iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47130

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Literature Review on Application of Natural Language Processing and Machine Learning Techniques for Risk Prediction of Mucormycosis

Kavitha.U[1], Babu Rao[2]
[1]*U-Research Scholar,* [2]*Professor in CSE & HOD IT/CTIS/CCE, School of Engineering and Technology, CMR University, Bengaluru, India*

Abstract: ***Mucormycosis is an infection caused by fungi capable of angio-invasion that is associated with high mortality and morbidity. Recently, An AI device was developed for detecting mucormycosis using Fourier Transform Infra-Red(FITR) sensors where human serum samples were tested using Fourier Transform Infra- Red Scanner. To obtain health outcome for various diseases, Clinical NLP has been extensively utilised. Word Embedding techniques and Feature Engineering Models in NLP convert clinical text data from EHR(Electronic Health Record) into useful format that serve as input for machine learning classification algorithms.***

Keywords: Mucormycosis, Natural Language Processing, Feature Engineering, Machine learning, Disease Risk Prediction.

I. INTRODUCTION

Mucormycosis, also known as Black Fungus is a rare infection more commonly found in moderate to severely immunocompromised people. During second wave of Covid-19 pandemic, a major surge was noted in cases of mucormycosis. Ever since the importance of incorporating NLP methods in healthcare are widely recognised over the past few years, there has been tremendous advances in health informatics. For disease risk prediction, application of clinical NLP along with Machine learning techniques prove to be highly beneficial.

Feature Engineering models along with Word Embedding techniques in Natural language processing converts the unstructured clinical text into a structured vector form that could be fed into machine learning algorithms to train them for classification and prediction purposes. Since unstructured text cannot be directly fed into machine learning algorithms, it is necessary to convert them into a structured format by using the feature engineering models.

In the following sections, Various feature engineering models and also the application of NLP and machine learning techniques for various disease diagnosis are reviewed.

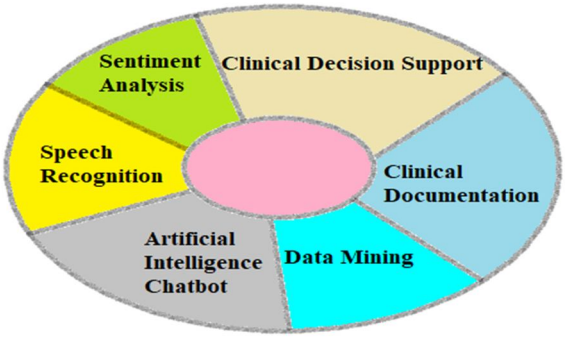


Figure 1 : Use Cases Of Natural Language Processing

Figure 1 Explanation: Natural language Processing in healthcare enhances the accuracy of clinical text records by transforming free text into standardized data.It helps to identify patients who need improved care and also addresses the need for higher quality healthcare.

II. FEATURE ENGINEERING MODELS

A. Bag-Of-Words(BOW) Model

Lei Wu et al describes that BOW model is an efficient technique for image representation and categorization. However, during codebook generation process which is an important step of BOW, most of the semantic information is lost and this is the critical disadvantage of existing BOW models. Therefore, a novel scheme was proposed aiming to map semantically related features to same visual words in order to learn optimized BOW models. Rui Zhao et al demonstrates that representing documents in form of numerical vectors is the main issue in NLP. However, this could be achieved by using a classical BOW model. Few limitations of BOW representation is its intrinsic extreme sparsity, inability to capture high semantic meanings behind text data and high dimensionality. Therefore, a new document representation method named Fuzzy Bag-Of-Words(FBOW) was proposed to overcome this limitation.

Teng Li et al demonstrates that on the basis of visual vocabulary in BOW model, an image can be represented by histogram of local patches. Bow has gained immense attention in visual categorization due to its good performance and flexibility. However, due to naïve Bayesian assumption conventional BOW neglects contextual relations between local patches. Therefore, a Contextual Bag-Of-Words(CBOW) model was proposed. This representation could model two kinds of typical contextual relations between local patches namely semantic conceptual relation and spatial neighboring relation.

Silva. F et al develops a BOW model that could encode the local structures of a digital object in form of graphs hence the model was named Bag-Of-Graphs(BOG). Two methods based on BOG were defined namely Bag-Of-Singleton Graph(BOSG) and Bag-Of-Visual Graph(BOVG).These methods were used to create vector representations for images and graphs. Josip. K et al introduces a representation which is derived using fisher kernel framework to encode spatial mean and variance of image regions that are associated with visual words. This BOW image representation is further extended by using gaussian mixture model for encoding spatial layout. Extensive experimental evaluation suggested that using fisher kernel to encode spatial layout resulted in efficient and compact image representation that also yields excellent performance while using linear classifiers.

