JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 646-653
https://publishoa.com
ISSN: 1309-3452

Feature Extraction based Text Classification: A review

Saif safaa shaker[1], Dhafer Alhajim[1], Ahmed Ali Talib Al-Khazaali[1],
Hussein aqeel hussein[1], Ali. F. Athab[2]
[1]Computer Technical Engineering Department, College of Technical Engineering, The Islamic University, Najaf, Iraq.
[2] Ministry of Communication, ITPC, Iraq.
saif9410@iunajaf.edu.iq, idcte202006@iunajaf.edu.iq, aata8803@yahoo.com,
Hussein_Aqeel@iunajaf.edu.iq, l.fadhel1983@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10.

Abstract.

This article reviews and discusses feature extraction techniques used for text classification as well as natural language processing in biomedical applications.This researchaims to analyze the similarities of techniques used as technology and algorithms that have become more sophisticated to optimize feature extraction. In feature extraction, a specific of words is taken out from text data. After that, they transform into a feature set to be usable by a classifier. Several algorithms have been identified for classification,including but not limited to SVM, deep neural networks as well as Naïve Bayes algorithms. Next, the natural language is processed and achieves better performance results as indicated by certain metrics like execution time, specificity, accuracy, specificity,and sensitivity.

Keywords; natural language processing, feature extraction, text classification.

Introduction.

Coupling machine learning concepts with Natural Language Processing (NLP) (Ofer et al., 2021) is important in digitalizing data. We are compelled to perform in-depth research of this data in various domains as data value keeps altering. For about ten years now, natural language processing has become more important (Shankar et al., 2022) as it shows much-unseen information in the texts. A huge volume of text data makes it difficult sometimes to see the information. Therefore, it is necessary to extract information based on computational text processing with the help of methods of real-time web servers that gather and distribute information (Shnain et al., 2021) and Improve Information Retrieval in a Digital Library Management System (Aliwy et al., 2021). The use of text could be for biomedical or clinical purposes. Feature extraction is an essential subset of data features that aims to enhance the classification task. Therefore, features can be extracted for the documents' classification (Fajardo et al., 2021).In a text, it is crucial to identify the related feature correctly. So, implementing NLP methods is helpful for better studying and understanding the data(Athab et al., 2020).

Feature Extraction.

Extracting information from text or feature extraction helps to analyze (Vijayakumar et al., 2021) the text data for many applications, namely automated terminology management, data mining, research subject identification, clinical records, and studying the effect of research on them.To extract the important features, feature Extraction techniques are proposed (Mutlag et al., 2020) withthe help of common feature selection and feature extraction methods. These techniques determine the effectiveness, achieve better performance of learning algorithms, and ultimately improve the prediction accuracy of classifiersin many fields, for example, security(Alhajim et al., 2022). In healthcare and biomedical domains, several text mining techniques and tasks, including clustering and classification and text data pre-processing.

In the literature, there are many feature selection methods such as information gain, document frequency, and feature selection techniques, such as latent semantic indexing (LSI) (Nagamura et al., 2021), principal component analysis (PCA) and Clustering uses NLTK dictionary and K-mean algorithm(Ibrahimet al., 2021).

Text Classification.

Texts and speeches are usually the most common form of unstructured data as they involve plenty but are hard to extract useful information. NLP allows us to mine information and analyze this data and perform tasks such as identifying fake news, spam filtering, cognitive assistant, sentiment analysis, and real-time language translation as it understands the parts or texts of speech. Using NLP, the text classification deals with classifying (Lavanya et al., 2021) the text into words group to classify and automatically analyze text. Based on its context, it assigns a set of predefined categories or tags.

There are mainly three text classification approaches which are Rule-based System (Favi et al., 2021) , Machine System (Gedara et al., 2021) and Hybrid System (Li et al., 2021). Using a set of handicraft linguistic rules, texts are separated into an organized group in the rule-based approach. The used rules consist of users determining a list of words groups characterized. For instance, Boris Johnson and Donald Trump are categorized into politics. On the other hand, Ronaldo and Missi should be categorized into sports.

Based on past data sets' observations, a machine-based classifier learns to make a classification. As test data, user data is prelabeled and stored. From the previous inputs, it collects the classification strategy and learns continuously. For feature extension, a machine-based classifier uses a bag of a word. Words frequency is represented by a vector in a predefined word list dictionary in a bag of words. NLP can be performed with machine algorithms, such as Deep Learning, SVM, and Naïve Bayer. For text classification, Hybrid Approach is the third approach. Its usage combines machine Based and rule-based approach. The hybrid-based approach uses the rule-based system to create a tag. Additionally, to train the system and create a rule, it uses machine learning. Next, the lists of machine-based rule with the rule-based rule are compared. The list shall be improved manually if (sometimes) the tags do not match.

