Sentiment Analysis, Visualization and Classification of Summarized News Articles: A Novel Approach

Siddhaling Urologin
Department of Computer Science
Birla Institute of Technology and Science, Pilani-Dubai,
Dubai, U.A.E.

***Abstract*—Due to advancement in technology, enormous amount of data is generated every day. One of the main challenges of large amount of data is user overloaded with huge volume of data. Hence effective methods are highly required to help user to comprehend large amount of data. This research work proposes effective methods to extract and represent the data. The summarization is applicable to obtain a brief overview of the text and sentiment analysis can obtain emotions expressed in the text computationally. The combined text summarization and sentiment analysis is proposed on BBC news articles. A pronoun replacement based text summarization method is developed and VADER sentiment analyzer is used to determine sentiment information. The 3-D visualization schemes have been provided to represent the sentiment information. The sentiment analysis and classification are performed on original BBC news articles as well as on summarized articles using classifiers, such as Logistic Regression, Random Forest and Adaboost. On original news articles highest classification rate of 84.93%, using summarization of ratio 25%, 50% and 75% highest classification rates of 78.73%, 83.06% and 83.23%, respectively are observed.**

***Keywords*—Summarization; sentiment analysis; 3-D visualization; sentiment classification**

I. Introduction

Huge amount of data is being generated every day in the form of social media data, various blogs, web sites, Wikipedia, online newspapers, etc. Due to wide spread usages of social media such as Facebook, Twitter, Yahoo! etc. have enormously increased the amount data that has been produced. The Wikipedia alone contain five million articles and thousands of new articles generated every day. There are various online web sites which are publishing newspapers on daily basis. One of the main challenges of huge data is that user gets over loaded with data and requires effective way to absorb the large volume amount of data. Effective data extraction and representation techniques are needed to help user to comprehend huge data. The text summarization is the technique intended to produce a brief overview of the input text and also reduces the amount of data. Moreover, the sentiment analysis is the computational technique, which deduce the user emotion expressed in the text. The sentiment analysis been effective applied to various fields such as product reviews [1], [2], news articles [3], political debate [4], twitter sentiment data analysis [5], [6], stock market [7], [8], etc.

The goal of the text summarization is to obtain a brief summary of the text [9]. This method of text summarizing can be utilized in different applications namely searching documents related to a particular subject and obtain an overview, gather headline from newspaper articles, assimilate emails, obtain summary of medical information, to produce brief of scientific articles [10], [11] etc. There are various steps involved in text summarization such as topic identification, interpretation and summary generation [12]. A notable work by Bennostein et al. on topic identification is presented in [13] with a frame work for topic identification and applications. Work on Wikipedia graph centrality method for topic identification is presented in [14]. During text interpretation, the meaning of the text is obtained. For text interpretation researchers have focused on various methods such as ontology based interpretation [15] and text interpretation [16]. The goal of text summarization is to generate an abstract or synopsis on single or multiple documents. J. Alan et al. [11] have presented a text summarization method based on novelty detection at the sentence level. Literature review presented in [17] by Lloret et al. have noted that there are two summarization methods: abstraction and extraction. Semantic representation are constructed from text to produce a brief overview in abstraction method [13], [18]. The extractive summarization methods discussed as in [19]–[21] are intend to choose words, sentences and phrases from the given text to obtain the summary. Forming summary based on frequency of words related to the topic has found suitable application in several area [22], [23] of text analysis. It is observed that in a given document the words that are occurring more frequently indicates the subject on which the text is pivoted. Rafael Ferreira et al. [24] have accessed the sentence scoring technique for text summarization. In their work, it was noted that obtaining the frequency of important words and extract sentence to prepare the summary is one of the effective methods. Pronouns are place the holders for proper nouns, which are often used in the text. In the process of filtering and stopword removing, pronouns are also eliminated affecting the frequency of proper nouns. In this research work the summarizing technique is proposed in which, pronouns are replaced at first with proper nouns and then the frequency of words are computed, thereby enhancing the frequency information related to proper nouns to generate an improved version of the text summary.

Sentiment analysis has found applications in healthcare [25], [26], tourism [27], fraud detection [28], finance [29], politics [30], business [31]. There are additional area of applications that are found in [32]. The sentiment analysis of online news articles is presented in [33]. The prediction of positive and negative sentiment on financial news is carried out in [34]. The opinion mining engine for news article is present in [35], which uses the knowledge from ConceptNet and SenticNet. The sentiment classification described in [36]

uses informatics and theoretic approach. A. Mudinas et al. have presented a notable work [37] on lexicon and concept-level sentiment analysis. T. H. A. Soliman et al. [38] have carried out mining of online customer reviews utilizing support vector machines and a similar work on sentiment analysis has been reported in [39] based on As-LDA model. There is an interesting work reported on sentiment analysis based on machine learning techniques in [40]. Sentence level sentiment analysis has been carried out using cloud machine learning techniques in [41]. Sentiment analysis using different types of lexicon dictionaries are listed in [42], [43] .

With motivation to help user to comprehend large volume of data, in this research work, summarization on news articles is performed then carried out sentiment analysis and representation. The extractive text summarization method is developed based on [21] to produce a brief overview of news articles. VADER [42] sentiment analyzer is used on original news articles and summarized news articles to deduce sentiment opinion from the text. By using VADER, various sentiment information has been collected as negative, neutral, positive, compound score and count related to sentiment words. Further sentiment information is represented using several visualizations schemes in three dimension such as column plots, surface plots, scatter plots etc. These 3-D visualization methods give a clear and better scheme to portray the sentiment information. Further the sentiment analysis and classification are carried out on original and summarized news articles using classifiers namely Logistic Regression, Random Forest and Adaboost classifiers. The experiments are carried out on BBC news articles and classification performance is tested on 10-fold cross validation. In Section 2 the method of text summarization with pronoun replacement is described, sentiment visualization and classification is presented in Section 3, an example on summarization and sentiment analysis is given in Section 4. Experiment results are presented in Section 5 followed by Section 6, which covers the conclusion.

II. Pronoun Replacement based Text Summarization

The text summarization involves in generating a brief summary of given text. Before generating summary, the preprocessing is carried out on the text. The preprocessing involves noise elimination, lowering text, tokenization, identify stopwords such as *that, a, the,* etc., and removal of them [44]. During preprocessing, pronouns which are place holders for proper nouns are also eliminated. In this research, a summarization technique is developed in which pronouns are replaced with proper nouns and then extractive summarization is carried. In the extractive methods [24] of text summary generation is to look for keywords or the most important words and their frequency in the text. The approach for identifying the important words is to eliminate stopwords and remaining words are taken as important words. As a part of stopword elimination, pronouns are also eliminated, thereby losing the frequency information. In this research the summarization method of [21] is developed as depicted in Fig. 1. For a given input text, Part of the Speech (POS) tagger of [45], [46] is employed to recognize various parts of a sentence. The pronouns are recognized and replaced with proper nouns. The proper noun that is occurring before pronoun and closer to pronoun is considered to replace the pronoun. However the original input text is used to produce the final text summary.

