analysis in airways system is methodically done with the help of feedback forms or online questionnaires, in their respective websites. The procedure is quite simple on an overview but demands much of a complex nature when one tries breaking it down. Collecting feedback forms from a mass public and then analysing each and every form is a difficult task, requiring manpower as well as cost. In case of online pooling maintaining site regulations and keeping a database while performing computations on the database is also a complex way of approach. As for existing algorithms for sentimental analysis one such being *Maximum Entropy (MaxEnt) Classifier* coverts labelled feature sets to vector using encoding.

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l)) for l in labels)} \quad [1]$$

The problem faced with this type of approach is that it works best with dependant features, meaning one event is related with another. But, doing sentimental analysis two events must be uniquely identified so as, we can differentiate between the mass tweets.

4. Proposed System

In the proposed system we are using one classifier for the sentimental analysis for the airlines system, namely being *Naive Bayes Algorithm*. We concluded to use these two because it's works best with independent features, which was our prime requirement for this analysis.

$$\text{P(positive|tweet)} = (P(tweet|positive)P(positive)/ P(tweet)) \quad [2]$$

4.1. System Model Design

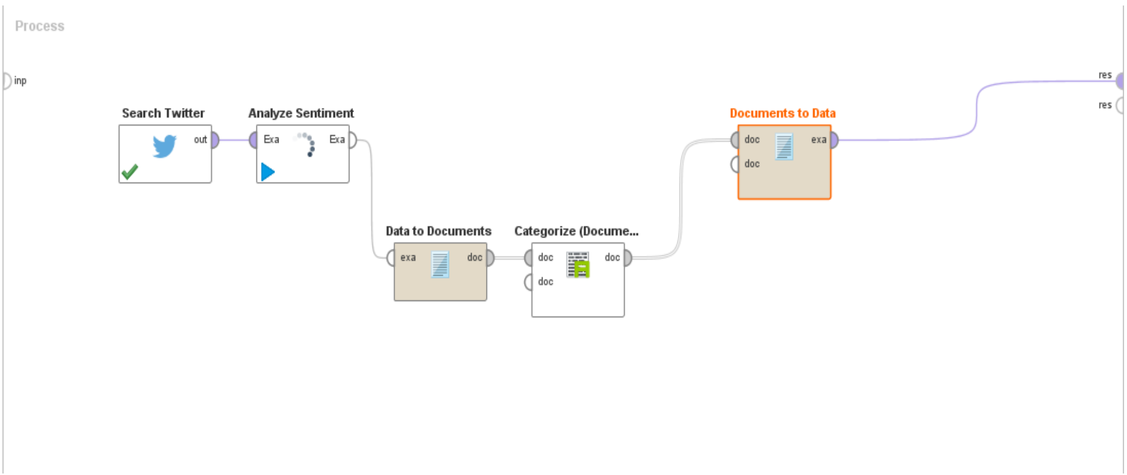

Fig.1. System Model Design

The fig 1. shows the system architecture overview for the sentimental analysis of twitter data and the figure below explains how the whole architecture flow is working.

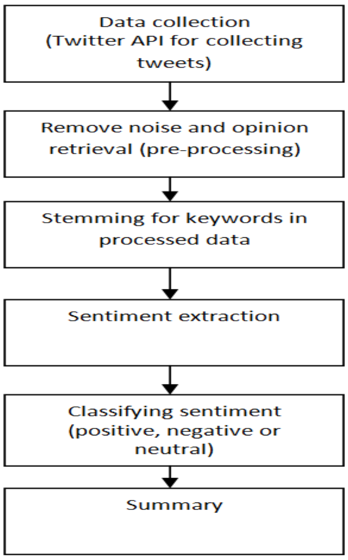

Fig.2. Flow of data in the system

4.2. Methodology

In the process for determining the sentimental values and analyzing them, we used the R studio software and its various package's [8] namely:

1. *twitteR*-providing an interface or the web API of Twitter.
2. *Rcurl*- A wrapper for 'libcurl' Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server. This provides a great deal of control over the HTTP/FTP/... connection and the form of the request while providing a higher-level interface than is available just using R socket connections.
3. *sentR*- package using bayesian classifiers for emotion classification
4. *tm*- Text Mining framework in R for applications
5. *NLP*-Natural Language Processing basic classes and methods [5]

The following code was used for the work done in R studio:

```
api_key <- "##############################" <twitter_account key>

api_secret <- "###############################################" <account secret key>

access_token <- "###############################################" <access token>

access_token_secret <- "###############################################" <account secret token>

setup_twitter_oauth (api_key, api_secret, access_token, access_token_secret)

united <- searchTwitteR ("#unitedairlines", n=4000, since ='2017-04-12', until = '2017-04-16', lang="en")

m <- strip_retweets (united, strip_manual = TRUE, strip_mt = TRUE)

tweetF21 <- twListToDF (m)

write.csv (tweetF21$text,"D:\\set\\12_04_EVE.csv")

result<-classify.naivebayes (tweetF21$text)

r <- as.data.frame(result)

table(r$SENT)

write.csv(r,"D:\\set\\result.csv")
```

Fig. 3. R Code for Sentimental Analysis

The bag of words on which analysis has been done is as follows (the number of words is quite extensive hence a nominal amount has been shows just for example):

Table 1. Bag of words

Positive Words	Negative Words
Work	Abort
Zeal	Abrupt
Zenith	2-faced
Yay	Disgust
Acclaim	Despicable

The fetched excel data spreadsheet from *twitter* was then applied through the above algorithm to get the results. An example of the fetched results is:

	x
1	Sorry,
2	Lil Donny #Trump is running the county like a company. The problem is, the company is #UnitedAirlines
3	LIVE: Realquotables ep 017 on #spreaker #alcohol #introvert #lasertag #live #unitedairlines https://t.co/xXQhoja6hP
4	#UnitedAirlines Changed Their Policy For Overbooked Flights https://t.co/UQb5tIXrZx #DISSunited
5	@GRRMspeaking The problem is the Doctor did not follow the procedure expressly outlined in the card available in th... https://t.co/DNMHsD7VND
6	I have my own trauma with United (and Expedia.) Wrongly charged! WTF? #unitedAIRLINES #biteme
7	<ed><U+00A0><U+00BD><ed><U+00B2><U+0080> Is #unitedairlines still in business...anybody know? The memes alone can take that company... https://t.co/oFXBVF9X9A
8	#legos
9	You're Not Mad at #UnitedAirlines You're Mad at America; You don't hate Mondays you hate #Capitalism #DraftBernie https://t.co/rdKhsZvho9
10	Am I a horrible person to fly United??? Cause like their tickets cost like $70 now #unitedAIRLINES #united https://t.co/iSR4wqa6An
11	United, would you rather be right, or would you rather be in business? #unitedAIRLINES https://t.co/6sX1V3l3Ja https://t.co/ViohnC6Ykx
12	CEO OscarMunoz was 2 have had a routine heart transplant as per #UnitedAIRLINES policy#4756 but apparently no heart was available #heartless
13	We don't
14	Bay Area
15	I dislike flying except with Southwest. No problems with them ever. #SouthwestAirlines #unitedAIRLINES https://t.co/XybS9YL9DG
16	#unitedAIRLINES not so United !!!
17	Barbie about to punch and kick out some passengers <ed><U+00A0><U+00BD><ed><U+00B2><U+0081><U+2708><ed><U+00A0><U+00BD><ed><U+00B1><U+008A> #Barbie #UnitedAirlines #ExpoBarbie @... https://t.co/VkDwwUQonZ
18	Pets!!??? Have you seen how they treat passengers? #PaxEx #unitedAIRLINES https://t.co/5xsicj1KFp
19	So USA #unitedAIRLINES are their shares a sell now. Anyone think they"ll last more than 6 months?
20	Lake Forest's Brooke Glanden hit a softball so far, #UnitedAirlines overbooked it #delhs https://t.co/6ChjiDR2LX
21	That's called #smart #opportunity #marketing #lol #united #airline #unitedairlines #flight #3411... https://t.co/xC2nq377oy
22	That's called #smart #opportunity #marketing #lol #united #airline #unitedairlines #flight #3411... https://t.co/pkYaycHV5u
23	#UnitedAi
24	This, hands down, has to be the best one. #unitedAIRLINES https://t.co/conpL1hSr4
25	@987Wo
26	#southwest That's called #smart #opportunity #marketing #lol #united #airline #unitedairlines... https://t.co/9kc8aP6a8F
27	New #video shows #UnitedAirlines passenger telling them to drag him off flight, take him to #jail https://t.co/kxN0tDDkQX
28	The No. 1 Way to Lose Business #unitedAIRLINES #nightmare #respect #customerexperience #Entrepreneur https://t.co/lhPJBwxQY2
29	The ugly
30	#UnitedAirlines changes crew booking policy in reaction to #DavidDao being dragged off plane https://t.co/cJAHrOTitO