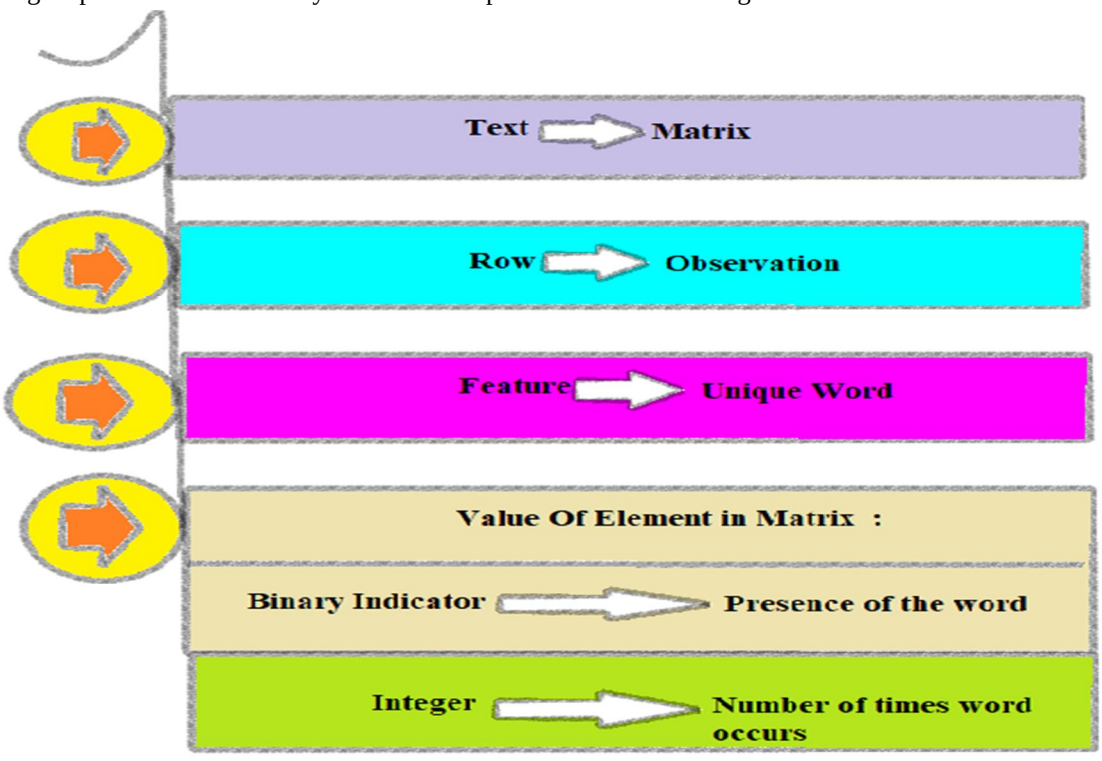


Figure 2 : Bag-Of-Words Representation

Figure 2 Explanation: BOW Model Converts text to a matrix where every row is an observation and every feature is a unique word. The value of each element in matrix is either a binary indicator based on the presence of the word or an integer indicating the number of times word appears in the document.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

B. Bag-Of-N-grams Model

Hajek. P et al proposed word-level n grams approach in order to find the similarity between text. Self-Organising Map(SOM) and Text Similarity Measures were the two techniques combined for this approach. Four measures were evaluated namely Cosine, Overlap, Extended Jaccard's and Dice. Text has been split into word -level n-grams to create a bag of n-grams in order to convert text into numerical expression. The filters used for creating the bag of n-grams were stemming algorithms, stop words and punctuation removers. Fusilier D.H et al proposed an approach to detect opinion spam wherein the character n-grams were used as features. Evaluation was performed on a corpus of 1600 hotel reviews consisting of both positive and negative reviews. Comparison was made between character n-grams and word n-grams. Results obtained showed that character n-grams were good features for opinion spam detection and they capture content of deceptive opinions and writing style of deceiver better than word n-grams.

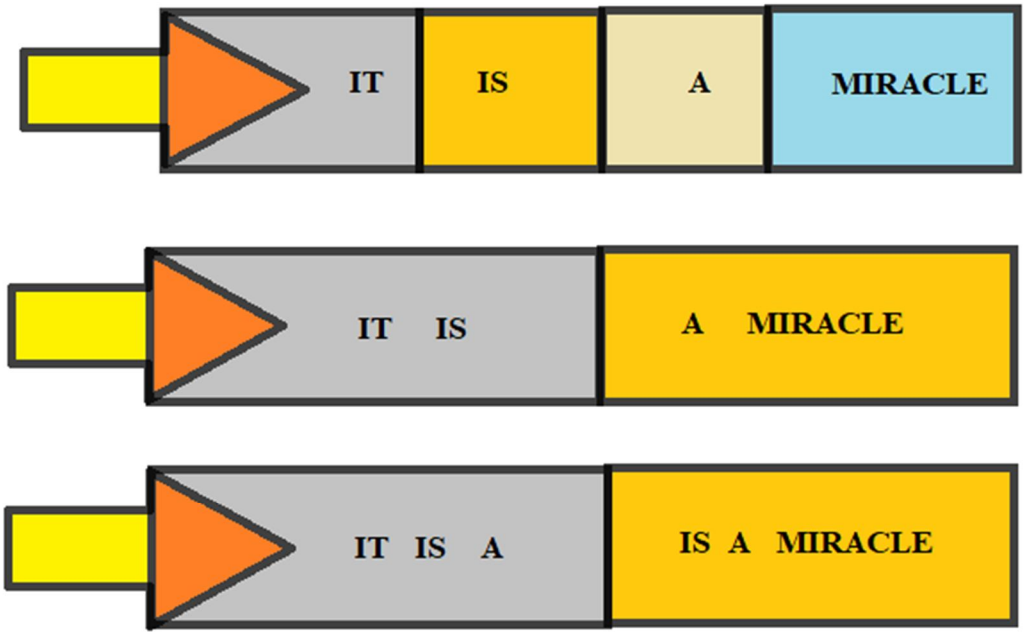


Figure 3 : Bag-Of-Words With N-grams

Figure 3 Explanation : Bag-Of-N grams Model is extensively used in text mining and NLP tasks. A contiguous sequence of n items from a given sample of text is referred as n-gram. Unigram is n-gram of size 1,bi-gram is of size 2 and tri-gram is of size 3.

Barushka. A et al proposed a novel content-based approach which considers both word context and bag-of-words. In order to build a vector model, this approach utilizes n-grams and skip-gram word embedding method and as an end result, it generates high dimensional feature representation. In the following step a deep feed-forward neural network handles the representation and classifies the review spam accurately. Two hotel review datasets with positive and negative datasets were used to verify the proposed system and the results showed that the proposed detection system outperforms other algorithms in terms of accuracy for review spam detection. Fatma. E et al proposed a novel bi-gram alphabet approach to construct feature terms to perform text classification. This approach resolves high dimensions of vector space and the need for language-dependent tools which is the main problem in BOW approach. The proposed approach has two main contributions to text classification namely reducing dimensions of vector space for large corpus and it does not require NLP tools. The work has also proved ability to classify Arabic/English documents collections successfully. Bongares.F et al describes that automation speech recognition system based on Driven Decoding Algorithm(DDA) involves use of 1-best hypothesis that is provided by auxiliary system in search algorithm of a primary system. A new method is proposed to manage auxiliary transcriptions in form of bag-of-n-grams representation without temporal matching. Different auxiliary systems provide several hypotheses, these combinations of hypotheses are made easier by modifications suggested in the proposed new method.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

C. TF-IDF Model

Dey.A et al demonstrates that the existing methods use only TF-IDF rating to represent unigram or n-gram feature vectors. Whereas, Some approaches stress on using the sentiment dictionaries and score of unigram sentiment word therefore, ignore TF-IDF rating. Hence, n-gram sentiment features are constructed by extracting sentiment words and their intensifiers from a review. Then its score is multiplied to TD-IDF rating to determine the feature score.

