Imbalanced Twitter Sentiment Analysis using Minority Oversampling

Kushankur Ghosh
Department of Computer Science & Engineering
University of Engineering and Management, Kolkata
Kolkata, India
kush1999.kg@gmail.com

Arghasree Banerjee
Department of Computer Science & Engineering
University of Engineering and Management, Kolkata
Kolkata, India
banerjeearghasree@gmail.com

Sankhadeep Chatterjee
A.K.Choudhury School of Information Technology
University of Calcutta
Kolkata, India
chatterjeesankhadeep.cu@gmail.com

Soumya Sen
A.K.Choudhury School of Information Technology
University of Calcutta
Kolkata, India
iamsoumyasen@gmail.com

***Abstract*—Micro-Blogging platforms have become one of the popular medium which reflects opinion/sentiment of social events and entities. Machine learning based sentiment analyses have been proven to be successful in finding people's opinion using redundantly available data. However, current study has pointed out that the data being used to train such machine learning models could be highly imbalanced. In the current study live tweets from Twitter have been used to systematically study the effect of class imbalance problem in sentiment analysis. Minority oversampling method is employed here to manage the imbalanced class problem. Two well-known classifiers Support Vector Machine and Multinomial Naïve Bayes have been used for classifying tweets into positive or negative sentiment classes. Results have revealed that minority oversampling based methods can overcome the imbalanced class problem to a greater extent.**

Keywords—Sentiment Analysis, Class Imbalance, Oversampling, SVM, SMOTE

I. INTRODUCTION

In the recent years Micro-Blogging platforms have become one of the most important tools for anyone to share their opinion related to any ongoing sensational topic. These platforms are the key source of data which can be then be processed for opinion mining. Since 2009 [1], Twitter has become the most widely networked Micro-Blogging platform with an ongoing increment in the number of users, where the messages written by the user are known as tweets. In the year 2001 [2], Sentiment analysis became a popular approach to analyze various opinions of the mass regarding any particular topic on any particular platform which makes it easier to predict the future actions regarding that topic in favor of the public and thus, made its place in the field of natural language processing [3].Sentiment of a tweet can provide useful indicators for many purposes valuable for organizations and brands [4] and be classified into a positive and a negative category. In [5], Desai and Mehta depicted a detailed study on the procedure to classify twitter data and also showed us various machine learning techniques to carry out the analysis. Machine Learning models which were developed to solve real life problems, can itself encounter with various issues which are gradually being discovered by various researchers. The Class-Imbalance problem is one of the most fatal among them. The problem in binary classification occurs when the sizes of the classes differ greatly. If such a problem occurs then the classifier is biased towards the majority class as a result of which the minority class prediction [6] is very poor. A study done by Japkowicz and Stephen [7] showed that the complexity of the model increases with the increase of the degree of the Class-Imbalance problem and the effect of this problem on the classifier also increases with the decrease in the size of the training dataset. The importance of this problem is growing and has been identified as one of the 10 challenges of data mining [8].

The biasing due to the class imbalance could lead to decisions which hamper the prediction and accordingly wrong business decisions may be taken. Moreover the new perception that is created among the people is not identified initially where future business scopes lie. Therefore analyzing the class imbalance problem is very much useful for identifying future business opportunities.

In this paper we analyze this Class-Imbalance problem in the context of Sentiment Analysis. We have proposed a technique in order to undertake a perfect comparison between different classification algorithms applied on the dataset with tweets and sentiments. The paper is based on supervised learning method [23] and presents a detailed illustration of the performance deflection after resolving the Class-Imbalance problem for each algorithm.

II. PROPOSED WORK

The current study is focused on the imbalanced class problem encountered in the sentiment analysis of tweets using machine learning based methods. It has been found that the sentiment of tweets collected from Twitter on majority of keywords related to social events, people, organization is highly imbalanced. Thus, the machine learning based methods that use classifiers to classify tweets into different sentiment classes could become biased towards the majority class. In the current study, first twitter data is collected and labeled using [24]. Later the labeled dataset is used to train two well-known machine learning

model namely Support Vector Machine and Multinomial Naïve Bayes classifier. The performance is measured in terms of accuracy, precision, recall and f-measure. Thereafter, an oversampling technique called Synthetic Minority Oversampling Technique (SMOTE) has been used to balance the dataset and the same classifiers are again trained. The performance is measured and compared with the performance of classifiers which were trained earlier with imbalanced dataset.