Fig. 4. Sample of Fetched Results

5. Results and Discussion

The data was obtained from tweets based on United Airlines controversy and then it was classified in Rapidminer software [version 7.4] and R Studio for stemming and cleaning of data, then the results drawn from the twitter data spreadsheet was that it was classified that *395* out of the 1298 tweets for United Airlines were classified to be negative, *187* neutral while *716* were classified as to positive with the help of Naive Bayes classification.

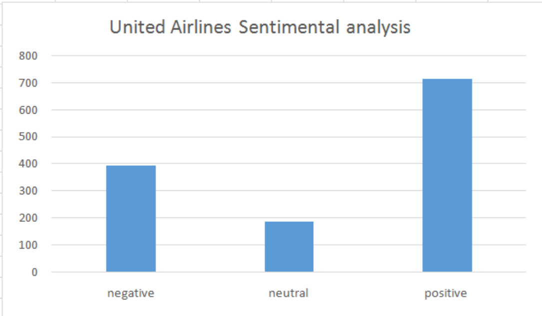

Fig. 5. Histogram for sentimental analysis

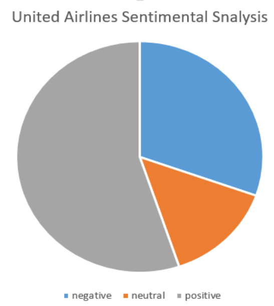

Fig. 6. Pie chart for sentimental analysis

In case of knowing how the mass public feel about Emirates airlines and Jet Airways another sentiment analysis was done on twitter data this time focusing on these two airlines specifically. The classification was in this case was not only the sentiment of the people but also the opinion of the passengers who have travelled by the respective airline. We took in the measure of *hash tags* as well since sometimes the users only tweet has tags and avoid using words and sentences as well. In this case we took in *200* tweets. The analysis had certain connections to be made in case of R Studio and Rapidminer namely twitter connection and AYLIEN [6-7] connection. Both, requiring their respective API's to be connected to the software. The tweet analysis is as follows:

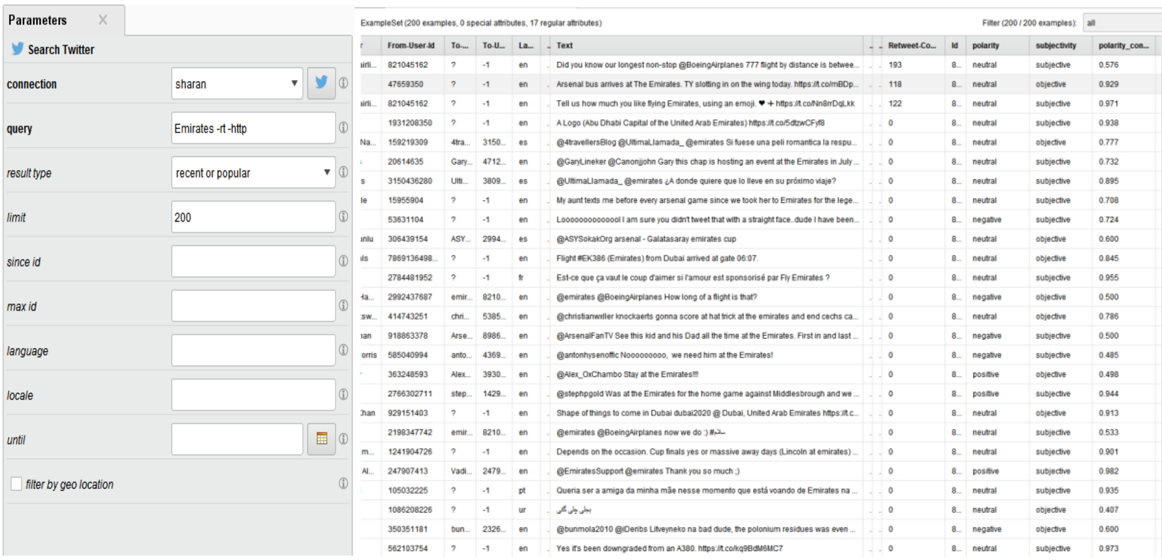

Fig.7. Rapid miner view of generated sentimental analysis

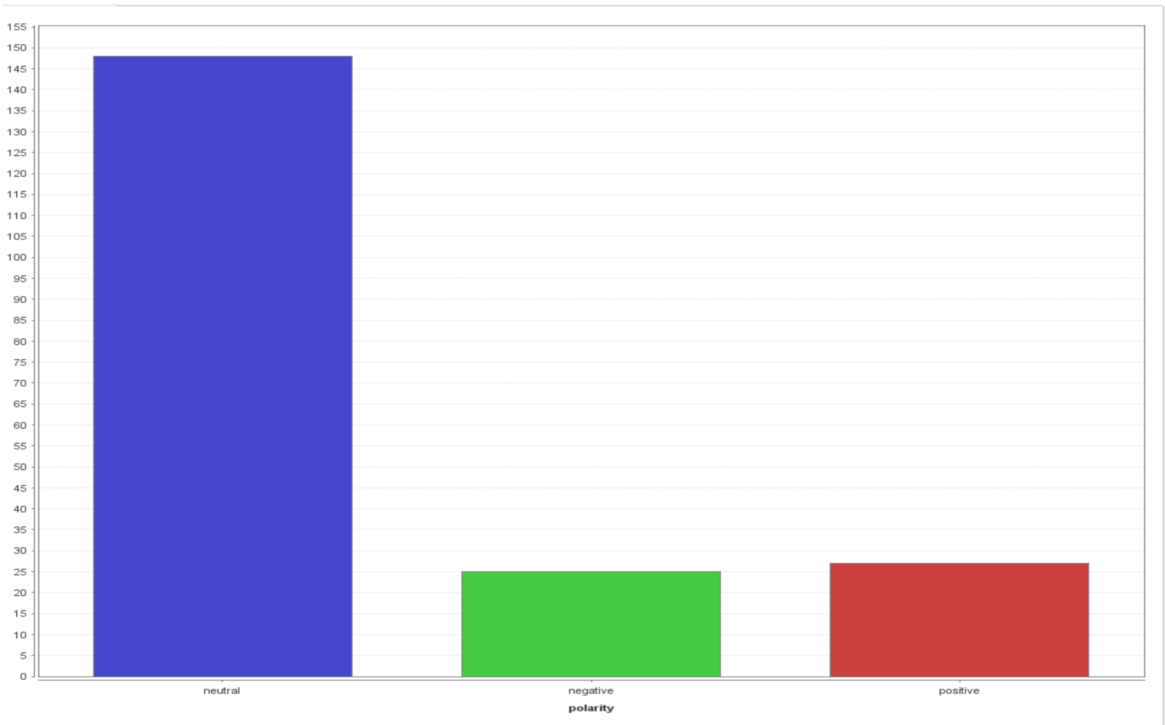