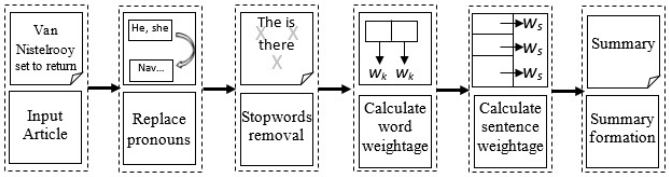


Fig. 1. Pronoun replacement based text summarization.

The next step is to eliminate the stopwords from the text and determine frequency of remaining words in the text. The computation of weightage of keywords or important words is as follows. Let n_k be the number of keywords and n_o be the number of stopwords, then a sentence has total $n_s = n_k + n_o$ words. Let f_k be the frequency of the k^{th} keyword. Also n_t be the number of keywords in the entire text. The weightage of k^{th} keyword can be calculated as

$$w_k = \frac{f_k}{n_t} \tag{1}$$

The sentence weightage is computed as summation of weightage of words given in (2). The sentences having important keywords with more weightage will have higher sentence weightage.

$$w_s = \frac{1}{n_s} \sum_{k=1}^{n_k} w_k \tag{2}$$

The priority order of the sentence is determined using sentence weightage, which indicates order to extract the sentences to form the summary. The user specifies the summary ratio S_r to decide the length of the summary required. For a text with m_t number of lines and S_r given summary ratio , the length of the summary m_s is calculated as

$$m_s = \frac{s_r}{100} \times m_t \tag{3}$$

The text summary is generated by extracting lines in priority ordered up to required length of m_s.

III. Sentiment Visualization and Classification

The VADER of [42] is a simple rule based sentiment analyzer. It consists of list of lexical features and associated sentiment measures. Based on grammatical and syntactical usage of the language, several rules are formed, which are used to determine the sentiment of the text. A lexicon basically is a list of words with each word assigned a semantic oriented values as positive value or negative value [47]. In VADER list of lexicons the features are assigned values between the range of -4 to +4, here -4 being extreme negative and +4 is extreme positive. In Table I, few words from VADER lexicon list are shown.

It is interesting to perform the sentiment analysis of the news article. Sentiment analysis on news articles are carried in various research such as [33]–[35]. The news article written by an author or journalist provides an opinion on the subject about which article was written. The sentiment analysis thus

TABLE I. EXAMPLE FROM VADER LEXICON

Word	Sentiment Score
Excel	2.0
Exhaust	-1.2
Favorable	2.1
Impatience	-1.8

provides a sentiment evaluation of the news articles. In this research VADER is utilized to perform the sentiment analysis of the BBC news articles. Schematic diagram for sentiment analysis is depicted in Fig. 2. The news articles are subjected to preprocessing such as word tokenization and stopwords removal. Then VADER is applied to compute sentiment score of the news article. The VADER utilizes lexicon list and computes sentiment information such as compound, neutral, negative and positive scores. Also it gives count of positive, negative and neutral words. In this research, novel 3-D visualizations of sentiment information obtained from VADER are presented. The visualizations schemes in terms of three dimensions column plots, surface, scatter plots etc., are developed. These 3D visualization provide better insight of sentiment information gathered from news articles.

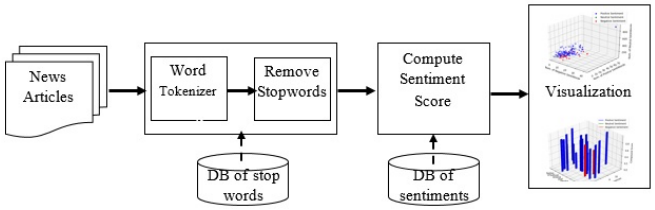


Fig. 2. Sentiment analysis of news articles.

Furthermore, the sentiment classification on summarized news articles as shown in Fig. 3 is performed. As significant amount of data being generated every day, it is becoming important to provide techniques which help user to effectively comprehend the data. The text summarization provide a brief overview of input text and effectively enable user to focus on reduced version of the text. When text summarization is applied to news articles it gives a brief overview of the news with inherent subjective information. Usually the news articles provide elaborated discussion on the subject and hence it is appropriate to perform text summarization to obtain important discussions in news. The sentiment analysis and classification on summarized version of news articles is introduced as shown in Fig. 3. The preprocessing of news articles is performed in which words are tokenized and stopwords are removed. The news articles are subjected to summarization to generate overviews. The sentiment classification is performed on the summarized version of the news articles. Feature vectors of N-gram size are created using a bag of words [48]. Next Logistic Regression, Random Forest and Adaboost classifiers are used for classification. Logistic Regression used as base classifier in Adaboost classifier.

IV. SUMMARIZATION AND SENTIMENT ANALYSIS EXAMPLE

The summarization and sentiment analysis is briefly explained with an example in this section. An input news article is shown in the Fig. 4, which is on Football from BBCSport

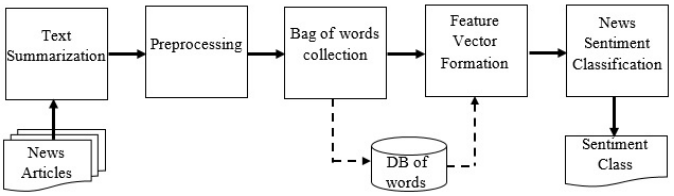


Fig. 3. News summarization and sentiment classification.

news dataset. The number of occurrences of various nouns in this article is determined. Top three nouns with maximum occurrences are 'Van', 5 times, 'Nistelrooy', 5 times and 'United', 4 times. The sentiment score for the Fig. 4 is computed using VADER. Positive score is 0.212, negative score is 0.083, neutral score is 0.705 and compound score is 0.9726 observed.

In this article, the pronoun such as 'he', 'it', 'they' etc., have been used several times as place holder for proper nouns. Text summarization is performed on this text using pronoun replacement method as described in Section 2. Once the pronouns are replaced, the stopwords are eliminated from the text to identify important words or keywords. Next, weightage of each keywords is computing by using (1). Table II shows the computed weightage for the few keyword words from input article.

