International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.183

Evaluation of the Speakers' Attitudes Expressed in their State of the Union

Sarthak Chawla
Department of Electronics and Communications Engineering,
SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203,
Chengalpattu District, Chennai TN, India

Harisudha Kuresan
Department of Electronics and Communications Engineering,
SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203,
Chengalpattu District, Chennai TN, India
harisudk@srmist.edu.in

Nandan Nishad Gopuram
Department of Electronics and Communications Engineering,
SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203,
Chengalpattu District, Chennai TN, India

Manas Sharma
Department of Electronics and Communications Engineering,
SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203,
Chengalpattu District, Chennai TN, India

Subashree Guruchandar
Department of Electronics and Communications Engineering,
SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203,
Chengalpattu District, Chennai TN, India

***Abstract* — The implications and implementations of the data science and machine learning frenzy as it sweeps the country are broad. Each of them is examined in this study through the prism of a national issue, at least in the United States. The words used by past Presidents of the United States are the subject of this discussion. The focus of this research is on Natural Language Processing and its applications in important historical speeches. Natural Language Processing serves as an effective method for the analysis of text-based data. Using Natural Language Processing, sentiment analysis was used on the collection of different Addresses to achieve a better understanding of the tone used by various Presidents over the course of history. This sentiment analysis provided a collection of sentiment-based scores about important topics and issues in the United States. A feeling can be expressed via natural language processing (NLP). The Addresses were analyzed with the purpose to attain a deeper and broader understanding of the tone used throughout history by presidents.**

Keywords — Natural Language Processing (NLP), Latent Semantic Indexing (LSA), Bidirectional encoder representation from the transformer (BERT), Graphical User Interface (GUI), Hypertext Markup Language (HTML), Extensible Markup Language (XML)

I. Introduction

Elections contribute a significant amount to proper democratic governance. In light of the fact that a directly democratic form of governance in which political decisions are decided directly by the entire assembly of qualified people is unachievable in most present environments, ballot government should be carried out through intermediaries. Democratic institutions give the public the ability to pick their leaders and hold them accountable for their actions while in office. Elections contribute to the soundness and legitimacy of the social and political system by creating an enabling environment for participation.

All of the Presidential State of the Union addresses were gathered and processed as the core textual data. We obtained the text from a Presidential Address Repository. To make processing easier, the text was divided into individual text files for each individual address from a big text file that comprised every speech. The previous works are listed so that readers can quickly see what's been done.

Vader et al., in his work discussed the practical applications of sentiment analysis face significant obstacles due to the fundamental nature of social media information. The author explores various techniques to understand the behavior of the sentiments using the features of data points from text, moreover, the high frequency of words in the data corresponding to adjectives or helper verbs affects the model on a larger scale than expected. Researchers build and try to validate a top-tier collection of language features that are specifically tailored to sentiment in blog-type contexts using a merging of qualitative and quantitative methodologies. Then, consider five broad pillars that include grammatical and language patterns for expressing and highlighting sentiment intensity, in addition to these lexical aspects [6].

Helen Balinsky et al., proposed text summarization by computer is a difficult problem to solve. In recent times, the text available electronically has massively increased. As a result of this growth, the demand for automatic text summarizing methods and tools is increasing. Data compression can be regarded as a sort of automatic summarization. To get compression, improvement in modeling and implementation to reduce noise in the data and create networks with more depth [7].

Muhammad Fachrie et al., used data from the General Election Commission to investigate the use of Machine Learning (ML) to forecast candidate win in Indonesian Regional Elections (KPU). Each candidate's political party affiliations are included in the data. The researchers built a Machine Learning model-based data provided by official institutions to predict the winner of each candidate in a regional election in Indonesia, rather than using social media data as in previous studies. The forecast is a classification-type task with two categories: 'win' and 'loss.' [8]

II. Dataset Description

The data is collected from various online resources such as using a web crawling-based technique. In web crawling, data from a website is fetched using a script that renders HTML content and converts it to text-based content. After conversion to text-based content, it is then downloaded in the desired format such as pdf, XML, etc. After fetching the data from the source, it is subjected to a data cleaning process. In data cleaning undesired text, spaces are removed and converted into a table format so that it can be subjected to a model for training. This data is also segregated based on the parties i.e., democratic and republican as shown in Figure 1.

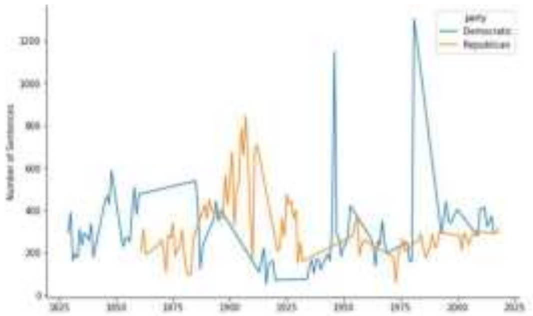


Fig. 1. Number of sentences used by each political party

Data is then analysed based upon the number of sentences and to which party they belong, in order to understand dataset composition of both parties Vs timeline as shown in Figure 2.

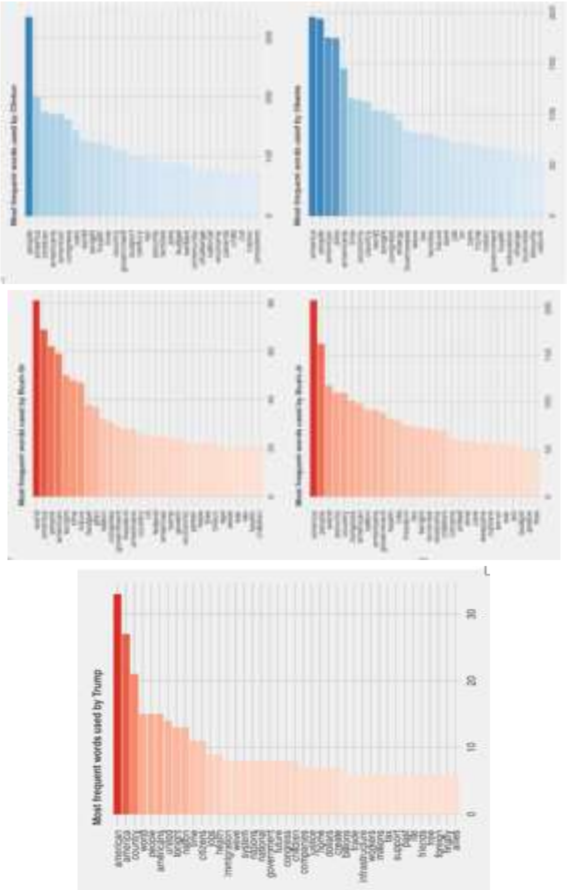


Fig. 2. Most frequent words used by each President

This figure shows how the number of sentences vary compared to the timeline of their delivery. Using this chart it can be concluded that the data is well balanced and does not have significant bias towards a particular label. The density of words Vs sentences is plotted in Figure 3, to check in which range our data lies. Only 91 words or fewer make up 99% of the data. Data padding is therefore done to this length to guarantee consistency throughout.

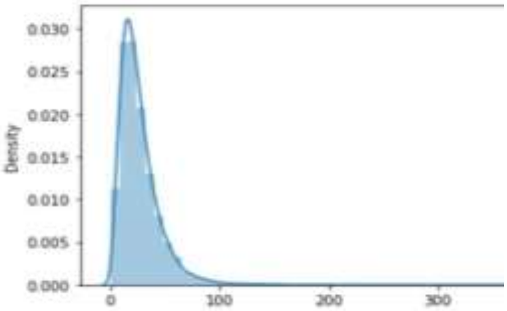


Fig. 3: 99% of all SOTU sentences are under 91 words

To classify the sentiment of a given text the following processes are used:

i. Data preparation

Data set is scraped from online resources using web scraping libraries in .py format. once data is extracted it is subjected to a data cleaning process where each data point is scanned and type errors, redundant spaces are removed, and these data points are made in the form as to map them into a pandas data frame so they can be easily interpreted by the model for use. Then each data point is subjected to frequency check, null imputation, and data process. Each process such as cleaning, stemming, and lemmatization / lemma formation. When web-scraping, we make use of a library called "When web-scraping, utilize a "regular expression." "to substitute and replace data in order to eliminate unwanted phrases The stop word is then eliminated from the sentence by combining all these techniques data is converted to form which can be used by the model, further techniques have been discussed below.