Fig. 8. The analysis for Emirates from the collected data

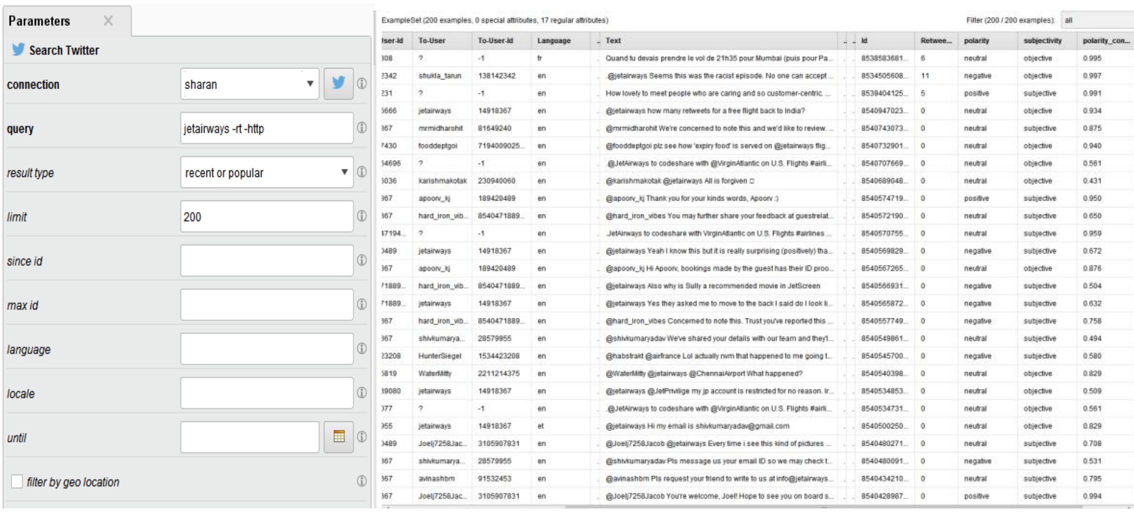

Fig. 9. The analysis obtained from Jet Airways are as follows:

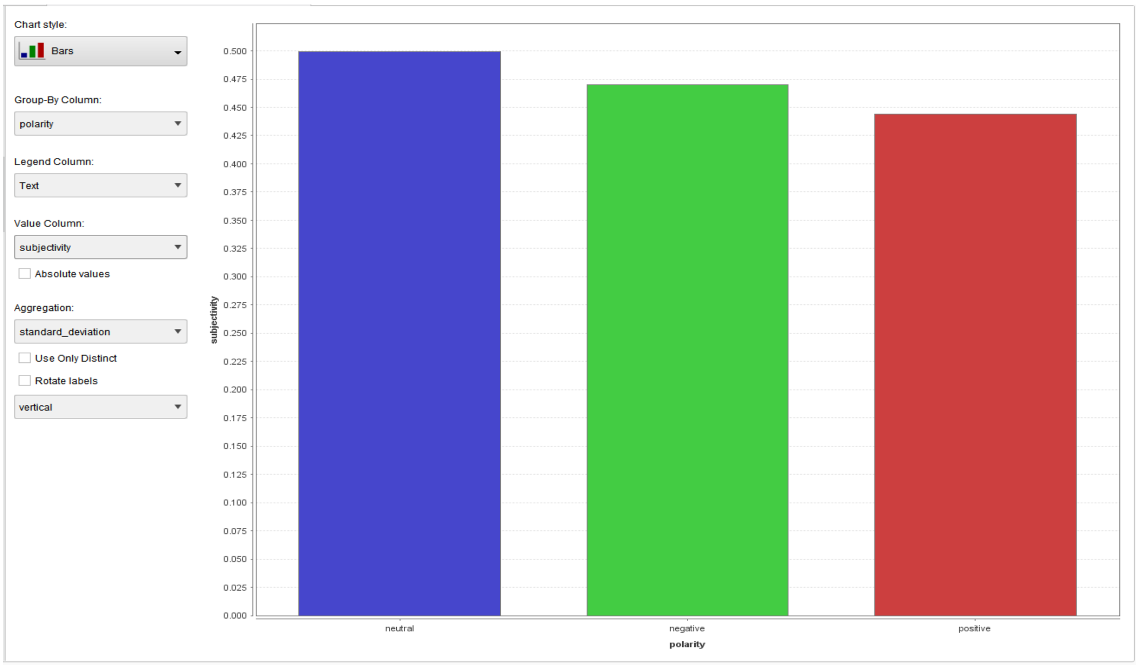

Fig. 10. The sentimental analysis results

6. Conclusion

Sentimental Analysis is a latest trend to understand the needs of the mass public; it's an easier and cost effective way to understand how the people are feeling about a particular subject of matter and the brand impact of micro-blogging [3]. In this scenario we had consider the sentiment of the people towards the airline industry and tackled the recent issues of United Airlines and how the public feels about it.

The analysis confirmed our assumption on how effective an approach twitter sentiment analysis is. The Naive Bayes classifier used in the algorithm, along with two software for better results depict clearly the sentiment of the mass crowd and thus the airlines could easily interpret the data and benefit from it by trying to improve on the aspects that seem negative or is disliked by the targeted audience.

There is still scope for improvement in this analysis since it is very new and yet has not been tested on many other classifying models. And the major setback is the limit in the number of tweets to be analysed using AYLIEN in Rapid Miner being 1000 tweets a day for a free user otherwise one has to opt for plans [10]. So in the future we are planning to further expand our research and analysis by gather a huge number of data and expanding the process of data mining involved in this analytical approach.

References

[1] D. O. Computer, C. wei Hsu, C. chung Chang, and C. jen Lin. A *practical guide to support vector classification* chih wei hsu, chih-chung chang, and chih-jen lin. Technical report, 2003.
[2] N. Cristianini and J. Shawe-Taylor. An *Introduction to Support Vector Machines and Other Kernel-based*
Learning Methods. Cambridge University Press, March 2000.

[3] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. *Micro-blogging as online word of mouth branding*. In CHI EA ’09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pages 3859–3864, New York, NY, USA, 2009. ACM.

[4] T. Joachims. *Making large-scale support vector machine learning practical*. In B. Scholkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in kernel methods: support vector learning, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[5] C. D. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[6] *https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId*

[7] *https://www.aylien.com*

[8] *https://www.rstudio.com/products/packages/&grqid=54b7IBH&hI=en-IN*

[9] *https://dev.twitter.com/basics/counting-characters*

[10] https://newsapi.aylien.com/docs/rate-limit&grqid=oK8gjwfP&hl=en=IN

[11] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." *Icwsm* 11.538-541 (2011): 164.

[12] Gao, Wei, and FabrizioSebastiani. "Tweet sentiment: From classification to quantification." *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, 2015

[13] Torunoğlu, Dilara, et al. "Wikipedia based semantic smoothing for twitter sentiment classification." *Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on*. IEEE, 2013.

[14] Saif, Hassan, et al. "On stopwords, filtering and data sparsity for sentiment analysis of twitter." (2014): 810-817.

[15] Vo, Duy-Tin, and Yue Zhang. "Target-Dependent Twitter Sentiment Classification with Rich Automatic Features." *IJCAI*. 2015.

[16] D. S. Rajput, Praveen K. Reddy, D. P. Shrivastava, “Mining Frequent termset for Web Document Data using Genetic Algorithm”, Published in International Journal of Pharmacy and Technology, Vol. 8 (2), pp. 4038-4054 2016.

[17] Sanchita Gupta, AkashKataria, Shubham Rathore, Dharmendra Singh Rajput, “Information Security issues in big data: Solution using PPDM (Privacy Preserving Data Mining)” Published in International Journal of Pharmacy and Technology, Vol. 8 (4), pp. 25551-25568, 2016.

[18] D. S. Rajput, Neelu Khare, “FSS Decision Making Model for Social Networking Sites”, published in International Journal of Social Network Mining, Inderscience, Vol 2(3) pp. 256-266, 2016.