TABLE II. KEYWORDS AND WEIGHTAGE

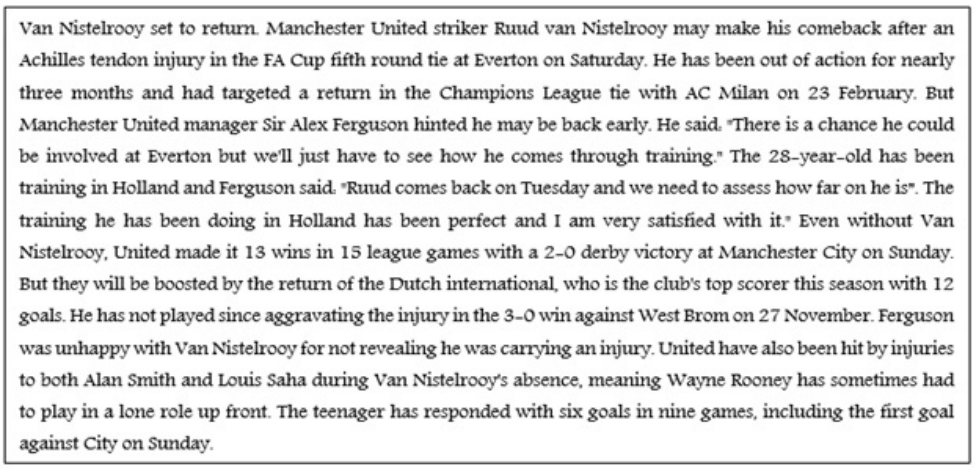
Sl. Num.	Keyword	Weightage
1.0	Nistelrooy	0.05625
2.0	United	0.03125
3.0	League	0.0125
4.0	Champions	0.00625
5.0	Fifth	0.00625

The sentence weightage is computed using (2) based on the weightage of keywords present in that sentence. Also each sentence is assigned a priority number based on its sentence weightage. Lower priority number is assigned for the sentence with higher weightage. In Table III sentence (partially depicted), its weightage and priority number for few sentences are shown.

TABLE III. SENTENCE WEIGHTAGE AND PRIORITY NUMBER

Sl. Num.	Sentence	Weightage	Priority Number
1.0	Van Nistelrooy set to return.	0.0225	1.0
2.0	Manchester United striker Ruud van Nistelrooy may make his comeback after an Achilles tendon.	0.0127	5.0
3.0	He has been out of action for nearly three months and had targeted a return in the Champions?	0.0063	12.0
4.0	But Manchester United manager Sir Alex Ferguson hinted he may be back early.	0.0159	4.0
5.0	He said: "There is a chance he could be involved at Everton but we'll just have to see how he comes?	0.0106	6.0

User provides the summarization ratio, using which the summary is generated. The number sentences to be included in the summary can be found by equation (3) using. The summary is formed by extracting the sentences from the article in priority order. Summarized text for Fig. 4 is collected with



Van Nistelrooy set to return. Manchester United striker Ruud van Nistelrooy may make his comeback after an Achilles tendon injury in the FA Cup fifth round tie at Everton on Saturday. He has been out of action for nearly three months and had targeted a return in the Champions League tie with AC Milan on 23 February. But Manchester United manager Sir Alex Ferguson hinted he may be back early. He said. "There is a chance he could be involved at Everton but we'll just have to see how he comes through training." The 28-year-old has been training in Holland and Ferguson said. "Ruud comes back on Tuesday and we need to assess how far on he is". The training he has been doing in Holland has been perfect and I am very satisfied with it." Even without Van Nistelrooy, United made it 13 wins in 15 league games with a 2-0 derby victory at Manchester City on Sunday. But they will be boosted by the return of the Dutch international, who is the club's top scorer this season with 12 goals. He has not played since aggravating the injury in the 3-0 win against West Brom on 27 November. Ferguson was unhappy with Van Nistelrooy for not revealing he was carrying an injury. United have also been hit by injuries to both Alan Smith and Louis Saha during Van Nistelrooy's absence, meaning Wayne Rooney has sometimes had to play in a lone role up front. The teenager has responded with six goals in nine games, including the first goal against City on Sunday.

Fig. 4. Input news article.

ratio as 25%, 50% and 75% and results of summary are shown in Table IV.

TABLE IV. NEWS SUMMARIZATION AND SENTIMENT ANALYSIS

S_r	Summarized Text	Num. of sentences in summary	Negative Score	Neutral Score	Positive Score	Compound Score
0.25	Van Nistelrooy set to return. But Manchester United manager Sir Alex Ferguson hinted Ferguson may be back..	4.0	0.168	0.749	0.084	-0.4215
0.5	Van Nistelrooy set to return. But Manchester United manager Sir Alex Ferguson hinted Ferguson may be back..	8.0	0.084	0.774	0.142	0.6908
0.75	Van Nistelrooy set to return. But Manchester United manager Sir Alex Ferguson hinted Ferguson may be back..	12.0	0.079	0.726	0.195	0.9618

Further, sentiment analysis using VADER is performed for each summarized text. The VADER computes sentiment information such as negative, neutral, positive and compound score which are given in Table IV.

V. EXPERIMENTAL RESULTS

The experiments are conducted on news article collected from [49], which are BBC articles. The BBCSport dataset includes 737 documents about articles on five topical areas as Athletics, Cricket, Football, Rugby and Tennis from BBC sport web site between the years 2004 to 2005. It is interesting to perform the sentiment analysis on news articles. VADER sentiment analyzer is applied on the news articles on dataset BBCSport. Moreover the POS tagger of [45], [46] is utilized to determine various parts of sentences. Proper nouns and their occurrences in article are gathered. In Table V, top three nouns having maximum occurrence in the articles with their frequency are shown. The VADER sentimental analyzer gives various scores such as negative, neutral, positive and compound score which are given in columns 3, 4, 5 and 6 respectively in Table V. The count of negative, neutral and positive words are given in column 7, 8 and 9, respectively.

Sentiment Visualization:

A novel 3-D visualization of sentimental information obtained from VADER is presented in Fig. 5. Twenty news article on Football and Athletic from BBCSport dataset are considered. For each article, the number of occurrences of proper nouns is determined. In Fig. 5(a) negative sentiment score versus positive sentiment score for each article is represented. In this figure, along x-axis the proper noun with maximum frequency, along y-axis the negative score and along z-axis the positive sentiment score are shown. Fig. 5(b) shows maximum occurring proper noun and count of that noun along x-axis and y-axis against compound sentiment score along z-axis. Fig. 5(b) highlights the compound score on an article with respect to the noun having maximum occurrence and its count, hence showing the importance of the noun as a subject in that article. Fig. 5(c) provides 3D visualization of negative score versus positive score for Athletic articles. In Fig. 5(d) noun occurrences versus compound score is represented for 20 Athletic articles.

The novel 3-D visualizations are developed to represent the compound sentimental score as shown in Fig. 6. In these figures, compound score versus count of positive and negative sentiment words are shown. Fig. 6(a) show the 3-D representation for compound sentiment of all Football articles from the BBCSport dataset. In Fig. 6(a) highest compound score of 2.927 having the number of negative words 6 and number of positive words 14 is observed. Fig. 6(b), 6(c) and 6(d) represent the compound scores for articles on Cricket, Athletic and Rugby respectively are shown. These 3-D visualizations signifize the changes in compound score that can occur when count of positive or negative sentiment words vary.

