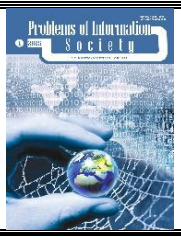

www.jpis.az

Determining the effectiveness of medications based on patient reviews collected on medical social media

Nargiz Shikhaliyeva

Institute of Information Technology, B. Vahabzade str., 9A, AZ1141, Baku, Azerbaijan

shikhaliyeva.nara@gmail.com

orcid.org/0000-0002-5427-0765

ARTICLE INFO

http://doi.org/10.25045/jpis.v16.i1.07

Article history:
Received 04 September 2024
Received in revised form
06 November 2024
Accepted 23 January 2025

Keywords:
Medical social media
Sentiment analysis
Medical decision making
Machine learning
Drug review
Classification
Opinion Mining

ABSTRACT

This article highlights solution of the problem of determining the medications' effectiveness based on sentiment analysis of patient reviews collected in the medical segment of social media. Public opinion about media subjects (physicians, nurses, clinics, pharmaceutical companies, etc.) can be determined based on the information collected in medical social media. One of the most discussed topics in medical social media is related to medications (drugs), their effectiveness, and determining public opinion based on collected user comments is one of the current issues. To analyze patient reviews about drugs, the Kaggle platform drugsComTest_raw.csv medical database is used, lexicon-based sentiment analysis, statistical methods and machine learning algorithms are applied. The problem is solved in stages on the patient-disease, disease-drug and patient-drug segments, the issues of which diseases are most often used for drugs, and which drugs are most often used for each disease are resolved. Based on the integration of the results obtained from the problem solutions, a mechanism for forming public opinion on the effectiveness of drugs is developed. The proposed approach takes into account not only positive but also negative opinions when determining public opinion about the effectiveness of drugs. Such results can be used to support appropriate decision-making in the healthcare sector, specifically in pharmaceutical companies.

1. Introduction

Improving the quality of medical services and ensuring patient satisfaction are among the priority issues in the healthcare system. In recent years, data collected on digital platforms has become an important source of information for determining public satisfaction. Patient reviews collected on medical social networks, websites, blogs, and various online platforms regarding physicians, nurses, clinics, medicines, etc. represent public opinion about media subjects, support medical decision-making based on their analysis, and act as one of the factors shaping public health (Mammadova et al., 2023; Tang, J. E. et al., 2023; Yadav, S. et al., 2018). With this approach, an objective assessment of medical services is realized. Analysis of reviews allows for increased transparency in the healthcare system and improved service quality. Analysis of reviews not only assesses the quality of medical services, but also identifies potential problems and areas for further development. Thus, regular analysis of patient reviews is becoming an important tool for increasing transparency in the healthcare sector and further improving services (Issam. et al., 2021; Mammadova et al., 2020). In order to make more objective and transparent decisions taking into account public opinion, a mechanism for evaluating the performance of doctors and medical institutions is proposed (Mammadova et al., 2019; Mammadova et al., 2020; Mammadova et al., 2022). Recently, researchers focus on the analysis of patient reviews in medical decision-making, assessment of drug efficacy, and improvement in this healthcare segment

(Cavalcanti et al., 2017; Gräßer et al., 2018; Panda et al., 2022).

Patients share their experiences with drug use writing on various platforms, posting important information about the effectiveness and side effects of treatment methods in these reviews. However, since this information is numerous and in various formats, its analysis is complex and difficult. Moreover, since these reviews contain unstructured information and emotional expressions, their correct processing requires distinct approaches. In this regard, natural language processing (NLP) and machine learning (ML) technologies are used (Alexander et al., 2022). NLP methods extract meaning and structure it by analyzing text. Machine learning algorithms achieve accurate and automated results. Through these technologies, it is possible to analyze emotions and thoughts in user reviews, and determine whether the reviews are positive, negative, or neutral. This process helps to understand public opinion more clearly and accurately. Therefore, determining the rating of drug effectiveness based on data collected in the patient-drug segment of medical social media is a relevant issue.

2. Related works

(Vijayaraghavan et al., 2020) analyze reviews of various drugs. The study's dataset is obtained from the UCI machine learning archive. The drug ratings are evaluated on a 10-point scale based on the reviews presented in text format in the database. The database is divided into 75% training and 25% test data. Drug reviews are classified into 3 groups as positive (10-7), negative (1-4), and neutral (4-7). The solution to the problem of determining drug ratings is based on reviews of drugs related to both similar and different medical conditions.

The study mainly uses various positioning methods such as term frequency–inverse document frequency (TFIDF) and count vectorization (CV). Results are obtained in the database for medical conditions such as "Birth control", "Depression", and "Pain".