A. Data Extraction

Live tweets are extracted and are stored in a single data-frame. A tweet may contain various irregular expressions and symbols which cannot help in determining the polarity of the statement but may result in wrong analysis. During the extraction process each and every tweet is cleaned to get rid of these symbols and expressions and then they are fitted into the data-frame to get a proper hazardless polarity of each statement. The polarity value of each tweet is calculated and is replaced by 1 for values greater than or equal to 0 and by -1 for values lesser than 0 representing positive and negative tweets. Figure 1 depicts the distribution of tweets of two polarities. The negative tweets have formed the minority class.

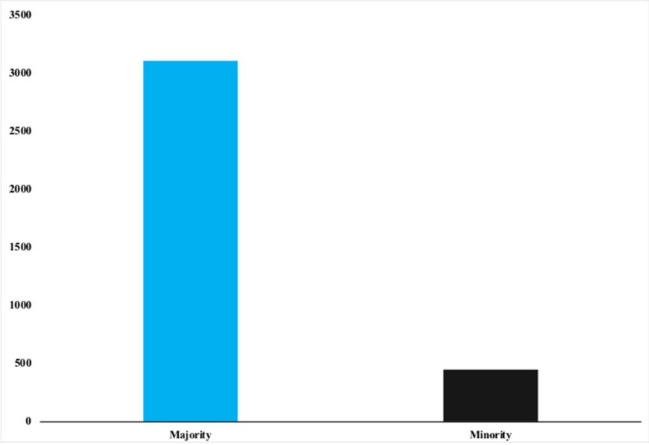

Figure 1. The imbalanced classes of the dataset under current study

B. Minority Oversampling

In order to remove the class imbalance problem, various methods have been proposed. The most popular among them is the oversampling technique. The approach is basically to oversample the minority class by introducing different ratios of synthetic samples [9]. In 1997, Kubat and Matwin [14] proposed an approach of under-sampling the majority class by keeping the minority class constant. Then in 1998, [15] the SHRINK technique was proposed which could classify the overlapping regions as the minority class. The research regarding oversampling of dataset basically started with the works of Ling and Lee in 1998 [16]. Oversampling by SMOTE [9] algorithm, resembles a duplication of minority samples by making the decision regions larger and more specific. Based on this algorithm, various algorithms have been proposed [9-13]. The main problem with the under-sampling technique was the data loss. In order to propose an efficient algorithm, the most crucial point is to undertake a minimum amount of data loss. Oversampling is the most efficient way to remove this problem.

C. Classification Algorithms

For our experiment, we have implied the Multinomial Naïve Bayes (MNB) and the Support Vector Machine (SVM) algorithms on our testing dataset. The basic structure of the SVM algorithm is based on Structural Risk Minimization Algorithm [18]. One of the most important features of the algorithm is its independent way to learn the dimensionality of the feature space [19]. The algorithm falls under the category supervised learning which performs classification on the basis of a *Hyperplane*. The *Hyperplane* or *Decision Boundary* is a line which divides the two-dimensional space into two parts where a particular class resides at each side of the line and makes the decision for any new data [20].