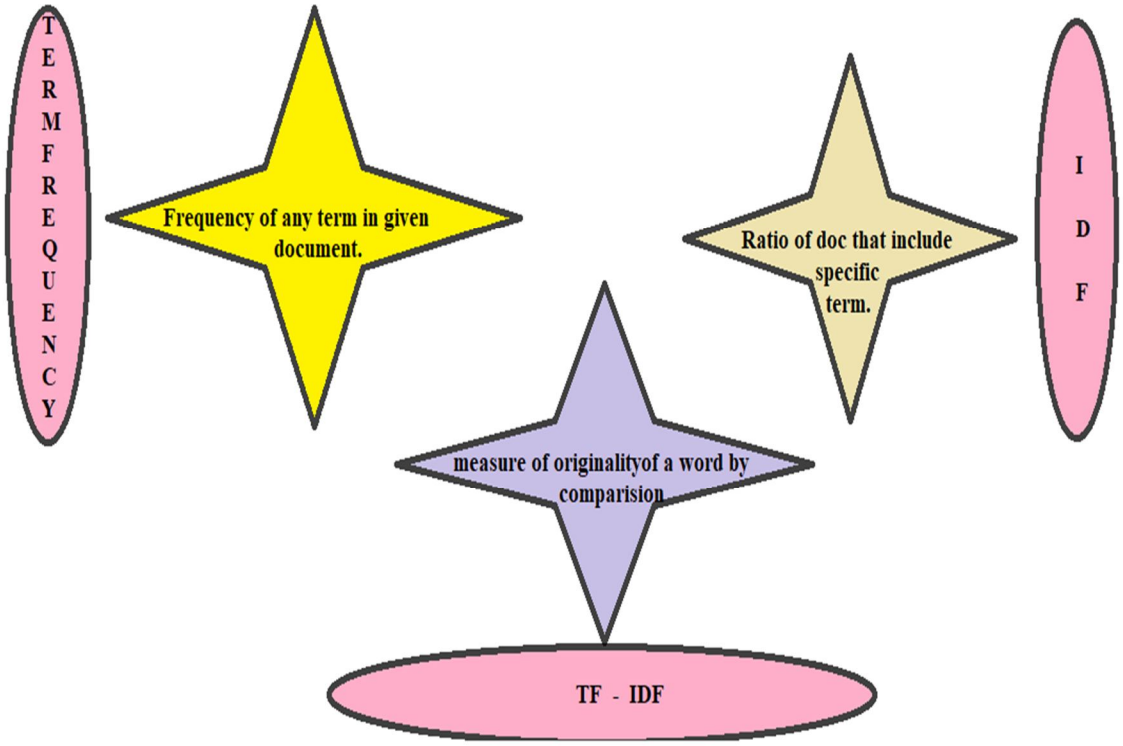


Figure 4 : Term Frequency-Inverse Document Frequency In NLP

Figure 4 Explanation : The Frequency of any term in a given document is termed as Term-Frequency. The ratio of document that include specific term is termed as Inverse Document Frequency. The measure of originality of a word by comparing number of times word appears in a document versus number of documents the word appears is termed as TF-IDF.

Yun-Tao et al proposed a new improved TF-IDF approach to enhance the recall and precision of text classification by using confidence, support and characteristic words. Improved TF-IDF approach processes lexicons that define synonyms. Experimental results suggest that new TF-IDF approach provided improved precision and recall for text classification when compared to conventional or traditional TF-IDF approach.

Wen.Z et al conducted a comparative study of LSI,TF*IDF and multi-word for text classification. The three methods were evaluated on Chinese and English document collections for information retrieval and text categorization. Experimental results suggested that LSI has better performance than other methods in text categorization. LSI has also proved to be best in retrieving English documents. Bafna.P et al demonstrates that in order to analyse the research papers, there is a need to classify the repositories according to the topic. Experiments were conducted on various real and artificial datasets. Hierarchy algorithm, fuzzy k-means and TF-IDF algorithm were applied on small dataset and cluster analysis was performed. Further, the best algorithm was applied on the extended dataset. Bruno T.et al demonstrates the possibility of using KNN algorithm with TF-IDF based framework for text classification. The classification was enabled by framework according to various parameters, measurement and analysis of results. Based on the speed and quality of classification, the framework was evaluated. Experimental tests conveyed both good and bad features algorithm, which led way to further development of similar frameworks.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

D. Word2Vec Model

Lilleberg.J et al demonstrates that word2vec has a new approach on text classification by converting words and phrases into vector representation. The proposed work demonstrates that TF-IDF and word2vec combined can outperform TD-IDF alone because word2vec provides complementary features. The conclusion was that the combination of two can outperform the performance when used individually. Long.M et al demonstrates that word2vec model which was proposed and supported by google consists of two learning models namely Continuous-Bag-Of-Words(CBOW) and skip-gram. The text data is given as input to any one of the above learning models, the output from word2vec would be word vectors that could be represented as a large piece of text. In the proposed work, data is trained and then its word similarity is evaluated. Similar words are clustered together and the generated clusters are used to fit into new data dimension in order to decrease data dimension.