ii. Text Input

Any data that is entered into a model must be in vector format. The following transformations will be carried out:

- Remove all numbers and punctuation.
- Change all of the words to lowercase.
- Remove any stop words (for example, the, a, that, this, it, etc.).
- Texts are tokenized.
- Using a bag-of-words format, convert the sentences into vectors.

iii. Tokenization

Each word in the sentence has meaning and tries to interpret something that might or might not be important to the context. Tokenization helps us in breaking sentences into words also known as tokens which represent one keyword in the data point. Each token has some meaning associated with it and hence affects model training according to the weight assigned to them.

iv. Stop Word Filtering

In all human languages, there are a plethora of stop words. These words add unnecessary overhead while training the model. They don't impact the model but increase the size of the dataset. Removal of stop words is very important to make the model more robust and faster. Many libraries such as sklearn python library help in performing this task after the removal of stop words unnecessary data are removed and model accuracy and training become more robust.

v. Negation Handling

Sentiment Analysis includes a sub-task called negation handling. In written literature, negatives play an important role. Contrary terms in sentences frequently shift the sentiment of the sentence while the opposite makes it more aligned so both things have equal importance. Negation handling is important so that the context is well balanced and the model does not become biased towards negation.

vi. Stemming

Stemming is a method that involves converting the words which have similar meanings, such as eat, ate, eaten, eating, and will eat, but are represented differently in the dataset. So here stemming is a method of mapping each of these words to eat so that they correspond to similar things and the model can understand that they have the same meaning in the context which was different earlier as in data representation.

vii. Classification

Using a plethora of machine-learning or deep-learning models to classify sentiment.

III. Methodology

3.1 Text Summarization

The process of shortening the text so that it is presented in a summarized fashion is known as text summarization. Text summarization is done in such a way as to ensure the summarized part can describe the whole text in a brief manner. Text summarization becomes very important when the given text is very long, for companies that show news updates, summarizing text is important for adding more users to the platform more features need to be added that can interact with users and their emotions by interpreting sentiment in a concise manner. To automate the process of text summarization, Machine learning and deep learning-based techniques involve natural language processing. Using these techniques, automatic text summarization can be performed. There are two major types of text summarization:

- Abstractive Text Summarization
- Extractive Text Summarization

3.2 Abstractive Text summarization

In this technique, data is interpreted using advanced natural language processing-based approaches and then it tries to extract information and summaries it in such a way that the summarized part may or may not appear in the text but presents a clearer summarized text as shown in Figure 4. The part presented in the summarized version consists of the most critical information around which the whole context revolves. This technique takes advantage of sequence-to-sequence models which helps them understand the context line by line. Recent developments have helped this technique to improve its accuracy to a greater extent but this technique is still far from reaching the level where it can be used in a real-time production-based environment. Lots of researchers have recently taken a keen interest in this field and improvements are being made incrementally

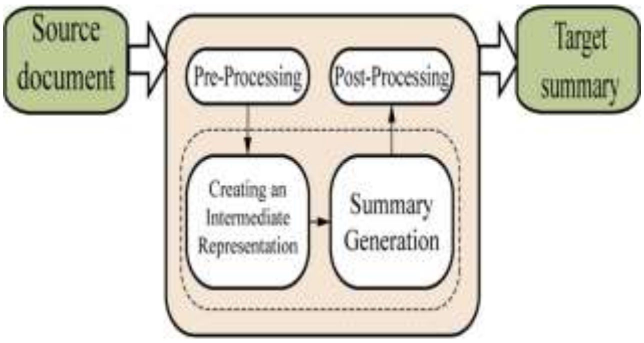


Fig. 4. Abstractive Text summarization

3.3 Extractive Text Summarization

According to this technique, direct sentences are chosen from the text. In this method, scores from the text content are directly correlated with the scoring algorithm to provide a summary based on those results. It contains the most important information from the text and can include pagination. A high probabilistic score ensures that the sentence will be chosen for the summary. Probabilistic scores are used in the scoring process. Summary is generated from top s sentences where s is the hyper parameter that can be adjusted based on the model performance and accuracy as shown in Figure 5.

