The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications

Mahmoud Elbattah[1], Émilien Arnaud[2], Maxime Gignon[2] and Gilles Dequen[1]
[1]*Laboratoire MIS, Université de Picardie Jules Verne, Amiens, France*
[2]*Emergency Department, Amiens-Picardy University, Amiens France*

Keywords: Text Analytics, Natural Language Processing, Unstructured Data, Healthcare Analytics.

Abstract: The implementation of Data Analytics has achieved a significant momentum across a very wide range of domains. Part of that progress is directly linked to the implementation of Text Analytics solutions. Organisations increasingly seek to harness the power of Text Analytics to automate the process of gleaning insights from unstructured textual data. In this respect, this study aims to provide a meeting point for discussing the state-of-the-art applications of Text Analytics in the healthcare domain in particular. It is aimed to explore how healthcare providers could make use of Text Analytics for different purposes and contexts. To this end, the study reviews key studies published over the past 6 years in two major digital libraries including IEEE Xplore, and ScienceDirect. In general, the study provides a selective review that spans a broad spectrum of applications and use cases in healthcare. Further aspects are also discussed, which could help reinforce the utilisation of Text Analytics in the healthcare arena.

1 INTRODUCTION

> *"Most of the knowledge in the world in the future is going to be extracted by machines and will reside in machines",* (LeCun, 2014).

The above-mentioned statement describes the ever-rising abundance of data-driven knowledge, which continuously calls for further utilisation of Machine Learning (ML). By the same token, healthcare is delivered in data-rich environments where a broad variety of data sources can be created at the individual and population levels. The format of heath data ranges from Electronic Health Records (EHR) to images, time series, or unstructured textual notes.

Data Analytics has been increasingly considered as an enabling artefact to leverage health data for competitive advantage. Using a diversity of ML techniques, analytics has been widely utilised to summarise, explain, and get insights into the interrelationships underlying complex datasets in novel ways. Such insights can play a positive role in various medical and operational aspects including diagnosis, health monitoring and assessment, healthcare planning, and management of hospitals and health services.

However, one of the key challenges for healthcare analytics is to deal with huge data volumes in the form of unstructured text. Examples include nursing notes, clinical protocols, medical transcriptions, medical publications, and many others. In this respect, the use of Text Analytics has increasingly come into prominence in order to deliver benefits for health organisations in a wide range of applications.

Text Analytics, or Text Mining, is generally defined as the methodology followed to derive quality and actionable insights from textual data (Sarkar, 2019). Text Analytics represents an overarching field of techniques and technologies including Natural Language Processing (NLP), ML, and Information Retrieval. The power of Text Analytics is to extract information that could allow for forming and exploring new facts or hypotheses from unstructured textual data (Hearst, 1999).

Compared to conventional tasks, the obvious challenge of Text Analytics is to extract patterns from natural-language text, rather than well-structured databases. Textual data are largely stored in an unstructured form, which does not adhere to any pre-defined schema or data model. Further, standard ML algorithms were genuinely crafted to deal with numeric data. As such, Text Analytics need to apply

Elbattah, M., Arnaud, É., Gignon, M. and Dequen, G.
The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications.
DOI: 10.5220/0010414508250832
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 5: HEALTHINF*, pages 825-832
ISBN: 978-989-758-490-9

especially designed techniques and transformations to effectively operate over textual data.

The potentials of NLP have been constantly discussed in the healthcare literature (e.g. Demner-Fushman, Chapman, and McDonald, 2009; Jensen, Jensen, and Brunak, 2012; Spasić, Uzuner, and Zhou, 2020). In this respect, the main motivation for this study was to explore the recent developments and applications in this context. The study provides a selective review that spans a broad spectrum of the applications and use cases of Text Analytics in the healthcare domain particularly.

