COMPARATIVE SENTIMENT ANALYSIS: GREAT BRITAIN VERSUS THE UNITED STATES OF AMERICA USING ENSEMBLE METHODS

Szerző:

Sirmad Mahmood Hashmi
University of Debrecen

Szerző e-mail címe:
Sirmadhashmi@gmail.com

Lektorok:

Márton Ispány (PhD)
University of Debrecen

Csilla Csukonyi (PhD)
University of Debrecen

…és további két anonim lektor

Absztrakt

ÖSSZEHASONLÍTÓ ÉRZELEMELEMZÉS ENSEMBLE MÓDSZEREK HASZNÁLATÁVAL: NAGY-BRITÁNIA VERSUS AMERIKAI EGYESÜLT ÁLLAMOK

A szociális média közvéleménykutatás gyors és hatékony módjává vált az üzleti, politikai, sport stb. élet terén egyaránt. Ugyanakkor a YouTube egyedi jellemzői olyan új problémákat okoznak a jelenlegi közösségi média elemzések során, mint például a különböző földrajzi régiókról érkező videókkal kapcsolatos véleménykülönbség. Ez a kutatás az ensemble gépi tanulási megközelítéseket használó mondatszintű érzelem osztályozás összevetésére törekszik. Ebben a dokumentumban bemutatásra kerül a kísérleti eredmények részletes összehasonlítása a javasolt technikákkal és azok jellemzőivel.

Kulcsszavak: Ensemble Machine Learning, AdaBoost, Extra Tree Classifier, Random Forest Classifier, érzelemelemzés, YouTube, földrajzi alapú adatbányászat

Diszciplinák: informatika, nyelvészet, pszichológia

Abstract

Social media has become a rapid and effective way of gauging public opinion for business, politics, sports, etc. However, YouTube's unique characteristics give rise to new problems for current social media analysis such as the difference of opinion on videos from different geographical regions. This research is devoted to the comparison of sentence-level sentiment classification using ensemble machine learning approaches. In this

document, the detailed comparison of experimental results is shown with the proposed techniques and their performances.

Sentiment analysis is a dedicated field for investigating people's opinions, sentiments, and emotions (Medhat, Hassan and Korashy, 2014). It is also known as text analysis, opinion mining and text classification. It analyzes sentiments based on different multimedia data e.g. text, videos, images, and voice recordings. It has been widely studied in the field of data mining and natural language processing. In the current era of information, businesses are in need of sentiments analysis to capture most market shares. Measuring people's opinions for application development is leading to success in social media platforms. Facebook has recently introduced five new emoticons in their application for analyzing sentiments (Tian et al., 2017). Many people would agree that the preferences of people are dissimilar according to geographical locations. A car that has a higher preference in the United States of America may not be suitable for Great Britain. Thus, there is a need to analyze the sentiments of people based on different geographical locations for better product development, medications, academic curriculums, etc. In this research, a comparison of ensemble machine learning methods for sentiments classification is presented using the YouTube data set. YouTube is the most unique social media platform among all the other platforms. It is the leading video-sharing based social media platform. It is considered the second largest search engine (YouTube: Mushroomnetworks.com, 2016) from the time it was bought over by "Google" in 2006. YouTube user account is considered as a channel. The owner of the channel can share, edit, prompt discussions, and create a play-list. These channels can also be monetized with the advertisements. Recently two music videos on YouTube reached three billion views with an estimated earning of 75 thousand - 12 million dollars per musician (McIntyre, 2017). Furthermore, the data generated through YouTube daily is increasing the value of

Big Data. Few recent facts (Aslam, 2018) of YouTube are the following:

- In a total of 1.57 billion, active monthly users are on YouTube.
- In total 30 million-plus active daily users are on YouTube.
- In total 300,000 Paying Subscribers for YouTube TV.
- In total 5 billion-plus videos have been shared on YouTube to date.
- In total 300 hours of video content is uploaded on YouTube per minute.

In the past decade, several approaches have been developed for the classification of opinion from textual data. These approaches are mainly divided into three categories: Lexicon-Based Analysis, Mchine Learning Classification, and Deep Learning Classification. Furthermore, a hybrid approach of combining the lexicon-based method with machine learning is also discovered (Zhang et al, 2011). The lexicon-based analysis makes use of a dictionary of opinionated synonyms like TextBlob, NRC or Bing Liu's (Potts, 2011, Mohammad, 2018) for aggregating sentiments. It extracts words from the dataset and compares it with a dictionary to categorize them in classes: Positive, Negative, and Neutral. In this research, a lexicon label-based approach is proposed, with a sentence-level sentiment technique for YouTube opinion mining. This proposed technique is labeling the dataset using TextBlob (AdvancedUsage:, 2019) and training it with ensemble machine learning techniques using Scikit-Learn.