(Das et al., 2021) develops a training model to predict the type of disease based on the name of the drug and its reviews. Various machine learning-based prediction methods are applied to solve the problem. The performance of these methods is compared based on metrics such as precision, recall, F1-Score, and accuracy.

In (Shiju et al., 2021), various classification models are built to classify drugs based on collected patient reviews. In this regard, machine learning and deep learning methods are used, and supervised machine learning models such as Random Forest and Naive Bayes are built. In addition, transformer-based neural network models that use raw text as input for classification, including BERT, BioBERT, Bio_ClinicalBERT, RoBERTa, XLNet, ELECTRA, and ALBERT, are also built, and the Bio_ClinicalBERT model is determined to outperform other models with an accuracy of 87%.

(Mishra, 2021) performs sentiment analysis of patient reviews after drug use through machine learning boosting algorithms. The database includes patient reviews of a number of specific drugs, patient diseases, and ratings (from 1 to 10) representing patient satisfaction with the drugs. The study builds LightGBM, XGBoost, and CatBoost models to classify reviews as positive and negative, with LightGBM performing best with an accuracy of 88.89%.

(Colón-Ruiz et al., 2019) develop a model to predict patient satisfaction with each drug based on patient reviews. A CNN model is developed to classify reviews, and the precision, recall, and F1-score indicators of the built model are shown to be better than support vector machine (SVM) classification methods.

In (Basiri et al., 2020), two deep fusion models are proposed using machine learning and deep learning traditional methods to analyze patient reviews about drugs in the drugs.com dataset. As a result of the study, the accuracy indicators accuracy and F1-Measure increased by 4%.

(Jiménez-Zafra et al., 2019) analyze reviews about drugs and doctors extracted from internet sources using machine learning and lexicon-based sentiment analysis approaches, and the classification of reviews about drugs is reported to be more difficult.

In (Nasrullah Makhdom et al., 2024), the results of the analysis based on a comprehensive lexicon-based model on the *UCI_Drug* database are presented, and the classification of drugs according to patient reviews is demonstrated.

Studies show that in the analysis of patient reviews about drugs, supervised machine learning algorithms, LASA, and deep learning methods are often used for classification and comparison, and the problem is solved based on

information only in the patient-drug segment of medical social media. Based on these traditional methods of analysis of information collected in medical social media, this article highlights the solution to the problem of assessing the effectiveness of drugs in the integration of patient-drug, patient-disease (or patient-medical condition) and disease-drug relationships.

Problem statement. The purpose of the present study is to determine public opinion about the effectiveness of drugs based on reviews about drugs. Therefore, patient reviews (text-type comments) of patients about the effectiveness, side effects and use of drugs are analyzed.

3. Materials and methods

3.1. Problem solution stages

To determine the effectiveness of drugs based on patient reviews of drugs, the following sub-problems are intended to be solved:

1. Collection and processing of reviews in the patient-drug segment.
2. Sentiment analysis of patient reviews and verification of the results accuracy.
3. Arrangement of reviews for various medical conditions in the patient-disease segment.
4. Arrangement of drugs for medical conditions.
5. Determination of the drugs rating for the same medical condition.

3.2. Problem solution

The problem is solved in the following stages:

Stage 1: An open database called *drugsComTest_raw.csv* from Kaggle is selected for the reviews analysis.

The database includes the following main columns (fig.1):

- Drug name: Drugs used by patients in various medical conditions, i.e., diseases. There are many drugs that can be used for each condition.
- Condition: Describes the medical condition for which the drug is used.
- Review: Includes text-based comments from patients about the drug.
- Rating: A metric set to represent the overall satisfaction or effectiveness of the drug.
- Useful Count: Indicates how many users have used the drug.

At this stage, Data Pre-Processing is applied to analyze the raw text data, and the data is converted into a usable form. The following operations are performed at this stage (Kamran Kowsari et al., 2019):

1. Converting text into lowercase: All uppercase letters in the text are converted to lowercase, thus ensuring that different spellings such as “Quick” and “quick” represent the same word;
2. Removing punctuation marks: Periods, commas, and other punctuation marks in the text are removed from the text to avoid creating obstacles during analysis.
3. Number removal: numbers in the text are removed as the analysis is based on word-based data.
4. Stop word removal: in this stage, words such as “and”, “the”, “is” that are often used but are not useful for sentiment analysis are removed from the text. This helps to make the text more concise and suitable for analysis.

```
df = pd.read_csv('/content/drugsComTest_raw.csv',low_memory=False)
print(df.shape)
```

```
(53766, 7)
```

```
df.head()
```