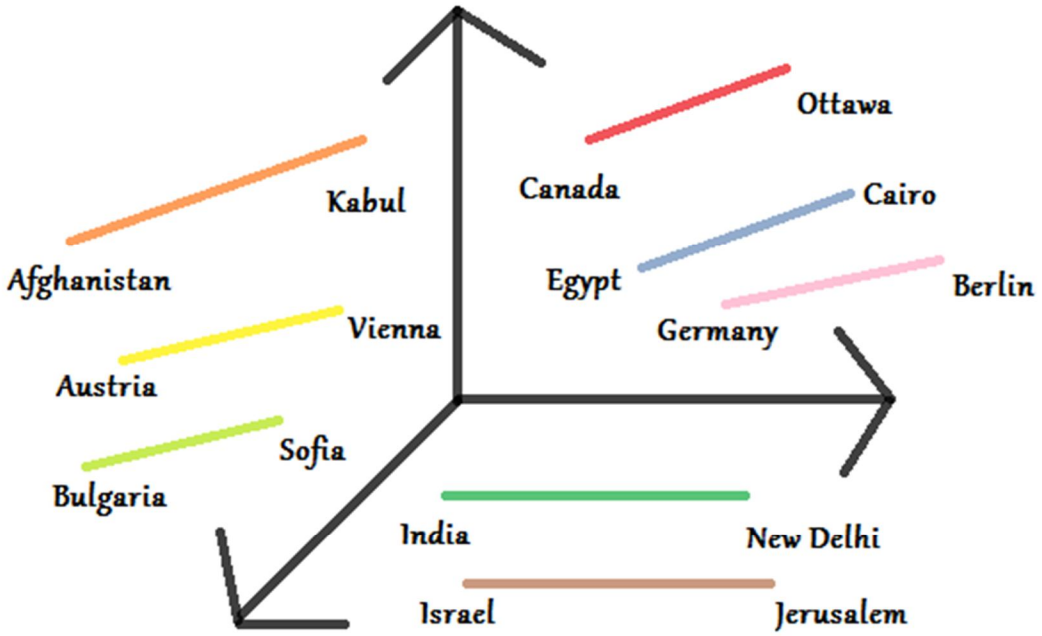


Figure 5 : Country and Capital in Word2Vec Representation

Figure 5 Explanation: Word2Vec takes as its input a large corpus of text and produces a vector space of several hundred dimensions with each unique word in corpus being assigned a corresponding vector in the space. Words that share common contexts in corpus are located close to one another in vector space.

Dongwen.Z et al demonstrated a sentiment classification method based on word2vec and SVM^{Perf}. Similar features clustering was achieved using word2vec, and also deep semantic features were extracted. It was evaluated that SVM^{Perf} trains faster and predicts more accurate than other SVM packages. More than 90% accuracy was achieved by the classification result. Derry. J et al demonstrates in the proposed work, the calculation of the similarity between words in English were done by using word representation techniques. Word2vec model was formed using 3,20,000 articles in English Wikipedia as corpus and then similarity value is determined by cosine similarity calculation method. The model was tested by the test set gold standard wordsim-353 and Simlex-999.Pearson correlation was used to find out accuracy of the correlation. Bofang Li et al demonstrates that in the proposed work,word2vec is scaled on a GPU cluster. Reducing dependencies inside a large training batch is one main challenge. A variation of word2vec is pdesigned heuristically to ensure that each word-context pair contains uniformly sampled contextual word and a non-dependent word. This variation also fixes the number of samples thereby controlling the training iterations and treats both high and low frequency words equally.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

E. Glove Model

Yash. S et al demonstrates that one of the application of NLP is sentiment analysis. It enables in understanding the common language of people. The sentiments of the users are deciphered to understand their liking and disliking. Unsupervised technique Glove that can represent words in form of vectors is very effective in interpreting both the meaning and sentiments. Ru Ni et al demonstrated that deep learning can build a sentence sentiment analysis system based on word vector space model and recurrent neural network. To make the best use of global and local information for training corpus, Glove word model could be used. A neural network model combining both LSTM and GRU was proposed in the work. Experimental results suggested that LSTM-GRU model performs the best.

Flora. S et al demonstrates that Glove embeddings have been widely used for various NLP tasks and text mining due to their high quality as textual features. In the proposed work, an efficient algorithm has been introduced that produces word embeddings enhanced by semantic information. The proposed algorithm utilizes semantic information during training or post processing steps and outperforms other related approaches. Experiments validate that the proposed model improved the quality of word vector representations. Seyed.M et al demonstrates that word embedding methods are mostly used for sentiment classification. Among which Word2vec and Glove are most accurate methods for converting words into meaningful vectors. The proposed method in the work is based on part-of-speech tagging techniques, lexicon based approaches and Word2vec/Glove methods. The proposed method named Improved Word Vectors(IWV) increases the accuracy of existing pre-trained word embeddings. The experiment results show that IWV is very effective for sentiment analysis.

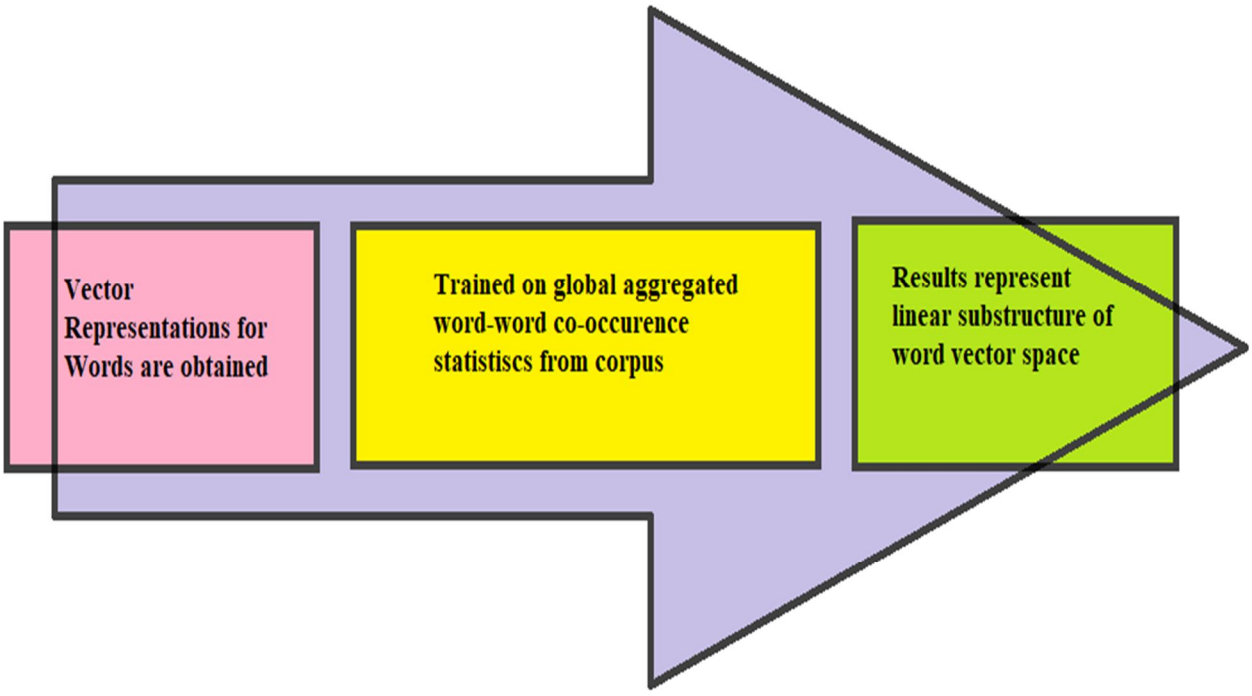


Figure 6 : Glove In Natural Language Processing

Figure 6 Explanation: Glove is an unsupervised learning algorithm that obtains vector representations for words. The training is performed on global aggregated word-word co-occurrence statistics from a corpus. The resulting representations showcase interesting linear substructure of word vector space.

