

Sentiment Analysis and Prediction in Social Media

Ajay S, Department of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune
Purvash Bhawe, Department of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune
Prannay Deshpande, Department of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune
Archana Chaudhari, Department of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune

Article Info

Volume 83

Page Number: 2129 - 2136

Publication Issue:

March - April 2020

Article History

Article Received: 24 July 2019

Revised: 12 September 2019

Accepted: 15 February 2020

Publication: 18 March 2020

Abstract:

The work provides an overview of sentiment analysis in social media. Sentiment analysis has become important with the use of social media and technology. Several methods for sentiment analysis have been proposed in literature. Data from social media websites such as Instagram, Facebook and YouTube have been used by researchers for prediction. The work proposes to predict sentiment and classifies the comments as positive, negative or neutral. The work also attempts to fit a linear regression model to predict the subscriber/follower count for any social media account, say a YouTube channel or an account on Instagram.

Keywords: Prediction, Classification, Linear regression, Sentiment analysis.

I. INTRODUCTION

Machine learning is a quintessential component of modern technology. Its applications range from a vast spectrum: from social media, business analytics and medical examinations to government security systems. Machine learning focuses on building computer programs that can access data so that the model can learn by itself without being explicitly programmed. It is a component of artificial intelligence that tries to provide systems the capability to automatically learn from experience and improve without human intervention.

One major application where machine learning is essential is in social media. Social media relies on machine learning predictions to gain insights in order to improve marketing strategies. In order to improve social media marketing, we need to know more about the audience. We need to know what kind of content the audience prefers, what kind gets mixed opinions and what kind of content draws negative views. Once we get an idea of the sentiments of the audience we can schedule or plan our content accordingly. Social media heavily relies on big data and machine learning for this purpose. Machine learning lets us scale our analysis to any

huge number of data, which would mean millions of posts in a considerable amount of time. Machine Learning aims to develop computer programs and models that can access some data and use it to experience and learn for themselves. The learning by the model begins with observation of data, for instance examples and direct experience, or a set of rules, so that we can look for patterns in data and make better decisions in the future based on the past examples we have provided.

II. LITERATURE SURVEY

Machine Learning enables the analyzing and manipulation of huge quantities of data. While it usually delivers at a faster pace and provides more accurate results, it may also require more time and resources to train it properly. Machine Learning has become a tool for turning information into knowledge [1]. In the past 50 years, there has been an explosion of data. This mass of data is useless unless we analyze it and find the patterns hidden within.

Machine Learning techniques can automatically find the patterns that are underlying within complex data that we may not have discovered. The hidden

patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making.

Social media analysis is of great importance for marketing strategy and related research. For instance, Facebook uses machine learning algorithms to personalize each user's feed. One such metric which we can use to measure sentiment of audience is the comments [2]. The comments give us a direct word of what the audience thinks about the user's videos or photos. We can write a bot that reads these comments and segregates them into positive, negative and neutral comments and throw the values onto a histogram to see how the overall audiences have reacted to the post.

In this work a linear regression model is used to predict where the user will be in terms of subscribers/followers by feeding the old data into the model. The model uses the old (past) data to learn and predicts the subscriber count for the future. For conducting experiments two datasets in CSV format are used. One of them contains the comments data which includes the video ID, comments on the video, liked or disliked comments and so on. The other dataset contains the date and number of subscribers gained on that day. This data will be used by the model to predict the number gained in the future

There are several important libraries that are used for this model. Each library is a collection of functions that allow us to write programs using them without having to write our own codes which is a tedious task. The entire theory of "intelligent" machines involves understanding humans in order to interact with them. Natural Language Processing (NLP) is a rapidly growing area which meets the above demand. NLTK is a growing and leading platform for processing and working with human language data [3]. It is being used for several key activities performed by technology users across the world including search engine optimization, information

extraction, pos tagging, noun phrase extraction, text and speech recognition and sentiment analysis. NLTK has the ability to understand, manipulate and even generate natural language. Textblob is a Python library that uses a simple API and performs basic NLP tasks.

