Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET 2022)
20-22 July 2022, Prague-Czech Republic

2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-7087-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECET55527.2022.9872849

Social Network Influence: Making the Case for Semantic Analysis

David Oriedi
Department of Computer Science Dedan Kimathi University of Technology Nyeri, Kenya
david.opondo@dkut.ac.ke

George Musumba
Department of Computer Science Dedan Kimathi University of Technology Nyeri, Kenya
george.musumba@dkut.ac.ke

Morris Kaburu
Department of Computer Science Dedan Kimathi University of Technology Nyeri, Kenya
morris.kaburu@dkut.ac.ke

James Iraya
Department of Computer Science Dedan Kimathi University of Technology Nyeri, Kenya
james.iraya@dkut.ac.ke

David Gichohi
Department of Computer Science Dedan Kimathi University of Technology Nyeri, Kenya
david.gichohi@dkut.ac.ke

Sammy Kamundi
Department of Computer Science Dedan Kimathi University of Technology Nyeri, Kenya
sammy.kamundi@dkut.ac.ke

***Abstract*—Influence on the social network platform has become an interesting research area. A typical social network influencer is known by the amount of reactions that he/she attracts based on the posts made. The amount of reactions received is usually a reflection of the overall visibility of the influencer on the social network arena. This visibility has attracted commercial benefits to such influencers through word of mouth marketing, political endorsements as well as ambassadorial appointments. However, existing methods in literature that are used to scientifically quantify the amount of social influence that can be attributed to an influencer, tend to lump the amount of social influence in a single basket labelled either as positive or negative influence. The effect is a binary classification of influencers as either negative or positive. In this article, we make the case that this is not necessarily accurate. The reason being that in a collection of comments that a post from an influencer attracts, it it seldom that all the comments would be absolutely positive or negative. There is a fuzzy space defined by comments that do not necessarily belong to either of these categories.**

In this work, we have used data harvested from Facebook comments and grouped into four different categories - those in full support of the opinion expressed, those in full opposition of the opinion, those that somewhat support and those that somewhat oppose the opinion expressed. Results have shown that the fuzzy space created by the comments in between the full support and full opposition negates the assumption that an influencer can be a positive or negative influencer just because the number of positive or negative comments are the majority.

***Index Terms*—Social Influence, Influencer, Fuzzy Space, Multiclass Classifier, Social Network**

I. Introduction

Recently, millions of people are increasingly interconnected through online platforms, such as Twitter, Facebook and Google, resulting in huge volumes of data generated from these networks. Researchers in big data research have shown a lot of interest in social network analysis with a view to getting knowledge and trends through the analysis of opinions and sentiment, as well as the consequent influence of actors in a social network. Among the issues of research in this regard are multi-objective optimization (maximization or minimization) of influence and its application for brand marketing or brand campaigns. Social network applications have provided a unique way for people all over the world to socially interact and exchange opinions over different subject matters. In particular, this interaction takes place through mechanisms of social actions such as likes, comments and sharing in Facebook or replying and retweeting in the Twitter micro blogging social media application. Invariably, influence on the social network platform is associated with how much reaction in form of social actions a user attracts based on the post made. Therefore, a user whose post attracts numerous reactions in likes, comments or shares is viewed as being influential while users who do not attract as much by their posts may be considered non influential. The question for social network analysis researchers have been whether the influence attributed to a user under these circumstances can be exclusively categorized as either positive or negative. Indeed there is evidence in literature of works concentrating on extracting either positive, negative or both types of influence on the social network platform [1], [2]. This categorization is based on the assumption that when a majority of the reactions are positive, then overally the owner of the post is associated with positive influence and vice versa for negative influence. However it is also a true observation that such influence may not always be absolutely positive or negative [3]. In other words, in between the obviously positive or negative reactions, there exists a fuzzy space of linguistic variables that are neither absolutely positive nor absolutely negative. In this paper, we make the claim that this fuzzy space of linguistic variables accounts for a certain proportion of alternative view that deserves to be paid attention to. For example, consider an individual who makes a post that attracts over 10,000 comments. Hypothetically, if about 8,000 of these comments are positively in support of the post made, it would be reasonable to assume that the owner of the post - or in some cases the post itself - has exhibited positive influence. This is because a majority of the comments will have expressed positive concurrence with the post made. It would also be possible to identify comments that are obviously in opposition to the view expressed in the post and therefore should be easily labelled as expressing contrary feelings to the content of the post. However, there will be some comments that will neither be explicitly positive nor negative. Our work is interested in this category of comments. This is because they are usually unaccounted for and yet there is a sense in which they dilute the otherwise positive or negative influence attributed to an influencer. In this paper, we have used data from Facebook comments. We have categorized the comments into four different categories namely those in full support of the post, those showing some support to the post, those completely opposed to the post and those showing some opposition to the post. We have then trained a multiclass classifier with the intention of going through the comments and putting them in the respective pre-labelled categories. In this way, we have been able to account for the comments that fall within the fuzzy set and in particular we have been able to bring to the fore interesting insights that can be used to initiate debate on whether conventional methods of computing social influence should be considering the fuzzy component of the attributes used to approximate the amount of social influence attributed to a particular user. As our contribution to the body of knowledge, we have shown the significance of incorporating the fuzzy category of comments in the computation of social influence for online influencers. This has been done through successfully training our model to classify not only the positive and the negative comments but also those comments that are not explicitly positive or negative into the right categories. In addition, the outcome of our experiment has shown that natural language processing should be a major part of analysing especially textual interactions on the social network platforms in order to get a near accurate of what constitutes positive or negative influence. This is majorly because of the contextual, figurative and metaphorical nature of written speech which some classification algorithms may not accurately categorize and yet should be considered in calculating social influence. To the best of our knowledge, no work exists that has investigated the identification of fuzzy sets of comments and their impact in the overall approximation of social influence that is attributed to an individual on the social online space.