Aytug.O et al demonstrates that web consists of rich information source where many text documents are judged with opinions and reviews. This sentiment recognition can be useful for decision maker's government and business organizations. A deep learning based approach for sentiment analysis on product reviews obtained from twitter were presented. A combination of TF-IDF weighted Glove word embedding with CNN-LSTM architecture was presented. Predictive performance of different word embedding schemes with several weighting functions were evaluated along with deep neural network architectures. Results show that proposed deep learning architecture outperforms conventional deep learning method.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

F. Fast Text Model

Igor. S et al demonstrates that recent work shows that for Natural Language Processing, CNN Performs well. The main concept is to gather embeddings as an image. In the proposed work, the task of sentiment analysis is performed by using recently released Facebook Fast Text word embeddings. The results proved that the proposed approach performs better than the baseline models, and similar performance like state-of-art models. Aziz. A et al demonstrates that to achieve good results in NLP tasks, quality of word representation plays important role. Better sentiment analysis results could be achieved using CBOW and Skip-gram models by building sentiment-specific embeddings. The results showed that FastText embedding models are excellent alternative to extract semantic and syntactic information of Arabic and dialectal language.

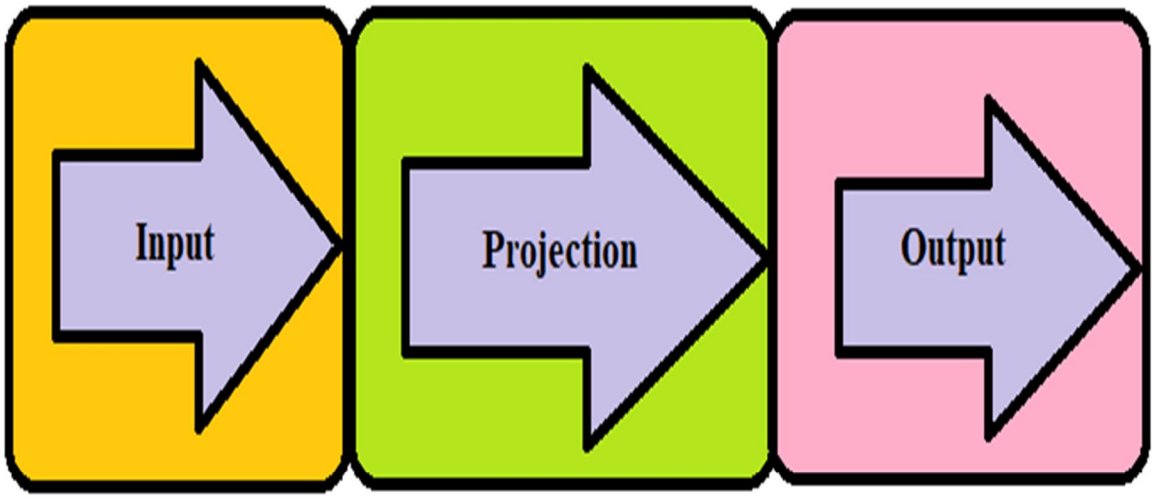


Figure 7 : Layout of FastText Model

Figure 7 Explanation: Fast Text Library enables efficient learning of word representations and text classification. FastText supports supervised classifications and unsupervised embeddings representation of words and sentences. These representations can be used for applications such as data compression, candidate selection, features into additional models and as initializers for transfer learning.

Sajeetha T et al demonstrates that SentiwordNet is the sentiment lexicon for determining the sentiment of texts. Grouping sentiment words that are not in Sentiword Net is a tough task. Hence a sentiment lexicon expansion method using Word2vec and FastText along with rule-based sentiment analysis method is used. Two steps were followed to expand sentiment lexicon from initial seed list, it is by gathering related words using Word2vec and also then by gathering lexical similar words using FastText. The final results showed that the accuracy of 88% was obtained. Lal.K et al demonstrates that the primary objective of the study is to develop a benchmark dataset for resource-deprived Urdu language for sentiment analysis and evaluate various machine and deep learning algorithms for sentiment analysis. Two modes of text representation are compared namely Count based and FastText pretrained word embedding. Experiments were conducted with a set of machine learning classifiers for all feature types. Results signify that the combination of word n-gram features with LR outperformed other classifiers for sentiment analysis. Salur.M et al proposed a novel hybrid deep learning model that combines Word2vec,FastText and character-level embedding along with different deep learning methods, where features are extracted and combined and texts are classified in terms of sentiment series of experiments were performed by several deep learning models called basic models for verifying the performance of the proposed model. Experimental results showed that through comparison, the proposed model offered better sentiment classification.

G. BERT Model

Li et al proposed a method based on BERT, an automatic text classification method along with feature fusion. BERT model transforms text-to-dynamic character-level embedding. BiLSTM and CNN output features are combined and merged in order to make complete use of CNN and extracts advantages of local features and BiLSTM has memory to link extracted context features for better text representation and improves accuracy of text classification. Experimental results signify proposed method performed better than state-of-art methods in accuracy. Lu et al Proposed a model which combines capability of BERT with vocabulary Graph Convolutional Network(VGCN) hence the model was named VGCN-BERT. A final representation for classification was built by allowing mutual influence of local information and global information interacting through different layers of BERT. Experimental results show that the proposed approach outperformed BERT and GCN alone when the proposed model was applied on several text classification datasets.

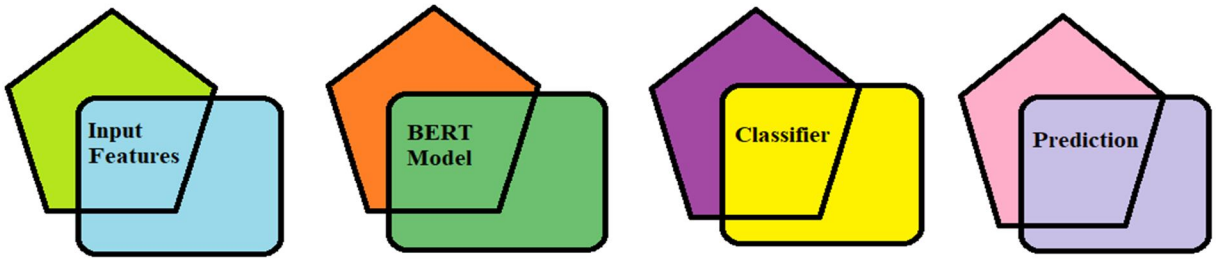


Figure 8 : BERT Model For Prediction

Figure 8 Explanation: BERT Models extract high quality language features from the input text data and these models are fine-tuned to perform classification task with the data to produce state-of-art predictions.