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	163740	Mirtazapine	Depression	"I've tried a few antidepressants over th...	10	28-Feb-12	22
1	206473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has done ...	8	17-May-09	17
2	159672	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9	29-Sep-17	3
3	39293	Contrave	Weight Loss	"Contrave combines drugs that were used for al...	9	5-Mar-17	35
4	97768	Cyclafem 1 / 35	Birth Control	"I have been on this birth control for one cyc...	9	22-Oct-15	4

Fig. 1. Access to Kaggle’s open database drugsComTest_raw.csv for sentiment analysis

5. Lemmatization: words are reduced to their root form, for example, the words "running" and "runs" are normalized to "run". This ensures that different forms with the same root are accepted as the same concept in the analysis process.

Stage 2. In this stage, a lexicon-based sentiment analysis approach is applied. This approach determines the tonality (positive, negative or neutral) of each opinion using the *TextBlob* library. *TextBlob's* sentiment analysis function calculates the polarity (positive or negative) value of that opinion by analyzing the emotional expressions in the text. Based on the polarity value, it is determined whether the opinion is positive, negative or neutral. Then, the data is divided into train and test parts and the texts are vectorized with TF-IDF. The Logistic Regression model is trained and predictions are made on the test data. The results are evaluated with the accuracy and evaluation of the model (accuracy_score and classification_report) (fig. 2). This stage provides the creation of a sentiment analysis model for text-type reviews and their evaluation.

```
Model Accuracy: 87.97%

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.82      0.86      3586
           1       0.88      0.95      0.91      6613
           2       0.83      0.40      0.54       555

    accuracy                           0.88     10754
   macro avg       0.87      0.72      0.77     10754
weighted avg       0.88      0.88      0.87     10754
```

Fig. 2. Accuracy result of the Logistic Regression model

As a result of the analysis, it is determined that out of a total of 53766 reviews, 32920 are positive, 17956 are negative, and 2890 are neutral. A visual representation of the results in percentages is illustrated in fig. 3.

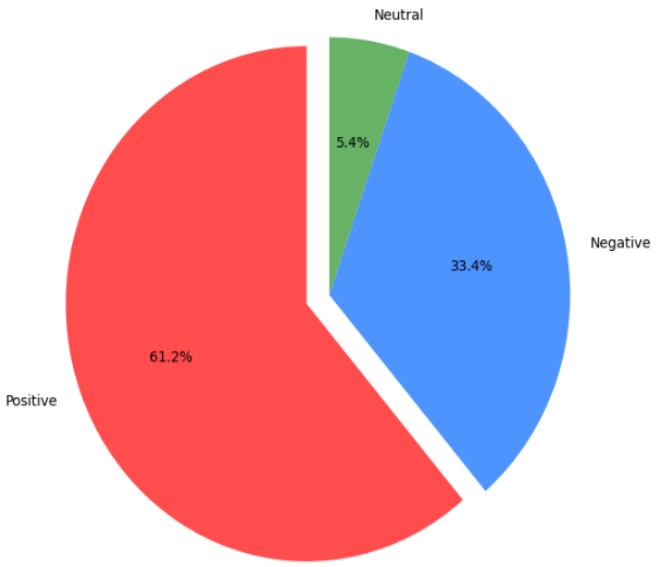

Fig. 3. Visual representation of the results obtained from the sentiment analysis of reviews in the Kaggle's open database *drugsComTest_raw.csv*

Stage 3. This stage explores which medical condition (disease) the reviews of patients using the drug are related to, and the study continues in the patient-disease segment. In this stage, the 5 medical conditions with the highest number of reviews are selected (Acne, Anxiety, Birth Control, Depression, and Pain), and SA of the reviews is performed on them. The distribution of the results into negative, neutral, and positive categories is given in Table 1, and a visual representation of these results is presented in fig. 4.

Table 1: Results of sentiment analysis across five medical conditions

Condition	Total	Positive	Negative	Neutral
Birth Control	9648	5696	3835	117
Depression	33095	2043	936	116
Pain	2100	1369	537	194
Anxiety	1908	1290	539	79
Acne	1847	1210	608	29