The news articles are subjected to sentiment analyzer VADER, which provides various sentiment score also it gives count of negative, positive and neutral words in the articles. The sentiment of the article is positive for compund score greater than zero, neutral for compound score of zero otherwise it is negative. Fig. 7 provides 3-D visualization of count information obtained from VADER. In Fig. 7(a) the 3-D scatter

TABLE V. NEWS ARTICLE WITH NOUN FREQUENCY IN THE ARTICLE ALONG WITH SENTIMENT SCORES

Article	Noun and frequency	Negative Score	Neutral Score	Positive Score	Compound Score	# Negative Sentiment Words	# Neutral Sentiment Words	# Positive Sentiment Words
1.0	('Everton', 9) ('United', 8) ('Martyn', 8)	0.102	0.676	0.221	0.995	19.0	292.0	32.0
2.0	('Van', 5) ('Nistelrooy', 5) ('United', 4)	0.083	0.705	0.212	0.9726	6.0	132.0	14.0
3.0	('Moyes', 5) ('Beattie', 5) ('Gallas', 5)	0.084	0.805	0.111	0.4504	8.0	204.0	13.0
4.0	('Ronaldo', 3) ('United', 3) ('Manchester', 1)	0.044	0.688	0.268	0.9661	2.0	72.0	11.0
5.0	('Home', 5) ('Smith', 4) ('Scotland', 4)	0.069	0.749	0.182	0.9456	4.0	121.0	11.0

plot is depicted for news articles of Football. More positive sentiments are observed in Fig. 7(a) than negative or neutral. The 3-D scatter plots for Cricket shown in Fig. 7(b), Athletic in Fig. 7(c) and Rugby in Fig. 7(d).

In Fig. 8, ten words with positive sentiment and in Fig. 9, ten words with negative sentiments are depicted. In each graph, the word with its sentiment score and its percentage contribution are shown. In Fig. 8(a) the graph shows positive sentiment words for Football articles. In Fig. 8(b), 8(c) and 8(d) showing words with positive sentiment for Cricket, Athletic and Rugby news articles. Fig. 9 depicts top ten negative sentiment words for news articles.

Summarization and Classification: The sentiment classification is carried out on news articles. The BBCSport news article dataset consists of 737 article related to Football, Cricket, Athletic, Rugby, and Tennis. Later each article is subjected to summarization with ratio of 25%, 50% and 75% hence dataset consists of 2948 articles. The sentiment analysis is performed on each article using VADER. The Logistic Regression, Random Forest and AdaBoost classifiers are used for sentiment classification. Feature vectors of N-gram size are constructed from news articles by preparing bag of words as given in [48]. From the BBCSport dataset of articles occurrences of words are collected and bag of words is prepared by taking 'N' most frequent words. Here 'N' is taken as 2000, 3000 and 4000. Table VI shows 10-fold cross validation results on the BBCSport dataset of articles without summarization. A maximum classification rate of 84.93% is observed for AdaBoost classifier with N as 3000.

Next, the news articles are subjected to summarization using method described in Section 2. The summarization ratio of 25%, 50%, 75% is applied on each article. Using the sentiment analyzer VADER, the sentiment type of each article is determined. The 10-fold cross validations are performed on three classifiers Logistic Regression, Random Forest and AdaBoost classifier with varying 'N' as 2000, 3000 and 4000. In Table VII, the 10-fold cross validation results are presented. It is observed that as summarization ratio increases better sentiment classification rates are obtained. When summarization ratio is 25%, a maximum classification rate of 78.73% for AdaBoost classifier with 'N' equal to 4000 is observed. For summarization ratio of 50%, maximum classification rate 83.06% with 'N' 4000 on AdaBoost classifier is obtained. A maximum classification rate of 83.23% for AdaBoost classifier with 'N' 3000 is observed for 75% text summarization.

VI. CONCLUSION

In the recent years we are witnessing significant amount of data being generated in numerous forms such as social media, web blogs, web sites, Wikipedia, news articles and many more. Due this the end user is overloaded with data and there is a greater need for effective methods to help user to absorb the data. Data extraction and representation methods are highly desirable to assist user to comprehend the huge data. One of the effective methods to obtain brief overview is using text summarization. Also sentiment analysis and classification being used to determine opinion expressed in the text. In this research, the text summarization and sentiment analysis on BBC news articles is combined. BBC news articles are collected from [49], which consists of 737 news articles on various sports topics. Extractive based text summarization method is developed in this research which involves pronoun replacement with proper noun and form text summary. The sentiment analysis of BBCSport news articles is carried by VADER. The VADER provides various evaluated information including positive, compound, negative and neutral score along with count of neutral, negative and positive words in the text. Novel three dimensional visualizations are provided to depict sentiment information obtained on BBCSport. Later, using the summarization ratio of 25%, 50% and 75% the text summarization is carried out on news articles. On the dataset of news articles, the feature vector is formed using bag of words of N-gram size. The sentiment classification is carried out on news articles at first without summarization and later on summarized text of 25%, 50% and 75% ratio. Three classifiers are employed to perform sentiment classification such as Logistic Regression, Random Forest and Adaboost classifier with varying N as 2000, 3000 and 4000. When classification is carried out without summarization highest classification rate of 84.93% observed. For 25%, 50% and 75% summarized text a maximum classification rate of 78.73%, 83.06% and 83.23% are respectively obtained.

ACKNOWLEDGMENT

Author is grateful to Dr. Sunil Thomas, Department of Electrical and Electronics Engg, Birla Institute of Technology and Science, Pilani-Dubai, Dubai for his suggestions to improve the manuscript.

REFERENCES

[1] Xing Fang and Justin Zhan, Sentiment analysis using product review data, in Journal of Big Data, 2:5, 2015.