Sun. C et al demonstrates that language model pretraining is useful for learning universal language representation. BERT has achieved amazing results in many language understanding tasks. Exhaustive experiments were conducted to investigate different fine-tuning methods of BERT on text classification task and a general solution for BERT fine-tuning was provided. New state-of-art results were obtained when proposed solution was applied on eight widely studied text classification datasets.

Zheng. S et al presented a model for text classification based on BERT-CNN. Information about important fragments in text can be obtained from the proposed model by adding CNN to task specific layers of BERT model. Finally to get representation of whole text through transformer layer, local representation is given as input along with output of BERT into transformer encoder. Experimental results show that proposed model performs better than state-of-art baselines when applied to four datasets. Yu.S et al demonstrates that BERT model Pre-trained on text corpus, Book corpus and Wikipedia achieves excellent performance on a couple of NLP tasks, however it lacks task specific knowledge and domain related knowledge. Therefore, for improving the performance of BERT model, fine-tuning strategy analysis is necessary. A BERT-based text classification model BERT4TC was proposed which aimed at addressing limited training data problem and task-awareness problem. The architecture and implementation details of BERT4TC were also presented along with post-training approach for addressing domain challenge of BERT. Experimental results showed that the proposed model outperforms both feature-based and fine-tuning methods.

1) Application of NLP and Machine Learning Techniques for Disease Diagnostics

Qingcai. C et al proposed a hybrid system to automatically identify heart disease risk factors. According to the descriptions, different types of risk factors are divided into three categories .Evaluation results showed that the proposed hybrid system achieved a F-score of 92.86% on 2014 i2b2 corpus, which is top-ranked. Surabhi.A et al demonstrates that alzheimer's disease is a slow progressive disease that affects the cognitive abilities of people. Memory deteriorates in alzheimer's disease patients and as disease progresses ,speech is completely impaired. stopwords are most commonly used words by alzheimer's patients. In the proposed work, the stopwords used in capturing linguistic information of alzheimer's patients are discussed. Evaluation of learning algorithms are done by comparing them during pre-processing stage through inclusion and dropping of stopwords. Ravi. G et al demonstrates that the study used natural language processing of electronic health records in combination with machine learning methods to automate IS Subtype classification.

Analysis of unstructured text-based EHR data of neurology progress notes and neuroradiology reports using natural language processing. Several Feature selection methods were performed to reduce high dimensionality of features and overfitting is minimized. Andres. A et al presented an innovative recommendation system for diagnosis of breast cancer using patient's medical histories. NLP techniques were implemented on real datasets, medical histories from a hospital for breast cancer,cysts,other cancers, surgeries and diagnosis. Another dataset from MIMIC III word embedding techniques-word2vec's skip gram, BERT and ML techniques were used to design a recommendation system to support physician's decision-making. Results signified that use of word embedding defined a good quality recommendation system.

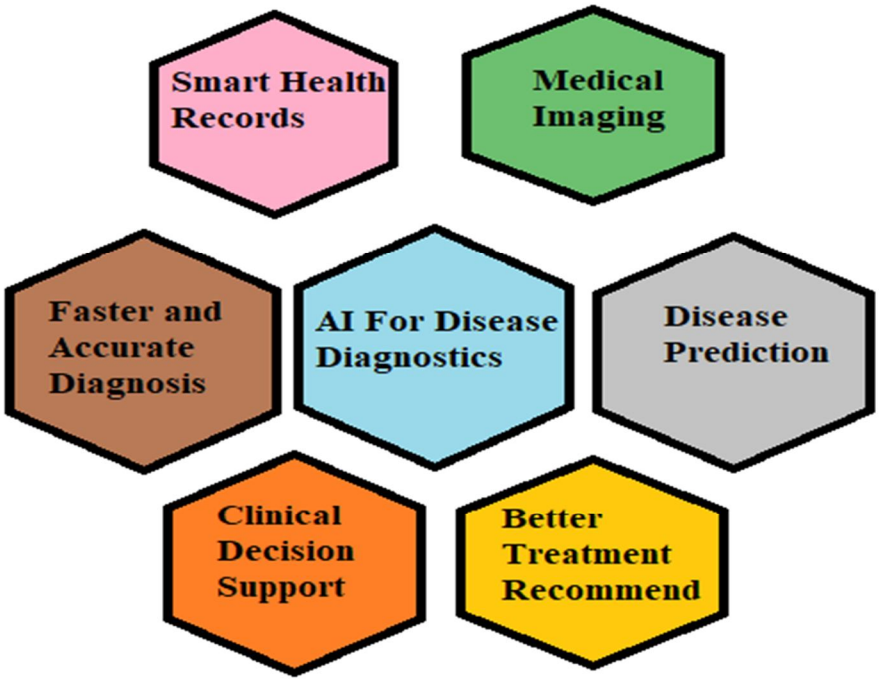


Figure 9 : Artificial Intelligence For Disease Diagnostics

Figure 9 Explanation: Artificial Intelligence in healthcare is a term that describes the use of machine-learning algorithms to imitate human cognition in analysis, presentation and comprehension of healthcare data. Artificial Intelligence in healthcare includes drug development, personalised medicine, patient monitoring and care.

Ashkan. E et al proposed many models that were built by several machine learning algorithms. Multiple data sources like PubMed and Arxiv were used for the study. The latent topics were identified and temporal evolution of extracted research themes of Covid-19 were analysed, along with publications similarity, sentiments within time frame of Jan-May 2020.The results signified that PubMed exhibited greater diversity in Covid-19 related issues and Arxiv focused more on intelligent tools to predict or diagnose Covid-19.Special attention was given to high risk groups and people with complications.