2 REVIEW METHODOLOGY

The review aimed to explore the state-of-the-art approaches and applications of Text Analytics in the healthcare context. We were generally motivated by a set of exploratory questions as below:

- What are the potential data sources for applying Text Analytics in healthcare?
- What are the recent technological advances in implementing Text Analytics in this context?
- How could Text Analytics help healthcare providers make better decisions?
- What are the challenges of integrating NLP tools into healthcare systems?
- What are the key limitations of Text Analytics in the healthcare domain?

The review incorporated two main stages. The initial stage included the screening and selection of studies retrieved from the search results. Subsequently, we analysed a set of representative studies to be included in the literature review. The study sought to largely follow the procedures of a systematic literature review as informed by (Booth, Sutton, and Papaioannou, 2011).

The search of literature was conducted to find relevant studies in two major digital libraries including: i) IEEE Xplore, and ii) ScienceDirect. It is acknowledged that other relevant studies could have been published in other conferences or journals, but we believe that the selected venues generally provided excellent representative studies. The review timeframe stretched through the past 6 years (i.e. 2015-2020).

The inclusion of studies was conducted over a three-step process for screening and classifying studies. First, potential studies were screened based on the title. Second, the abstracts were initially inspected to confirm the suitability for full-text review. Eventually, the final decision of inclusion was made based on the full-text inspection. Figure 1 sketches a flowchart of the review process. Table 1 summarises the search strategy.

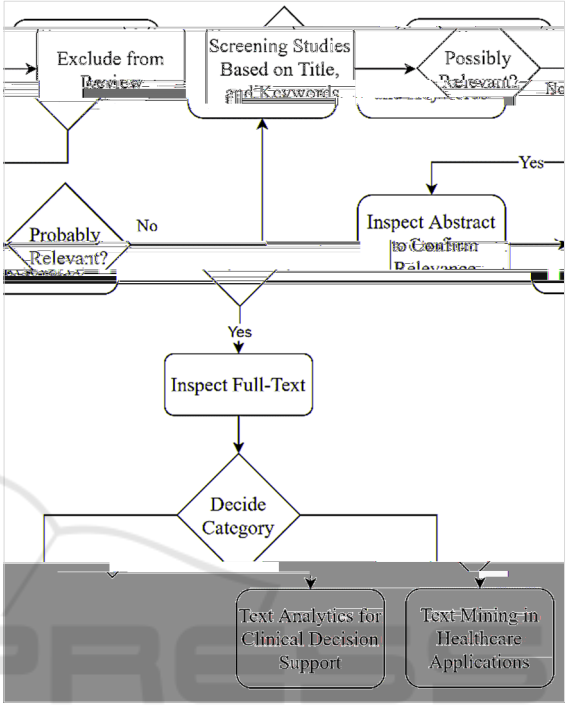

Figure 1: The process of screening and selecting studies in the review.

Table 1: Summary of search strategy.

Digital Libraries	IEEE Xplore, ScienceDirect
Search Terms	Text Analytics Healthcare, Text Mining Healthcare, NLP Healthcare
Search Items	Title, Abstract, Keywords
Types of Document	Conference Proceedings, Journal Articles
Timespan	2015-2020
Language	English

3 REVIEW ANALYSIS

This section aims to provide an analysis of the studies reviewed. The search results included about 200 publications overall. Eventually, a set of 35 studies were included in the review based on the process of screening and analysis as described before.

The review is organised into two broad categories of Text Analytics. On one hand, the first part presents

selective studies that applied Text Mining in the context of healthcare. On the other hand, the second part describes Text Analytics in a diversity of predictive applications to support the clinical decision making. The review is unavoidably selective rather than exhaustive. However, it is believed that the study could adequately provide representative studies in each category.

3.1 Text Mining Applications in Healthcare

Text Mining consists of two phases as follows. The initial phase typically includes the application of text refining procedures, which transform free-text documents into another intermediate form. Subsequently, the process of knowledge extraction, which attempts to learn patterns or insights from that intermediate form (Tan, 1999). This section provides selective studies that applied Text Mining with different modalities and for various purposes in the healthcare context.