Related work

The lexicon-based approach is one of the most commonly used methods to investigate sentiments. Various methods have been developed for lexicon-based analysis, it uses opinionated dictionary such as SentiWordNet, Synsets, and Bing Liu's for aggregating sentiments (Vu, 2017). On the other hand, the machine learning approach trains a model based on the features e.g. uni-grams, bi-grams, and bi-gram combines with part of speech (Pang, Lee and Vaithyanathan, 2002). The classification models such as Naive Bayes, Support Vector Machine, and Maximum Entropy are most known for sentiment analysis. The major drawback of Machine learning methods lies in the required manual labeling for training the sentiment classifiers . Our research is divided into three parts: automatic labeling, dataset division and training the classifier for comparative analysis.

Methodology

In this section datasets, algorithm architecture, and machine learning classifiers are discussed in detail.

A. YouTube Dataset

The datasets (M. J., 2017) used in this research are downloaded from kaggle.com. These datasets contain YouTube videos statistics of 200 trending videos from September 2017 based on two geographical regions Great Britain and The United States of America. Each region has two files in which one file is dedicated to YouTube Video statistics such as

numbers of likes, dislikes, and comments, etc. the second file contains the textual comments of videos. The following is the description of each dataset: Comments datasets: It contains four attributes: video id, comments text, likes, and replies. The file types of comments datasets are in CSV format. The USA Comments data file size 64.9 MB has 4 columns and 691464 rows. Great Britain Comments data file size 73.4 MB has 4 columns and 718458 rows.

B. Handling Comments Dataset

The comment dataset (M. J., 2017) files are combined and consist of around 1.3 million unstructured comments (textual data). These comments are the most important aspect of this research as they are to be used for training the multiple ensemble models for comparison. The scheme that has been developed here to process comments data for learning models can be divided into three parts: Cleansing and cleaning, automatic labeling, and Dataset divisions. A detailed description of each part of the scheme is presented as the following:

B.1) Cleansing and Cleaning: In this phase, all the comments are cleansed by removing all the punctuation marks, symbols, unnecessary white spaces, etc. In the cleaning processes, all the comments are cleaned from the stop words and non-English words.

B.2) Automatic labeling: In this phase, comments datasets were labeled automatically using the TextBlob python package. These labels were categorized in the following sentiments: - 1(negative), 0(neutral), 1(Positive). After automatic labeling, a random sample of 4000 comments was picked to verify their labels manually. This verification process assured that the automatic labeling process correctly classified comments 99.9 percent with an accurate label.

Figure 1 represents the total number of comments from each class (negative, neutral, positive) in the comments dataset of Great Britain. Figure 2 illustrates the total number of comments from each class (negative, neutral, positive) in the comments dataset of the United States of America.

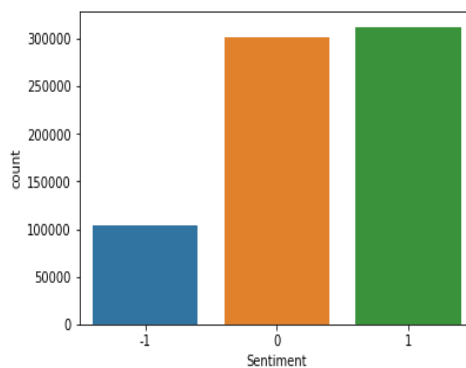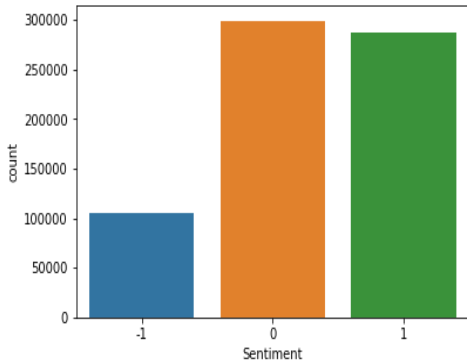*Figure 1: Labels count bar chart for Great Britain comments*

Figure 2: Labels count bar chart for The United States of America comments

B.3) Dataset Division: The comments dataset for the United States of America contains 691373 records and Great Britain contains 718452. These datasets have a different number of labels from each category. In order to understand why the divisions of datasets are necessary, look into Figure 3 and Figure 4 below.