III. CONCLUSION

There has been plenty of research done in natural language processing of extracting information from clinical text of electronic health records. However, there is not much research done on obtaining health outcome for mucormycosis using NLP and machine learning techniques. There is scope for applying deep learning techniques using neural networks for mucormycosis risk prediction. Future research work suggested could be applying a word embedding model with deep learning neural network concept to produce a diagnostic prediction model for mucormycosis.

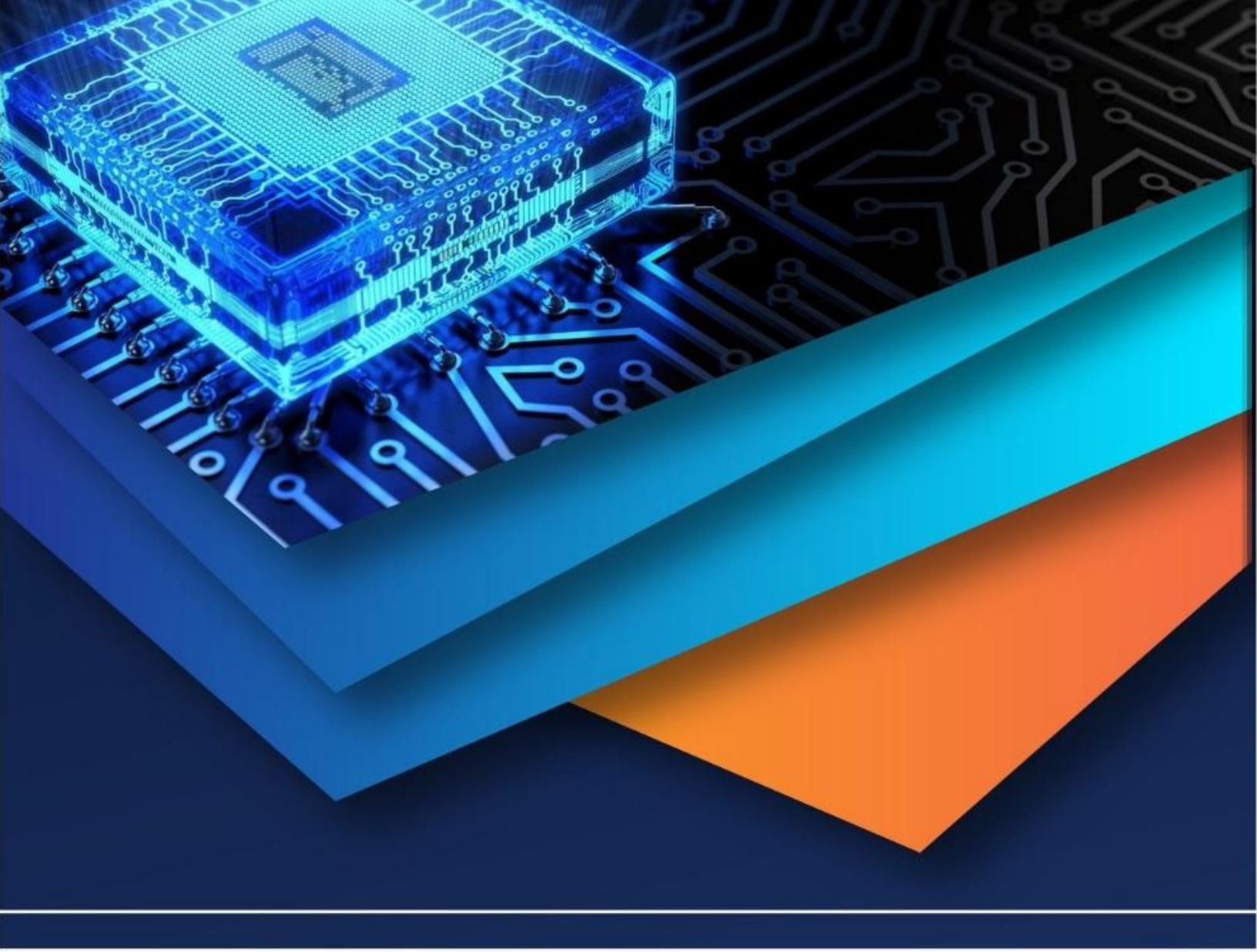
IV. ACKNOWLEDGEMENTS

Authors thank Directorate Of Research & Innovation (DORI), CMR University for training and support. The Research is funded by CMRU Student Research & Innovation Fund.