Natural Language Processing with classification

Natural language processing has been widely used for many applications such as smartphones, speakers, cars, computers, websites, and health applications (Bera et al., 2021). As an NLP system, a machine translator is utilized by Google Translator. NLP supports Google Translator, utilizing spoken natural language users seek to translate and supports removing extra noises, building CNN to understand native voice, and understanding words in context. Because it decreases the human work of asking what needs of customers, NLP is also popular in very useful chatbots. NLP chatbots ask sequential questions,such as what the user problem is and where to find the solution. In their systems, a robust chatbot is available on AMAZON and Apple (Locke et al., 2021). In case of any questions, the chatbot converts them into understandable phrases in the internal system using a token. Then, to get the users' ideas are asking, the token goes into NLP.

In information retrieval (IR) the NLP is utilized as well. The IR is a software program that involves massive storage. Additionally, it evaluates information from large text documents from repositories.Many researchers attempted to improve IR systems by using that either named entity recognition (NER) methodology or the meaning of words (word sense) and then implementing the improvements in a specific language (Aliwy et al., 2021). Only relevant information is retrieved. For example, to trim unnecessary words, it is utilized in Google voice detection. NLP is used majorly in machine translation, i.e. question answering, information retrieval, Google Translator, mining large data, social media analysis, sentiment analysis, summarization, and ChatBot.

A Literature Review.

The results of the literature are shown in Table 1

Ref.	Data Set	Algorithm Type	Features	Results
(Fu et al., 2020)	The data set consists of contains protected health information and real clinical notes. So, it cannot be available in public.	word2vec, tf-idf. For classification, the algorithms invested were Naive Bayes, Random Forest and SVM.	Window size, stop word removal, stemming	Compared with spaCy packages and parsedatetime, the recall of date extraction and accuracy is improved using the timex.py-adapted regex function
(Alfattni et al., 2020)	Two publicly available corpora only	Deep Neural Networks and SVM and Conditional Random Fields	PubMed and the DBLP computer science bibliography TLINK tasks	Rule-based methods were effective when used for extracting; they quickly achieve baseline performance, and when high-quality text with consistent grammatical the form is available.
(Kumar et al., 2021)	Lung molecular event-level extraction dataset	Radial Belief Neural Network	Automatic feature selection	The Radial Belief Neural Network (RBNN) is suggested to achieve higher performance outcomes, such as Execution time, F-measure, Specificity, Sensitivity, and Accuracy.
(Han et al., 2022)	Clinical notes	Bidirectional Encoder Representations from Transformers (BERT), long short-term memory (LSTM) network, convolutional neural network (CNN), and Deep neural network (DNN).	Automate extraction of Social determinants of health (SDOH)	From clinical notes in the HER, improved performance for efficiently identifying a systematic range of SDOH categories. Additionally, Healthcare outcomes may further be improved by improving the identification of patient SDOH.
(Peterson et al., 2020)	Medical records	A neural network classification	Fast Healthcare Interoperability Resource (FHIR) models	an F1 score of 0.95.
(Ambalavanan et al., 2020)	Known as Clinical Hedges, manually annotated dataset of ~49 K MEDLINE abstracts	Neural network	Ensemble architectures	In screening scientific articles, pre-trained neural contextual language models (e.g., SciBERT) performed well.

Based on the table above, with natural language processing tools and techniques, Fu et al., (2020) attempted to extract dates and classify them as relevant diagnoses or not. The overall pipeline performance is affected by incomplete and inaccurate data interpretation and extraction. The performance of these tools was plagued by limitations of inaccuracy and incomplete extraction of information, which would directly affect the interpretation and ultimately impact the performance of the overall process. In the clinical text, according to Fu et al., 2020 existing software packageslike Python have low specificity capabilities for interpreting and identifying partial and relative dates. Therefore, they proposed a rules-based regular expression (regex) approach that managed to achieve a recall of 83.0% on manually annotated diagnosis dates as well as 77.4% on all annotated dates. Besides, initial MPN diagnoses are represented with only 3.8% of annotated dates. To alleviate noise and class imbalance, additional methods target candidate date instances. Comparing the performance of classification algorithms and feature extraction methods is a relevant practice. The preprocessing steps operated on stopword removal, window size, stemming and the feature extraction used word2vec, tf-idf. For classification, the algorithms invested were Naive Bayes, Random Forest, and SVM. In the pipeline, the tests demonstrated better configurations and provided F1 scores of 39.1 (SVM), 13.1 (Naive Bayes), and 0.3 (Random Forest). The analysis of Subsequent errors revealed that it was largely because of inaccurate data extraction. Compared to the parsedatetime and spaCy packages, the timex.py-adapted regex function improved the recall and accuracy of date extraction.