(Han, Nandan, and Sun, 2015) presented a rule-based system for question retrieval. The goal was to search for similar questions in a large corpus of questions posted on online health forums. The system was mainly based on the RAKE algorithm (Rose, Engel, Cramer, and Cowley, 2010) to perform the automatic extraction of keywords. Additional NLP methods were applied using the popular NLTK library (Bird, Klein, and Loper, 2009).

In another application of Text Mining, a study aimed to develop automated methods for extracting information from the application webpages on the iTunes App Store (Paglialonga, Riboldi, Tognola, and Caiani, 2017). The study considered around 86K applications under the categories of Medicine, and Health/Fitness. They used the NLP capabilities provided by the IBM Watson API to identify the medical specialty (e.g. cardiology, nutrition, neurology, etc.), and the type of sponsor (e.g. industry manufacturer, or government organisation). Likewise, (Paglialonga et al., 2017) applied Text Mining to automate the extraction of meaningful information about health apps on the web.

(Lieder et al., 2019) developed a system that could mine millions of public business webpages to extract a multi-faceted representation of customers. In addition, the extracted data were enriched with external information collected from Wikipedia. In this respect, a large-scale knowledge graph was constructed including millions of inter-connected entities, which could be continuously enriched and connected to new entities. The system could be applied to industry use cases, such as healthcare, to support insight discovery in real time.

In addition, several studies applied Text Mining to extract information or insights from online forums or discussions. For instance, (Sutar, 2017) presented an interesting application of Text Mining to extract healthcare-related information from the user-generated content on social media. Using a dataset from a cancer-related forum, they developed a system that could be used to extract practical information such as treatments, medication names, and side effects. The dataset included a set of unstructured and semi-structured textual fields. Similarly, (Deng, Zhou, Zhang, and Abbasi, 2019) proposed a framework to support the analytics of online discussions. The framework was named as Discussion Logic-based Text Analytics (DiLTA). The DiLTA framework attempted to extract features that could reveal the discussion logic underlying online forums. The framework was experimented using a case study related to healthcare forums.

(Martínez et al., 2016) discussed exploiting the health-related online content into actionable knowledge using Text Mining. To this end, they developed an approach to help monitor online user-generated streams on social Media. An NLP-based processing pipeline was applied to extract and transform information stemming from real-time streams of social media. The system could not only extract the mention of diseases and drugs, but also it could identify useful relationships among medications, indications, and adverse drug reactions.

(James, Calderon, and Cook, 2017) analysed unstructured textual feedback of physicians. They aimed to extract sentiments and topics pertaining to the quality of healthcare service. Specifically, they attempted to identify the tones and topics that could shape the service ratings. In this regard, more than 20K patient reviews of more than about 4K physicians were analysed using the Latent Dirichlet Allocation (LDA) method. Further, a dictionary-based text analysis was applied to determine the tone elements in the physician reviews.

(Pendyala, and Figueira, 2017) explored the potentials of Text Mining for automating the medical diagnosis. They study applied the Bag-of-Words representation to medical documents. To simplify the text representation, the Bag-of-Words model builds a histogram of the words, while each word count is considered as a feature (Goldberg, 2017). As such, each document can be simply represented as a “bag” of words, while disregarding the order, sequence, and grammar of text. Though using a small dataset, their experiments demonstrated promising results for that

application. More recently, (van Dijk et al., 2020) applied Text Mining to EHR data to validate the screening eligibility of trial patients. The study was based on a multi-centre, and multi-EHR systems as well. The accuracy of the Text-Ming approach was compared to the standard process produced by research personnel. The accuracy of the automatically extracted data was about 88.0%.

(Chang et al., 2016) developed a workflow using Text Mining to search, extract, and synthesise information about Comparative Effectiveness Research (CER) in healthcare. The study included the development of an NLP-based pipeline to extract information from unstructured CER data sources. The Text-Mining solution could allow for the generation of timely alerts, and the collection of systematic reviews as well. Their approach was experimented using trial data from multiple sources including ClinicalTrials.gov, WHO International Clinical Trials Registry Platform (ICTRP), and Citeline Trialtrove.