REFERENCES

[1] Prakash, H., & Chakrabarti, A. (2019). Global Epidemiology of Mucormycosis. Journal of fungi (Basel, Switzerland), 5(1), 26. https://doi.org/10.3390/jof5010026
[2] Jetty, Dilipkumar Varma & Mohammed, Ishaq Azhar. (2021). AI Device to Identify Black Fungus disease -Post Covid-19 Pandemic Infections. SSRN Electronic Journal. 10. 1-8.
[3] Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., & Dutta, R. (2018). Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. Journal of biomedical informatics, 88, 11–19. https://doi.org/10.1016/j.jbi.2018.10.005
[4] Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. JMIR medical informatics, 8(3), e17984.shttps://doi.org/10.2196/17984
[5] Lei Wu, Hoi, S. C., & Nenghai Yu. (2010). Semantics-preserving bag-of-Words models and applications. IEEE Transactions on Image Processing, 19(7), 1908-1920. https://doi.org/10.1109/tip.2010.2045169
[6] Zhao, R., & Mao, K. (2018). Fuzzy bag-of-Words model for document representation. IEEE Transactions on Fuzzy Systems, 26(2), 794-804. https://doi.org/10.1109/tfuzz.2017.2690222
[7] Li, T., Mei, T., Kweon, I., & Hua, X. (2011). Contextual bag-of-Words for visual categorization. IEEE Transactions on Circuits and Systems for Video Technology, 21(4), 381-392. https://doi.org/10.1109/tcsvt.2010.2041828
[8] Silva, F. B., Werneck, R. D., Goldenstein, S., Tabbone, S., & Torres, R. D. (2018). Graph-based bag-of-words for classification. Pattern Recognition, 74, 266-285. https://doi.org/10.1016/j.patcog.2017.09.018
[9] Krapac, J., Verbeek, J., & Jurie, F. (2011). Modeling spatial layout with Fisher vectors for image categorization. 2011 International Conference on Computer Vision. https://doi.org/10.1109/iccv.2011.6126406
[10] Hajek, P., Barushka, A. & Munk, M. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Comput & Applic 32, 17259–17274 (2020). https://doi.org/10.1007/s00521-020-04757-2
[11] Fusilier, D. H., Montes-y-Gómez, M., Rosso, P., & Cabrera, R. G. (2015). Detection of opinion spam with character N-grams. Computational Linguistics and Intelligent Text Processing, 285-294. https://doi.org/10.1007/978-3-319-18117-2_21
[12] Barushka, A., & Hajek, P. (2019). Review spam detection using word embeddings and deep neural networks. IFIP Advances in Information and Communication Technology, 340-350. https://doi.org/10.1007/978-3-030-19823-7_28
[13] Elghannam, F. (2021). Text representation and classification based on bi-Gram alphabet. Journal of King Saud University - Computer and Information Sciences, 33(2), 235-242. https://doi.org/10.1016/j.jksuci.2019.01.005
[14] Bougares, F., Esteve, Y., Deleglise, P., & Linares, G. (2011). Bag of n-Gram driven decoding for LVCSR system harnessing. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. https://doi.org/10.1109/asru.2011.6163944
[15] Dey, A., Jenamani, M., & Thakkar, J. J. (2017). Lexical TF-IDF: An n-Gram feature space for cross-domain classification of sentiment reviews. Lecture Notes in Computer Science, 380-386. https://doi.org/10.1007/978-3-319-69900-4_48
[16] Zhang, Y., Gong, L., & Wang, Y. (2005). An improved TF-IDF approach for text classification. Journal of Zhejiang University SCIENCE, 6A(1), 49-55. https://doi.org/10.1631/jzus.2005.a49
[17] Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765. https://doi.org/10.1016/j.eswa.2010.08.066
[18] Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. Procedia Engineering, 69, 1356-1364. https://doi.org/10.1016/j.proeng.2014.03.129
[19] Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). https://doi.org/10.1109/icci-cc.2015.7259377
[20] Ma, L., & Zhang, Y. (2015). Using Word2Vec to process big text data. 2015 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata.2015.7364114
[21] Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 42(4), 1857-1863. https://doi.org/10.1016/j.eswa.2014.09.011
[22] Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2Vec model analysis for semantic similarities in English words. Procedia Computer Science, 157, 160-167. https://doi.org/10.1016/j.procs.2019.08.153
[23] Li, B., Drozd, A., Guo, Y., Liu, T., Matsuoka, S., & Du, X. (2019). Scaling Word2Vec on big corpus. Data Science and Engineering, 4(2), 157-175. https://doi.org/10.1007/s41019-019-0096-6
[24] Sharma, Y., Agrawal, G., Jain, P., & Kumar, T. (2017). Vector representation of words for sentiment analysis using glove. 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT). https://doi.org/10.1109/intelcct.2017.8324059
[25] Ni, R., & Cao, H. (2020). Sentiment analysis based on glove and LSTM-GRU. 2020 39th Chinese Control Conference (CCC). https://doi.org/10.23919/ccc50068.2020.9188578
[26] Sakketou, F., & Ampazis, N. (2020). A constrained optimization algorithm for learning glove embeddings with semantic lexicons. Knowledge-Based Systems, 195, 105628. https://doi.org/10.1016/j.knosys.2020.105628
[27] Onan, A. (2020). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. Concurrency and Computation: Practice and Experience, 33(23). https://doi.org/10.1002/cpe.5909
[28] Santos, I., Nedjah, N., & De Macedo Mourelle, L. (2017). Sentiment analysis using convolutional neural network with fastText embeddings. 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). https://doi.org/10.1109/la-cci.2017.8285683
[29] Altowayan, A. A., & Elnagar, A. (2017). Improving Arabic sentiment analysis with sentiment-specific embeddings. 2017 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata.2017.8258460
[30] Thavareesan, S., & Mahesan, S. (2020). Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. 2020 Moratuwa Engineering Research Conference (MERCon). https://doi.org/10.1109/mercon50084.2020.9185369

[31] Khan, L., Amjad, A., Ashraf, N., Chang, H., & Gelbukh, A. (2021). Urdu sentiment analysis with deep learning methods. IEEE Access, 9, 97803-97812. https://doi.org/10.1109/access.2021.3093078
[32] Salur, M. U., & Aydin, I. (2020). A novel hybrid deep learning model for sentiment classification. IEEE Access, 8, 58080-58093. https://doi.org/10.1109/access.2020.2982538
[33] Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., & Li, W. (2019). The automatic text classification method based on BERT and feature union. 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). https://doi.org/10.1109/icpads47876.2019.00114
[34] Lu, Z., Du, P., & Nie, J. (2020). VGCN-BERT: Augmenting BERT with graph embedding for text classification. Lecture Notes in Computer Science, 369-382. https://doi.org/10.1007/978-3-030-45439-5_25
[35] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? Lecture Notes in Computer Science, 194-206. https://doi.org/10.1007/978-3-030-32381-3_16
[36] Zheng, S., & Yang, M. (2019). A new method of improving BERT for text classification. Intelligence Science and Big Data Engineering. Big Data and Machine Learning, 442-452. https://doi.org/10.1007/978-3-030-36204-1_37
[37] Chen, Q., Li, H., Tang, B., Wang, X., Liu, X., Liu, Z., Liu, S., Wang, W., Deng, Q., Zhu, S., Chen, Y., & Wang, J. (2015). An automatic system to identify heart disease risk factors in clinical texts over time. Journal of Biomedical Informatics, 58, S158-S163. https://doi.org/10.1016/j.jbi.2015.09.002
[38] Garg, R., Oh, E., Naidech, A., Kording, K., & Prabhakaran, S. (2019). Automating ischemic stroke subtype classification using machine learning and natural language processing. Journal of Stroke and Cerebrovascular Diseases, 28(7),2045-2051. https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004
[39] Adhikari, S., Thapa, S., Singh, P., Huo, H., Bharathy, G., & Prasad, M. (2021). A comparative study of machine learning and NLP techniques for uses of stop words by patients in diagnosis of Alzheimer's disease. 2021 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn52387.2021.9534449
[40] Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R., & Wong, A. (2020). Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. Scientometrics, 126(1), 725-739. https://doi.org/10.1007/s11192-020-03744-7
[41] Magna, A. A., Allende-Cid, H., Taramasco, C., Becerra, C., & Figueroa, R. L. (2020). Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. IEEE Access, 8, 106198-106213. https://doi.org/10.1109/access.2020.3000075



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY