Predicting the Emotions of Research Articles on Twitter

Murtuza Syed
Computer Science
Northern Illinois University
DeKalb Illinois U.S.A
z1836478@students.niu.edu

Mariam Abbaz
Computer Science
Northern Illinois University
DeKalb Illinois U.S.A
z1816567@students.niu.edu

Godwin Richard Thomas
Computer Science
Northern Illinois University
DeKalb Illinois U.S.A
z1838366@students.niu.edu

ABSTRACT

Online social media has a great potential in sharing a lot of content and discussing about many topics. Through online discussion, people have some sentiments or emotions associated with what they post or discuss. Other users who view such posts might also tend to have sentiments for the content posted. In this study, we come up with a novel area of research where we intend to know how these sentiments are observed for scholarly articles shared online. Understanding the sentiments of tweets would give us an insight of how the people have emotions on research articles on online platform.

We also used a creative and unique approach in using Automation for Web Scraping to extract the desired feature. We then performed sentiment analysis on tweets, title and abstracts using Natural Language Processing and then performed statistical analysis on all the subjects that have been shared on Twitter. Using machine learning models, we built models that would predict the sentiments of tweets related to research articles and then we evaluated them in terms of precision, recall, accuracy and f1-score.

KEYWORDS

Sentiment Analysis, Twitter, Natural Language Processing, Emo Emotional impact, Machine Learning, Automated Web Scraping

1 Introduction

Twitter is a very powerful and influential medium playing an important role in influencing the current trends in most of the different aspects of the society. Kwak et al. (2010) mentions that Twitter as a microblogging service as emerged as a relatively new medium in spotlight. The medium which emerged as a spotlight has a current daily active usage of 126 million users according to Twitter (2019). In addition to that, the 140 character limit is allows us to acquire the perfect balance in terms of content. Based on the above reasons, we decided to use Twitter as our medium.

The next step in our agenda was to perform preprocessing on the Altmetrics dataset. We also performed some data analysis on the dataset in general to get a gist of what we were dealing with. While looking at the list of features provided in the feature selection process, we were not completely convinced with the set of features. From the list of original features, we were able to derive some new features and append that to the dataset. But, we noticed that we were missing an important ingredient in our recipe. It was the total number of citations which mentions how many different papers have cited the particular paper. It was missing in our arsenal and we decided to not proceed further without it.

We had to use a technique called Web-Scraping which is defined by Vargiu and Urru (2013) as a software technique aimed at extracting information from websites. The traditional method would be to scrape the data manually from the web but we overloaded the process by adding an automation aspect to it. Using the selenium framework, we wrote a script to automatically scrape the required information and store it into a temporary file which will then be appended as a feature to our main dataset. We faced a problem of being blocked by Google because of sending a lot of data in a short time but we overcame it with the help of proxy rotation and delayed scraping.

We needed to perform Sentiment Analysis on the dataset to acquire the sentiment scores of the respective features. To get started on Sentiment Analysis, we need to go back to the definition of Natural Language Processing (NLP). Chowdhury (2003) explains that NLP is an area of research and application that explains how computers can be used to understand and manipulate natural language text or speech to do useful things. The toolkit which allows us to do NLP is the NLTK (Natural Language ToolKit) framework. As stated by Madnani (2007), NLTK is a collection of modules and corpora, released under an open-source license, that allows students to learn and conduct research in NLP. After the initial preprocessing, we were able to use the NLTK to estimate the sentiment scores of the given features. These scores were appended to the main dataset as features.

Now that we had all the necessary features in our belt, we were ready to dive into Machine Learning. We were really interested by the frameworks present in Python for Machine Learning. "Scikit-learn harnesses this rich environment to provide state-of-the-art implementations of many well known machine learning algorithms, while maintaining an easy-to-use interface tightly integrated with the Python language." (Pedregosa et al., 2011, p. 2). Using the Scikit-learn framework, we were able to try out different Machine Learning algorithms. First, we experimented with the regression algorithms and found out that the accuracy was not really high. We then proceeded with the classification algorithms and found better results. The algorithms were evaluated in terms of their accuracy initially. Support Vector Machines (SVM) had the highest accuracy out of all the models. One thing to be mentioned is that we used Grid Search and performed Hypertuning on the parameters of the algorithms to achieve better accuracy.