The rest of this paper is organized as follows: Section I provides an introduction to the paper and Section II summarizes related works. In Section III, we detail the description and the working of our model while experiments and results are discussed in Section IV. Finally the paper concludes in Section V with suggestions on open issues for future research.

II. Related Works

The subject of text classification is widely researched and there are numerous text classification algorithms that can be applied when analysing the textual content of documents. However, the application of text categorization algorithms in social network conversations is not as widespread. In view of this fact, this section of the article will address two aspects of related work that is relevant to the theme of this work. Firstly we will review works that have investigated the computation of social influence with a bias to user interactions that involve exchange via text. In this case we are interested in finding out if there are works that have considered fuzzy set segment of social actions in calculating social influence. Secondly we will review works that have applied text classification algorithms with a view to identifying a meeting point for both the text classification algorithm and parsing of textual information that is exchanged among users while interacting online.

A majority of research works dedicated to the investigation of social influence tend to be thematic. This means that most authors set out to investigate influence in a specific area or as influenced by a particular need. For example, in [4], where the target of the investigation is on sentiment analysis. The need in this case is to establish the overall feeling of clients with respect to a service provider.Analysis of sentiment can reveal patterns of negative or positive online user behaviour and therefore help in the detection of malicious or terrorist activities [5]. Determination of opinion leaders [6], [7] involves identifying users on the social network that have substantial following on a subject matter. Such people can be experts in a particular field, religious leaders or just politicians. Opinion leaders can easily sway public opinion and therefore promote national dialogue in a particular direction [8]. Kempe $et\ al.$ [9] identified two major information diffusion models, namely, the Independent Cascade Model and the Linear Threshold Model. The Independent Cascade Model traces information propagation as it proceeds from one node to another in the social network without semantically analysing conversations between the users. In the Linear Threshold Model, a node is influenced if a minimum threshold of its neighbors have already acted upon the information being passed, and thus is indirectly influenced. Again, this happens without any reference to the content of the exchanges. In fact, estimation of the strength of user relationships in this model is determined through probabilistic estimation. Azaouzi and Lofti [10] proposed a model that estimates the influence of a user by putting together the likes, replies and retweets that the post of a user attracts. However, they did not incorporate any semantic analysis to determine whether the said influence would be positive, negative or indeterminate through analysis of the contents exchanged. In other works like [11], [12] and [13], determination of influential users has been done through centrality measures. While widely applied in identifying influential users on social network, centrality measures generally do not involve an analysis of the content of the information exchanged among users. Li $et\ al.$ [14]

used a combination of retweets, comments, mentions and keyword similarity to model node to node relationships. While using social actions as a basis for the definition of node influence, the work does not include any form of analysis of the text either for classification or interpretation of meaning. There are also research works that emphasised on semantically analysing content that define user relationships. These include recommender systems [15], [16], inferring user interests based on followees [17] or modelling of topics [18]. However, none of these works have addressed the case for fuzzy set of respective influence metrics.