According to Alfattni et al., (2020), it is challenging to extract temporal relations from free text data since it lies between temporal reasoning, temporal representation, and medical natural language processing. For extracting temporal relations, they surveyed existing methods involving the DBLP computer science bibliography and a systematic search in PubMed for papers published over 13 years (from 2006 to 2018); the relevant papers were selected by examining the titles and abstracts. After that, the full text of chosen papers was analyzed in-depth, and information was gathered on TLINK types, TLINK tasks, methods used, features selection, reported, data sources, and performance. Two thousand eight hundred thirty-four publications were the outcome, which were identified to screen their title and abstract. Fifty-one papers from these publications were chosen. The papers used as follows: 32, 15, and 4 used for machine learning approaches, hybrid approaches, and a rule-based approach, respectively. For extracting TLINKs, deep Neural Networks and Machine learning classifiers (such as Conditional Random Fields and SVM) were among the best-performing methods. However, on two publicly available corpora only, nearly all the work has been carried out and tested. The field would benefit from more public available datasets, high-quality, annotated clinical text corpora.

To feed the machine learning classifier, the authors of Kumar, et al. (2021) proposed a model that used a rich set of extracted features to get a better extract of the events. For optimal molecular biomedical event detection, the authors used a classification model and an automatic feature selection using Radial Belief Neural Network (RBNN). To give accurate disease detection results, the RBNN was implemented as the classifier.So the algorithms were implemented to improve the scalability and generalization performance scalability of detecting the molecular event triggers. The authors used the RBNN model with a lung molecular event-level extraction dataset to validate the cystic fibrosis event trigger based on the gene ontology biosystem. The extensive computation demonstrated that the RBNN achieved higher performance in terms of Execution time, F-measure, Specificity, and Accuracy.

Natural language processing was used by Ha et al., (2022) to automate the extraction of Social determinants of health (SDOH) from the text. Usually, it is emphasized on an ad hoc selection of SDOH, and the latest advances in deep learning are not used. Their objective was to advance the automatic extraction of SDOH from a clinical text by (a) Creating a set of SDOH systematically, which is based on psychiatric ontologies and standard biomedical and (b) Extracting mentions of these SDOH from a clinical note by training state-of-the-art deep neural networks are conducted. They designed a framework for the automated classification of multiple SDOH categories. In the MIMIC-III Clinical Database, their dataset comprised narrative clinical notes under the “Social Work” category. They systematically curated a set of 13 SDOH categories and created annotation guidelines for these using standard terminologies, DSM-IV, and SNOMED-CT. for automated detection of eight SDOH categories, and after manually annotating the 3,504 sentences, theydesigned and examined the Bidirectional Encoder Representations from Transformers (BERT), long short-term memory (LSTM) network, convolutional neural network (CNN), and three deep neural network (DNN) architectures. The performance of identifying a systematic range of SDOH categories has enhanced by the framework-based DNN models. Consequently, healthcare outcomes may be further improved by improving the identification of patient SDOH.

Discussion of Results Section.

Given the techniques discussed in this review article for clinical concept extraction, the applications to clinical data and records all strive to optimize the feature extraction as well as the language processing of clinical data. As seen in the results obtained in the experiments and research, more accuracy was achieved by using classification algorithms like the deep neural network, SVM as well as Naive Baines algorithms. For most of the experiments, the F1 score achieved was above 90%. However, the actual methods adopted for a specific task were impacted by five factors which are namely data and resource availability, domain adaptation, model interpretability, system customizability, and practical implementation.

Conclusion - Limitation - Recommendation - Future Work.

In conclusion this study has reviewed the techniques used to achieve better accuracy and optimization when it comes to feature extraction (for text classification and natural language processing. Finding and reordering and sifting information or specific scientific data in a large collection is an important natural language processing challenge in the biomedical domain. Systematic searches and classification algorithms which help with screening information for a combination of selection criteria prove to give better results compared to manual and traditional methods of feature extraction. While machine learning has been harnessed to make the process of classifying and retrieving specific information easier in biomedical applications. The review identified number of limitations predominantly on the scarcity of datasets, basically additional datasets would broaden the results. Secondly using a limited number of datasets or just one could make the research susceptible to bias as there is little to none cross referencing of the results.