Finally, we were able to deploy a model which takes the list of features and trains itself itself to predict the sentiment scores of the research articles based on their tweets. In the future, we wanted to expand our work to Artificial Neural Networks (ANN) and also increase the range of the automated web scraping.

2 Related Work

Narr et al. (2012) has used a language independent sentiment analysis model on the twitter data to estimate the polarity of the tweets. They collected tweets in different languages manually with the help of Amazon mTurk and then used Naive Bayes classifier on the n-gram features to classify the sentiments. The results were again evaluated manually and compared between the different languages. Similarly, the polarity and sentiment of tweets of celebrities who had a follower count of more than 1 million were taken by Bae and Lee (2012). The sentiment score was calculated after performing Lexical Sentiment Analysis. The results were then subjected to three types of correlational analysis (Parson, Spearman and Granger) to determine the strength of independence. The results proved that the celebrities influenced the tweeters by a high margin. Kharde and Sonawane (2016) tested out the various classification methods such as Support Vector Machines (SVM), Naive Bayes and Maximum Entropy on the twitter dataset. They provided insights individually on each of the classification methods and then finally evaluated them based on the metrics of precision and accuracy. In addition to the above classification methods, the classification algorithm Naive Bayes was done separately as a Unigram Multinomial Bayes and a Multigram Multinomial Bayes by Parikh and Movassate (2009).

There were different approaches and methods applied towards sentimental analysis on Twitter. One of them was by Zaman et al. (2010) suggesting that a probabilistic collaborative filter model that predicts future retweets thereby showing the spread of information. Another interesting approach was analysing using Hadoop cluster to process larger amount of data in real-time as pointed out by Mane et al. (2014). The paper focused more emphasis on the processing speed rather than accuracy. They implemented the Map-Reduce algorithm to achieve faster processing speed. There was also a visual based approach compared to the previous text-based approaches. Hao et al. (2011) used various visual based algorithms like pixel-cell based algorithm to perform the analysis along with the help of the tools Pixel Sentiment Calendar and Pixel Sentiment Geo Map.

The models discussed above have done pre-processing on the data by the tradition methods. However, Da Silva et al. (2014) suggested that bag of words and lexicons using feature hashing can be used for preprocessing the data and then feed into the classification algorithms. They compared this with the baseline where the tradition method of preprocessing was followed. On a similar note, Saif et al. (2014) raised the notch by using several different methods to remove the stop-words in preprocessing step. They used Zipf's Law, Term-Based Random Sampling (TBRS), Mutual Information (MI) and the classic (Pre-Compiled) approaches. They compared the results with the baseline model where the stop words were not removed.

Likewise, Pak and Paroubek (2010) performed linguistic analysis initially on the collected tweets. After preprocessing, feature extraction was done and the features were then used in the multinomial Naive Bayes classifier algorithm. Instead of taking the tweets as an whole, Kouloumpis et al. (2011) suggested taking only the hashtags present in the tweet and called the dataset to be the Hashtagged dataset. This was taken a step further above by Wang et Al. (2011) by not only taking the hashtags but by using the graph-based classification algorithms Loopy Belief Propagation (LBP), Relaxation Labeling (RL) and Iterative Classification Algorithm (ICA). The model was compared with the baseline where the traditional classification algorithms were taken.

Next, we wanted to do the analysis on specific applications. Gerber et al. (2012) used the twitter data and extracted event-based tweets using Semantic Role Labeling (SRL). From that, they used Latent Dirichlet Allocation (LDA) to identify the salient topics in the events. Using these topics, they built a predictive model that predicts future criminal occurrences. In a similar attempt to predict crimes, Chen et al. (2015) wanted to use weather as a feature in addition to the twitter sentiment to predict the time and location of a crime. They used Kernel Density Estimation (KDE) along with lexicon based methods with weather for the prediction. They observed that temperature was an important feature to influence high behaviors of aggressiveness and ultimate leading to an increase in crime rate. On the context of weather, "Hurricane Irene" was chosen as the topic and whether the tweet statistics were found for the particular topic by Mandel et. al (2012). They found out that the number of tweets related to the Hurricane directly affected the region peaks at the time of hurricane and that the level of concern was dependant on that particular region. In addition to crime and weather a lot of different areas were analyzed and one of them was flu and illness analyzed by Achrekar et al. (2011) who used the Social Network Enabled Flu Trends (SNEFT) to perform analysis and predict flu trends from the tweets. They performed correlation analysis and were able to narrow out the illness associated from the tweets. Twitter sentiments could also be used to analyse stock market trends as suggested by Mittal and Goel (2012). To classify the public sentiment, Opinion Finder and Google Profile of Mood States (GPOMS) algorithm were used.

We also wanted to touch on an important topic on whether the election results are influenced by the tweets. Wang et al. (2012) used a real-time data processing infrastructure on the IBM's InfoSphere Streams platform for writing the visualisation modules and analysis. Naive Bayes classification algorithm was used on the features provided. It displayed the trending words associated with each candidate by an AJAX based HTML dashboard. A special case study was done on the Irish General Election 2011 by Bermingham and Smeaton (2011). The uniqueness of this paper is that it differentiates the emotions used in Inter-Party polls and Intra-Party polls by having a separate measure. The results showed that during the weeks before election, the sentiments seemed to be pretty close to each other but on the day before the election, the sentiments were polarised. On the contrary, Gayo-Avello (2012) said that the election results cannot be predicted using twitter data.

He performed predictive modelling and arrived at the prediction that Barack Obama would be winning the Presidential Election in the US 2008 and that he would win all the states including Texas. But in reality, Obama did not win in the state of Texas proving his point that the tweet sentiments cannot entirely predict the outcome of an election.

Machine Learning was an important aspective we wanted to combine with the sentiment analysis. After performing Natural Language Processing (NLP) on the preprocessed tweets, they were divided into training set and test set and then Naive Bayes and Maximum Entropy algorithms were used in Machine Learning. They were compared and evaluated using the metrics of precision and recall. On a similar note, we found that by using Support Vector Machines (SVM) there is an higher accuracy and irrelevant data are filtered out as pointed by Neethu and Rajashree (2013). SVM was used to train the machine into calculating and predicting the sentiment scores of an electronic product. Machine Learning using the sentiment scores were carried out on movie reviews by Amolik et al. (2016). They used a unigram approach to preprocess the data and ensured hashtags were also removed and then stored in the feature vectors. They used both SVM and Naive Bayes to predict the sentiments of movie reviews. The scores were then compared to the baseline model on the metrics of precision and recall.

3 Dataset

The dataset we worked on was provided by altmetric.com. Altmetrics deal with the metrics of scholarly articles shared on online social media like Facebook, Twitter, Wikipedia and online reference managers like Mendeley. The data has details of research articles including but not limited to title of an article, author details, subject of the article, Tweets on Twitter and counts of shares of research articles on social media.

For the purpose of our project, we deal with the research articles shared on the social media 'Twitter'. We wanted to observe the sentiments of the tweets shared on Twitter. For this, we come up with a random sample of 150,000 research articles. This dataset was trimmed and all the values not available were discarded. The final dataset had XXXXXXX research articles details.

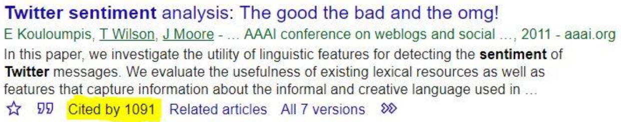
In order to predict the sentiment of tweets, we wanted to take into account those features that would aid in knowing how the Twitter users react and have emotion on their tweets related to research articles. Some of these features include:

4 Methodology

4.1 Automated Web Scraping

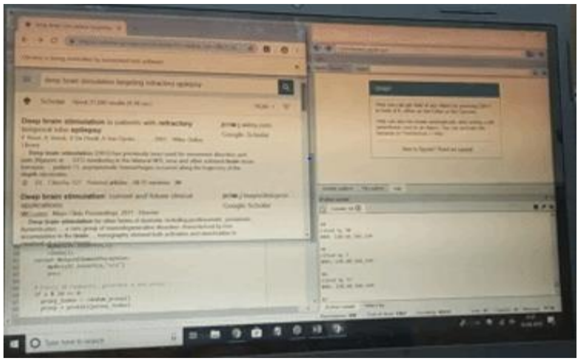
Initially, when we looked at all the different features found in the dataset, we noticed that there was an important feature called "Total Number of Citations" missing. The feature gives the number of times the paper has been cited by other papers. We estimated that this would be a valuable feature and wanted to make sure that we acquire it.