Sentiment analysis is basically trying to understand the attitude and emotion of the writer: whether it is positive, negative or neutral. The sentiment function of Textblob returns two properties namely polarity and subjectivity. Polarity is a float that lies in the range [-1, 1] where 1 indicates positivity and -1 indicates negativity. Subjective sentences refer to whether the statement is a personal opinion or emotion whereas an objective statement generally denotes factual information [4].

The literature on sentiment analysis focuses on different domains such as media and content to marketing, management to computer science, social sciences and business and so on. Its functions are solely focused on certain aspects due to its importance to society, such as: sentiments of words, objective and subjective sentences etc. Machine learning based approach uses a classification process to classify text. it consists of two sets of documents: training and a test set. The training set is used for learning the differentiating characteristics of a document, while the test set is used for checking how well the classifier performs. It applies the ML algorithms and uses linguistic features [5].

The main advantage of this method is the capability to adapt to and create trained models for specific customized purposes and contexts but its main disadvantage is the low applicability of the method on new data because it requires the availability of labeled data that may be expensive. It can use supervised and unsupervised methods. In proposed method a supervised approach is used as there is a finite set of classes (positive and negative). The method needs labeled data to train classifiers. In a machine learning based classification a training set is

used by an automatic classifier to learn the different characteristics of documents, and a test set is used to validate the performance of the automatic classifier [6]. The unsupervised methods are used when it is difficult to find labeled training documents. Unsupervised learning does not require prior training in order to mine the data.

III. PROPOSED SYSTEM ARCHITECTURE

The sentiment analysis is a complex process that involves 5 different steps to analyze sentiment data [7]:

a) **Data collection:** the first step of sentiment analysis consists of collecting data from content contained in the social networks. These data collected are not organized as they may be expressed in different contexts, slangs etc. hence we cannot analyze it manually. Hence, we have to use natural language processing and text analytics to extract and classify.

b) **Text preparation:** This involves cleaning the data before analysis. Any data that is non textual or containing irrelevant information may be identified and discarded.

c) **Sentiment detection:** The extracted opinions are examined and classified as objective and subjective. Objective statements are discarded and subjective statements are retained.

d) **Sentiment classification:** in this step, subjective sentences are classified in positive, negative and neutral.

e) **Presentation of output:** the main objective of sentiment analysis is to convert unstructured text into meaningful information. The analysis can be graphically displayed to convey the required information of classified statements.

The workflow diagram of the proposed architecture is shown below in Fig. 1.

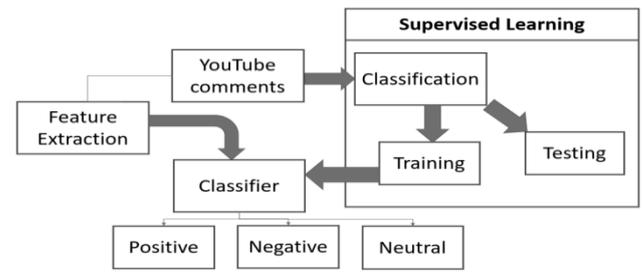


Figure 1. Workflow diagram of the proposed architecture

The architecture for the prediction model that uses linear regression is:

Input: The prediction for the number of subscribers requires for the model to be fed with a range of past data of subscriber gain.

Data Acquisition: The model is fed with a dataset containing the number of subscribers gained in a day for a month.

Data Processing: By using the past data the model has to predict the number for the next month to the month which was fed, of the values were valid and correct [8].

Fig.2 represents the architecture proposed for the prediction.

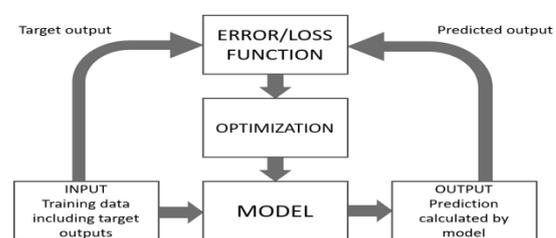


Figure 2. Work flow of prediction architecture

IV. ALGORITHMIC OVERVIEW

Most of the practical and real-world machine learning applications use supervised learning having input variables (x) and an output variable (Y) and an algorithm to study the mapping function, or, in other words, the relation from the input to the output.

$$Y = f(X). \quad (1)$$

The proposed work aims to approximate the relation in Eq. (1) in such an optimized manner that for new input data, one can predict the output variables for that data.

Regression: A regression problem is one which yields a continuous output variable, such as an integer or a floating-point variable. Many different regression methods can be used but the simplest one is linear regression. Linear regression tries to best-fit the data with the best hyper plane passing through the points, as shown in Fig. 3.

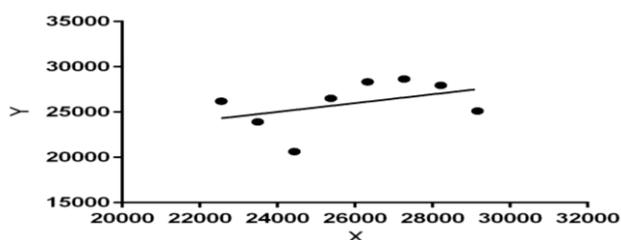


Figure 3. Linear Regression

Linear regression assumes a linear relationship between the input variables (x) and the single output variable (y). In other words, it is a linear model that assumes that y can be calculated from a linear combination of the input variables (x).

In simple linear regression, there is a single input variable (x) whereas in multiple linear regression there are multiple input variables. This is shown in the Fig. 4.

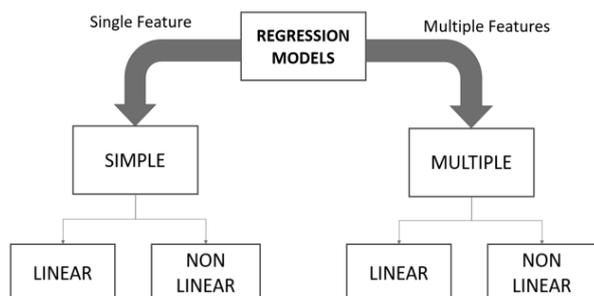


Figure 4. Regression models

The simple linear regression equation is graphed as a straight line. The simple linear regression equation is represented as Eq. (2)

$$E(y) = (\beta_0 + \beta_1 x) \quad (2)$$

β_0 is the y intercept of the regression line. β_1 is the slope.

$E(y)$ is the expected value or mean of y for a given value of x.

Classification: A classification problem is when we get a discrete output variable i.e. when the output is categorical. A classification model tends to draw some proper conclusion from the observed values, for example, good or bad, yes or no etc. In proposed work the classification is done into positive negative or neutral comments.

Textblob provides in-built classifiers using which a custom classifier can be created [5]. Here the sentiment function is used that returns two properties namely subjectivity and polarity.

V. ALGORITHMIC IMPLEMENTATION

The following are the steps for the sentiment classification problem.

Dataset Description: For prediction, a dataset containing more than 692,000 comments from different YouTube videos across the platform is used. This data was collected from the 200 listed YouTube videos that are contained in the trending category every day in the US and the UK. An image of this dataset is shown in Fig. 5.