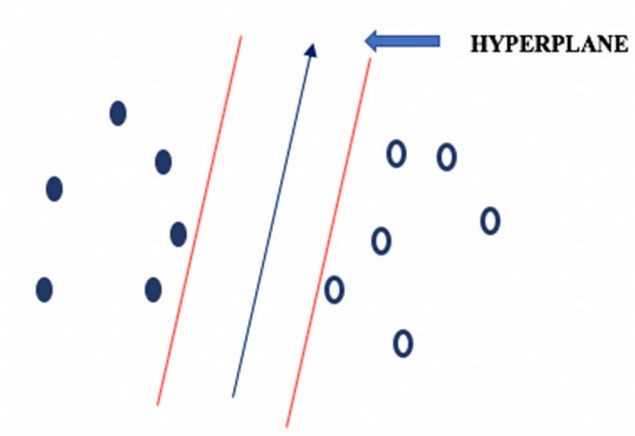

Figure 2. Pictorial representation of Hyperplane in SVM Algorithm

The MNB algorithm is one of the most popular techniques to undertake text classification. The MNB algorithm originates from the classic Naïve Bayes algorithm and was proposed by McCallum and Nigam (1998) [21] and can be used to understand the frequency of any particular word in any tweet document [21]. The working procedure of the algorithm is basically to calculate the probability of each class present in the whole given dataset. The algorithm is based on the following equation:

$$z_{mnb}(\emptyset) = \frac{\sigma}{\prod_{x=1} \alpha_x!} \prod\nolimits_{y=1} {z_y}^{\alpha_y}$$

such that, z_y is the probability of a word to occur in a text document and the feature vector $\emptyset = \alpha_1 \ldots . \alpha_n$. The α_x denotes the count of a word x in any tweet and σ corresponds to the factorial of the summation of all the outcomes α.

III. Experimental Analysis

The minority class of our dataset is oversampled by injecting synthetic values in our testing dataset. From Figure 3, the gradual oversampling of our dataset is visible with the constant rise of the minority samples.

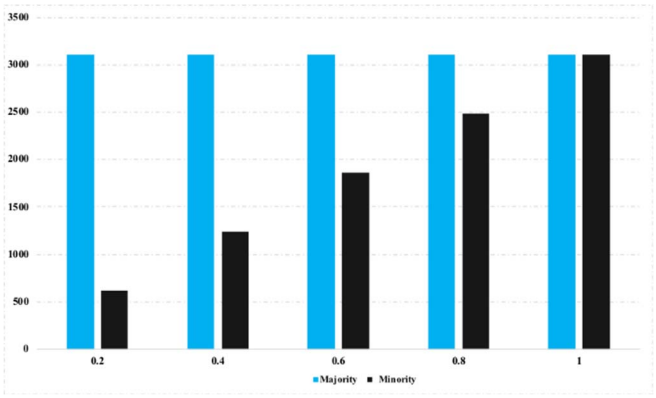

Figure 3. Pictorial representation of over sampling on training dataset

TABLE I. COMPARISON OF ACCURACY BETWEEN MNB AND SVM ALGORITHM FOR DIFFERENT RATIOS OF SYNTHETIC OVERSAMPLING

Algorithm	Without SMOTE	SMOTE ratio=0.2	SMOTE ratio=0.4	SMOTE ratio=0.6	SMOTE ratio=0.8	SMOTE ratio=1.0
MNB	0.86	0.88	0.88	0.85	0.84	0.79
SVM	0.85	0.85	0.85	0.85	0.85	0.86

TABLE II. COMPARISON OF PRECISION BETWEEN MNB AND SVM ALGORITHM FOR DIFFERENT RATIOS OF SYNTHETIC OVERSAMPLING

Algorithm	Without SMOTE	SMOTE ratio=0.2	SMOTE ratio=0.4	SMOTE ratio=0.6	SMOTE ratio=0.8	SMOTE ratio=1.0
MNB	0.88	0.9	0.88	0.85	0.85	0.84
SVM	0.73	0.73	0.73	0.73	0.73	0.85

TABLE III. COMPARISON OF RECALL BETWEEN MNB AND SVM ALGORITHM FOR DIFFERENT RATIOS OF SYNTHETIC OVERSAMPLING

Algorithm	Without SMOTE	SMOTE ratio=0.2	SMOTE ratio=0.4	SMOTE ratio=0.6	SMOTE ratio=0.8	SMOTE ratio=1.0
MNB	0.87	0.89	0.89	0.86	0.84	0.8
SVM	0.86	0.86	0.86	0.86	0.86	0.87

Table I tabulates the performance of MNB and SVM classifiers for different SMOTE oversampling ratio. In the current study, we start with no over sampling (Denoted by Without SMOTE in table) and gradually increase the oversampling ratio of minority class from 0.2 to 1. For each oversampling ratio, the performances of the classifiers have been observed. We see that using MNB algorithm the accuracy is the highest between smote ratio 0.2 and 0.4 whereas it starts dropping since the oversampling ratio of 0.4 and it reaches a minimum of 0.79 for a 100% oversampled minority class. For the SVM algorithm the accuracy remains constant throughout the experiment but experiences a minor increase for an oversampling ratio of 1.0. Figure 4(a) and 4(b) reveal that for 100% oversampling, the accuracy is least for the MNB algorithm and highest for SVM algorithm. It is observed that MNB experiences a constant fall for any synthetic ratio more than 40%.

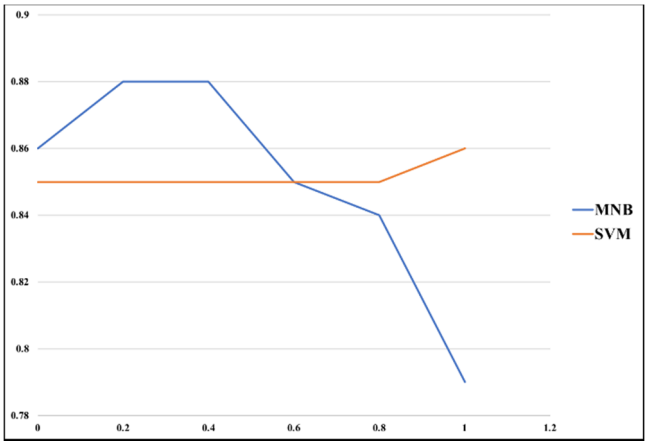

Figure 4(a). The line chart comparison between the smote ratios and the performance of the algorithms in terms of accuracy

Table II reports the precision of classifiers for different oversampling ratios. For MNB algorithm the precision lies between 0.80 and 0.90 while for SVM it is remains constant at 0.73 and it improves a little for an oversampling ratio of 1.0. Figure 5(a) and 5(b) reveals that for SVM the highest precision can be reached by a ratio of 1.0 and for MNB the highest can be reached by a smote ratio of 0.20.

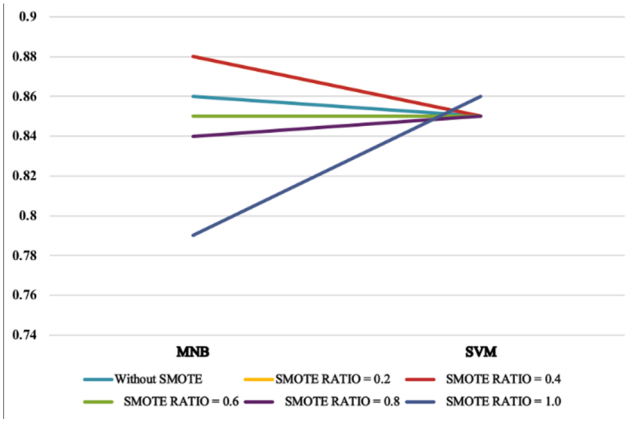

Figure 4(b). The line chart comparison performance of the algorithmsfor the smote ratios in terms of accuracy

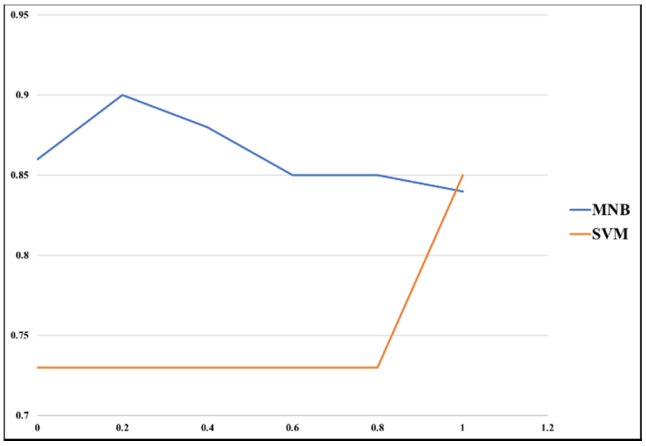

Figure 5(a). The line chart comparison between the smote ratios and the performance of the algorithms in terms of precision