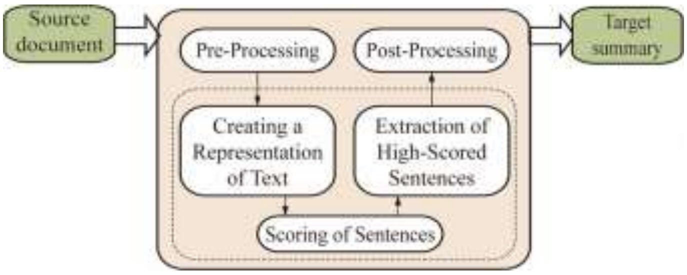


Fig. 5: Extractive Text Summarization

3.4 Lex rank

It is a graph-based approach that is used for summarizing text. It takes in any kind of text and converts it to summarized text. It detects relationships in the graph using clauset Newman algorithm. Its working can be classified into 7 stages.

- Input Document
- Document Cosine similarity score

DOI: rp-9788770040723.183

- Sentence converted to word embedding
- Adjacency matrix score
- Eigenvector Score
- Connectivity matrix
- Output document

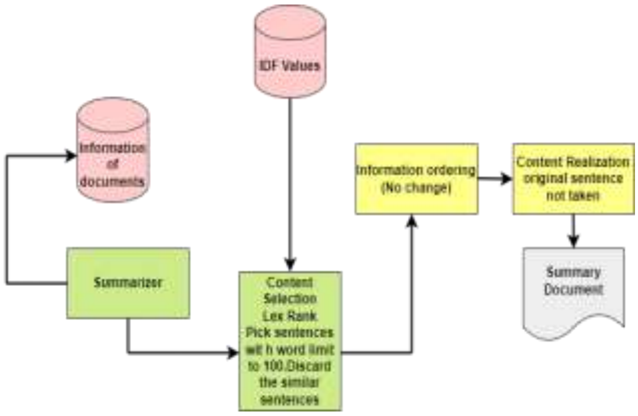

IDF Values
Information of documents
Summarizer
Content Selection Lex Rank Pick sentences wit h word limit to 100.Discard the similar sentences
Information ordering (No change)
Content Realization original sentence not taken
Summary Document

Fig. 6: Lex Rank summarizer

3.5 *Luhn*

Luhn text summarization is a method that aids in text summarization and is based on the TF-IDF approach. This method summarizes sentences based on the significance of the root words present in the sentence. After scoring these sentences then only the summarization is done. After scoring the sentences, ranking is done, and based on ranking text summarization is concluded.

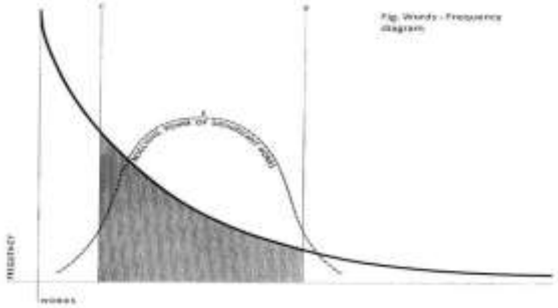


Fig. 7. Luhn summarizer

3.7 *Latent Semantic Indexing (LSA)*

LSA is Latent Semantic Indexing. It relies on extractive-based summarization. This technique tries to identify relationships between different words in sentences using a technique known as SVD i.e. Singular value decomposition. Using this it tries to weigh the importance of each word and then produce the desired ranking of words and summarizes the top ones.

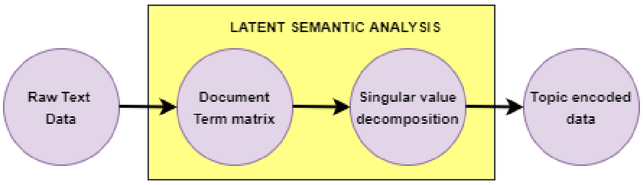


Fig. 8. LSA Summarizer

3.8 *Text Rank*

As the name implies, it is a ranking-based text summarization technique. It relies on extractive-based summarization. Using a graph-based unsupervised method, it breaks each sentence down into its constituent words and rates them. This technique produces superior results for the current problem statement.