References

1. Fu, J. T., Sholle, E., Krichevsky, S., Scandura, J., & Campion, T. R. (2020). Extracting and classifying diagnosis dates from clinical notes: a case study. Journal of Biomedical Informatics, 110, 103569.
2. Alfattni, G., Peek, N., & Nenadic, G. (2020). Extraction of temporal relations from clinical free text: A systematic review of current approaches. Journal of Biomedical Informatics, 108, 103488.
3. Kumar, R. D., Arvind, C., & Srihari, K. (2021). Extraction of the molecular level biomedical event trigger based on gene ontology using radial belief neural network techniques. Biosystems, 199, 104313.
4. Han, S., Zhang, R. F., Shi, L., Richie, R., Liu, H., Tseng, A., ... & Tsui, F. R. (2022). Classifying Social Determinants of Health from Unstructured Electronic Health Records Using Deep Learning-based Natural Language Processing. Journal of biomedical informatics, 103984.
5. Peterson, K. J., Jiang, G., & Liu, H. (2020). A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. Journal of Biomedical Informatics, 110, 103541.
6. Ambalavanan, A. K., & Devarakonda, M. V. (2020). Using the contextual language model BERT for multi-criteria classification of scientific articles. Journal of Biomedical Informatics, 112, 103578.
7. Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal, 19, 1750-1758.
8. Shankar, V., & Parsana, S. (2022). An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. Journal of the Academy of Marketing Science, 1-27.
9. Fajardo, J. M., Gomez, O., & Prieto, F. (2021). EMG hand gesture classification using handcrafted and deep features. Biomedical Signal Processing and Control, 63, 102210.
10. Athab, A.F. and Daghal, A.S., 2020, November. IoT system to limit speed rate by using gps and rf devices. In IOP Conference Series: Materials Science and Engineering (Vol. 928, No. 3, p. 032077). IOP Publishing.
11. Vijayakumar, T., Vinothkanna, R., & Duraipandian, M. (2021). Fusion based feature extraction analysis of ECG signal interpretation–a systematic approach. Journal of Artificial Intelligence, 3(01), 1-16.
12. Mutlag, W. K., Ali, S. K., Aydam, Z. M., & Taher, B. H. (2020, July). Feature extraction methods: a review.

In Journal of Physics: Conference Series (Vol. 1591, No. 1, p. 012028). IOP Publishing.

13. Zhang, X., Jiang, X., Jiang, J., Zhang, Y., Liu, X., & Cai, Z. (2021). Spectral–Spatial and Superpixelwise PCA for Unsupervised Feature Extraction of Hyperspectral Imagery. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-10.
14. Nagamura, Y., Arima, K., Arai, M., & Fukumoto, S. (2021). Layout Feature Extraction Using CNN Classification in Root Cause Analysis of LSI Defects. IEEE Transactions on Semiconductor Manufacturing, 34(2), 153-160.
15. Favi, C., Garziera, R., & Campi, F. (2021). A rule-based system to promote design for manufacturing and assembly in the development of welded structure: Method and tool proposition. Applied Sciences, 11(5), 2326.
16. Millagaha Gedara, N. I., Xu, X., DeLong, R., Aryal, S., & Jaberi-Douraki, M. (2021). Global Trends in Cancer Nanotechnology: A Qualitative Scientific Mapping Using Content-Based and Bibliometric Features for Machine Learning Text Classification. Cancers, 13(17), 4417.
17. Lavanya, P. M., & Sasikala, E. (2021, May). Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey. In 2021 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 603-609). IEEE.
18. Li, X., Cui, M., Li, J., Bai, R., Lu, Z., & Aickelin, U. (2021). A hybrid medical text classification framework: Integrating attentive rule construction and neural network. Neurocomputing, 443, 345-355.
19. Bera, A., Ghose, M. K., & Pal, D. K. (2021). Sentiment Analysis of Multilingual Tweets Based on Natural Language Processing (NLP). International Journal of System Dynamics Applications (IJSDA), 10(4), 1-12.
20. Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: a review. Trends in Anaesthesia and Critical Care, 38, 4-9.
21. Alhajim, D., Akbarizadeh, G. and Ansari-Asl, K., 2022, February. FFDR: Design and implementation framework for face detection based on raspberry pi. In *2022 International Conference on Machine Vision and Image Processing (MVIP)* (pp. 1-4). IEEE.
22. Ibrahim, R.K., Zeebaree, S.R., Jacksi, K., Sadeeq, M.A., Shukur, H.M. and Alkhayyat, A., 2021, July. Clustering Document based Semantic Similarity System using TFIDF and K-Mean. In 2021 International Conference on Advanced Computer Applications (ACA) (pp. 28-33). IEEE.
23. Shnain, A.H., Hussain, A., Mohammed, W.A., Ghanimi, H.M., Shaheed, S.H. and Sabri, M.I., 2021, September. Real Time Web Server Aggregator to Collect Fresh Information Based on Multi-Services. In 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA) (pp. 175-178). IEEE.

24. Aliwy, A., Abbas, A., & Alkhayyat, A. (2021). NERWS: Towards Improving Information Retrieval of Digital Library Management System Using Named Entity Recognition and Word Sense. Big Data and Cognitive Computing, 5(4), 59.