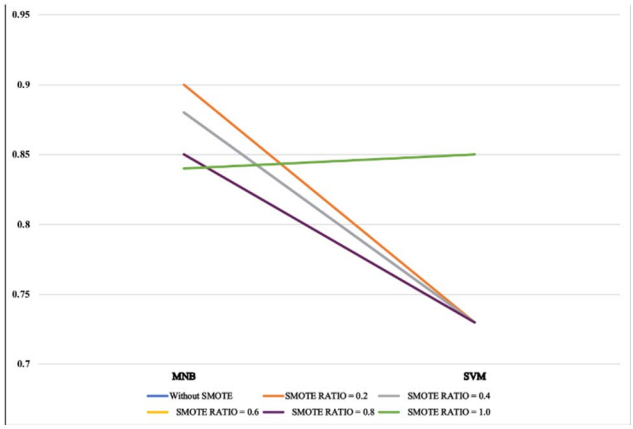

Figure 5(b). The line chart comparison performance of the algorithms for the smote ratios in terms of precision

Table III reports the recall of classifiers for different oversampling ratios. For the MNB algorithm, the recall is constant and the maximum for the oversampling ratio of 0.2 and 0.4 and decreases to 0.8 for a ratio of 1.0. For SVM algorithm it remains constant and increases to 0.87 for an oversampling ratio of 1.0. Figure 6(a) shows accurately that for MNB algorithm the lowest recall can be reached for 100% oversampling and the highest recall can be reached by the same ratio for SVM algorithm. Figure 6(b) reveals that the highest recall for MNB algorithm is achieved at an oversampling ratio of 0.4.

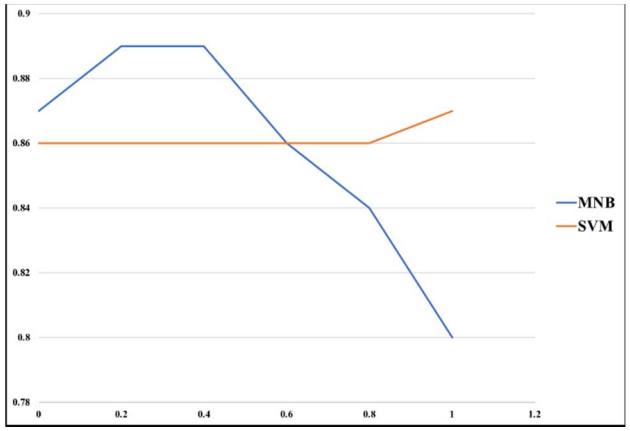

Figure 6(a). The line chart comparison between the smote ratios and the performance of the algorithms in terms of recall

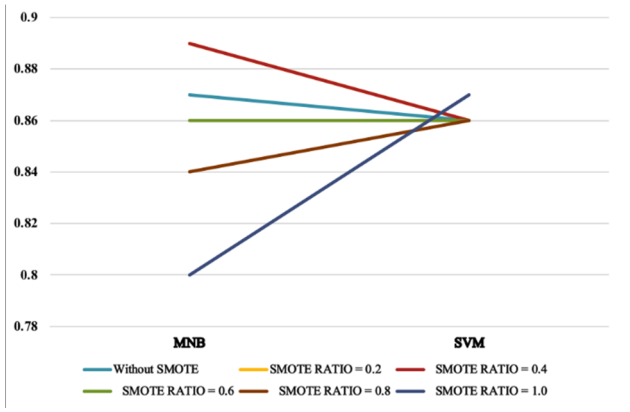

Figure 6(b). The line chart comparison performance of the algorithms for the smote ratios in terms of recall

CONCLUSION

The current work proposed a detailed study on the effect of imbalanced class problem observed in tweet-based sentiment analysis. The study compared well known machine learning classifiers and found that the minority oversampling based method significantly improved the classifier performance as revealed by precision and recall values. Further the ratio of oversampling in SMOTE has been varied with a range to identify a suitable oversampling ratio. Oversampling over 60% improved SVM's performance, however below 60% MNB performed better. Future study can include other machine learning models to establish the improvement performance further.

REFERENCES

[1] Liang, Y., S. Kang, and C. Zhang. "The effects of soil moisture and nutrients on cropland productivity in the highland area of the Loess Plateau." ACIAR MONOGRAPH SERIES 84 (2002): 187-194.

[2] H. Kwak, C. Lee, H. Park, S. Moon, "What is Twitter, a social network or a news media?", In Proceedings of the 19th international conference on World wide web, Pages 591-600, 2010

[3] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", In Foundations and Trends in Information Retrieval, Volume 2, Issue 1-2, Pages 1-135, 2008

[4] Lei Zhang, Shuai Wang, Bing Liu, "Deep Learning for Sentiment Analysis: A Survey", In Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 8, Issue 4, 2018

[5] M. Annett, and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," Conference on web search and web data mining (WSDM). University of Alberia: Department of Computing Science, 2009.

[6] Mitali Desai, Mayuri A. Mehta, "Techniques for sentiment analysis of Twitter data: A comprehensive survey", In 2016 International Conference on Computing, Communication and Automation (ICCCA), IEEE, Pages 149-154, 2016

[7] He, H., Ma, Y : Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley-IEEE Press (2013)

[8] Nathalie Japkowicz, S. Stephen, "The class imbalance problem: A systematic study1", In Journal of Intelligent Data Analysis, vol. 6, Issue. 5, pp. 429-449, 2002

[9] Q. Yang, X. Wu, "10 challenging problems in data mining research," Int. J. Inf. Tech. Decis., vol. 5, no. 4, pp. 597–604, 2006

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", In Journal of Artificial Intelligence Research 16, Pages 321–357, AI Access Foundation and Morgan Kaufmann Publishers, 2002

[11] Hui Han, Wen-Yuan Wang, Bing-Huan Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", In International Conference on Intelligent Computing, Springer, Pages 878-887, 2005

[12] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem", In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Pages 475-482, Springer, 2009

[13] Tomasz Maciejewski, Jerzy Stefanowski, "Local neighbourhood extension of SMOTE for mining imbalanced data", In 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Pages 104-111, IEEE, 2011

[14] Enislay Ramentol, Yailé Caballero, Rafael Bello, Francisco Herrera, "SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and under sampling for high imbalanced data-sets using SMOTE and rough sets theory", In Knowledge and Information Systems, Volume 33, Issue 2, Pages 245-265, Springer, 2011

[15] Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186 Nashville, Tennesse. Morgan Kaufmann.

[16] Kubat, M., Holte, R., & Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning, 30, 195–215.

[17] Ling, C., & Li, C. (1998). Data Mining for Direct Marketing Problems and Solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) New York, NY. AAAI Press.

[18] Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. Decision Analytics, 2(1).

[19] Vladimir N. Vapnik. The Nature o] Statistical Learning Theory. Springer, NewYork, 1995

[20] Joachims T. (1998) Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998.

[21] Rudy Prabowo, Mike Thelwall, "Sentiment Analysis: A Combined Approach", In Journal of Informatics, Volume 3, Issue 2, Pages 143-157, Elsevier, 2009

[22] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, "Tackling the Poor Assumptions of Naïve Bayes Text Classifiers", In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), ACM, pages 616-623, Washington DC, 2003.

[23] Ghiassi, M. and Lee, S., 2018. A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. Expert Systems with Applications, 106, pp.197-216.

[24] Loria, S., 2018. textblob Documentation (pp. 1-73). Technical report.