video_id	comment_text	likes	replies
1	XpVt6Z1Gjjo	Logan Paul it's yo big4	0
2	XpVt6Z1Gjjo	I've been following y 3	0
3	XpVt6Z1Gjjo	Say hi to Kong and m3	0
4	XpVt6Z1Gjjo	MY FAN, attendance3	0
5	XpVt6Z1Gjjo	trending 4Y% 3	0
6	XpVt6Z1Gjjo	#1 on trending AYYE13	0
7	XpVt6Z1Gjjo	The end though 4Y -4	0
8	XpVt6Z1Gjjo	#1 trending!!!!!!! 3	0
9	XpVt6Z1Gjjo	Happy one year vlog 3	0
10	XpVt6Z1Gjjo	You and your shit br 0	0
11	XpVt6Z1Gjjo	There should be a m 0	0
12	XpVt6Z1Gjjo	Dear Logan, i really v 0	0
13	XpVt6Z1Gjjo	Honestly Evan is so a 0	0
14	XpVt6Z1Gjjo	Casey is still better th 0	0
15	XpVt6Z1Gjjo	aw gezz rick this guy 0	0
16	XpVt6Z1Gjjo	He happy cause he ir 0	0
17	XpVt6Z1Gjjo	Ayyyyoooo Logan w 1	0
18	XpVt6Z1Gjjo	Bro y didnt u give me 0	0
19	XpVt6Z1Gjjo	It's been fun watchin 3	0
20	XpVt6Z1Gjjo	Made a lot of people 0	0
21	XpVt6Z1Gjjo	NO HONEY NOOO ca 0	0
22	XpVt6Z1Gjjo	Jake Paul is the faste 1	0
23	XpVt6Z1Gjjo	lol PEWDIEPIE accide 0	0
24	XpVt6Z1Gjjo	You should do a prar 0	0
25	XpVt6Z1Gjjo	I love Logan and jake 1	0
26	XpVt6Z1Gjjo		0

Figure 5. Dataset containing statements

The model classifies comments into three categories. The headers in the comments file are:

- video_id
- comment_text
- likes
- replies

The sentiment function in the Textblob library is deployed now for the sentiment classification. After importing the dataset into the model, 1000 random samples are extracted from it and sentiment polarity for each sample is computed.

The sentiment function segregates the statements by polarity and subjectivity. Objective comments are excluded as mentioned as earlier. The sentiment polarity column is added to the dataset and then it is display. For classification, the polarity is converted from continuous to categorical. The positive comments fall under polarity of 1 as shown in Fig. 6, where 10 comments are displayed.

video_id	comment_text	likes	replies	pol
8	XpVt6Z1Gjjo Happy one year vlogiversary	3	0	1.0
11	XpVt6Z1Gjjo Dear Logan, I really wanna get your Merch but I don't have the money. We don't even have a Car. It would really make my day to have any of your merch	0	0	1.0
13	XpVt6Z1Gjjo Casey is still better then logan	0	0	1.0
15	XpVt6Z1Gjjo He happy cause he in a movie	0	0	1.0
18	XpVt6Z1Gjjo It's been fun watching you grow. I'm at 42 days straight and can't seem to grow. Any advice?	3	0	1.0
22	XpVt6Z1Gjjo lol PEWDIEPIE accidentally played song with words nig%\$r during his stream, u can watch that part on my channel, I've just uploaded. Like so more ppl could see this	0	0	1.0
24	XpVt6Z1Gjjo I love Logan and Jake so much and they are so amazing and I look up to them so much 🥰🥰🥰🥰	1	0	1.0
27	XpVt6Z1Gjjo if you get allot of diss likes do you get on the top comments? lets see diss like this please	0	0	1.0
29	XpVt6Z1Gjjo I love you so much I love to meet you But I live in Israel You really inspire me I really like your channel Keep doing what you do nBecause everyone loves it and will be Unique and Pashan will be ...	0	0	1.0
30	XpVt6Z1Gjjo 🐦 watch by clicking here you can see people's are entertaining	0	0	1.0

Figure 6. Displaying positive comments.