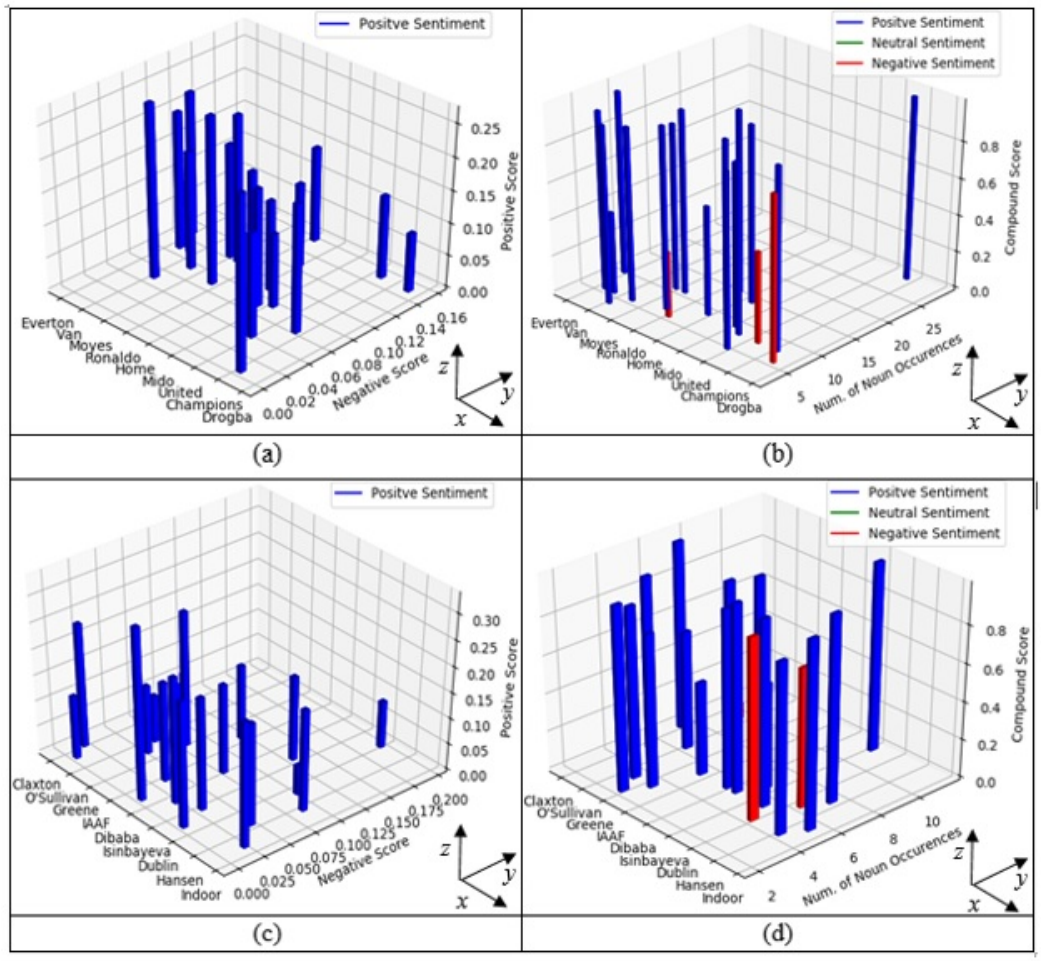


Fig. 5. The 3-D representation of sentimental information for 20 news articles. (a) Negative versus positive score for Football articles. (b) Noun occurrences versus compound score for Football articles. (c) Negative versus positive score for Athletics articles. (d) Noun occurrences versus compound score for Athletics articles.

TABLE VI. PERFORMANCE OF SENTIMENT CLASSIFICATION

	Logistic Regression			Random Forest			AdaBoost		
N	2000	3000	4000	2000	3000	4000	2000	3000	4000
Classification Rate in %	84.3	84.73	84.3	81.81	81.33	81.02	84.43	84.93	84.63
Misclassification Rate in %	15.7	15.27	15.7	18.19	18.67	18.98	15.57	15.07	15.37

[2] Himmat M., Salim N., Survey on Product Review Sentiment Classification and Analysis Challenges In: Herawan T., Deris M., Abawajy J. (eds) Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Lecture Notes in Electrical Engineering, vol 285, Springer, Singapore, 2014.

[3] Khoo, C.S.G., Nourbakhsh, A., & Na, J.C., Sentiment analysis of online news text: A case study of appraisal theory Online Information Review, 36(6), 2012.

[4] Yu Wang, Tom Clark, Jeffrey Staton, Eugene Agichtein Towards Tracking Political Sentiment through Microblog Data in Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 8893, Baltimore, Maryland USA, 27 June 2014.

[5] Aliza Sarlan, Chayanit Nadam, Shuib Basri, Twitter sentiment analysis, in International Conference on Information Technology and Multimedia (ICIMU), Malaysia November 18 20, 2014.

[6] Youngsub Han, Kwangmi Ko Kim, Sentiment analysis on social media using morphological sentence pattern model, in IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, 2017.

[7] Smailovic J., Grcar M., Lavrac N., nidaric M., Predictive Sentiment Analysis of Tweets: A Stock Market Application, In: Holzinger A., Pasi G. (eds) Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science, vol 7947. Springer, Berlin, Heidelberg, 2013

[8] Rajat Ahuja, Harshil Rastogi, Arpita Choudhuri, Bindu Garg Stock market forecast using sentiment analysis, in 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pages 1008-1010, 2015.

[9] Fang Chen, Kesong Han and Guilin Chen, An Approach to sentence selection based text summarization, in Proceedings of IEEE TENCON02, pp. 489-493, 2002.

[10] Amari, S.-I. and Nagaoka, H. Methods of Information Geometry, Translations of Mathematical Monographs, in Oxford University Press, 2001.

[11] J. Allan, C. Wade, and A. Bolivar, Retrieval and novelty detection at the sentence level, in Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 314321, 2003.

[12] Hovy, E. and C.-Y. Lin, Automatic Text Summarization in SUMMARIST, in I. Mani and M.Maybury (eds), Advances in Automatic Text