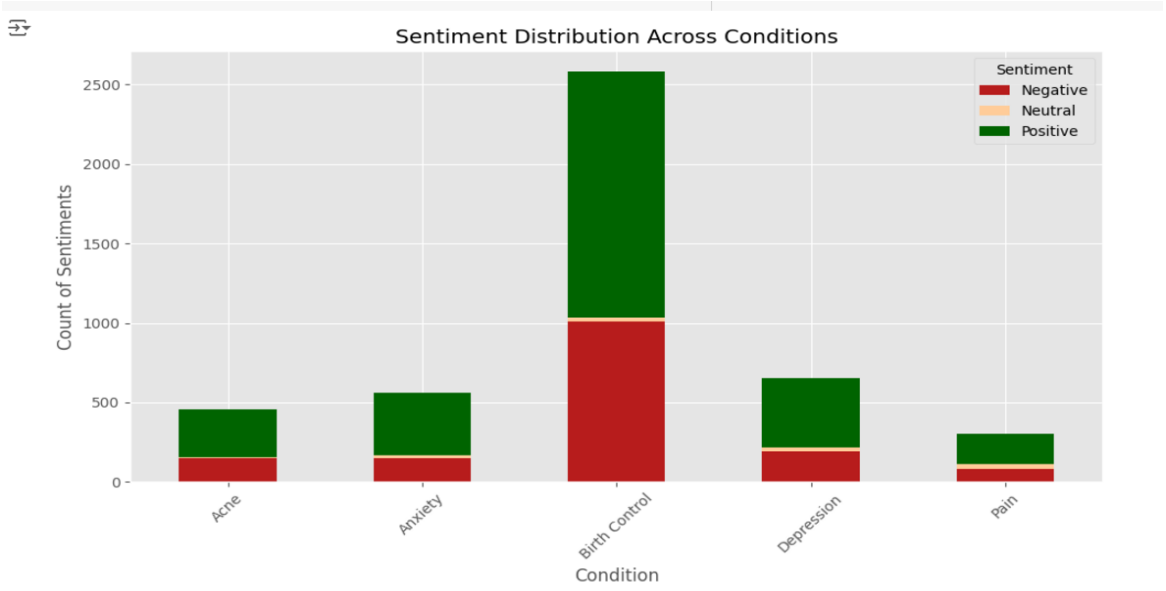

Fig. 4. Visual representation of sentiment analysis of patient reviews of medications across five medical conditions

Stage 4. In this stage, a sentiment analysis of reviews of medications used for the 5 medical conditions considered is performed, and a ranked list of medications is obtained according to the number of positive reviews for each condition (Table 2).

Stage 5. Based on the integrity of the results obtained in stages 3 and 4, the public opinion regarding the effectiveness of the drugs is interpreted as follows:

1. Acne: Among the drugs for acne, Accutane has the highest number of positive reviews and outperforms other drugs in terms of effectiveness. Accutane, in addition to having the highest number of positive reviews, is also the drug with the lowest percentage of negative reviews among those with the highest number of positive reviews for the acne condition. According to the results, it is found that there are no neutral reviews about this drug.

Table 2. Sentiment analysis results of the top three drugs across five medical conditions

Condition	*Drug Name*	*Positive (%)*	*Neutral (%)*	*Negative (%)*
Acne	Accutane	74.57	0	25.43
	Doxycycline	66.42	3.47	30.11
	Isotretinoin	64.36	0.93	34.71
Anxiety	Bupropion	75	0	25
	Gabapentin	74.67	1.33	24
	Escitalopram	72.65	1.68	25.67
Birth Control	Etonogestrel	64.53	0.75	34.72
	Ethinyl estradiol / norethindrone	61.11	0.86	38.03
	Levonorgestrel	57.83	1.55	40.62
Depression	Bupropion	71.46	3.35	25.19
	Sertraline	70.77	2.93	26.30
	Escitalopram	67.88	3.85	28.27
Pain	Oxycodone	64.88	12.70	22.42
	Tramadol	63.79	8.60	27.61
	Gabapentin	60.53	10.53	28.95

2. Anxiety: Among the drugs for anxiety, Bupropion is the drug with the most positive reviews. It is also found that there are no neutral reviews about this drug.

3. Birth Control: Among the drugs for birth control, Etonogestrel and Etonogestrel have the highest percentage of positive reviews. This drug also receives fewer negative reviews than others. According to the results obtained, it can be noted that it has the lowest percentage of neutral and negative reviews among the drugs with the most positive reviews.

4. Depression: Among the drugs for depression, Bupropion has the best results and receives more positive reviews from users. According to the results obtained, it can be noted that it has the lowest percentage of negative reviews among the drugs with the most positive reviews for depression. The results of the drugs Sertraline and Bupropion are very close to each other in terms of the percentage of positivity of patient reviews. However, Sertraline is the drug with the most positive reviews for depression, but it is also the drug with the most negative reviews.