Figure 3, and Figure 4 represent the statistics of the actual labels from each category.

Figure 3: Datasets statistics of GB

	count	mean	std	min	25%	50%	75%	max
Sentiment								
-1	104036.0	0.670071	0.458543	0.0	0.0	1.0	1.0	1.0
0	301911.0	0.023981	0.146841	0.0	0.0	0.0	0.0	1.0
1	312505.0	0.692405	0.446224	0.0	0.0	1.0	1.0	1.0

Figure 4: Datasets statistics of USA

	count	mean	std	min	25%	50%	75%	max
Sentiment								
-1	105395.0	0.674520	0.456317	0.0	0.0	1.0	1.0	1.0
0	298879.0	0.024545	0.148162	0.0	0.0	0.0	0.0	1.0
1	287099.0	0.689861	0.446293	0.0	0.0	1.0	1.0	1.0

So, I have divided the datasets into 3 different small datasets having the same number of labeled data (first labeled data). As both datasets have a huge difference between -1 (negative) and the other two labels count. So, it is a better approach to divide dataset with an equal number of labels from each category so that the learning algorithm will not be biased. According to the dataset statistics, the dataset is divided in the following manner: Table 1 and Table 2 represent the divisions of datasets. In this research, three datasets were created for each continent (GB, USA) with the same number of labels for each category. The idea behind creating three separate datasets is to understand whether the size of the training set can make a difference in the accuracy of the learning models in the case of dealing with textual data.

C. Proposed Architecture

Figure 5 is representing the proposed architecture of the research. This architecture is divided into three sub-modules: pre-processing, aggregating sentiments, and ensemble learning models. Each of these modules works independently. The following are the working details of each module.

Table 1: Datasets divisions for training models USA

The United States of America Labels	Dataset 1	Dataset 2	Dataset 3
1 (Positive)	4000	24000	50000
0 (Neutral)	4000	24000	50000
-1 (Negative)	4000	24000	50000
Total	12000	72000	150000

Table 2: Datasets divisions for training models GB

Great Britain Labels	Dataset 1	Dataset 2	Dataset 3
1 (Positive)	4000	24000	50000
0 (Neutral)	4000	24000	50000
-1 (Negative)	4000	24000	50000
Total	12000	72000	150000

Figure 5: Proposed Architecture:
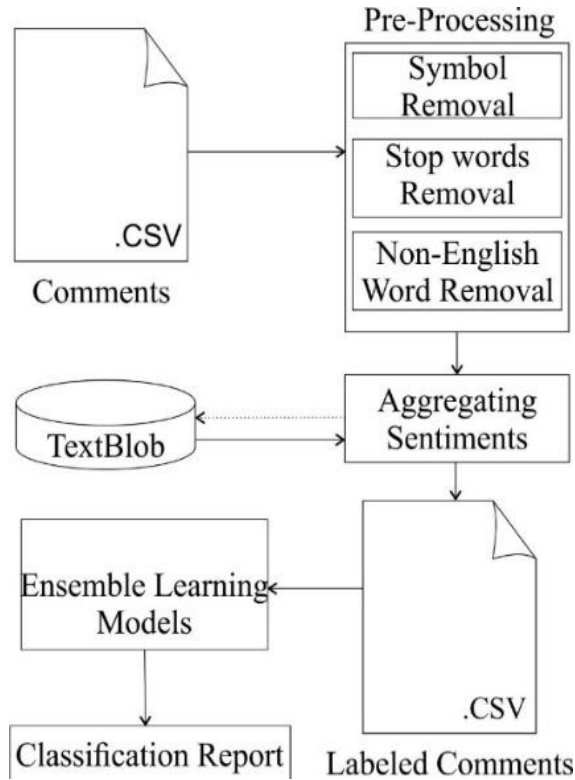
C.1) Pre-Processing: This module is performing the cleansing and cleaning of the unstructured comment datasets. To know the detail of cleaning and cleansing scheme please refer to section B.1. This module is taking data from.csv comment file and cleaning it and storing data to a data frame for the next module to perform its functionality.