While other contributions focused on exploiting Text Mining techniques for extracting concepts and association rules from the scholarly literature. For instance, (Kumari, and Mahalakshmi, 2019) applied Text Mining to a subset of the biomedical literature on PubMed. They aimed to discover information related to the phytochemical properties of medicinal plants. In another application, (Ji, Tian, Shen, and Tran, 2016) developed a scalable approach to extract associations among biomedical concepts in scientific articles. Biomedical concepts were derived by matching the text elements with the Unified Medical Language System (UMLS) thesaurus. A MapReduce-based algorithm was used to calculate the strength of associations. The experimental dataset included a large set of about 34K full-text articles. Their results generally demonstrated that meaningful association rules were highly ranked.

Recent studies considered more sophisticated implementations based on the Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art NLP model (Devlin, Chang, Lee, and Toutanova, 2019). The BERT approach brings the advantage of allowing pre-trained models to tackle a broad set of NLP tasks. In this regard, (Peterson, Jiang, and Liu, 2020) developed a framework for transforming free-text descriptions into a standardised form based on the Health Level 7 (HL7) standards. They utilised a combination of domain-specific knowledgebases in tandem with the BERT models. It was demonstrated that the BERT-based language representation contributed significantly to the model performance. Likewise, the literature includes recent contributions that made use of the BERT approach for a variety of Text Mining tasks such as (Fan, Fan, and Smith, 2020), (Liao et al., 2020), and (Vinod et al., 2020).

Furthermore, a major part of the recent contributions has been positioned in the COVID-19 context. For instance, (Jelodar, Wang, Orji, and Huang, 2020) used Text Mining to extract the COVID-19 discussions from social media. They applied topic modeling of public opinions to gain insights into the various issues pertaining to the COVID-19 pandemic. In addition, they implemented an LSTM model for the sentiment classification of comments. While (Bharti et al., 2020) developed a Multilingual conversational bot to provide primary healthcare education, information, and advice to chronic patients. Using NLP methods, the chatbot was aimed to act as a personal virtual doctor to interact with patients like human beings.

3.2 Text Analytics for Clinical Decision Support

(Tvardik et al., 2018) developed a Text-Analytics solution for the automatic detection of medical events using EHR data. The textual records included data collected from three University hospitals based in France over the period October 2009 to December 2010. The dataset spanned a variety of medical surgical specialities including neurosurgery, orthopaedic surgery, and digestive surgery. The system performance was compared with standard methods. The overall sensitivity and specificity were about 84%. The study generally confirmed the feasibility of using NLP-based methods to automate the detection and monitoring of healthcare-associated events in hospital facilities.

In another interesting application, (Brown, and Marotta, 2017) developed a set of classification models to predict the protocol and priority of MRI brain examinations. They used the narrative clinical information provided by clinicians. The models were trained to make predictions on three tasks including: i) Selection of examination protocols, ii) Evaluation of the need for contrast administration, and iii) Estimation of priority. The dataset consisted of about 14K MRI brain examinations over the period of January 2013 to June 2015. The empirical results largely demonstrated that the models could be effectively employed to assist the clinical decision support in this regard.

In the context of radiology, several studies sought to explore the application of NLP methods to extract information from the mammography reports. For example, (Castro et al., 2017) developed a system to

automate the annotation and classification of the Breast Imaging Reporting and Data System (BI-RADS) categories. Specifically, the system tackled two tasks including: i) Annotation of the BI-RADS categories, and ii) Classification of the laterality for each BI-RADS category. The study included about 2K radiology reports collected from 18 hospitals of the University of Pittsburgh from 2003 to 2015. While (Miao et al., 2018) applied Deep Learning to extract the BI-RADS categories from breast ultrasound reports in Chinese. The experiments included a dataset of 540 manually annotated reports. The model accuracy could achieve F1-score of 0.904.