The traditional way to tabulate this feature would be to go to google scholar manually and then input the title name of the paper taken from the dataset for a particular column into the text-box and then search for it. In the next page, the value would be noted down and filled in the dataset. This process is called web-scraping. This has to be done manually for more than hundreds of data and this would require a lot of human effort and time.



Twitter sentiment analysis: The good the bad and the omg!
E Kouloumpis, T Wilson, J Moore - … AAAI conference on weblogs and social …, 2011 - aaai.org
In this paper, we investigate the utility of linguistic features for detecting the **sentiment** of **Twitter** messages. We evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in …
☆ 99 Cited by 1091 Related articles All 7 versions »

Considering the above difficulties, we wanted to automate the process and wanted the machine to be programmed in such a way that after a click of a button, the machine would take all the titles, search for it, extract information and then store it. For the automation, we imported the Selenium framework in Python and used the chromedriver for the web scraping. After the implementation, the program ran well for the first 25 entries but at the 26th entry, the program stopped working and we were sent a captcha by Google since we were sending a lot of requests in a short period of time from the same IP address.



The problem was Google assumed us to be a spambot since we were causing a lot of traffic from the same machine. To overcome this, we had to use proxy rotation and delayed scraping. First, we used BeautifulSoup to periodically extract 100 proxies from an online website every 60 seconds and store it. From the list of stored proxies, we setup a proxy rotation module where a proxy (IP Address and Port Number pair) is selected randomly out of the 100 proxies and used for scraping. Every 5 seconds, the proxy is deleted and another proxy is selected from the proxy pool. This is done for every 60 seconds after which the pool if refreshed by another set of proxies. We also had to sleep the program for a couple of seconds in various stages of execution. Finally, we had to use an exception handling to catch and handle the ElementNotFound instance and StaleElement instance.

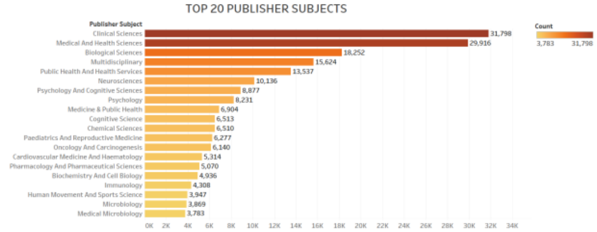
Using this method, we were able to acquire the total number of citations and then append it to our dataset. Initially, we calculated the values for a range of 1000 papers.

4.2 Data Analysis

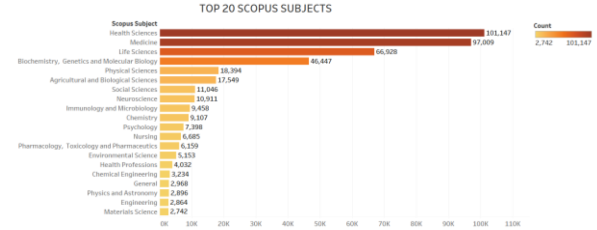
Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.mData Analysis library used was Pandas which is an open-source library. We used it to load, manipulate, analyze, and visualize cool datasets.

We extracted all the subjects ie general, publisher and scopus subjects from the data set so as to represent the trends and get an insight on the user's interest on Twitter.

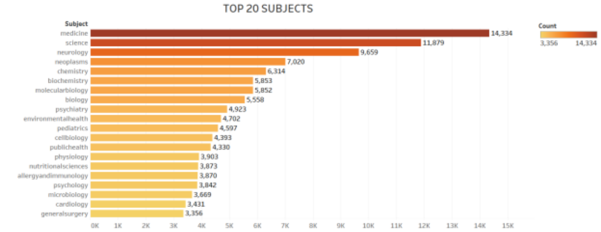
We obtained 30 scopus subjects, 2005 publisher subjects and 107 general subjects which we graphed using Tableau to show top 20 statistics.



Sum of Count for each Publisher Subject. Color shows sum of Count. The view is filtered on Publisher Subject, which has multiple members selected.

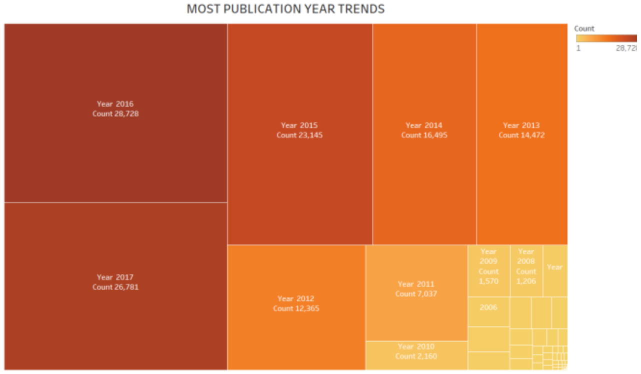


Sum of Count for each Scopus Subject. Color shows sum of Count. The view is filtered on Scopus Subject, which keeps 20 of 30 members.



Sum of Count for each Subject. Color shows sum of Count. The view is filtered on Subject, which keeps 20 of 107 members.

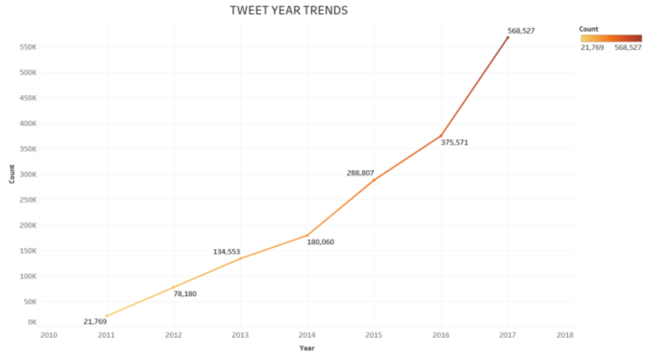
Similarly, for the publisher year trends and twitter year values, we fetched the data from the data set and obtained the year count in a similar approach.



Year and sum of Count. Color shows sum of Count. Size shows sum of Count. The marks are labeled by Year and sum of Count.

Here, in the picture above, year 2016 has the most number of the count which states the publication papers widely attained the twitter users interest.

Twitter year trends shows the increase of articles shared on Twitter over the years and the data is from 2011 to 2017.



The trend of sum of Count for Year. Color shows sum of Count.

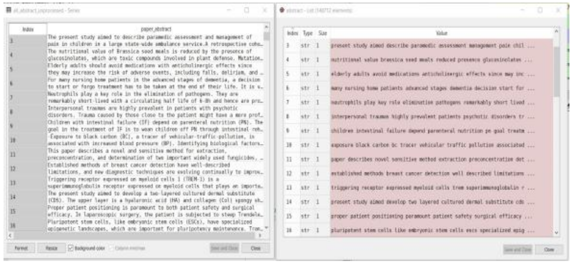
4.3 Feature Selection:

For building the machine learning models, we needed the sentiments of the texts in our dataset. The texts were in Title, Abstract and Tweets. These features were given a sentiment scores.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information.

Tweet being the target variable, we did the preprocessing on it by cleaning the tweets which was done by importing the class stopwords from NLTK library. The tweet was first converted to lower-case. All the links, hashtags and non english words were being removed and checked for any stop words in the tweet which was eliminated later from the final tweet. Once the tweets were cleaned and processed, we imported Sentiment Intensity Analyzer from NLTK library to evaluate the scores. The scores were compounded and later appended to the data set for the further machine learning procedure.

Likewise , for the title and abstract sentiment analysis, we followed the same protocol of cleaning the title and abstract and obtaining a cleaned, organized data and evaluated the scores which again were appended to the final dataset.



Index	paper_abstract
3	The present study aimed to describe paramedic assessment and management of pain in children in a large state-wide ambulance service.A retrospective coho…
4	The nutritional value of Brassica seed meals is reduced by the presence of glucosinolates, which are toxic compounds involved in plant defense. Mutation…
5	Elderly adults should avoid medications with anticholinergic effects since they may increase the risk of adverse events, including falls, delirium, and …
6	For many nursing home patients in the advanced stages of dementia, a decision to start or forgo treatment has to be taken at the end of their life. It is v…
7	Neutrophils play a key role in the elimination of pathogens. They are remarkably short-lived with a circulating half life of 6-8h and hence are pro…
8	Interpersonal traumas are highly prevalent in patients with psychotic disorders. Trauma caused by those close to the patient might have a more prof…
9	Children with intestinal failure (IF) depend on parenteral nutrition (PN). The goal in the treatment of IF is to wean children off PN through intestinal reh…
10	Exposure to black carbon (BC), a tracer of vehicular-traffic pollution, is associated with increased blood pressure (BP). Identifying biological factors…
11	This paper describes a novel and sensitive method for extraction, preconcentration, and determination of two important widely used fungicides, …
12	Established methods of breast cancer detection have well-described limitations, and new diagnostic techniques are evolving continually to improv…
13	Triggering receptor expressed on myeloid cells 1 (TREM-1) is a superimmunoglobulin receptor expressed on myeloid cells that plays an importa…
14	The present study aimed to develop a two-layered cultured dermal substitute (CDS). The upper layer is a hyaluronic acid (HA) and collagen (Col) spongy sh…
15	Proper patient positioning is paramount to both patient safety and surgical efficacy. In laparoscopic surgery, the patient is subjected to steep Trendele…
16	Pluripotent stem cells, like embryonic stem cells (ESCs), have specialized epigenetic landscapes, which are important for pluripotency maintenance. Tran…

Index	Type	Size	Value
3	str	1	present study aimed describe paramedic assessment management pain chil ...
4	str	1	nutritional value brassica seed meals reduced presence glucosinolates ...
5	str	1	elderly adults avoid medications anticholinergic effects since may inc ...
6	str	1	many nursing home patients advanced stages dementia decision start for ...
7	str	1	neutrophils play key role elimination pathogens remarkably short lived ...
8	str	1	interpersonal traumas highly prevalent patients psychotic disorders tr ...
9	str	1	children intestinal failure depend parenteral nutrition pn goal treatm ...
10	str	1	exposure black carbon bc tracer vehicular traffic pollution associated ...
11	str	1	paper describes novel sensitive method extraction preconcentration det ...
12	str	1	established methods breast cancer detection well described limitations ...
13	str	1	triggering receptor expressed myeloid cells trem superimmunoglobulin r ...
14	str	1	present study aimed develop two layered cultured dermal substitute cds ...
15	str	1	proper patient positioning paramount patient safety surgical efficacy ...
16	str	1	pluripotent stem cells like embryonic stem cells escs specialized epig ...

Above picture gives a before-and-after image of the data preprocessing on the abstract.

Additionally, we have 3 types of subjects: Subject, Scopus subject and Publisher subject. Since the unique subjects count for Scopus Subject was XXXXXX, which could be used as a decent number for categorical data, we wanted to use only Scopus subjects as a feature.

The final set of features are:

5.4 Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

We did not want to get into much detail on Logistic Regression as the recall was the lowest out of all the other classifiers (0.02). We wanted to upgrade Logistic Regression by concatenating it with SMOTE.

5.5 Logistic Regression[SMOTE]

With SMOTE the objective is to find a new balanced dataset which includes all the majority class examples and a synthetic over-sampled replica of the minority class examples, such that the new set is balanced [12]. This classifier had the highest recall of 0.46. Therefore, for this project, the chosen classification algorithm is Logistic Regression with SMOTE.

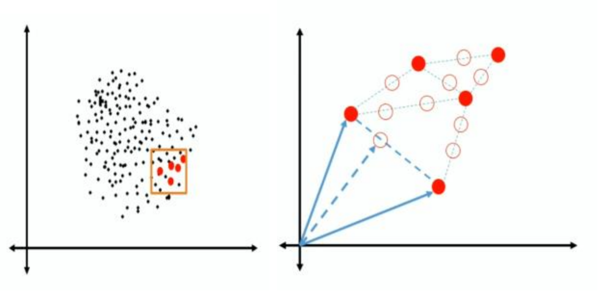


Figure 2: SMOTE

6 Metrics

Precision and Recall were used as the metrics for the classifiers. Machine learning algorithms were typically evaluated using predictive accuracy but this was not appropriate when the data was imbalanced and/or the costs of different errors very markedly [6]. Considering the above statement, we proceeded to take recall as the evaluation metric. The classifier with the highest recall was Logistic Regression with SMOTE and we decided to use that as our classifier.

7 Performance Measurement

Logistic regression model is a modeling procedure applied to model the response variable Y that is category based on one or more of the predictor variables X, whether it is a category or continuous [7]. Logistic Regression was used along with SMOTE to deal with the class imbalance problem.