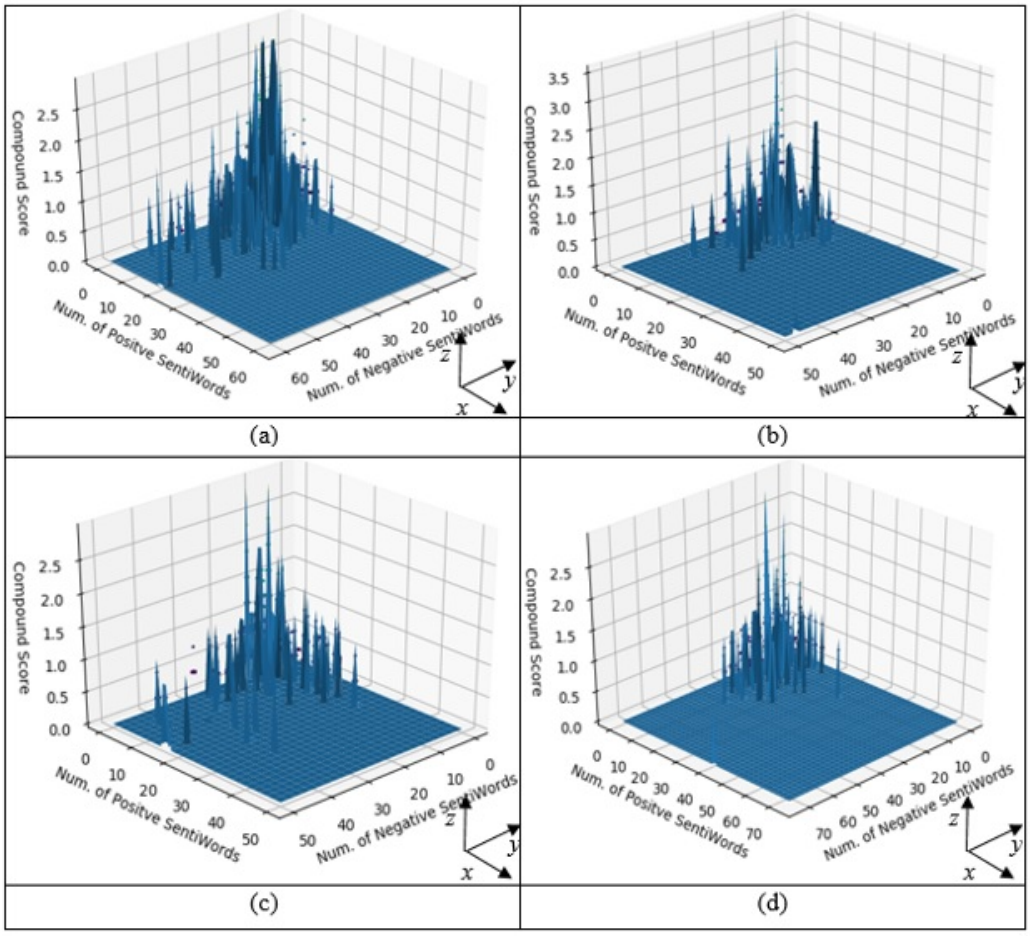


Fig. 6. The 3-D representation of compound score versus count of positive and negative sentiment words. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

TABLE VII. PERFORMANCE OF SENTIMENT CLASSIFICATION ON SUMMARIZED TEXT

		Logistic Regression			Random Forest			AdaBoost		
S_r	N	2000	3000	4000	2000	3000	4000	2000	3000	4000
0.25	Classification Rate in %	78.5	78.49	78.33	77.24	75.96	76.61	77.57	77.41	78.73
	Misclassification Rate in %	21.5	21.51	21.67	22.76	24.04	23.39	22.43	22.59	21.27
0.5	Classification Rate in %	82.73	81.06	81.31	79.76	78.98	80.22	82.1	82.06	83.06
	Misclassification Rate in %	17.27	18.94	18.69	20.24	21.02	19.78	17.9	17.94	16.94
0.75	Classification Rate in %	82.16	82.43	82.12	81.65	79.74	79.77	82.71	83.23	82.2
	Misclassification Rate in %	17.84	17.57	17.88	18.35	20.26	20.23	17.29	16.77	17.8

Summarization, pp.81-94, MIT Press, 1999.

[13] Benno Stein, Sven Meyer zu Eissen, Topic Identification: Framework and Application, in Proceedings of International Conference on Knowledge Management, pp 522-531, 2004.

[14] Kino Coursey, Rada Mihalcea, Topic Identification Using Wikipedia Graph Centrality, in Proceedings of NAACL HLT 2009, pages 117120, Boulder, Colorado, June 2009.

[15] Irma Sofia Espinosa Peraldi, Atila Kaya, Sylvia Melzer, Ralf Moller, On Ontology Based Abduction For Text Interpretation, in Proceedings of 9th International Conference Computational Linguistics and Intelligent Text Processing, Israel, pp 194-205, 2008.

[16] Marti A. Hearst, Direction-Based Text Interpretation as an Information Access Refinement, in Text-Based Intelligent Systems, Lawrence Erlbaum Associates, 1992.

[17] Lloret, E. & Palomar, M, Text summarisation in progress: a literature review, in Artificial Intelligence Review, vol. 37, issue 1, pp 1-41, January 2012.

[18] Kazuo Sumita, Seiji Miike, Kenji Ono, Tetsuro Chino, Automatic abstract generation based on document structure analysis and its evaluation as a document retrieval presentation function, in Systems and Computers in Japan, vol. 26, issue 13, 2007.

[19] Zhang Pei-ying and LI Cun-he Automatic text summarization based on sentences clustering and extraction, in 2nd IEEE International Conference on Computer Science and Information Technology, pp. 167-168, 2009.

[20] Daniel Gayo-avello , Daro lvarez-gutirrez , Jos Gayo-avello, Naive Algorithms for Key-phrase Extraction and Text Summarization from a Single Document inspired by the Protein Biosynthesis Process, in First International Workshop Biologically Inspired Approaches to Advanced Information Technology, BioADIT 2004, Lausanne, Switzerland, pp. 440-455, January 29-30, 2004.

[21] Siddhaling Urolagin, Jagadish Nayak, Likitha Satish A method to generate text summary by accounting pronoun frequency for keywords weightage computation, in Proceedings of 2017 International Conference on Engineering & Technology (ICET'2017) Akdeniz University, Antalya, Turkey, pages 1-4, 21-23 August, 2017.