The negative comments fall under the polarity of -1 as shown in the Fig.7.

video_id	comment_text	likes	replies	pol
9	XpVt6Z1Gjjo You and your brother may have single handedly ruined YouTube.....thanks...	0	0	-1.0
12	XpVt6Z1Gjjo Honestly Evan is so annoying. Like its not funny watching him try to be famous he's trying way to hard and I don't like it	0	0	-1.0
16	XpVt6Z1Gjjo Ayyyooooo Logang what up . This was a hard vlog to watch Logan how dare are you to destroyed that YouTube bag . Logang Army check my covers and share them can Logang help me to hit 1,000 Subscri...	1	0	-1.0
19	XpVt6Z1Gjjo Made a lot of people hate youtube - GJ	0	0	-1.0
28	XpVt6Z1Gjjo Evan is a horrible human being he also looks so jealous when you open it like he should be happy for you	0	0	-1.0
34	XpVt6Z1Gjjo Gotta love Youtube for giving morons the ability to earn a buck from other morons.	0	0	-1.0
36	XpVt6Z1Gjjo Can the Pauls please stop saying they're the fastest growing YouTube channels?! That's PewDiePie's position!! He made managed to get all the subscribers you guys have in 6 MONTHS!!! Why is that so...	1	0	-1.0
42	XpVt6Z1Gjjo Not a hater but you	0	0	-1.0
43	XpVt6Z1Gjjo Where is the other dog	0	0	-1.0
44	XpVt6Z1Gjjo I go out of my way to dislike every single one your videos	0	0	-1.0

Figure 7. Displaying negative comments

video_id	comment_text	likes	replies	pol
0	XpVt6Z1Gjjo Logan Paul it's yo big day !!!!!	4	0	0.0
1	XpVt6Z1Gjjo I've been following you from the start of your vine channel and have seen all 365 vlogs	3	0	0.0
2	XpVt6Z1Gjjo Say hi to Kong and maverick for me	3	0	0.0
3	XpVt6Z1Gjjo MY FAN . attendance	3	0	0.0
4	XpVt6Z1Gjjo trending 🤔	3	0	0.0
5	XpVt6Z1Gjjo #1 on trending AYYEEEEEE	3	0	0.0
6	XpVt6Z1Gjjo The end though 🥰🥰🥰	4	0	0.0
7	XpVt6Z1Gjjo #1 trending!!!!!!!	3	0	0.0
10	XpVt6Z1Gjjo There should be a mini Logan Paul too!	0	0	0.0
14	XpVt6Z1Gjjo aw geez rick this guy is the face of YouTube.	0	0	0.0

Figure 8. Displaying neutral comments

Similarly, 10 neutral comments which fall under polarity of 0 are also displayed as shown in Fig. 8. Hence the steps we covered for comment segregation are:

1. Reading the data
2. Data preprocessing
3. Calculating polarity for each sample
4. Adding the polarity column to the data
5. Converting the polarity values from continuous to categorical
6. Displaying positive, negative and neutral comments
7. Calculating the count for each sentiment

A count of the categorized comments is computed, and the values as displayed onto a graph in Fig. 9.