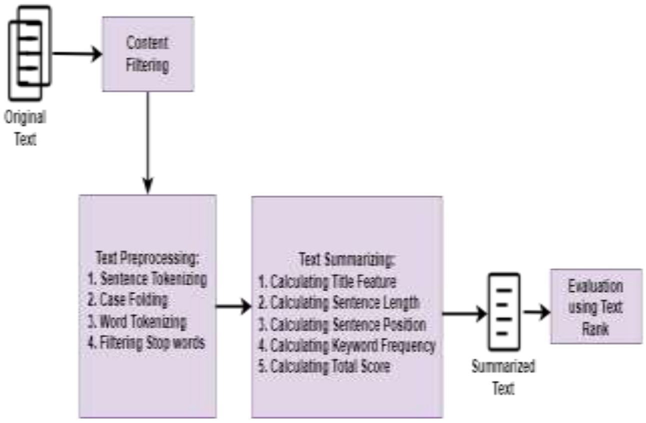

Original Text
Content Filtering
Text Preprocessing:
1. Sentence Tokenizing
2. Case Folding
3. Word Tokenizing
4. Filtering Stop words
Text Summarizing:
1. Calculating Title Feature
2. Calculating Sentence Length
3. Calculating Sentence Position
4. Calculating Keyword Frequency
5. Calculating Total Score
Summarized Text
Evaluation using Text Rank

Fig. 9. Text Rank

3.9 *Bidirectional encoder representation from the transformer (BERT)*

BERT stands for Bidirectional encoder representation from the transformer and is a pre-trained model developed by Google. This model comprises an encoder with the same structure as the transformer. It has 12 multi-heads, an embedding dimension of 784, and 12 repeated blocks. An even bigger configuration is available for this model known as BERT-Large. This model is used for training purposes and produces state-of-the-art results. The model learns on the training dataset using a transfer learning-based technique. All the data points are preprocessed using natural language processing and then passed in as input to the model. The model tries to learn the features of the data and then can be tested in real-time to produce desired results.

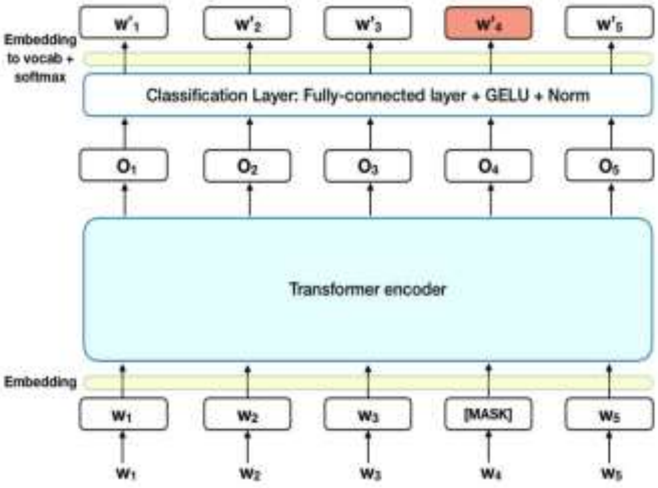

W'1
W'2
W'3
W'4
W'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
W1
W2
W3
[MASK]
W5
W1
W2
W3
W4
W5

