National College of Ireland

Multi scale context aware drug review sentiment analysis using pretrained med-BERT.de

MSc Research Project
MSc Data Analytics

Mohammad Farooque Azam
Student ID: x21198926

School of Computing
National College of Ireland

Supervisor: Dr. Anh Duong Trinh (Senja)

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mohammad Farooque Azam…………………………………………………………
Student ID:	21198926…………………………………………………………………………….……
Programme:	MSc Data Analytics………………………………… **Year:** 2022-23………..
Module:	MSc Research Project…………………………………………………………….………
Supervisor:	Dr. Anh Duong Trinh (Senja)……………………………………………………….………
Submission Due Date:	14/08/2023…………………………………………………………………………….………
Project Title:	Multi scale context aware drug review sentiment analysis using pretrained MedBERT….………………………………………………………………………………....
Word Count:	……6476…………………………… **Page Count**…………20…………………………….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Mohammad Farooque Azam……………………………………………………………

Date: 14/08/2023……………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multi scale context aware drug review sentiment analysis using pretrained medBERT.de

Mohammad Farooque Azam
21198926

Abstract

A massive amount of drug-related data is available online in the form of reviews and comments. This data could be leveraged in analyzing the opinion and sentiments of users related to the drug. The written medical texts of patients and carers, in particular, have a considerable influence on how people, physicians, and drug developers regard drug users. Sentiment analysis approaches have grown from fundamental concepts to advanced methods of machine learning such as deep learning, which has emerged as a recent development in many NLP applications. The emergence of transformer-based models has revolutionized the field even further.. However, most of them have worked on the dataset from a single source with a limited number of rows and a general-purpose model for analysis, making them incapable of capturing the contextual meaning of medical terms used in the text. This defect affects the performance of the pharmaceutical recommendation system. Hence, this study proposes sentiment analysis of drug reviews using the medical domain-specific pre-trained transformer-based model medBERT.de. The proposed model achieved the highest F-score of 83%, a precision of 84% for optimal hyperparameter values of epochs=3, learning rate= 2e-5, batch size=32 and decay_weight= 0.1 , suggesting the model works fine in capturing the medical terms used in the dataset.

1 Introduction

Understanding the emotions portrayed by a given source—whether it's a tweet, a paper, a report, a blog, a piece of a politician's speech, or anything else—is a critical task for humans. Our understanding of the real world, as well as our choices in life, are highly impacted by how other people perceive and understand the world. Natural Language Processing(NLP) is a field of computing which bridges the gap between machine and human intelligence, by automating sentiment extraction from large and complex data.

There has been a gradual evolution in Natural Language Processing(NLP) in the past decade. It has emerged as a computational subfield and ultimately finds itself at the crossroads of growing online vocabulary, especially in the English language, advancement in computer science, and the emergence of artificial intelligence technologies. Despite the term's broad use, NLP applications tend to understand human language using computational analysis. Sentiment analysis is the study of people's views, sentiments, and emotions regarding things such as products, services, organisations, subjects, and their qualities.

With the introduction of social media platforms and advancements in internet access, along with the exponential rise in the number of hand-held devices, users express their thoughts and ideas on anything and everything that may affect them. This has resulted in a vast amount of data being generated and one such data source is drug reviews posted on different platforms. Opinion mining or sentiment analysis on drug reviews plays a critical part in pharmacovigilance (Noferesti et al., 2015), which is concerned with the collection, identification, assessment, monitoring, and prevention of adverse medication reactions. Pharmacovigilance seeks to investigate issues and any negative effects or Adverse Drug Effects(ADEs) that a medicine may produce. According to current data, ADEs account for over 3.5 million physician visits, 1 million emergency department visits, and over 2 million injuries, hospitalization, and fatalities. Along with having a major impact on wellness, ADEs have a broad financial effect, costing more than $75 billion (Fan et al., 2020). This health and economic damage stems from the fact that ADEs are not properly considered during medication. The unreported ADEs are not in the knowledge of the manufacturers and ultimately doctors keep on prescribing the medicine resulting in a surge in hospitalization. In recent years, research in pharmacovigilance using collective intelligence has been gradually developing and has shown effectiveness in the identification of adverse medication events (Korkontzelos et al., 2016).

Users review contain vast information regarding a drug and sentiment analysis on such information can assist healthcare practitioners assess the efficacy and adverse effects of a certain medicine by identifying the general impression of a medicine rating(Fan et al., 2020). However, the reviews are in the form of unstructured data, and hence cannot be directly used for sentiment analysis before processing them. Once preprocessed, data is good for sentiment analysis techniques and can be used to determine the underlying opinion of drug reviews. The results are pooled by drug name and condition to rank the medicines for a condition, eventually high-ranked drugs being recommended by the system.

NLP methodologies are well-suited to analysing para-linguistic components such as acronyms or emojis as well as labelling terms according to parts of speech to signal polarity based on their grammatical role. Various research has found that sentiment analysis is domain-specific, with the same text having various connotations depending on the domain of the dataset utilized. Furthermore, because of the somewhat informal language used in online forums, user-expressed medical conceptions are typically nontechnical, descriptive, and difficult to separate, and thus traditional lexicon-based analysis is of little value in sentiment analysis of medication evaluations. Hence, sentiment Analysis is a classification problem in Natural Language Processing (NLP) in which a sentence or a document is evaluated to determine the sentiment and polarity of that text from the writers' perspective. Sentiment analysis may be done at several levels, including document, phrase, and aspect. The document-level analysis provides an opinion on the entire document, whereas sentence-level analysis provides sentiment for each sentence in the document. Aspect-level sentiment analysis is conducted at opinion targets in a phrase and has the benefit of doing medical term analysis and extracting hidden sentiment. Drug reviews contain medical terminology that may seem trivial in the common lexicon but have more weight in medical science, so

deducing feelings based on such phrases may be more valuable for doctors, drug manufacturers, and, ultimately, patients.
Sentiment extraction from text is a challenging task given a word used at different places in a sentence can have a different meaning.For example,

“The apples are nice”
“Apple products are robust and reliable”
“The boy is the apple of her mothers’ eyes”

The word “Apple” used in the above sentences have different contextual meaning and catching this contextual meaning correctly is a daunting task. The problem is exacerbated by a lack of labeled data, noise, and inconsistent content. Furthermore, labelled data is difficult to come by and, when found, is domain-specific, since patients tend to report only on particular ailments and not others.

Traditional embedding models, when paired with networks capable of learning dependencies, such as Bag of Words (BOW), Glove, and Word2Vec may capture context to a limited extent (Rasmy et al., 2021). However, traditional approaches, owing to their incapability of taking word positioning into account, are incapable of collecting out-of-vocabulary terms such as social media trending words, noisy data, and domain-specific phrases. These limitations are handled in cutting-edge transformers up to a large extent. Capturing the contextual meaning of negative medical terms in drug review data, on the other hand, remains problematic and the state-of-the-art topicT-AttNN model(Job et al., 2023) did not address this issue which needs to be researched.

To capture the contextual meaning and its effect on sentiment classification of the drug review, we propose Multi-Scale context-aware sentiment analysis of drug review using medical domain-specific pre-trained medBERT.de (Bressem et al., 2023).

Research Question. The above research problem motivates the following research question:

The research hypothesis is that the suggested model with domain-specific pre-trained medBERT.de transformer model outperforms existing strategies and serves as the state of the art in sentiment mining of drug review. The study tries to solve below-mentioned problems:

- Does the application of a domain-specific pre-trained model helps in better classification of drug reviews loaded with medical connotations?
- How does the model perform on large datasets from multiple sources?
- Does hyperparameter tuning help improve the performance of the proposed model?
- Is the proposed model feasible and could it be generalized for similar tasks of different domains?
-

This research is focused on performing sentiment analysis on drug review data secured from open data sources Drugs.com, DrugLib and WebMEd and seeks to incorporate characteristics from drug reviews in the above-specified dataset for drug recommendation.The research challenge is stated as follows:

Let D denote a collection of papers, each having a collection of phrases: W, where W = {w1, w2,...,wn}, d D. A rating field with a numeric score linked to each dataset's comments. W is *converted to a vector representation and put into a classification algorithm as a matrix, which is transformed from probabilities to give a three-level classification outcome as good, neutral, and bad for each review in the dataset.*

.

2 Related Work

2.1 Gradual evolution of drug review sentiment analysis work

Web 2.0 has spurred the expansion of user-generated material on the internet, and since then, sentiment analysis has grown in popularity. One industry where sentiment analysis is used is pharmacovigilance, where online user evaluations are analysed to produce useful conclusions that aid medication producers, doctors, and patients in making decisions. However, much research in the field of sentiment analysis has concentrated on data from social media platforms such as Twitter, and data from the medical profession has not found much place in the academic arena. This lack of research in health-related data could be due to the lack of availability of domain-specific data. Nevertheless, limited research which was carried out on drug reviews was primarily focused on the detection of Adverse Drug Reactions(ADRs). With the improvement in technology and exponential surge in internet speed and handheld devices, numerous domain-specific platforms have come into being. Users, patients, caregivers and doctors use these platforms to register their reviews and opinion about the use and efficacy of a drug for a particular condition. Furthermore, with the growing prevalence of web-based pharmacy and the accessibility of critical consumer input on medicine, there is enormous opportunity in using this information to recommend treatments.

The following are some of the pioneer works on this topic, along with critical analyses:

- The majority of the work in this area is about detecting adverse events and predominantly lexicon-based analysis, in which each word of user evaluations is mapped to a specific collection of vocabulary from other users (Leaman et al., 2010). This research yielded 78.3% accuracy and 69.9% recall, for an f-measure of 73.9%. However, the drugs used in this paper annotated work largely on the central nervous system and hence have different adverse response profiles than other treatments with very different mechanisms. This strategy has disadvantages due to phonetic and spelling errors in user evaluations, which are solved in future articles by applying rule-based strategies.
- To fix the issues related to phonetic and spelling errors in the data, (Nikfarjam et al., 2012). used a set of annotated comments, and association rule mining to find the underlying patterns of colloquial phrases conveying unfavourable effects. The rules are developed based on the extracted semantic roles in sentences specified by the Stanford parser output. Despite dealing with informal material for the research, the study discovered that user reviews were consistent. The limitations of the lexicon-based technique in capturing misspelt words were solved by an association rule-based approach. The current pattern-based method eliminates the need for string matching by looking for

ADRs using templates rather than the precise mention. Although the suggested method performed well with colloquial data, it had worse accuracy and recall than the lexicon-based approach. These limitations were taken care of by the machine learning classifiers in subsequent research.

- Ali et al., (2013) used several lexicon-based characteristics, such as pronoun phrase categorization of opinionated and non-opinionated sentences, in conjunction with rule-based features. The study used a subjective lexicon and machine learning approaches to analyse the sentiment of postings on hearing loss-specific forums, as well as different combinations of features and classifiers, such as Nave Bayes, SVM, and Logistics-R, which shows gradual improvement in evaluation matrices such as F-score. The study discovered that, when compared to the other feature sets, the BOW representation produced superior results. However, with improved machine learning techniques such as neural networks and deep neural networks, the performance of sentiment analysis systems might be improved even more.

- Gopalakrishnan, V. and Ramaswamy, C. (2017). demonstrated that in terms of accuracy, recall, and f-score, the neural network-based opinion mining strategy surpasses the support vector machine method. In terms of the performance measurements utilised, it is also demonstrated that the radial basis function neural network approach outperforms the probabilistic neural network method.PNN and RBFN, two separate neural network-based techniques, were employed as function approximation tools in this paper. When compared to other methodologies, RBFN can discern the maximum number of positive, negative, and neutral judgments. However, in professions such as medicine, people regularly use indirect communication to explain themselves. Instead of openly expressing a viewpoint, patients in the drug domain often write about their experiences with the effectiveness or bad effects of medications. Further work was done to derive opinions based on finer information to recognize the implicit significance of the opinions provided by users. Aspect Based Sentiment Analysis (ABSA) is a pioneering study in this field that extracts intrinsic opinion from user language.

- Introduction attention mechanism by Vaswani et al., (2017), revolutionized the Natural Language Processing techniques in which a self-attention layer connects all positions with a constant number of sequentially executed operations, whereas a recurrent layer requires O(n) sequential operations (Vaswani et al., 2017).The study suggested that as an added bonus, self-attention may result in better interpretable models.Individual attention heads not only clearly learn to do diverse tasks, but many appear to demonstrate behaviour connected to sentence syntactic and semantic structure.

- Transfer learning helps in sentiment analysis as a model trained on a large corpus could be used for evaluating similar but small data. Gräßer et al., (2018) examined customer review data posted on online platforms. It has been observed that there is a dearth of annotated drug review data, and if they are available, they are condition specific. To address the paucity of annotated data, the paper studies the transferability of trained classification models between domains, i.e. circumstances and data sources. Online user reviews in this discipline offer information on a variety of topics, including therapeutic efficacy and side effects, making automatic analysis both exciting and difficult. Gräßer et al.,(2018) applied cross domain and cross data sentiment analysis.The model trained on data from one condition was reviewed on data from another condition, and similarly, the model trained on data from one domain was tested on data from another domain to assess the model's performance. The results of the tests conducted in this paper show that moving a model trained on relatively big data to tiny test data performs very well, but the outcome of the vice-versa strategy is less promising. Furthermore, this research increases the automated extraction of aspect-related experiences from patient drug ratings, promotes pharmacovigilance, and offers up new options for future research. Abuka, G., (2023) suggested a model that combines a fine-tuned Text-to-Text Transfer Transformer (T5) model and Pre-training with extracted gap sentences for abstractive summarization (PEGASUS) models with a Long Short-Term Memory (LSTM) model. The model was found to be an effective BBC news summaries dataset and achieved average ROUGE1, ROUGE2, and ROUGEL scores of 72.20, 63.59, and 57.42 respectively.

- Deep learning techniques were found to be helpful in sentiment analysis tasks and to exploit the same, Colón-Ruiz, C. and Segura-Bedmar, I. (2020) did a comparative research of several deep learning approaches on online medication review and employed different combinations of Convolutional Neural Network(CNN), Long Short Term Memory(LSTM), Bi-LSTM to obtain the findings.LSTM followed by the CNN model performs incredibly well with the greatest micro-F1 and macro-F1 scores, but the results were not better than a hybrid model for the same objective. The findings of the research demonstrate that, regardless of the model used, the performance of each class is notably dependent on the number of instances of each in the training set. Overall, the LSTM+CNN hybrid classifier outperforms all other classifiers, with micro-F1 values of 69.28% and macro-F1 values of 64.09%. Because of the revolution brought about by the transformer model BERT in the field of NLP, it was used to represent medication review and Bi-LSTM as a classifier, resulting in considerable cost savings. According to the findings, using variational autoencoders (VAE) in the drug review system would lessen the reliance on annotated corpora and yield higher classification results.

- Fan et al., (2020) used a manually labeled dataset of 10,000 reviews from WebMD and Drugs.com and proposes a deep learning-based approach for ADE detection and extraction based on Bidirectional Encoder Representations from Transformers (BERT) models and compares results to standard deep learning models and current state-of-the-art extraction models. The findings demonstrate that a BERT-based model delivers new state-of-the-art outcomes on both the ADE detection and extraction tasks. The model suggested in this study produces cutting-edge results and may be used for a variety of additional healthcare and information extraction tasks, such as medical entity extraction and entity recognition. This work also shows that deep learning is a feasible choice for the pharmacovigilance job and that innovative natural language processing techniques using BERT sentence and word embeddings outperform previously applied approaches.

- A semi-supervised technique based on a Semi-Supervised Generative Adversarial Network (SSGAN) was developed by Colón-Ruiz, C. and Segura-Bedmar, I., (2021).to address the lack of labelled data for sentiment analysis of pharmaceutical assessments and improve the outcomes provided by supervised algorithms. The semi-supervised technique uses the supervised approach architecture, but it also employs an adversarial network to improve the internal representation of the classification data. A discriminative (D) and generative (G) model creates artificial instances while learning how the data are distributed, with D predicting the likelihood of the instance created being synthetic and G maximizing the likelihood that D makes a mistake, resulting in a competition between D and G models. Because of the competition, both models improve until they are almost equal to the distributions created by the original training data. The suggested model was tested by comparing it to the supervised model, and the findings reveal that the proposed SSGAN model performed better in terms of macro-F1 scores, but not in terms of micro-F1 scores.

- Apart from conventional rule-based and machine learning-based models for sentiment analysis, a study by Li et al., (2021) implemented a hybrid model sentiment analysis task. For sentiment analysis of restaurant reviews, the researchers merged Word2vec word embeddings with an RNN-based attention network. Because of this feedback mechanism, recurrent neural networks may dynamically learn sequence features, increasing the efficacy of sentiment analysis. Because of this feedback mechanism, recurrent neural networks may dynamically learn sequence features, increasing the efficacy of sentiment analysis. The number of positive and negative phrases influences traditional sentiment research methodologies, which rely on domain dictionaries. The Word2vec+Bi-GRU+Attention approach is used in this work to analyse sentiment in online restaurant reviews. The results suggest that the hybrid model outperforms traditional machine learning approaches in categorizing feelings.

- Microblog data have become more complex with growing dimensions and features in due course of time. Hence, most research on machine learning models for sentiment classification focuses on developing valuable handcrafted features to improve prediction accuracy. Wu et al., (2021) suggested a sentiment classification approach (SC-ABiLSTM)for large-scale microblog content based on the attention mechanism and the bidirectional long short-term memory network. By recognizing sentimental words from large-scale microblog material, the study proved that the proposed approach can more precisely characterize the links between emotional words and their surrounding phrases. The study also showed that a traditional machine learning model enhanced with an attention mechanism increased classification performance. Although the suggested model improves classification task accuracy, it can only partially predict feelings in specific settings, such as long texts without emotional words or sentiments at various levels with comparison terms, which necessitates more research.

- There is a basic link between a consumer's total empirical evaluation and the text-based explanation of their opinion and Alantari et al., (2021) investigated the empirical trade-off between predictive and diagnostic skills using a variety of methodologies to evaluate this fundamental connection. According to the study, neural network-based machine learning approaches, particularly pre-trained ones, provide the most accurate forecasts, whereas topic models such as Latent Dirichlet Allocation provide deeper diagnoses. The study tries to rectify potential limitations in ordered logistic regression's diagnostic capacity in comparison to topic discovery approaches by applying Local Interpretable Model-Agnostic Explanations (LIME) on BERT, one of the methods with the best predictive ability, to obtain diagnostics. It also employed Pointwise mutual information (PMI), a popular measure in NLP, to compute relationships between words used by customers to describe their product experience, hence determining the relevance of each characteristic.LIME is a promising strategy for mitigating neural network models' lack of diagnostic capabilities, however, the resulting diagnostics are limited by a few limitations.

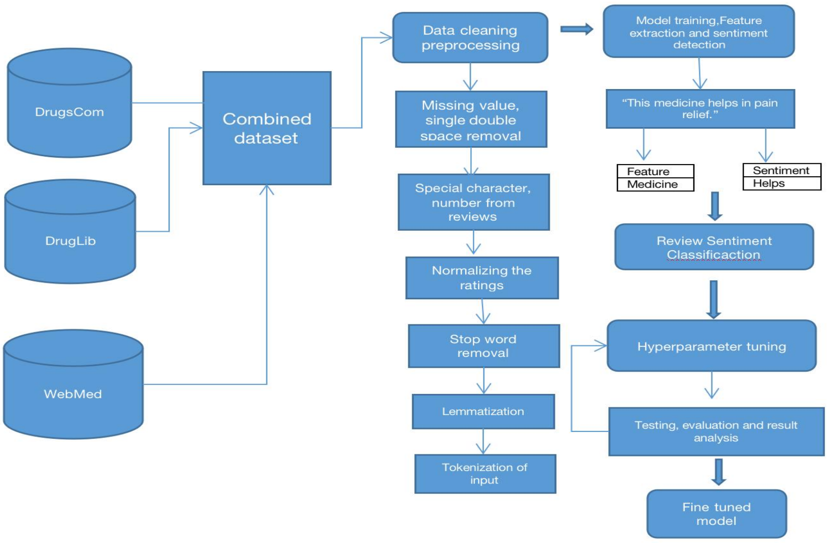
- Topic modelling techniques have proven to be useful in enhancing the performance of the sentiment classifier models. Job et al., (2023) used a transformer-based LSTM-Attention layered network and developed a model topicT-AttNN for drug review dataset classification and deemed it to be state of the art in drug review analysis. The study's findings clearly show that a transformer alone cannot increase classification performance; thus, enhancing it with LSTM-Attention layers and integrating it with a topic feature improves classification results even more. The topicT-AttNN model worked admirably, with F1 scores that are nearly perfect for all three sentiment components. The suggested model performed remarkably well in one element, side effects; however, the findings in another aspect, efficacy, are less hopeful.Furthermore, the dataset employed in the study contains negative medical phrases, the contextual meaning of which the model was unable to capture effectively, necessitating additional research in this area.

2.2 Summary

The extensive literature analyses on the results of a sentiment analysis of medicine evaluations suggest that the topic of research requires more investigation due to some of the main challenges related to the subject. Although the state-of-the-art transformer model did well in sentiment analysis, the mechanism's context-awareness technique misses the importance of prominent subjects in text analysis. This has given rise to customized transformer models, which are trained on a domain-specific large corpus of data. The state-of-the-art technique proposed by Job et al., (2023) lacks the capability of capturing the effect of medical terms in the drug review dataset analysis. This issue, to the best of my knowledge, has not been captured by any studies in the past. This paper suggests using a pre-trained medBERT.de (Bressem et al., 2023), transformer-based model to perform a robust multi-level drug review classification task.

3 Research Methodology

This section mentions all the necessary methods applied and steps taken to do sentiment assessment on medicine reviews using pre-trained medBERT.de (Bressem et al., 2023) transformer model. Three datasets used in the research are explained in detail. Data preprocessing steps, data cleaning, exploratory data analysis conducted and statistical tests performed are mentioned. The overall research methodology is shown in Fig.1



DrugsCom
DrugLib
WebMed
Combined dataset
Data cleaning preprocessing
Missing value, single double space removal
Special character, number from reviews
Normalizing the ratings
Stop word removal
Lemmatization
Tokenization of input
Model training,Feature extraction and sentiment detection
"This medicine helps in pain relief."
Feature
Medicine
Sentiment
Helps
Review Sentiment Classificaction
Hyperparameter tuning
Testing, evaluation and result analysis
Fine tuned model

Figure 1: Research methodology overview.

3.1 Data Pre-processing

Before assessing the received text information, information cleansing, and preliminary processing of information operations were carried out to achieve good results. These include standard letter case, punctuation, digits, and the elimination of special characters. In this research we are using data from three different sources namely, Drugs.com, DrugLib and WebMed. Two of the datasets, Drugs.Com and DrugLib are obtained from UCI ML Repository, which was originally sourced by Gräßer et al., (2018) from Drugs.com and Druglib.com and is publicly available for the research purposes The Drugs.com dataset has over 0.25 million rows and six columns

:

Table 1: A Drugs.com

Column	Description
drugName	name of the drug
condition	medical issue of the patient
review	comments posted by users
rating	rating on the scale of 1-10 given by users to the drug
date	date of review posted
usefulCount	indicating if the review was helpful

The second dataset DrugLib has over 3000 rows and 9 columns. The dataset has individual review comments for each aspect, such as benefits review, side effects review, along with overall comments review. These reviews are concatenated to generate a new column "review" for evaluating the overall opinion of the comment. The third dataset webmd has over 0.36 million rows and 12 columns. The dataset has no specific column for "rating", hence we are using the "satisfaction" column in the dataset as a review. However, the satisfaction values are in the range of 1-5 while the ratings in the other two datasets are in the range of 1-10, hence the satisfaction column was normalized in the 1-10 range and renamed as "rating" for the ease of analysis. The overall rating and columns from each dataset were extracted and merged into one DrugReview dataset. The distribution of ratings in the dataset is heavily imbalanced towards higher rating values. This will result in less number of rows for negative or neutral comments. As a result, for the value ranges of 1-3, 4-7, and 8-10, this study categorises the value as Bad-Neutral-Good. as well as in the range of 1-5,5 and 5-10 for different case studies. Multiple experiments will be run to test the efficacy of the model with different rating classifications.

The resultant dataset is having over 0.4 million rows and two columns, "review" and"rated".For text preparation, the following activities were completed using Pandas, NumPy, and the NLTK package.

- Stop words removal: Common English terms such as a, an, the, and so on are stop words that appear far too frequently.These stop words were eliminated since they aren't necessary for the classification system, and eliminating them decreases the amount of characteristics while also making the classification model more efficient.
- HTML tags and special character removal: Special characters like @,#,$,% etc and HTML tags were removed from the input data.
- Lemmatization: Lemmatization was applied to the input data in order to extract the root words known as lemma.Words like infectious, infecting, and infected, for example, were converted to the dictionary term infect.Lemmatization is recommended over stemming since stemming simply trims the end of the word and is a primitive technique presuming it will create the root word.
- Space removal: Space has its own ASCII value and hence computer treats it as a value.More than one space between two words can hamper model performance.Hence all single space after full stop and double space between two words are removed.

3.2 Feature extraction

The extraction of pertinent data from written input via feature extraction is of the utmost importance in sentiment classification as it impacts the model's efficacy and reliability inexorably. Feature extraction aims to choose meaningful material that includes the most essential textual characteristics (Wankhade et al., 2022). This method aims to choose meaningful material that encompasses the text's most essential aspects. For feature extraction, we are using TF-IDF, one of the most prominent and commonly used text mining approaches for extracting. Only term frequency (TF), i.e. the number of occurrences of a term in a document, is not a useful metric for determining what a document is about since common terms like a, an, the, is, and so on appear too frequently in a text and will have a larger weight. We can still create a dictionary of stop words and delete them, but doing so may disrupt the context of the text. This problem will be handled by using inverse document frequency (IDF), which reduces the weight of the most often used terms in the document, and combining it with term frequency, TF-IDF, which delivers the adjusted frequency of infrequently used terms. The goal of this strategy is to locate the most relevant or important terms in a document within a large corpus of documents. The following is the general mathematical formula for calculating TF-IDF.

.

$$W(t,d) = TF(t,d)*log(N/DF(t))$$

where:

W(t,d) is the total weight of the phrase *t* in document *d*
TF(t,d) is the number of occurrences of the term *t* in document *d*
DF(t) gives the number of document in which term *t* appeared
N is total number of documents in the corpus

3.3 Feature selection.

Feature selection is a prerequisite task when working with a dataset with multiple columns for the classification task. Unnecessary columns when fed into the model may affect performance significantly. However, according to our question, we are concerned about review and rating columns from the dataset as we are trying to find out the overall sentiment of the comment. Moreover, *benefits* and *sideeffect* reviews from the *DrugLib* dataset have already been concatenated into overall review comments in the *data cleaning and preprocessing* phase. Chi-squared test was performed between dependent and independent variables that is review and ratings. The *p-value* for the test was lower than the significance interval of 0.05. This suggests a strong association between dependent and independent variables. Following that, it is important to convert textual materials into a computer-readable format and for this purpose, we are employing medBERT.de(Bressem et al., (2023).) tokenizer. Tokenization entails breaking the input text down into individual words or subword units, transforming these units into numerical IDs, and adding special tokens that the model recognizes.

3.4 Sentiment classification.

Sentiment classification forms an important part of the sentiment analysis task which requires identifying and categorizing emotions represented in textual data such as reviews and tweets. In this study, the dataset used is already labelled with rating values. However, the ratings given are in the range of 1-10. The sentiment classification task is carried out by classifying the rating value as Bad-Neutral-Good for values 1-3, 4-7, and 8-10 respectively as well as with 1-5,5 and 5-10 and multiple experiments were run to check the efficacy of the system. This data is further tokenized and fed into pre-trained MedBERT.de (Bressem et al., (2023)) model for sentiment analysis.

3.5 Evaluation metrics.

We are going to evaluate the model performance based on precision, recall and F1-score values. The major performance measure used in this study is the F1 - measure, that integrates the accuracy and recall assessments into a single metric. Accuracy alone cannot be used to evaluate the model given the imbalanced nature of the dataset. The F-score is essentially the weighted harmonic mean of accuracy and recall, two measures that represent the level of accuracy and volume of predictions. Given the research problem, false-positive classification of a drug review could be costly, hence F1-score will be the ideal metric to judge the performance of the model. We are also using a confusion matrix to visualize the results.

4 Design Specification

Fig2. Shows the design architecture of the proposed model in the study. The input text is tokenized using a pre-trained medBERT.de tokenizer, which is then processed using the WordPiece tokenization approach, which breaks the content down into "subword" chunks to effectively catch strange or domain-specific phrases. MedBERT.de is built on the industry-standard BERT architecture as described in the original BERT paper (Fig.3) (Devlin et al., 2018). The model includes a multi-layer bidirectional transformer encoder, allowing it to collect context-relevant data in the input phrase from both the left-to-right and right-to-left directions. medBERT.de contains 12 layers, 768 hidden units per layer, and 8 attention heads in each layer, with the ability to handle up to 512 tokens in a single input sequence (Bressem et al., 2023).

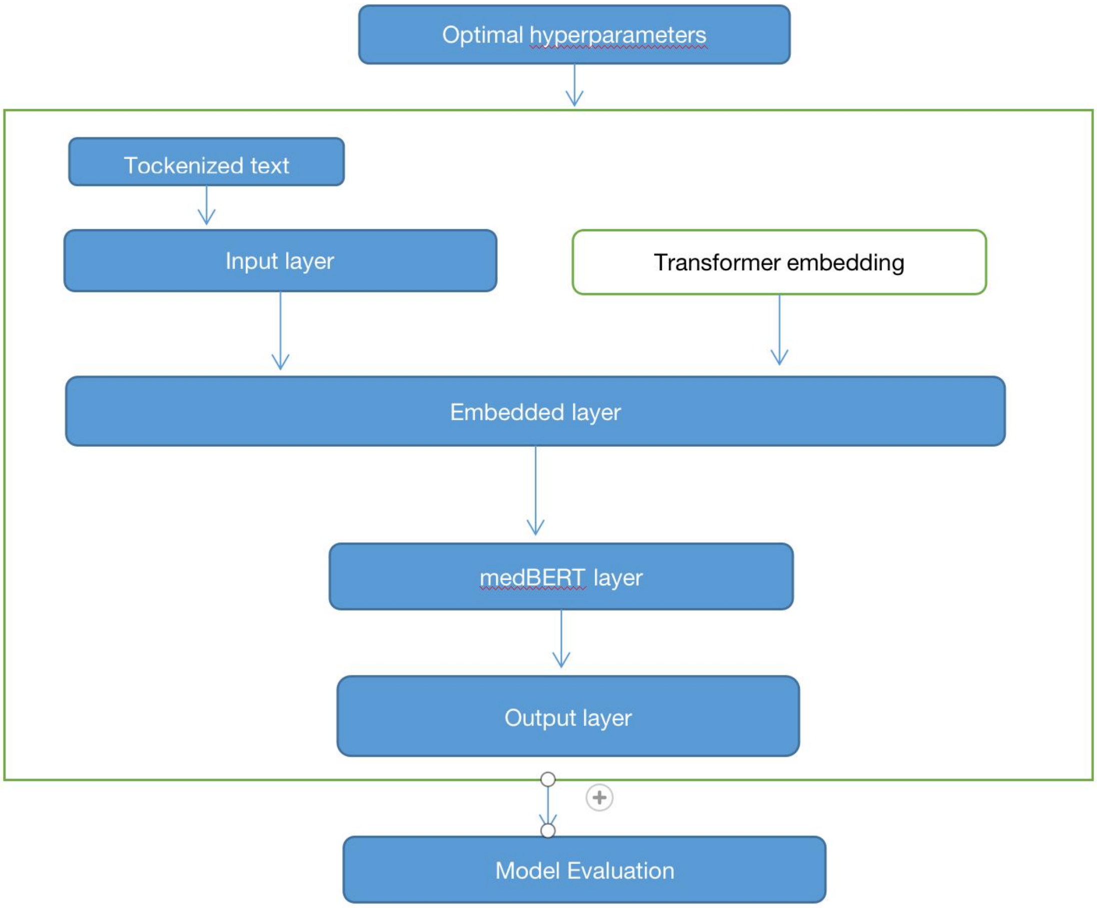


Figure 2: Experiment design of the proposed model.

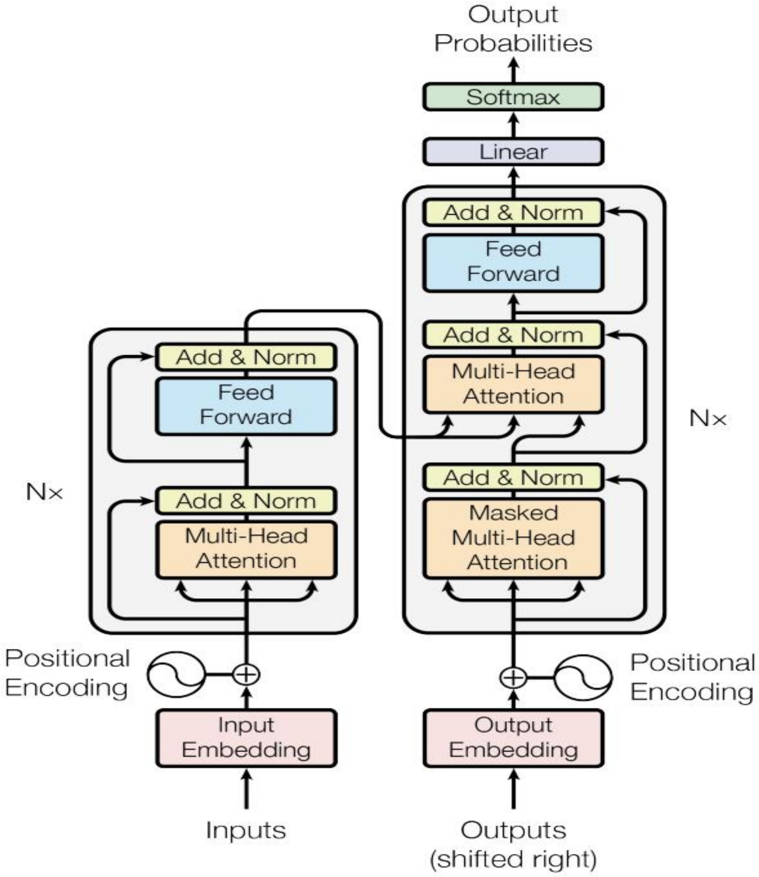


Figure 3: Transformer model architecture. (Source:Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert)

5 Implementation

The study uses the "python" language for implementation purposes. Once the data was preprocessed the cleaned data was loaded for further analysis. Necessary Python libraries and dependencies were imported and installed. We are using both the local machine and google collab for the execution of our code components. Basically, data loading, cleaning, exploratory data analysis, statistical tests and other preprocessing steps are performed on a local machine. The preprocessed data is exported in a "DrugReview.tsv" file to the local machine to be used in the Google Collab environment. Given the complex computational operations associated with the analysis, a collab environment is set up with GPU and high RAM requirements. Necessary Python libraries like transformers, *PyTorch*, *sklearn*, NLTK, "*Charangan/MedBERT*" a medBERT.de tokenizer and model were downloaded from " the Hugging Face Hub" and fine-tuned for the purpose of our analysis. The combined dataset "DrugReview.tsv" has over 0.4 million rows and hence, the complexity and time consumption are huge. We have taken random samples from the dataset for our experiment ranging from 5000-40000 and records of each run were recorded. The research question demands multi-class classification hence, we are using "Gaussian Error Linear Unit"(GELU) as the hidden layer activation function and softmax as the output layer activation function and a customized medBERT.de classifier was developed for the analysis. Training arguments are supplied with *learning_rate*=2e-5, *batch_size*={16,32} *weight_decay*={0.1,0.01}. Hyperparameter *epoch* was chosen in the range of {3-15} and results of each experiment were recorded. Hyperparameter tuning was done by changing the number of epochs and batch size per device to arrive at optimal values for the model. The results of each run were recorded and F1-score vs epochs graph was plotted for visual analysis of the results.

6 Evaluation

The dataset was split into train, and test in the ratio of 80:20. and validation sets for training and model evaluation purposes. Extensive experiments were carried out with varying numbers of records, different rating classification label values and hyperparameter values. The model was trained with the train data and evaluated on test and evaluation data. Table 2. gives the evaluation data of the medBERT.de model and compares it with the state-of-the-art technique.

Table 2: Results

Models	**Precision**	**Recall**	**F1-Score**
BERT	0.9250	0.8588	0.8157
topicT-AttNN model	0.9340	0.8556	0.9105
medBERT.de	0.8473	0.823	**0.83488**

6.1 Case Study 1: Fine tuning hyperparameters

The experiments were carried out on data with multilevel categorical values for rating columns by categorizing the ratings as positive, neutral and negative for rating values greater than 5, equal to 5 and less than 5 respectively. The hyperparameter values chosen were batch size 16, number of *epochs 10*, *learning rate 2e-5*, *decay_weight 0.1* and 10000 sample data from DrugReview. The results of the record were plotted on the epoch versus F1-score graph. It was found that under the above-given conditions, the best *F1-score of 0.828064* was achieved for 2 epochs. The decay_weight was on the higher side with a 0.1 value, meaning strong regularization and a higher penalty for the wrong predictions. However, for higher epochs, the model started converging and eventually, F-score declined afterwards.

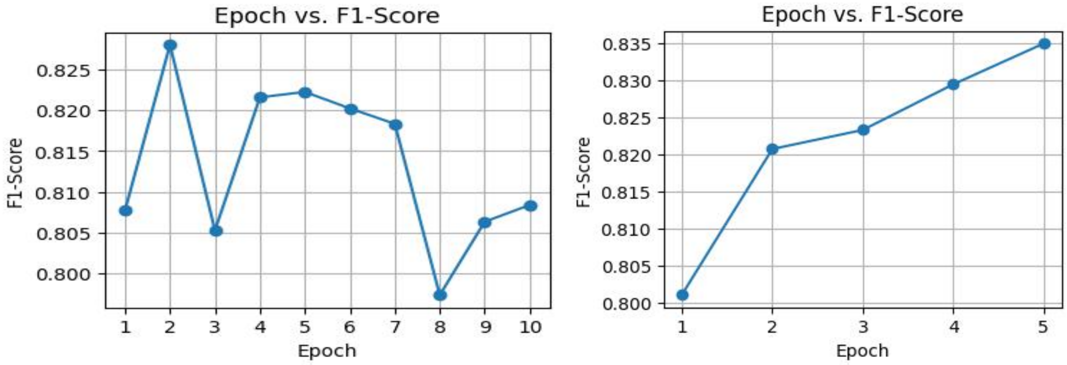


Figure 4: F1-score with 10 epochs and decay_weight 0.1 and with 5 epochs and decay_weight 0.01 respectively

6.2 Case Study 2: Optimal values of hyperparameters for the model

In this case study, rating categorization was kept similar to that of the first case and the number of epochs chosen was 5, for 10000 sample data. The decay_weight parameter was changed to 0.1 while keeping the learning rate at 2e-5. *Batch size =32* was chosen for this case study. It was observed that with slightly lower decay_weight, F1-score gradually increases suggesting a positive effect of hyperparameter value *batch size.* Fig.5 visualizes the confusion matrix suggesting the model was successful in capturing positive and negative sentiments, however, for neutral comments it struggled to classify correctly. However, the best part of this case study is, despite the review being "neutral", the model has largely classified the review as "bad" which has different meanings for different users.

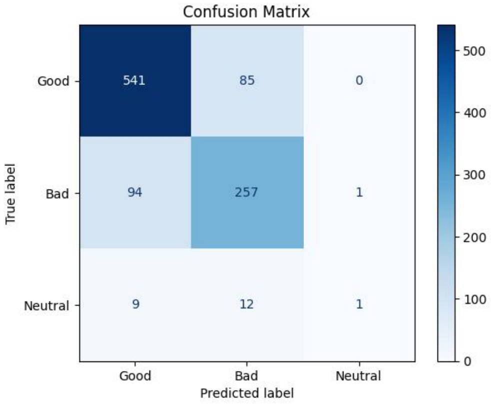


Figure 5: Confusion metrics for Case study2

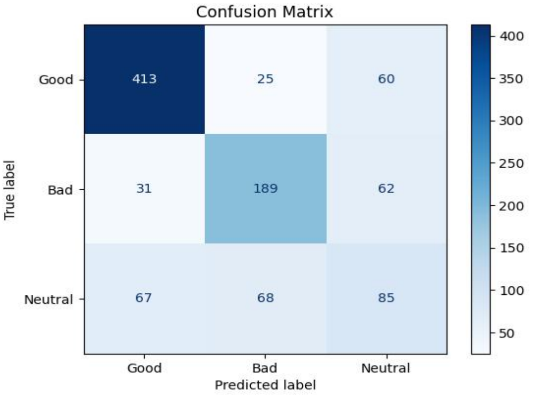


Figure 6: Confusion metrics for Caste study 3

6.3 Case Study 3 : Effect of rating categorization on model performance

In the third case, the rating categorization was changed to **1-3**, **4-7** and **8-10** as *bad, neutral* and *good* while keeping the number of records and other hyperparameters the same as in Case Study 2.Fig.6 shows the confusion matrix for case study 3 and it clearly shows that the rating categorization is leading to a surge in false positives and classification of *bad* comments as *neutral* and eventually *F1-score of the* experiment was around **73**% suggesting the categorization of rating was not fruitful.

6.4 Discussion

Comprehensive testing with different data and hyperparameter values was conducted. As earlier mentioned, we are using data from three different data sources with varying comments about drugs. The results achieved were satisfactory with an F1 score of almost 84%. The results achieved are less than the state-of-the-art technique *topicT_AtNN* model(Job et al., 2023) which achieved an F1-score of 0.9105 for the sentiment analysis of the overall review. However, we achieved a better F-score than the basic BERT (Job et al.,2023) model which had an F-score of 0.8157. The *topicT_AtNN model* was trained and tested on significantly lesser data with just over 3000 rows. Whereas, we have trained and tested our model by taking samples from a corpus of over 0.4 million rows and testing different chunks of data from the source. The difference in F-score could be because of variation in the number of rows in the data. Moreover, the medBERT.de(Bressem et al., 2023) model that we are using from the Huggingface Hub is trained in German medical literature.

In the beginning, the plan was to use medBERT (Rasmy et al.,2021) which was trained on over of over 20 million patients records from Electronic Health Records(EHR) database that consists of over 600 hospitals and clinics in the United States . The pre-trained model was not publicly available and despite writing multiple correspondences for authorization to use the pre-trained medBERT, no reply was received. Hence, we had to use medBERT.de trained on a German corpus.

7 Conclusion and Future Work

Transformer models and their variants are extensively being used in the area of sentiment analysis. There are various variants of transformer models trained on a different corpus of data sources. This study makes use of pre-trained medBERT.de from Hugging Face Hub for the overall sentiment analysis of the drug reviews from three different data sources. The model shows improvement in F-score and accuracy from the base BERT model in the classification task. The study tried to pick the medical terms from the reviews posted and based on the sentiment associated with the medical term, the overall review was classified into “Positive”, “Neutral” and “Negative”. An F1-score of 84% and precision of over 83% indicates that the model was successful in capturing negative medical terms from the review. The hyperparameter tuning approach has resulted in different result figures indicating that this has a positive effect on the performance of the model. Different values for *decay_weight*, *epochs,* and *batch size* were chosen for the experiments and have helped in arriving at the conclusion that the model performs best for *decay_weight 0.1,* and *5 epochs* with *batch size 32.*

Despite achieving a fairly good F1-score, the model could still be improved. The chosen medBERT.de has linguistic and location constraints as it was trained on the German medical literature corpus. This would have resulted in overlooking or overweight assignment to any feature in the review, resulting in incorrect classification. In future, a transformer model trained on English corpus like medBERT (Rasmy et al., 2021) could be more helpful for the task. Future research should widen the study to incorporate additional healthcare information and concentrate on improving the model's performance outcomes by generalising it throughout the medical domain.

References

Job, S., Tao, X., Li, Y., Li, L. and Yong, J. (2023). Topic Integrated Opinion-Based Drug Recommendation With Transformers. IEEE Transactions on Emerging Topics in Computational Intelligence, pp.1–11. doi:https://doi.org/10.1109/tetci.2023.3246559.

Bressem, K.K., Papaioannou, J.-M., Grundmann, P., Borchert, F., Adams, L.C., Liu, L., Busch, F., Xu, L., Loyen, J.P., Niehues, S.M., Augustin, M., Grosser, L., Makowski, M.R., Hugo J.W.L. Aerts and Löser, A. (2023). MEDBERT.de: A Comprehensive German BERT Model for the Medical Domain. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2303.08179.

Abuka, G., 2023. Text Summarization and Sentiment Analysis of Drug Reviews: A Transfer Learning Approach (Doctoral dissertation, Middle Tennessee State University).

Lin, H., Zhu, J., Xiang, L., Zhai, F., Zhou, Y., Zhang, J. and Zong, C., 2023. Topic-Oriented Dialogue Summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Suhartono, D., Purwandari, K., Jeremy, N.H., Philip, S., Arisaputra, P. and Parmonangan, I.H. (2023). Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews. Procedia Computer Science, [online] 216, pp.664–671. doi:https://doi.org/10.1016/j.procs.2022.12.182.

Sultana, A., Chowdhury, N.K. and Chy, A.N., 2022, October. Csecu-dsg@ smm4h'22: Transformer based unified approach for classification of changes in medication treatments in tweets and webmd reviews. In Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task (pp. 118-122).

Gawich, M. and Alfonse, M. (2022). A Proposed Model for Drugs' Review Analysis and Adverse Drug Reaction Discovery. 2022 7th International Conference on Mathematics and Computers in Sciences and Industry (MCSI). doi:https://doi.org/10.1109/mcsi55933.2022.00028.

Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55. doi:https://doi.org/10.1007/s10462-022-10144-1.

Alantari, H.J., Currim, I.S., Deng, Y. and Singh, S. (2021). An Empirical Comparison of Machine Learning Methods for Text-based Sentiment Analysis of Online Consumer Reviews. International Journal of Research in Marketing. doi:https://doi.org/10.1016/j.ijresmar.2021.10.011.

Li, L., Yang, L. and Zeng, Y. (2021). Improving Sentiment Classification of Restaurant Reviews with Attention-Based Bi-GRU Neural Network. Symmetry, 13(8), p.1517. doi:https://doi.org/10.3390/sym13081517.

Wu, P., Li, X., Ling, C., Ding, S. and Shen, S., 2021. Sentiment classification using attention mechanism and bidirectional long short-term memory network. Applied Soft Computing, 112, p.107792.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C. and Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, 4(1). doi:https://doi.org/10.1038/s41746-021-00455-y.

Colón-Ruiz, C. and Segura-Bedmar, I., 2021. Semi-Supervised Generative Adversarial Network for Sentiment Analysis of drug reviews. Institute of Electrical and Electronics Engineers (IEEE).

Fan, B., Fan, W., Smith, C. and Garner, H. "Skip" (2020). Adverse drug event detection and extraction from open data: A deep learning approach. Information Processing & Management, 57(1), p.102131. doi:https://doi.org/10.1016/j.ipm.2019.102131.

Liu, B., 2020. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press.

Colón-Ruiz, C. and Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics, [online] 110, p.103539. doi:https://doi.org/10.1016/j.jbi.2020.103539.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Gräßer, F., Kallumadi, S., Malberg, H. and Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. Proceedings of the 2018 International Conference on Digital Health. doi:https://doi.org/10.1145/3194658.3194677.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Gopalakrishnan, V. and Ramaswamy, C. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. Journal of Applied Research and Technology, 15(4), pp.311–319. doi:https://doi.org/10.1016/j.jart.2017.02.005.

Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S. and Gonzalez, G.H., 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of biomedical informatics, 62, pp.148-158.

Noferesti, S. and Shamsfard, M. (2015). Resource Construction and Evaluation for Indirect Opinion Mining of Drug Reviews. PLOS ONE, 10(5), p.e0124993. doi:https://doi.org/10.1371/journal.pone.0124993

Ali, T., Sokolova, M., Schramm, D. and Inkpen, D., 2013, September. Opinion learning from medical forums. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 18-24).

Nikfarjam, A., Emadzadeh, E. and Gonzalez, G. (2012). A Hybrid System for Emotion Extraction from Suicide Notes. Biomedical Informatics Insights, 5s1, p.BII.S8981. doi:https://doi.org/10.4137/bii.s8981.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J. and Gonzalez, G. (2010). Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. [online] Association for Computational Linguistics, pp.117–125. Available at: https://aclanthology.org/W10-1915.pdf.