(Afzal et al., 2018) applied NLP for the automatic identification of Critical limb ischemia (CLI). The dataset included narrative clinical notes retrieved form the EHR database. The model performance was validated compared to the human abstraction of clinical notes. Specifically, a physician reviewed and interpreted the information in the EHR data for each patient in the dataset. Overall, the method could achieve an excellent F1-score of about 90%.

Using a Text-Analytics approach, (Carchiolo et al., 2019) proposed a system for the automatic classification of medical prescriptions (i.e. grantable or not). Initially, the textual data were scanned from medical prescription documents. They could develop an effective classifier based on the data about patient/doctor personal data, symptoms, pathology, diagnosis, and suggested treatments. Their results reported that only 5% of the prescriptions could not be automatically classified.

Another recent study developed a framework to realise scalable Text Analytics (Ge, Isah, Zulkernine, and Khan, 2019). The framework aimed to support real-time analytics for decision support in a variety of domains such as healthcare for example. Deep Learning was applied for NLP tasks including language understanding and sentiment analysis. The framework utilised a set of open-source tools including Spark Streaming for real-time text processing along with Zeppelin and Banana for data visualisation. In addition, an LSTM model was trained for the sentiment analysis. They practically demonstrated the functionality of the framework using a scenario with Twitter data.

(Kidwai, and Nadesh, 2020) discussed the application of diagnostic chatbots in healthcare. They developed a chatbot that makes use of NLP methods to understand the user queries. After collecting the initial symptoms, the chatbot would guide the user through a sequence of questions towards making the appropriate diagnosis. The system uses decision trees and follows a top-down approach to conclude the diagnosis. The chatbot was experimented using a medical database of about 150 diseases.

While plentiful studies sought to develop predictive models to help streamline hospital admissions. Increasing contributions attempted to utilise unstructured data such as free-text notes made by nurses or physicians at the Emergency Department (ED). For instance, (Sterling, Patzer, Di, and Schrager, 2019) utilised the bag-of-words representation of triage free-text notes. Using a dataset of over 250K ED visits, neural network models were trained to predict hospital admissions. They could achieve a promising accuracy with ROC-AUC≈0.74. Further, (Chen et al., 2020) aimed to compare the performance of ML models with the inclusion of textual elements. They applied Deep Learning along with Word Embeddings using clinical narratives. They practically demonstrated that the model accuracy generally improved with the addition of free-text fields.

Similarly, (Arnaud, Elbattah, Gignon, and Dequen, 2020) presented an approach based on integrating structured data with unstructured textual notes recorded at the triage stage. The key idea was to apply a multi-input of mixed data for training a classification model to predict hospitalisation. On one hand, a standard Multi-Layer Perceptron (MLP) model was used with the standard set of features (i.e. numeric and categorical). On the other hand, a Convolutional Neural Network (CNN) was used to operate over the textual data. Their empirical results demonstrated that the classifier could achieve a very good accuracy with ROC-AUC≈0.83.

The use of ontologies has also drawn attention in a variety of medical and healthcare applications. To name a few, (Chakrabarty, and Roy, 2016) used ontology alignment for the personalisation of cancer treatment. A patient ontology was mapped to the disease ontology to dynamically transform general treatment options into individual intervention plans, personalised for the patient. In another application, (Comelli, Agnello, and Vitabile, 2015) proposed an ontology-based indexing and retrieval system for the mammography reports. Using an improved radiological ontology, medical terms were organised in a hierarchy, which could measure the semantic similarity between unstructured reports. The system was tested using a dataset of 126 mammographic reports in the Italian language, provided by the University Hospital of Palermo Policlinico.

Furthermore, part of the recent efforts explored the applicability of Text Analytics to predict the International Classification of Diseases (ICD) codes. The manual encoding process is usually time-

consuming, and prone to various errors as well. In this regard, (Teng et al., 2020) applied medical topic mining and Deep Learning to automatically predict the ICD codes from free-text medical records. The study used the MIMIC-III dataset, which provides a large freely accessible repository of ICU records (Johnson et al. 2016). The reported results indicated that their method could increase the F1-score approximately by 5% compared to earlier work. Similarly, (Gangavarapu et al., 2020) developed an approach to help predict the ICD-9 code groups based on unstructured nursing notes. They applied vector space and topic modeling to structure the raw clinical data, which allowed for capturing the semantic information in the free-text notes.