Text classification algorithms fundamentally classify text into different labels and there are many options to go for. While the Recurrent Neural Network (RNN) algorithm is suitable for short text [19] most of which is found on online interactions, it does not parallelize well. In addition, it takes longer time to train on long text [20]. On the other hand, the Conventional Neural Network(CNN) algorithm parallels very well and is suitable for both long text and different filter of of text. However, both of these algorithms rate comparatively lower on accuracy compared to the naive Bayes algorithm [21]. In addition, naive Bayes algorithm has a faster learning process because its classification model can make a single pass over training documents [20] and is less sensitive compared to other classification algorithms. Finally naive Bayes comes with the advantage of incremental update of the model due to its simplicity [22]. For these reasons we found naive Bayes classifier more appropriate and adaptable for our model. The gap in the literature is that there seems not to be evidence of classification comments in particular with a view to grouping them into positive, negative or other labels with a view to providing a new definition for user influence. This is the gap this work intends to address.

III. Our Model

A. Model Description

Our model is a text classification model with a bias to text that is exchanged among users as they interact on social networks. In particular, the model is designed to help in accounting for Facebook comment reactions to a post that do not necessarily fall into clear positive or negative categories and . For reasons that have already been highlighted in Section II, this model uses the naive Bayes classification algorithm to achieve this objective. Naive Bayes classifier uses bayes theorem with some assumptions on conditional independence for features. Naive Bayes classifier uses probability to place a document to a target class with a maximum posterior probability. For purposes of generalization in line with the definition of the naive bayes classifier, each Facebook comment will be treated as a document. For a document D and a target class $c_i \in C$ with C being a set of target classes, naive bayes classifier is defined as:

$$\hat{y} = \underset{c_i \in C}{\arg\max} P(c_i|D) \tag{1}$$

TABLE I:
Confusion Matrix

Actual	**Predicted**	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative

where $P(c_i|D)$ is the posterior probability class c_i given document D. According to Bayes Theorem, we can express $P(c_i|D)$ as:

$$P(c_i|D) = \frac{P(D|c_i)P(c_i)}{P(D)} \tag{2}$$

Since by law of commutivity, $P(D|c_i)P(c_i) = P(c_i)P(D|c_i))$, and the independent nature of events involved means that the denominator eventually cancels out, equation 2 can also be written as:

$$P(c_i|D) = P(D|c_i)P(c_i) \tag{3}$$

The prior probability $P(c_i)$ represents the proportion of total instances of documents in class c_i with respect to total instances in the dataset. The document D is represented as a bag-of-word. The document D is therefore represented as a collection of words $w_1, w_2, w_3, ..., w_n$ with which the classifier will determine the relative frequency of a word in the document and consequently decide which class the document belongs. Because of the iterative computation of the posterior probability $P(D|c_i)$ involving the bag-of-word, this probability can be written as as shown in equation 4 to reflect the process of looping through the items in the bag of word. In order to avoid zero probability terms, Laplace's rule of succession [23] is applied to add a smoothing to $P(w_i|c_i)$ as shown in equation 5. Laplace's smoothing adds a pseudo count to every word count [24]. And so we have equations 4 and 5 respectively representing iteration through the bag of words and Laplace's smoothing included;

$$P(D|c_i) = \prod_{w_i \in D} P(w_i|c_i) \tag{4}$$

$$P(w_i|c_i) = \frac{count(w_i, c) + 1}{count(c_i) + |V|} \tag{5}$$

where $|V|$ is the size of the vocabulary, $count(w_i, c_i)$ represents the total term count in target class c_i, and $count(c_i)$ is the total terms in target class c_i.