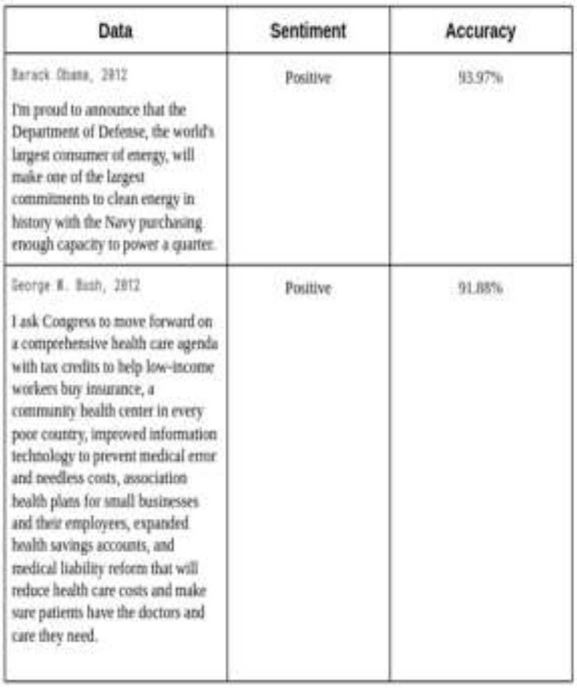
Fig. 10: BERT model

IV. Simulated Results

The significance and goal of the SOTU speech are important because it has a big impact on the content and message of presidential addresses. There are 26 Republicans and 16 Democrats, therefore it's important to note that the parties have 62 percent and 38 percent of the vote, respectively. The results here depict a text summarization of the president's speeches at the state of the union. All the speeches have been summarized using 4 major techniques: Luhn, lex rank, LSA, and Text rank. Although each technique for condensing speech has its own advantages, Text Rank yields the best results when used with this dataset.

Sentiment analysis is done on the president's speech and sentiment accuracy is depicted along with the label whether it is positive or negative. This sentiment accuracy shows how much the model is sure that a given label is valid or not. This functionality has been created for testing in real-time. A GUI has been developed to display the results. In fact, it takes in the input text and speech, produces it in real time with the aid of a trained model, and displays the outcomes as a pop-up.

TABLE 1: Simulated Results of Sentiment Analysis



Data	Sentiment	Accuracy
Barack Obama, 2012 I'm proud to announce that the Department of Defense, the world's largest consumer of energy, will make one of the largest commitments to clean energy in history with the Navy purchasing enough capacity to power a quarter.	Positive	93.97%
George W. Bush, 2012 I ask Congress to move forward on a comprehensive health care agenda with tax credits to help low-income workers buy insurance, a community health center in every poor country, improved information technology to prevent medical error and needless costs, association health plans for small businesses and their employers, expanded health savings accounts, and medical liability reform that will reduce health care costs and make sure patients have the doctors and care they need.	Positive	91.88%

V. Conclusions

Firstly, text analysis is done where the frequency of words, the density of sentences, and their importance were analyzed using natural language processing techniques. Understanding the dataset, and tradeoffs were covered. Secondly, text summarization was covered using different techniques such as Luhn, Text Rank, LSA, and Lex Rank. Both the summarization i.e. extractive and abstractive techniques were covered. Thirdly, Sentiment analysis was performed, and both machine learning and deep learning techniques were used. BERT model is used for conducting real-time sentiment analysis and UI is made for real-time testing of models. In the future, more datasets can be incorporated to train BERT Large versions. Multiple labels can be included such as very negative, negative, positive, and very positive. Other models can also be incorporated and reinforcement learning can also be applied by making sentiment analysis fun and interesting. Moreover, semi-supervised learning techniques can also be used to train the model on an unlabeled dataset.

References

[1] B. Pang, and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," Proceedings of the 42nd ACL, pp. 271-278, 2004.

[2] B. Pang, and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization concerning rating scales," Proceedings of the 43rd annual meeting on association for computational linguistics, pp. 115-124, 2005, DOI:10.3115/1219840.1219855

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Sentiment classification using machine learning techniques," In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79- 86, 2002, https://doi.org/10.3115/1118693.1118704

[4] R. L. Teten, "Evolution of the modern rhetorical presidency: Presidential presentation and development of the state of the union address," Presidential Studies Quarterly, vol. 33, no. 2, pp. 333–346, 2003, doi:10.1111/j.1741-5705.2003.tb00033.x.

[5] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds, in System Sciences (HICSS)," 2014 47th Hawaii International Conference on IEEE, pp. 1833–1842, 2014.