4 DISCUSSION

Over the past five years, there have been pronounced innovations in the NLP research including novel approaches and technologies, which in turn have resonated in the healthcare domain. Most remarkably, Deep Learning has been increasingly applied for developing large-scale language models. Deep architectures of CNNs have introduced a potent mechanism for learning feature representations from raw data automatically (LeCun et al. 1989; LeCun, Bottou, Bengio, and Haffner, 1998). Equally important, recent applications have started to adopt the BERT-based approach, which avails of Transfer Learning for NLP tasks. Furthermore, scalable analytics platforms have been utilised for real-time data processing. Examples include Apache Spark, and IBM Watson.

In terms of data sources, it appears that Text Analytics was applied against a broad variety of healthcare data. The datasets ranged from standard EHR datasets, medical reports, free-text notes, scientific literature, to user-generated content on online forums or social media. In this regard, Text Analytics was implemented for considerable problems including extracting evidence-based care interventions, and patient outcomes, or identifying the population at risk for example. To this end, NLP pipelines have been intensively developed for a variety of text-processing tasks such as: i) Named entity recognition, ii) Topic modeling, iii) Semantic labelling, iv) Relationship extraction, v) Question answering, vi) Text summarisation, vii) Sentiment analysis, and others.

Nevertheless, a set of hurdles stands in opposition to a widespread implementation of Text Analytics in the healthcare domain. A key challenge is the availability of quality data, which is a fundamental factor for building robust NLP models, and for ML in general. Beyond that, the underlying data biases pose multiple ethical concerns for the deployment of NLP models. Such ethical issues have been recently discussed in the literature (e.g. Davenport, and Kalakota, 2019; Baclic et al., 2020). While other technical challenges may relate to the integration of Text Analytics tools with existing healthcare systems. The conventional IT systems may not be well-poised to be integrated with sophisticated Text Analytics, which requires an advanced infrastructure and a highly technical skillset as well. Furthermore, the implementation of Text Analytics typically requires intensive development cycles.

In summary, it is conceived that the future holds many interesting opportunities for implementing Text Analytics in a multitude of healthcare applications. The need for leveraging unstructured textual data should bring up new practical areas for taking advantage of the Text Analytics potentials.

5 CONCLUSIONS

There is an obvious need to leverage unstructured textual data to support the operations of healthcare in many aspects. A large proportion of the clinical data is unavoidably stockpiled into unstructured, or semi-structured, documents or notes. Text Analytics should therefore play a key role in transforming textual data into actionable insights.

This study endeavoured to review the state-of-the-art applications of Text Analytics in healthcare. In this regard, the applications could be broadly summarised as follows:

- Information extraction from free-text data stored in EHR databases, clinical reports, nursing notes, scientific literature, and user-generated content.
- Applying vector-based representations to a variety of clinical documents, which transforms the textual data into an amenable form for ML.
- Sequence-based modeling to address tasks, such as sentiment analysis, using notes in clinical reports, or comments posted on online forums.
- Predictive analytics applications to support the clinical decision making.
- Implementations of Conversational AI technologies to use chatbots to interact with patients in a human-like way.

REFERENCES

Afzal, N., Mallipeddi, V. P., Sohn, S., Liu, H., Chaudhry, R., Scott, C. G., ... & Arruda-Olson, A. M. (2018). Natural language processing of clinical notes for identification of critical limb ischemia. *International Journal of Medical Informatics*, 111, 83-89.

Arnaud, E., Elbattah, M., Gignon, G & Dequen, G. (2020). Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text. *In Proceedings of the IEEE International Conference on Big Data*.

Baclic, O., Tunis, M., Young, K., Doan, C., Swerdfeger, H., & Schonfeld, J. (2020). Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*, 46(6), 161-168.

Bharti, U., Bajaj, D., Batra, H., Lalit, S., Lalit, S., & Gangwani, A. (2020). Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after COVID-19. *In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 870-875. IEEE.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Booth, A., Sutton, A., & Papaioannou, D. (2011). *Systematic approaches to a successful literature review*. Sage.

Brown, A. D., & Marotta, T. R. (2017). A natural language processing-based model to automate MRI brain protocol selection and prioritization. *Academic Radiology*, 24(2), 160-166.

Carchiolo, V., Longheu, A., Reitano, G., & Zagarella, L. (2019). Medical prescription classification: A NLP-based approach. *In Proceedings of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS),* pp. 605-609. IEEE.

Castro, S. M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, S., Bekhuis, T., & Jacobson, R. S. (2017). Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of Biomedical Informatics,* 69, 177-187.

Chakrabarty, A., & Roy, S. (2016). Personalizing healthcare services to support decision making in treatment of cancer patients using ontology alignment. *In Proceedings of the India International Conference on Information Processing (IICIP)*, pp. 1-6. IEEE.

Chang, M., Chang, M., Reed, J. Z., Milward, D., Xu, J. J., & Cornell, W. D. (2016). Developing timely insights into comparative effectiveness research with a text-mining pipeline. *Drug Discovery Today*, 21(3), 473-480.

Chen, C. H., Hsieh, J. G., Cheng, S. L., Lin, Y. L., Lin, P. H., & Jeng, J. H. (2020). Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. *International Journal of Medical Informatics*, 104146.

Comelli, A., Agnello, L., & Vitabile, S. (2015). An ontology-based retrieval system for mammographic reports. *In Proceedings of the 2015 IEEE Symposium on Computers and Communication (ISCC),* pp.1001-1006). IEEE.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2), 94.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. *Journal of Biomedical Informatics*, 42(5), 760-772.

Deng, S., Zhou, Y., Zhang, P., & Abbasi, A. (2019). Using discussion logic in analyzing online group discussions: A text mining approach. *Information & Management*, 56(4), 536-551.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Fan, B., Fan, W., & Smith, C. (2020). Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management*, 57(1), 102131.

Gangavarapu, T., Jayasimha, A., Krishnan, G. S., & Kamath, S. (2020). Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems*, Vol. 190, 105321.

Ge, S., Isah, H., Zulkernine, F., & Khan, S. (2019). A scalable framework for multilevel streaming data analytics using deep learning. *In Proceedings of the IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2, pp. 189-194). IEEE.

Goldberg, Y. (2017). Neural network methods for natural language processing. In Hirst, G. (Ed.). *Synthesis Lectures on Human Language Technologies*, 10(1), p. 69. Morgan & Claypool Publishers.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1989). Handwritten digit recognition with a back-propagation network. *In Proceedings of Advances in Neural Information Processing Systems (NIPS)* (pp. 396-404).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *In Proceedings of the IEEE*, 86(11), 2278-2324.

LeCun, Y. (2014). Chapter 3: Facebook. In Sebastian Gutierrez (Eds.). *Data Scientists at Work*. Apress.

Liao, Z., Liu, L., Wu, Q., Teney, D., Shen, C., van den Hengel, A., & Verjans, J. (2020). Medical Data Inquiry Using a Question Answering Model. *In Proceedings of the 17th IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1490-1493). IEEE.

Lieder, I., Segal, M., Avidan, E., Cohen, A., & Hope, T. (2019). Learning a faceted customer segmentation for discovering new business opportunities at Intel. *In Proceedings of the IEEE International Conference on Big Data*, pp. 6136-6138. IEEE.

Han, J., Nandan, N., & Sun, A. (2015). Did You Know? A Rule-Based Approach to Finding Similar Questions on Online Health Forums. *In Proceedings of the 2015 International Conference on Healthcare Informatics*, pp. 513-514). IEEE.

Hearst, M. A. (1999). Untangling text data mining. *In Proceedings of the 37th Annual meeting of the Association for Computational Linguistics* (pp. 3-10).

James, T. L., Calderon, E. D. V., & Cook, D. F. (2017). Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*, 71, 479-492.

Jelodar, H., Wang, Y., Orji, R., & Huang, H. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: NLP using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733-2742

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395-405.

Ji, Y., Tian, Y., Shen, F., & Tran, J. (2016). Leveraging MapReduce to efficiently extract associations between biomedical concepts from large text data. *Microprocessors and Microsystems*, 46, 202-210.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

Kidwai, B., & Nadesh, R. K. (2020). Design and development of diagnostic Chabot for supporting primary health care systems. *Procedia Computer Science,* 167, 75-84.

Kumari, B. N., & Mahalakshmi, G. S. (2019). A cloud based knowledge discovery framework, for medicinal plants from PubMed literature. *Informatics in Medicine Unlocked*, 16, 100226.

Martínez, P., Martínez, J. L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A., & Revert, R. (2016). Turning user generated health-related content into actionable knowledge through text analytics services. *Computers in Industry*, 78, 43-56.

Miao, S., Xu, T., Wu, Y., Xie, H., Wang, J., Jing, S., ... & Shan, T. (2018). Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. *International Journal of Medical Informatics*, 119, 17-21.

Paglialonga, A., Riboldi, M., Tognola, G., & Caiani, E. G. (2017). Automated identification of health apps' medical specialties and promoters from the store webpages. *In Proceedings of the E-Health and Bioengineering Conference (EHB)*, pp. 197-200. IEEE.

Paglialonga, A., Pinciroli, F., Tognola, G., Barbieri, R., Caiani, E. G., & Riboldi, M. (2017). e-Health solutions for better care: Characterization of health apps to extract meaningful information and support users' choices. *In Proceedings of the 3rd International Forum on Research and Technologies for Society and Industry (RTSI)* (pp. 1-6). IEEE.

Pendyala, V. S., & Figueira, S. (2017). Automated medical diagnosis from clinical data. *In Proceedings of the IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 185-190. IEEE.

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.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, *1*, 1-20.

Sarkar, D. (2019). *Text analytics with Python: a practitioner's guide to natural language processing*. Apress.

Spasić, I., Uzuner, Ö., & Zhou, L. (2020). Emerging clinical applications of text analytics. *International Journal of Medical Informatics*, Vol. 134.

Sterling, N. W., Patzer, R. E., Di, M., & Schrager, J. D. (2019). Prediction of emergency department patient disposition based on natural language processing of triage notes. *International Journal of Medical Informatics*, 129, 184-188.

Sutar, S. G. (2017). Intelligent data mining technique of social media for improving health care. *In Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS),* pp. 1356-1360. IEEE.

Tan, A. H. (1999). Text mining: The state of the art and the challenges. *In Proceedings of the 1999 PAKDD Workshop on Knowledge Discovery from Advanced Databases,* Vol. 8, pp. 65-70.

Tvardik, N., Kergourlay, I., Bittar, A., Segond, F., Darmoni, S., & Metzger, M. H. (2018). Accuracy of using natural language processing methods for identifying healthcare-associated infections. *International Journal of Medical Informatics*, 117, 96-102.

Teng, F., Ma, Z., Chen, J., Xiao, M., & Huang, L. (2020). Automatic medical code assignment via deep learning approach for intelligent healthcare. *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2506-2515.

van Dijk, W. B., Fiolet, A. T., Schuit, E., Sammani, A., Groenhof, T. K. J., van der Graaf, R., ... & Grobbee, D. E. (2020). Text-mining in electronic healthcare records can be used as efficient tool for screening and data-collection in cardiovascular trials: a multicenter validation study. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2020.11.014

Vinod, P., Safar, S., Mathew, D., Venugopal, P., Joly, L. M., & George, J. (2020). Fine-tuning the BERTSUMEXT model for Clinical Report Summarization. *In Proceedings of the 2020 International Conference for Emerging Technology (INCET)* (pp. 1-7). IEEE.