C.2) Aggregating Sentiments: This module is aggregating sentiments on comments. This aggregation is done with the help of automated TextBlob function for analyzing sentiments. TextBlob has categorized the sentences in three different labels -1 (Negative), 0 (Neutral), and 1 (Positive). After aggregating sentiments, the data frame is stored in the CSV file for further processing in the next module.

C.3) Ensemble Learning Model: This module is performing the classification of the machine learning model and it is generating the classification report for each model. The generated report for each dataset from this module is used for comparison between each ensemble model.

D. Ensemble Learning

It is a machine learning approach that combines multiple different learning models for the purpose of producing better predictive performance. Ensemble models belong to the family of supervised learning because their sole purpose is to train and make predictions. Ensemble methods are known to produce better results because of their diverse nature (Geurts, Ernst and Wehenke, 2006).

D.1) Ensemble Methods Type: The two most known ensembles constructing techniques are AdaBoost family and Boot-strap aggregation. These both techniques work by picking a base learning algorithm and run it multiple times with different training sets. In AdaBoost, a set of weights is maintained over the original training set and these weights are adjusted after each classifier is learned by the base learning algorithm. In bagging, all training sets are constructed by creating a bootstrap duplicate of the original training set.

D.2) Random Forest: It is an ensemble learning model that is used mainly for classification and regression. It works by constructing multiple decision trees while training and it outputs the classes in the mode of regression (mean prediction) or classification (classes) of the single trees.

D.3) Extremely Randomized Trees: The ExtraTreeForest or Extremely Randomized Trees is an ensemble learning model. This model is constructed on the unpruned regression trees or unpruned decision trees with the top-down approach. This algorithm has two significant differences in comparison with other ensemble tree-based models. The first difference is that it picks a cut-points randomly and uses them to split nodes. The second difference is that it uses the complete learning sample to construct or grows the trees rather than a bootstrap replication technique.

Results

In this section ensemble machine learning classifiers for sentiment analysis results are presented.

E. Train and Test

All the ensemble machine learning classifiers are trained on a 70:30 split ratio with 1000 trees (n_estimator = 1000). This means the training set size is 70 percent of the total dataset and the remaining 30 percent is test set. In this research, the Scikit-Learn package is used for the implementation of these models. Note: All the codes are uploaded on the following link: https://github.com/sirmad hashmi/ Sentiment-Analysis-of-Social-Media-Network

E.1) AdaBoost: This section represents the comparison of results based on the AdaBoost (SAMME.R) classifier between Great Britain and the United States of America for sentiment classification.

Table 3 illustrates the comparison of the AdaBoost sentiment classifier results on three different size datasets (refer to section B.3) of Great Britain. In which the dataset 3 has the highest accuracy of

0.8778. Table 4 illustrates the comparison of the AdaBoost sentiment classifier results on three different size datasets (refer to section B.3) of the United States of America. In which the dataset 3 has the highest accuracy of 0.86673.

E.2) Extra Tree Classifier: This section represents the comparison of results based on Extra Tree ensemble classifier between Great Britain and the United States of America for sentiment classification.

Table 3: Classification Report of AdaBoost on Great Britain Datasets

Great Britain				
Dataset 1	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.77	0.75	0.76	1200
0 (Neutral)	0.84	0.97	0.90	1200
1 (Positive)	0.86	0.75	0.80	1200
Accuracy	0.8277777			
Dataset 2	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.80	0.84	0.82	7200
0 (Neutral)	0.90	0.98	0.94	7200
1 (Positive)	0.91	0.78	0.84	7200
Accuracy	0.8663425			
Dataset 3	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.93	0.72	0.81	15000
0 (Neutral)	0.90	0.99	0.94	15000
1 (Positive)	0.82	0.92	0.87	15000
Accuracy	0.8778222			

Table 4: Classification Report of AdaBoost on The United States of America Datasets

The United States of America				
Dataset 1	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.67	0.74	0.71	1200
0 (Neutral)	0.80	0.94	0.86	1200
1 (Positive)	0.88	0.64	0.74	1200
Accuracy	0.77277777			
Dataset 2	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.77	0.80	0.78	7200
0 (Neutral)	0.87	0.98	0.92	7200
1 (Positive)	0.92	0.75	0.83	7200
Accuracy	0.84560185			
Dataset 3	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.79	0.86	0.82	15000
0 (Neutral)	0.90	0.98	0.94	15000
1 (Positive)	0.93	0.76	0.84	15000
Accuracy	0.86673333			

Table 5 illustrates the comparison of Extra Tree sentiment classifier results on three different size datasets (refer to section B.3) of Great Britain. In which the dataset 3 has the highest accuracy of 0.97775555

.