[6] C. Hutto, and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of social media Text," Proceedings of the International AAAI Conference on Web and social media, vol. 8, no. 1, pp. 216-225, 2014, Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[7] Helen Balinsky, Alexander Balinsky, and Steven J. Simske, "Automatic text summarization and small-world networks," In Proceedings of the 11th ACM symposium on Document engineering (DocEng '11). Association for Computing Machinery, New York, NY, USA, pp. 175–184, 2011, https://doi.org/10.1145/2034691.2034731

[8] "Machine Learning For Data Classification In Indonesia Regional Elections Based On Political Parties Support" Muhammad Fachrie.

[9] Jurnal Ilmu Komputer dan Informasi, Journal of Computer Science and Information, vol. 13, no 2, June 2022.

[10] P. P. Surya and B. Subbulakshmi, "Sentimental Analysis using Naive Bayes Classifier," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), pp. 1-5, 2019, doi: 10.1109/ViTECoN.2019.8899618.

[11] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data", In Proc. WLSM-11s, 2011.

[12] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," In Proc. EMNLP-09, 2009.

[13] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications.Computers and Electrical Engineering, 102, 108130.

[14] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," ICCL-10, 2010.

[15] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting Postpartum Changes in Emotion and Behavior via Social Media," In Proc. CHI-13, 2013.

[16] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," In Proc. ICWSDM-08, 2008.

[17] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system?" In Proc. CHI-10, 2010).

[18] A. Esuli, and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification", In Proc. ICIKM-05, 2005.

[19] C. Fellbaum, "WordNet: An Electronic Lexical Database," Cambridge, MA: MIT Press, 1998.

[20] J. T. Hancock, C. Landrigan, and C. Silver, “Expressing emotion in text-based communication,” In Proc. CHI-07, 2007.

[21] M. Hu, and B. Liu, “Mining and summarizing customer reviews,” In Proc. SIGKDD KDM-04, 2004.

[22] C. J. Hutto, S. Yardi, and E. Gilbert, “A Longitudinal Study of Follow Predictors on Twitter,” In Proc. CHI-13, 2013.

[23] J. Kamps, R. J. Mokken, M. Marx, and M. de Rijke, “Using WordNet to measure semantic orientation,” In Proc. LREC-04, 2004.

[24] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with Mechanical Turk,” In Proc. CHI-08, 2008.

[25] A. Kramer, An unobtrusive behavioral model of “grossnational happiness,” InProc. CHI-10, 2010.

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

[27] B. Liu, “Sentiment Analysis and Opinion Mining,” SanRa-fael, CA: Morgan & Claypool, 2012.

[28] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai, “Automatic construction of a context-aware sentiment lexicon: an optimization approach,” In Proc. WWW-11, 2011.

[29] F. A. Nielsen, “A new evaluation of a word list for sentiment analysis in microblogs,” In Proc. ESWC-11, 2011.

[30] B. Pang, and L. Lee, “A sentimental education: sentiment analysis using subjectivity summarization,” In Proc. ACL-04, 2004.

[31] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, “Recursive Deep Models for Semantic Compositionality Over Sentiment Treebank,” In Proc. EMNLP-13, 2013.

[32] Krupa, Abel Jaba Deva, Samiappan Dhanalakshmi, and R. Kumar. "An improved parallel sub-filter adaptive noise canceler for the extraction of fetal ECG." Biomedical Engineering/Biomedizinische Technik, vol. 66, no. 5, pp. 503-514, 2021.

[33] Pazhani. A, A. J., Gunasekaran, P., Shanmuganathan, V., Lim, S., Madasamy, K., Manoharan, R., &Verma, A. (2022).Peer–Peer Communication Using Novel Slice Handover Algorithm for 5G Wireless Networks.Journal of Sensor and Actuator Networks, 11(4), 82.

[34] K. Harisudha, and S. Dhanalakshmi et al., “Automated restricted Boltzmann machine classifier for early diagnosis of Parkinson's disease using digitized spiral drawings”, Journal of Ambient Intelligence and Humanized Computing, https://link.springer.com/article/10.1007/s12652-022-04361-3.

[35] S. Barui, S. Latha, D. Sammiappan, and P. Muthu, “SVM Pixel Classification on Colour Image Segmentation,” Journal of Physics Conference Series, 2018, 1000:012110. https://doi.org/10.1088/1742-6596/1000/1/012110.