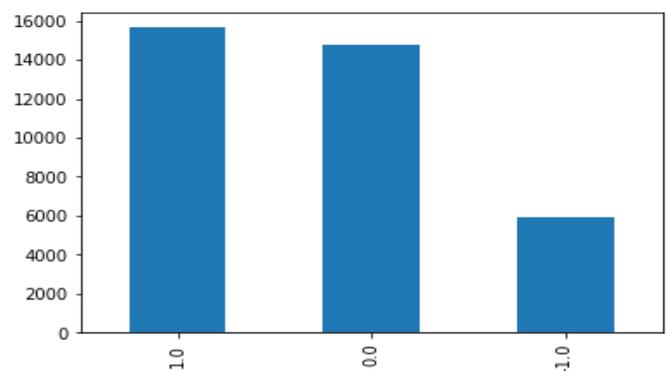


Figure 9. Graph of the count of categorized sentiments

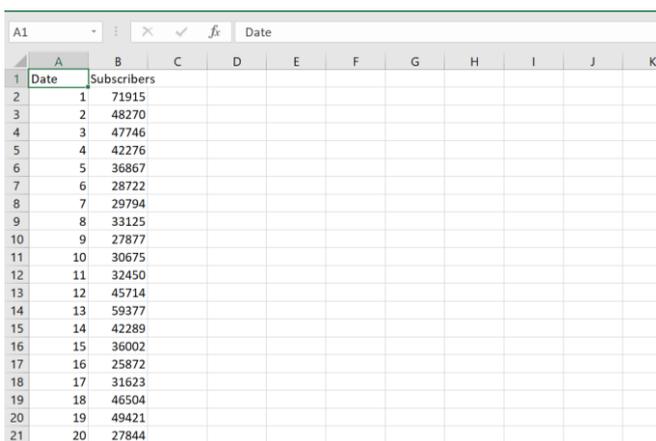
```

1.0    300796
0.0    284450
-1.0   106154
Name: pol, dtype: int64
    
```

Figure 10. Count of comments in each sentiment

The total number of comments in the dataset is 692273, but the total number of comments classified is 691400. The remaining comments would have been identified as objective and would have been ignored.

In the prediction part, dataset that contains the number of subscribers gained by the Swedish YouTuber PewDiePie in the month of May is used for experiments. The image of the dataset is shown as Fig. 11.



Date	Subscribers
1	71915
2	48270
3	47746
4	42276
5	36867
6	28722
7	29794
8	33125
9	27877
10	30675
11	32450
12	45714
13	59377
14	42289
15	36002
16	25872
17	31623
18	46504
19	49421
20	27844

Figure 11. Dataset containing subscriber information

The dataset shown above contains information of subscribers gained on each day starting from May 1 till May 20. After uploading this dataset into proposed model. The below image shows the predicted values, which display the predicted number of subscribers gained from May 21 to May 28:

```

Total number of increase in subscribers in May 21 ==> 29154
Total number of increase in subscribers in May 22 ==> 28210
Total number of increase in subscribers in May 23 ==> 27266
Total number of increase in subscribers in May 24 ==> 26322
Total number of increase in subscribers in May 25 ==> 25378
Total number of increase in subscribers in May 26 ==> 24434
Total number of increase in subscribers in May 27 ==> 23491
Total number of increase in subscribers in May 28 ==> 22547
    
```

Figure 12. Subscribers gain predicted using proposed model for the next 8 days.

The RMSE value calculated for these two columns is 0.17469. Root Mean Square Error (RMSE) is the

standard deviation of the residuals i.e. prediction errors. Prediction errors are a measure of how far the data points lie from the regression. RMSE is a measure of how spread out the residuals are. The RMSE value tells you how concentrated the data points are around the best fit. The smaller the RMSE values the better the prediction.

Fig. 13 shows the actual values of the number of subscribers gained from May 21 to May 28, obtained from Social Blade:



Date	Day	Change	Subscribers	Previous
2019-05-18	Sat	+46,504	95,888,098	21,500,643,435
2019-05-19	Sun	+49,421	95,937,519	21,515,600,792
2019-05-20	Mon	+27,844	95,965,363	21,525,769,515
2019-05-21	Tue	+25,129	95,990,492	21,535,298,312
2019-05-22	Wed	+27,963	96,018,455	21,545,907,665
2019-05-23	Thu	+28,663	96,047,118	21,553,261,633
2019-05-24	Fri	+28,345	96,075,463	21,563,668,727
2019-05-25	Sat	+26,532	96,101,995	21,568,281,297
2019-05-26	Sun	+20,640	96,122,635	21,585,621,484
2019-05-27	Mon	+23,921	96,146,556	21,596,709,721
2019-05-28	Tue	+26,212	96,172,768	21,607,185,737

Figure 13. Actual values of subscribers, obtained from Social Blade.

As seen in the Table I the results are almost the same except for a few deviations in the numbers. This is because of the smaller number of initial samples taken. Dataset of only 20 days is used for experiments. The accuracy of the proposed method can be improved by using the data for a longer period of time.

Hence, the steps followed to achieve this prediction analysis are:

1. Feeding the dataset
2. Read the dataset and store it in a data frame
3. Training and testing the model
4. Applying simple linear regression
5. Comparison of predicted values with actual values and calculating error i.e. RMSE

Figure 3 displays the scatter plot for the above linear regression problem. A comparison of the actual

values against the predicted values is summarized in Table I.

TABLE I. EVALUATION OF PREDICTED AND ACTUAL VALUES

DATE (MAY)	PREDICTED VALUES	ACTUAL VALUES
21	29154	25129
22	28210	27963
23	27266	28663
24	26322	28345
25	25378	26532
26	24434	20640
27	23491	23921
28	22547	26212

VI. CONCLUSION

In the proposed method sentiment analysis for dataset from YouTube was proposed using regression model. The work also proposed sentiments classification using regression. For the proposed method the RMSE values for the regressor model is observed to be around 0.17469. It was observed that some positive comments were classified under neutral due to the lack of keywords. Machine learning cannot be 100% accurate especially when it comes to regression and prediction, since it does not take into account several real-life problems. For example, it was observed in one of our predictions that the predicted values were very far varied from the actual values and the reason was that the YouTuber had posted a new song which had garnered a lot more views than predicted.

Another limitation is the lack of data. If we feed the model poor data, it will yield poor results. Some models require copious amounts of training data. The larger the model goes; the greater amount of data is needed to expect viable results. Keeping all this in mind we need to try and optimize the model as good as possible to provide a better classification and prediction. Most technologies these days rely heavily on machine learning for proper functioning. Machine learning acts as a bridge between humans

and technologies and that is why machine learning is considered to be the future of technology.

VII. REFERENCES

1. Y. Singh, P. K. Bhatia, and O.P. Sangwan, "A Review of Studies on Machine Learning Techniques," *International Journal of Computer Science and Security*, Volume (1) : Issue (1), pp. 70-84, 2007.
2. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis.
3. Nasukawa, T. & Yi, J. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, October 23–25, 2003. (pp. 70–77). Florida, USA.
4. Wiebe, J. & Riloff, E. 2005. Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, 2005, pp. 486-497.
5. R. Feldman, "Techniques and Applications for Sentiment Analysis", *Communications of the ACM*, Vol. 56 No. 4, pp. 82-89, 2013.
6. R. Joshi and R. Tekchandani, "Comparative analysis of twitter data using supervised classifiers," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 3, Aug 2016, pp. 1–6.
7. Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni and Tiziana Guzzo, "Approaches, tools and applications for sentiment analysis implementation", in *2015 International Journal of Computer Applications*, Vol. 125, No. 3, September 2015.
8. Prabowo, R.; Thelwall, M. Sentiment analysis: A combined approach. *J. Informetr.* 2009, 3, 143–157.
9. Pang, B., Lee, L., Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proc. of 7th EMNLP*, pp.79-86.
10. Jagdale, O.; Harmalkar, V.; Chavan, S.; Sharma, N. Twitter mining using R. *Int. J. Eng. Res. Adv. Tech.* 2017, 3,252–256.
11. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and

- applications: A survey. *Ain Shams Eng. J.* 2014, 5, 1093–1113.
12. A. Khan, B. Baharudin, “Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs,” *Processed on National Postgraduate Conference (NPC)*, pp. 1 – 7, 2011.
 13. Moro, P. Rita, and B. Vala, “Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach,” *Journal of Business Research*, vol. 69, no. 9, pp. 3341–3351, 2016.
 14. Cambria, E. *Affective computing and sentiment analysis. IEEE Intell. Syst.* 2016, 31, 102–107.