Table 6 illustrates the comparison of Extra Tree sentiment classifier results on three different size datasets (refer to section B.3) of the United States of America. In which the dataset 3 has the highest accuracy of 0.97311111.

Table 5: Classification Report of ExtraTreeClassifier on Great Britain Datasets

Great Britain				
Dataset 1	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.93	0.88	0.91	1200
0 (Neutral)	0.90	0.98	0.94	1200
1 (Positive)	0.93	0.90	0.91	1200
Accuracy	0.91888888			
Dataset 2	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.97	0.96	0.96	7200
0 (Neutral)	0.97	0.99	0.98	7200
1 (Positive)	0.97	0.96	0.96	7200
Accuracy	0.96986111			
Dataset 3	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.98	0.97	0.97	15000
0 (Neutral)	0.98	0.99	0.99	15000
1 (Positive)	0.97	0.97	0.97	15000
Accuracy	0.97775555			

Table 6: Classification Report of ExtraTreeClassifier on The United States of America Datasets

The United States of America				
Dataset 1	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.90	0.86	0.88	1200
0 (Neutral)	0.88	0.97	0.92	1200
1 (Positive)	0.92	0.86	0.89	1200
Accuracy	0.89861111			
Dataset 2	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.97	0.96	0.96	7200
0 (Neutral)	0.97	0.99	0.98	7200
1 (Positive)	0.97	0.96	0.97	7200
Accuracy	0.97032148			
Dataset 3	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.96	0.97	0.97	15000
0 (Neutral)	0.98	0.99	0.99	15000
1 (Positive)	0.98	0.96	0.97	15000
Accuracy	0.97311111			

E.3) Random Forest Classifier: This section represents the comparison of results based on the Random Forest ensemble classifier between Great Britain and the United States of America for sentiment classification.

Table 7 illustrates the comparison of the Random Forest sentiment classifier results on three different size datasets (refer to section B.3) of Great Britain. In which the dataset 3 has the highest accuracy of 0.97395555.

Table 7: Classification Report of RandomForestClassifier on Great Britain Datasets

Great Britain				
Dataset 1	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.94	0.84	0.89	1200
0 (Neutral)	0.85	0.98	0.91	1200
1 (Positive)	0.92	0.86	0.89	1200
Accuracy	0.89777777			
Dataset 2	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.97	0.95	0.96	7200
0 (Neutral)	0.96	0.99	0.97	7200
1 (Positive)	0.97	0.95	0.96	7200
Accuracy	0.96333333			
Dataset 3	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.97	0.96	0.97	15000
0 (Neutral)	0.98	0.99	0.98	15000
1 (Positive)	0.97	0.97	0.97	15000
Accuracy	0.97395555			

Table 8: Classification Report of RandomForestClassifier on The United States of America Datasets

The United States of America				
Dataset 1	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.91	0.83	0.87	1200
0 (Neutral)	0.84	0.97	0.90	1200
1 (Positive)	0.92	0.85	0.88	1200
Accuracy	0.88416666			
Dataset 2	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.96	0.95	0.95	7200
0 (Neutral)	0.96	0.99	0.98	7200
1 (Positive)	0.97	0.95	0.96	7200
Accuracy	0.96291666			
Dataset 3	**Precision**	**Recall**	**F1-Score**	**Support**
-1 (Negative)	0.96	0.96	0.96	15000
0 (Neutral)	0.97	0.99	0.98	15000
1 (Positive)	0.97	0.95	0.96	15000
Accuracy	0.96775555			

Table 8 illustrates the comparison of the Random Forest sentiment classifier results on three different size datasets (refer to section B.3) of the United States of America. In which the dataset 3 has the highest accuracy of 0.96775555.