B. Model Performance Metrics

Like most classification models, the performance of this model will be measured through its precision, recall and sensitivity. All these measures have their basis on the Confusion Matrix [25]. The Confusion Matrix is shown in Table I.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FB} \tag{6}$$

TABLE II:
DATA SUMMARY

Data Source/Label	Size	0	1	2	3
BBC Data	2225 News Items	Business	Entertainment	Politics	Sports
Facebook Data	2,000 comments	Fully Oppose	Fully Support	Some Opposition	Some Support

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (8)$$

IV. EXPERIMENT AND RESULTS

A. Experimental Data

For the experiment, we crawled 1666 Facebook comments attributed to a Facebook post[1] that President Barack Obama made about Russia-Ukraine war. At the time of crawling the comments, the post had attracted about 2,000 comments. Even though the majority of the comments were in support of the opinion expressed in the post, there were also comments that were completely opposed to his view about the war. In addition, we identified comments that we felt were somehow in support of his opinion but the support was not obvious. In the same way, we identified some comments that were somewhat in opposition to the opinion he expressed. Overally, we placed the crawled comments into four different categories namely Fully Oppose, Fully Support, Some Opposition and Some Support. Of the 1,666 comments crawled, 1249 have been used for training the classification model while 417 have been used to test the model.

B. Ground Truth Data

Machine learning models are best tested with ground truth data in order to evaluate their performance. Accordingly, we used one of popular ground truth data from the *BBC* news dataset [2] that has been used in other research works like in [26] The news groups in this dataset are *business, entertainment, politics, sport* and *technology*. The labels in this news dataset are clear since the nature of news, especially from reputable sources like *BBC*, usually have very little or no ambiguity unlike Facebook comments in response to a post which could bring ambiguities resulting from satire, metaphors or figurative speech. This is the main reason why we chose the *BBC* dataset for use as our ground truth data. Of the 2,225 total news items, we used 1,668 news items for training and the rest for testing the model. Table II shows a summary of both of these datasets.

C. Discussion of Results

The experiments were run on an HP Laptop with 16GB of RAM, an Intel Core i7 processor with a speed of 2.10GHz. The program was developed in Python through the Jupyter Notebook IDE. The metrics of model performance are precision, recall and f1-score. As can be seen in Table III, running the model on ground truth data (*BBC Dataset*) shows that

[1]https://www.facebook.com/barackobama/posts/510415667112523

[2]https://mlg.ucd.ie/files/datasets/bbc-fulltext.zip

TABLE III:
RESULTS FROM BBC DATA

Label	Precision	Recall	F1-Score
0	0.98	0.97	0.98
1	0.99	0.95	0.97
2	0.94	0.98	0.96
3	0.99	1.00	1.00

TABLE IV
RESULTS FROM FACEBOOK DATA

Label	Precision	Recall	F1-Score
0	0.53	0.63	0.58
1	0.59	0.81	0.68
2	0.37	0.24	0.29
3	0.27	0.18	0.22

the model is overally very accurate (97%) and performs very well on all metrics. We attribute this performance to the fact that news items, by definition, are very predictable. In other words, it is rare that a given news category will differ from the content of that news category especially if the news is from a reliable source like the *BBC*. The model therefore has no problem getting the classification right.

The model run on Facebook data as shown in Table , however shows slightly different results. The classification of the comments into the positive and negative labels had better outcome with respect to the metrics (above 50%) compared to those comments classified into the fuzzy sets whose performance outcome is below 50%. In particular, it is notable that in the labels that were not explicitly positive or negative i.e *Some Opposition* and *Some Support*, the model struggled to place the comments thereby scoring below 40% in all metrics. We attribute this poor performance to the largely satirical and metaphorical nature of comments in these categories of comments. Satire and metaphor in the English language have been known to cause confusion in everyday speech especially in cases where English is spoken as a second language [27] and in some cases making the message being communicated incomprehensible [28] and therefore incomplete. Moreover, a lot of social network users tend to employ unstructured short form style of communication [29] that a classification algorithm could easily misplace. We believe that these challenges have contributed in a big way in the under performance of the model when classifying fuzzy comment sets and by extension the overall accuracy score of the model which is only 48% against 98% that the model achieves with the ground truth data. Furthermore, the macro average as well as the weighted average is at 98% on the BBC data for the same reasons against very low values of below 40% on those same metrics on the Facebook data. Fig 1 shows a summary of results from the model. Because of the reasons already cited, the accuracy of our model is low. Even though there should be room for addressing classification dataset imbalance using techniques like Synthetic Minority Over-Sampling Technique (SMOTE) [30], the need to address the imbalance has not been considered here because there was a deliberate attempt when harvesting the data from Facebook to ensure that there was very little or no imbalance among the four identified labels of data. This was largely influenced more by the need to demonstrate the existence of the fuzzy linguistic set rather than the accuracy of the classification model. The reader should therefore remember that the main objective of this work was to show, with empirical evidence, that there exists a segment of comments from social network interactions that neither explicitly supports nor disagrees with the discourse of interest, and not to test how close our output is with reference to results from the ground truth data.

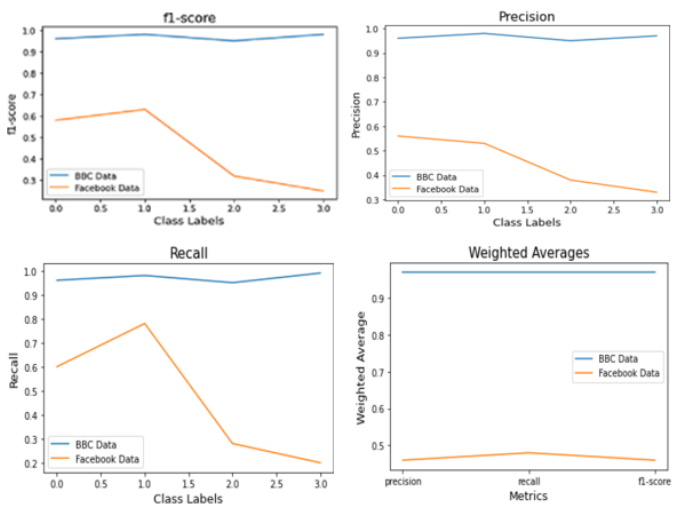

Fig. 1. Summary of Results

The main objective of this work is, therefore, to make the case for the need to do semantic analysis particularly when computing influence based on the number of comments that a user attracts based on a post made. We argue that the fuzzy comment set, however small, need to be handled separately in order to arrive at the true value of social influence that is due to an influencer. The output of this experiment has shown that the fuzzy set not only exists but also accounts for a portion of the total reactions that an influencer attracts after making a post on a social network platform.

V. Conclusion and Recommendations

In this work, our objective was to investigate the existence of what we are describing as fuzzy comments. These are comments that users make in response to an online post made by some other user. However, while most comments easily come across as being in support of the post made or against, the fuzzy category of comments are not easy to categorize. This is because they need further interpretation as a result of the possibility of involving satire, use of metaphors or just being unstructured and mixed -lingual. Our argument is that this category of comments should not be assumed but should be given own standing and their contribution - or lack thereof - to the overall influence power of a user online should be separately computed.

We have performed experiments that show that this category of comments is indeed in existence and our classifier has shown that for them to be correctly classified in appropriate labels, there may be need to improve the traditional text classification algorithms. Our results provide a good basis for including the fuzzy comment set in calculating social influence especially when using techniques that involve the use of user comments.

For further research, future investigations should consider developing an appropriate classifier for the social media comments that are usually unstructured, shortened, noissy and in some cases mixed-lingual. Successfull classification of such comments within the right context will be a good step in applying semantics in computing social influence for influencers.

References

[1] M. Talluri, H. Kaur, and J. S. He, "Influence maximization in social networks: Considering both positive and negative relationships," in *2015 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 479–480, IEEE, 2015.

[2] F. N. Abu-Khzam and K. Lamaa, "Efficient heuristic algorithms for positive-influence dominating set in social networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 610–615, IEEE, 2018.

[3] F. Belbachir and B. Le Grand, "Opinion detection: Influence factors," in *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*, pp. 522–523, IEEE, 2015.

[4] D. Li, X. Shuai, G. Sun, J. Tang, Y. Ding, and Z. Luo, "Mining topic-level opinion influence in microblog," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1562–1566, 2012.

[5] W. Yang, H. Wang, and Y. Yao, "An immunization strategy for social network worms based on network vertex influence," *China Communications*, vol. 12, no. 7, pp. 154–166, 2015.

[6] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 971–974, 2007.

[7] X. Tang and C. C. Yang, "Ranking user influence in healthcare social media," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, pp. 1–21, 2012.

[8] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: A survey," *Journal of Network and Computer Applications*, vol. 106, pp. 17–32, 2018.

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, 2003.

[10] M. Azaouzi and L. B. Romdhane, "An efficient two-phase model for computing influential nodes in social networks using social actions," *Journal of Computer Science and Technology*, vol. 33, no. 2, pp. 286–304, 2018.

[11] M. Kardara, G. Papadakis, A. Papaoikonomou, K. Tserpes, and T. Varvarigou, "Large-scale evaluation framework for local influence theories in twitter," *Information processing & management*, vol. 51, no. 1, pp. 226–252, 2015.

[12] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, "Modeling information diffusion over social networks for temporal dynamic prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1985–1997, 2017.

[13] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou, "Ranking in evolving complex networks," *Physics Reports*, vol. 689, pp. 1–54, 2017.

[14] X. Li, S. Cheng, W. Chen, and F. Jiang, "Novel user influence measurement based on user interaction in microblog," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 615–619, ACM, 2013.

[15] B. Deb, I. Mukherjee, S. N. Srirama, and E. Vainikko, "A semantic followee recommender in twitter using topicmodel and kalman filter," in *2016 12th IEEE International Conference on Control and Automation (ICCA)*, pp. 649–656, IEEE, 2016.

[16] D. P. Karidi, "From user graph to topics graph: Towards twitter followee recommendation based on knowledge graphs," in *2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)*, pp. 121–123, IEEE, 2016.

[17] C. Besel, J. Schlötterer, and M. Granitzer, "On the quality of semantic interest profiles for onine social network consumers," *ACM SIGAPP Applied Computing Review*, vol. 16, no. 3, pp. 5–14, 2016.

[18] K. Slabbekoorn, T. Noro, and T. Tokuda, "Ontology-assisted discovery of hierarchical topic clusters on the social web," *Journal of Web Engineering*, pp. 361–396, 2016.

[19] A. Bhavani and B. S. Kumar, "A review of state art of text classification algorithms," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1484–1490, IEEE, 2021.

[20] H.-j. Kim, J. Kim, and J. Kim, "Semantic text classification with tensor space model-based naïve bayes," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 004206–004210, IEEE, 2016.

[21] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.

[22] H.-j. Kim, K.-j. Hong, and J. Y. Chang, "Semantically enriching text representation model for document clustering," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 922–925, 2015.

[23] P. Kantor, "Foundations of statistical natural language processing," *Information Retrieval*, vol. 4, no. 1, p. 80, 2001.

[24] M. Poyraz, Z. H. Kilimci, and M. C. Ganiz, "A novel semantic smoothing method based on higher order paths for text classification," in *2012 IEEE 12th International Conference on Data Mining*, pp. 615–624, IEEE, 2012.

[25] S. Aggarwal and D. Kaur, "Naive bayes classifier with various smoothing techniques for text documents," *International Journal of Computer Trends and Technology*, vol. 4, no. 4, pp. 873–876, 2013.

[26] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on tf-idf and support vector machine for news classification," in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pp. 112–116, IEEE, 2016.

[27] J. H. Leigh, "The use of figures of speech in print ad headlines," *Journal of advertising*, vol. 23, no. 2, pp. 17–33, 1994.

[28] R. J. Kreuz and R. M. Roberts, "The empirical study of figurative language in literature," *Poetics*, vol. 22, no. 1-2, pp. 151–169, 1993.

[29] N. Fabian, W. Davis, E. M. Raybourn, K. Lakkaraju, and J. Whetzel, "Grandmaster: Interactive text-based analytics of social media," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1375–1381, IEEE, 2015.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.