[22] Jos M. Perea-Ortega, Elena Lloret, L. Alfonso Urea-Lpez, Manuel Palo-

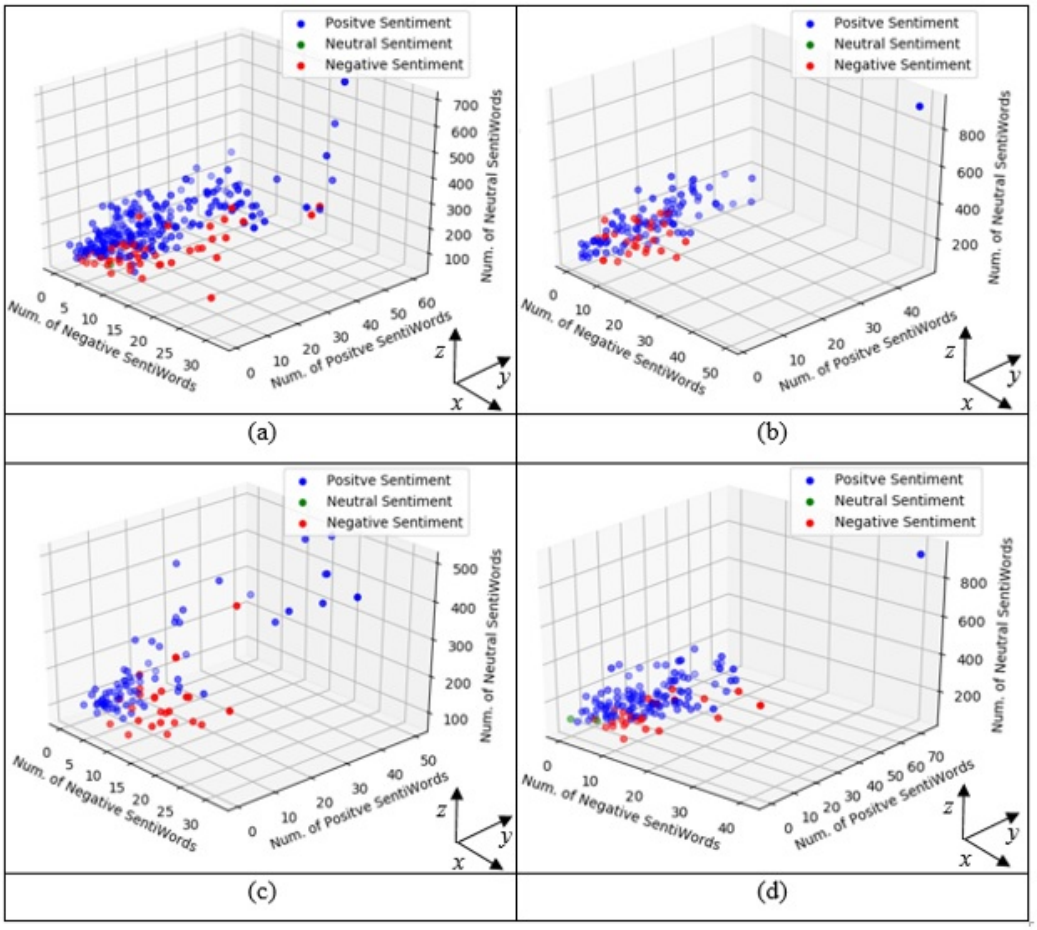


Fig. 7. Representing sentimental count information as 3-D visualization. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

mar, Application of Text Summarization techniques to the Geographical Information Retrieval task, in Expert Systems with Applications, vol. 40, issue 8, pp. 29662974, 2013.

[23] Nongnuch Ketui, Thanaruk Theeramunkong, Chutamanee Onsuwan, An EDU-Based Approach for Thai Multi-Document Summarization and Its Application, in Journal ACM Transactions on Asian and Low-Resource Language Information Processing TALLIP, vol. 14 issue 1, January 2015.

[24] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas,George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, Luciano Favaro, Assessing sentence scoring techniques for extractive text summarization, in Expert Systems with Applications, vol.40, issue 14, pp. 57555764, 2013.

[25] Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri, Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics, in Proceedings of the 6th International Conference on Information Technology in Bio- and Medical Informatics - Volume 9267,pp. 16-24, 2015.

[26] Denecke K, Sentiment Analysis from Medical Texts. In: Health Web Science. Health Information Science, in Springer, Cham. doi:https://doi.org/10.1007/978-3-319-20582-3-10, 2015.

[27] D Grabner, M Zanker, G Fliedl, M Fuchs, Classification of Customer Reviews based on Sentiment Analysis, in Information and Communication Technologies in Tourism pp 460-470. 2012.

[28] Gann W-JK, Day J, Zhou S, Twitter analytics for insider trading fraud detection system in Proceedings of the sencond ASE international conference on Big Data. ASE. May 27 - May 31, Stanford, CA, USA, 94305, 2014.

[29] Siddhaling Urolagin, Text Mining of Tweet for Sentiment Classification and Association with Stock Prices, in 2017 International Conference on Computer and Applications (ICCA), Doha, 2017, pp. 384-388.doi: 10.1109/COMAPP.2017.8079788.

[30] Kartik Singhal, Basant Agrawal, Namita Mittal, Modeling Indian General Elections: Sentiment Analysis of Political Twitter Data, Advances in Intelligent Systems and Computing, vol 339. Springer, New Delhi, pp 469-477, 2015.

[31] Van Looy A., Sentiment Analysis and Opinion Mining (Business Intelligence 1), In: Social Media Management. Springer Texts in Business and Economics. Springer, Cham, 2016.

[32] M.Walaa, A. Hassan, and H. Korashy, Sentiment Analysis Algorithms and Applications: A Survey, Ain Shams Engineering Journal,vol.5, no. 4,pp. 10931113, 2014.

[33] Simon Fong, Yan Zhuang, Jinyan Li, Richard Khoury, Sentiment Analysis of Online News Using MALLET, International Symposium on Computational and Business Intelligence (ISCBI), 2013.

[34] Bradley Meyer, Marwan Bikdash, Xiangfeng Dai, Fine-Grained Financial News Sentiment Analysis, in SoutheastCon, pages 1-8, At Charlotte, NC, USA, 2017.

[35] Prashant Raina Sentiment Analysis in News Articles Using Sentic Computing, in IEEE 13th International Conference on Data Mining Workshops (ICDMW), 2013.

[36] Lin Y, Zhang J, Wang X, Zhou A, An information theoretic approach to sentiment polarity classification, in Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, ACM, New York, NY, USA.pp 3540, 2012.

[37] A. Mudinas, D. Zhang and M. Levene, Combining lexicon and learning based approaches for concept-level sentiment analysis, in Proceedings of the First International Workshop on Issues of Sentiment Discovery