7.1 Confusion Matrix

In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or a contingency table. The confusion matrix has four categories: True Positives (TP) are examples correctly labeled as positives. False Positives (FP) are examples referring to negative values labelled incorrectly as positive. True Negatives (TN) correspond to negatives correctly labelled as negative. Lastly, False Negatives (FN) refer to positive examples incorrectly labeled as negative [20]. The confusion matrix for our project is given below:

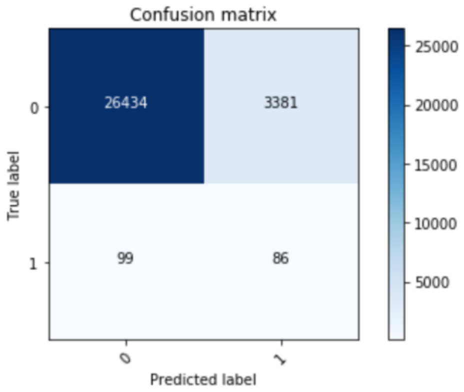


Figure 3: Confusion Matrix

7.2 ROC Curve

From the confusion matrix, we are able to plot the ROC curve. The False Positive Rate (FPR) is plotted on the x-axis and the True Positive Rate (TPR) is plotted on the y-axis. The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly labelled. The ROC curve is given below:

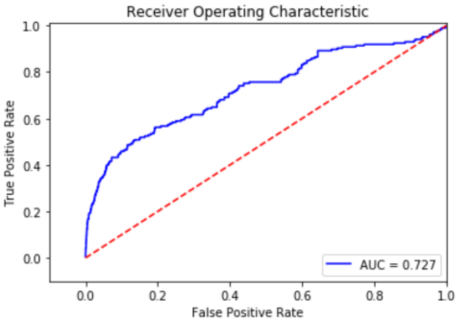


Figure 4: The ROC Curve

8 Future Work

Instead of taking only a portion of the Altmetrics dataset, the entire Altmetrics dataset could be taken as a whole. Though this would consume more time and space, the recall and precision could definitely vary. Also, under the different classifiers, Support Vector Machines (SVM) can be used to perform the classification. Different sampling techniques other than SMOTE could be experimented with for the class imbalance problem.

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using twitter data. In 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS) (pp. 702-707). IEEE.

Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, M. (2016). Twitter sentiment analysis of movie reviews using machine learning techniques. international Journal of Engineering and Technology, 7(6), 1-7.

Bae, Y., & Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. Journal of the American Society for Information Science and Technology, 63(12), 2521-2535.

Bermingham, A., & Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011) (pp. 2-10).

Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in twitter streaming data. In International conference on discovery science (pp. 1-15). Springer, Berlin, Heidelberg.

Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

Da Silva, N. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66, 170-179.

Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 437-442). IEEE.

Gayo-Avello, D. (2012). No, you cannot predict elections with Twitter. IEEE Internet Computing, 16(6), 91-94.

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L. E., & Hsu, M. C. (2011, October). Visual sentiment analysis on twitter data streams. In 2011 IEEE Conference on Visual Analytics Science and Technology (VAST) (pp. 277-278). IEEE.

Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In Fifth International AAAI conference on weblogs and social media.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). AcM.

Madnani, N. (2007). Getting started on natural language processing with Python. ACM Crossroads, 13(4), 5.

Mane, S. B., Sawant, Y., Kazi, S., & Shinde, V. (2014). Real time sentiment analysis of twitter data using hadoop. IJCSIT) International Journal of Computer Science and Information Technologies, 5(3), 3098-3100.

Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15.

Narr, S., Hulfenhaus, M., & Albayrak, S. (2012). Language-independent twitter sentiment analysis. Knowledge discovery and machine learning (KDML), LWA, 12-14.

Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

Parikh, R., & Movassate, M. (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report, 118.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Technologies (ICCCNT) (pp. 1-5). IEEE.

Twitter (2019) https://s22.q4cdn.com/826641620/files/doc_financials/2018/q4/Q4-2018-Selected-Company-Financials-and-Metrics.pdf

Vargiu, E., & Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artif. Intell. Research, 2(1), 44-54.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1031-1040). ACM.

Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in twitter. In Workshop on computational social science and the wisdom of crowds, nips (Vol. 104, No. 45, pp. 17599-601). Citeseer.