Discussion

This research is studying the behavior of ensemble machine learning classifiers on the textual datasets. This study is conducted to understand which ensemble machine learning model performs better for the classification of sentiments. This research used a comparative approach to measure the performance of ensemble machine learning model AdaBoost, Extra Tree classifier, and Random forest classifier for sentiment classification. The datasets used in this comparison are divided into three smaller datasets with an equal number of labels from the category of -1, 0, 1 (negative, neutral, positive). These datasets contain textual comments from YouTube based on the top 200 trading videos from September 2017. These textual comments extracted from YouTube based on two geographical regions the United States of America and Great Britain. The results of this research are very astonishing as ensemble machine learning models produced accuracy above 75 percent in all of the cases. This research has clearly shown that the performance of the sentiment classifier is depended on the size of the training data. Let us compare the performances of the model according to each dataset. In the case of classifying the "Dataset 1" (section B.3), ExtraTreeForest obtained higher accuracy in both region datasets followed by RandomForest and AdaBoost. In the case of classifying the "Dataset 2" (section B.3), ExtraTreeForest obtained higher accuracy in both region datasets followed by RandomForest and AdaBoost. In the case of Great Britain classification of "Dataset 2" (section B.3), ExtraTreeForest and RandomForest have very close results. In the case of classifying the "Dataset 3" (section B.3), ExtraTreeForest obtained higher accuracy in both region datasets followed by RandomForest and AdaBoost. In the case of Great Britain's classification of "Dataset 3" (section B.3), ExtraTreeForest and RandomForest have again very close results. These above-mentioned results are interpreting that ensemble approaches work well for a multiclass textual classification problem.

Conclusion

The major finding that we can conclude for this research is that ExtraTreeForest has outperformed both AdaBoost and RandomForest in sentiment classification. The other main important assumption that is drawn from this research is that a bigger size dataset produced higher accuracies from the smaller ones. According to these results of this research, the ensemble machine learning algorithm is one of the best techniques to solve the sentence level sentiment classification problems.

References

AdvancedUsage: OverridingModelsandthe-BlobberClassTextBlob0.15.2 documentation", Textblob.readthedocs.io, 2019. [Online]. Available :https://textblob.readthedocs.io/en/dev/advancedusage.htmlsentiment-analyzers. [Accessed: 24-Apr- 2019].

Aslam, S. (2018). "*YouTube by the Numbers: Stats, Demographics& Fun Facts",* Omnicoreagency.com, 2018. [Online]. Available:https://www.omnicoreagency.com/youtube-statistics/. [Accessed: 09- Nov-2018].

Geurts, P., Ernst, D. and Wehenke, L. (2006). Extremely randomised trees. *Machine Learning, 63, (1),* 3-42.

McIntyre, H. (2017). *"How Much Money Does 3 Billion YouTube ViewsActually Generate For A Musician?",* Forbes, [Online]. Available:https://www.forbes.com/sites/hughmcintyre/2017/09/18/how-much-money-does-3-billion-youtube-views-bring-in/#e4dabbc4aece.[Accessed: 01- Nov- 2018].

Medhat, W., Hassan, A. and Korashy, H. (2014). *'Sentiment analysis algorithms and applications: A survey'. Ain Shams Engineering Journal, 5, (4),* 1093-1113.

M. J, (2017). "*TrendingYouTubeVideoStatisticsandComments",*Kaggle.com,[Online].Available:https://www.kaggle.com/datasnaek/youtube.[Accessed:08-Jul-2018].

Mohammad, S. (2018). '*NRC Emotion Lexicon', Saifmohammad.co'.* [On-line]. Available: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Accessed: 09- Nov- 2018].

Pang, L., Lee, L. and Vaithyanathan, S. (2002). *"Thumbs up? SentimentClassification using Machine Learning Techniques,*" In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) '07, pp. 79-86.

Potts, C. (2011). *SentimentSymposiumTutorial:Lexicons",Sentiment.christopherpotts.net,* [Online].Available:http://sentiment.christopherpotts.net/lexicons.html. [Accessed: 09-Nov- 2018].

Tian, Y., Galery, T., Dulcinati, G. Molimpakis, E. and Sun, C. (2017). *FacebookSentiment: Reactions and Emojis,* in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media

YouTube: The2ndLargestSearchEngine (Infographic)",Mushroomnetworks.com,(2016) [Online]. Available:https://www.mushroomnetworks.com/infographics/youtube—the-2nd-largest-search-engine-infographic/. [Accessed: 05- Nov- 2018].

Vu, L. (2017). *A lexicon-based method for Sentiment Analysis using social network data*, In Intl Conf. Information and Knowledge Engineering, pp. 1016.

Zhang, L, Ghosh, R., Dekhil, M., Hsu, M. and Liu, B. (2011). *Combining lex-icon based and learning-based methods for twitter sentiment analysis*, HP Lab. Tech. Report, HPL-2011-89.