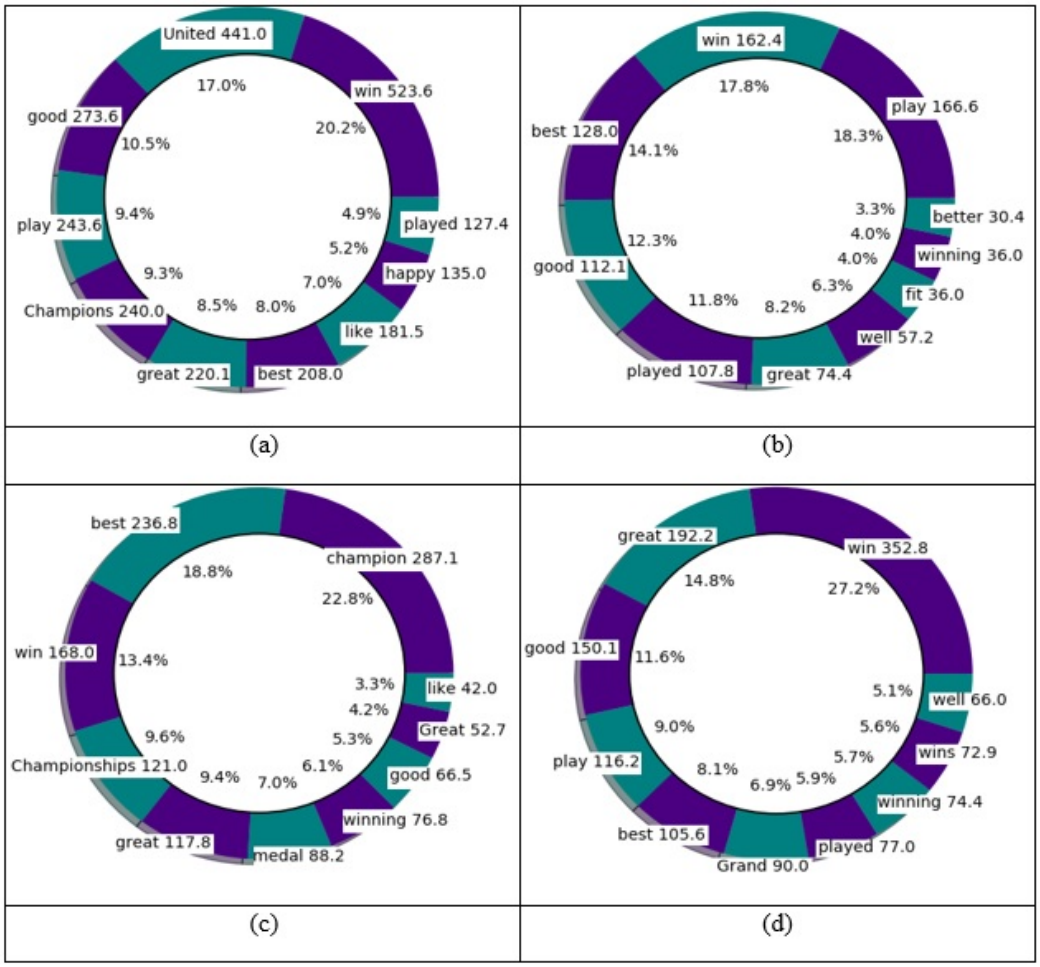


Fig. 8. Top ten positive sentiment words. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

and Opinion Mining, Beijing, China ugust 12 , 2012, Article No. 5, doi:10.1145/2346676.2346681, 2012.

[38] T. H. A. Soliman, M. A. Elmasry, A. R. Hedar and M. M. Doss, Utilizing support vector machines in mining online customer reviews, in 22nd International Conference on Computer Theory and Applications (ICCTA), Alexandria, 2012, pp. 192-197. doi: 10.1109/ICCTA.2012.6523568, 2012.

[39] J. Liang, P. Liu, J. Tan, and S. Bai, Sentiment Classification Based on AS-LDA Model, Procedia Computer Science, vol. 31, pp. 511516, 2014.

[40] A. P. Jain and P. Dandannavar, Application of Machine Learning Techniques to Sentiment Analysis, in 2nd International Conference on Applied and Theoretical Computing and Communication Technology, Bangalore, pp. 628632, 2016.

[41] R. Arulmurugan, K. R. Sabarmathi, and H. Anandakumar, Classification of sentence level sentiment analysis using cloud machine learning techniques, in pages 1-11, Cluster Computing, 2017, https://doi.org/10.1007/s10586-017-1200-1.

[42] Hutto, C.J. and Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, in Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[43] Y. Wang, Y. Zhang, and B. Liu,Sentiment Lexicon Expansion Based on Neural PU Learning, Double Dictionary Lookup, and Polarity Association, Conference on Empirical Methods in Natural Language Processing, Copenhagen, pp. 711, 2017.

[44] Alper Kursat Uysal and Serkan Gunal, The impact of preprocessing on text classification, in Information Processing & Management 50, 1, pages 104112, 2014.

[45] Kristina Toutanova and Christopher D. Manning, Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, in Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 63-70. 2000.

[46] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer, Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, in Proceedings of HLT-NAACL, pp. 252-259, 2003.

[47] Liu, B., Sentiment Analysis and Subjectivity, in N. In-durkhya & F. Damerau (Eds.), Handbook of Natural Language Processing (2nd ed.). Boca Raton, FL: Chapman & Hall, 2010.

[48] A. Deshwal and S.K. Sharma, Twitter sentiment analysis using various classification algorithms, 5th International Conference on Reliability, Infocom Technologies and Optimization, Noida, pp. 251-257, 2016.

[49] D. Greene and P. Cunningham. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering, in Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning Pages 377-384, 2006.

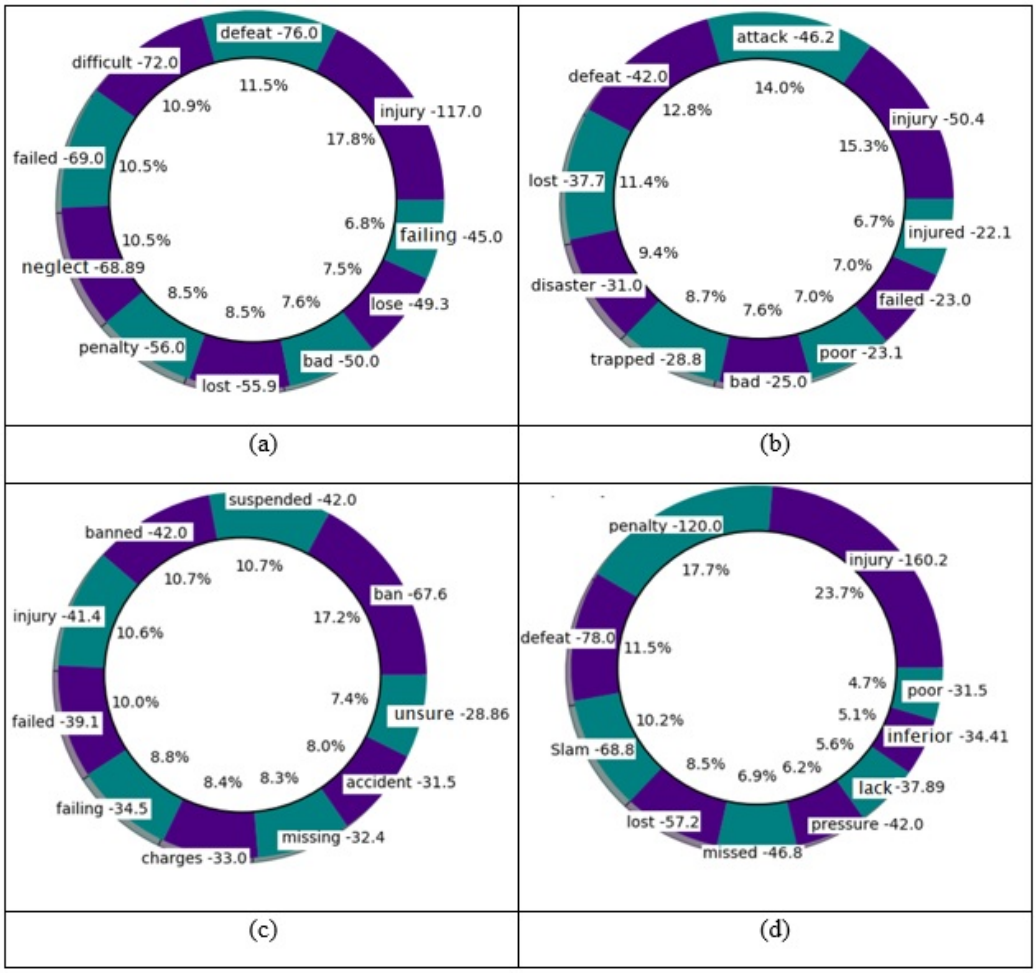


Fig. 9. Top ten negative sentiment words. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.