5. Pain: Oxycodone has the highest percentage of positive reviews among the drugs for pain treatment, and it is also the drug with the least negative reviews among the drugs with the most positive reviews for pain. Gabapentin is the drug with the highest percentage of negative reviews among the drugs with the most positive reviews for pain.

Thus, in the patient-drug segment of medical social media, issues such as which diseases patients turn to for the most drugs, which drugs they prefer for each disease, are resolved. Based on the results obtained from the integration of these issues, an interpretation of public opinion regarding the effectiveness of drugs is given.

4. Discussion

A large amount of information is collected on medical social networks, websites, blogs and various online platforms. Most of this information covers user reviews. This unstructured information allows us to take public opinion into account when making more objective and transparent decisions in solving medical problems. The article considers the solution of the problem of assessing the effectiveness of drugs based on patient reviews collected in the patient-drug segment of medical social media, and presents the solution of the problem step by step using the Kaggle's database *drugComTest_raw.csv* (fig. 1). 53766 patient reviews (text) about drugs included in the database are classified as "positive", "negative" and "neutral" based on LBSA, and then the accuracy of SA is checked using the Logistic Regression machine learning method (fig. 2). The verification confirms that the classification obtained based on SA provides high performance (0.88). Then, a visual representation of the classification results is depicted in fig. 3. This image illustrates the general picture of patient satisfaction in the patient-drug segment of medical media. In order to determine which diseases (or medical conditions) the drugs are used for and which diseases the patients mostly use the drugs for, the study is continued on the patient-disease and disease-drug segments. Therefore, in the next stage, the diseases are ranked according to the number of patient reviews in the patient-disease segment and the top five are selected (table 1). A visual representation of the results of the SA of reviews related to these diseases is given in fig. 4. The 3 drugs with the highest number of uses across these five main medical conditions are selected, and the results of the classification of reviews on these drugs are presented in table 2. Then, the results are interpreted to determine the effectiveness of the patient reviews of the drugs corresponding to the medical conditions, based on being positive, negative and neutral.

The solution of the problem is carried out in the python environment using the numpy, pandas, seaborn, TextBlob, etc. libraries.

The advantage of this article is that the solution to the problem in question is considered not only at the patient-drug level, but also at the patient-disease-drug level. The research algorithm is presented in a methodological sequence and step-by-step, clearly, based on the relevant scientific and theoretical basis.

The main problem with the study is that the information in medical social networks is not directly accessible, therefore, in order to obtain the necessary information, it is necessary to access open databases of companies that collect relevant information. Unfortunately, at present, such databases are few and most of them are inaccessible or can be obtained at a very high price. In this regard, the relevant database of the Kaggle company, which structures and collects in a single database the reviews written by patients on medical social networks about medicines

suitable for various diseases.

In further studies, it is planned to assess patient satisfaction in a wider and more detailed spectrum, and conduct a more detailed analysis taking into account a number of demographic characteristics of patients. Such approaches are aimed at understanding the needs of patients more accurately and offering personalized healthcare services. For example, by analyzing how positive and negative reviews vary between different age groups and genders of patients, more effective treatment methods and medicines aimed at the needs of specific groups can be identified. Moreover, the analysis of demographic data can help develop healthcare services and medicines in a more individual and needs-based manner.

5. Conclusion

The article presented the solution to the problem of assessing the effectiveness of drugs based on public opinion, and used the Kaggle's *drugComTest_raw.csv* database to analyze patient reviews about drugs on medical social media. The Logistic Regression machine learning method was used to determine the tonality of reviews about drugs, and sentiment analysis was performed.

In order to solve the problem of determining the effectiveness of drugs according to patient reviews, the study:

- solved the problem of which diseases patients most often resort to drugs, in solving this problem, the reviews in the patient-disease segment were referred to and five diseases (Acne, Anxiety, Birth Control, Depression and Pain) were selected;

- solved the problem of which drugs are most often resorted to by diseases, in solving this problem, the reviews in the disease-drug segment were referred to and 3 drugs were selected for each disease.

- considered the formation of public opinion on the effectiveness of drugs in the integration of the results obtained in the patient-disease and disease-drug segments and the interpretation of the results was given.

Based on the proposed approach for determining the effectiveness of drugs, it is also interesting to evaluate their possible side effects. Thus, through patient reviews, it is possible to obtain detailed information not only about the effectiveness of drugs, but also about their side effects and how they are taken by patients. On the other hand, despite the high percentage of positive reviews of some drugs used for certain diseases, it is possible to find information about serious side effects in negative reviews (even if they are few in number). Such information is an important point in the development of personalized medicine, since drugs taken with high effectiveness by one patient may cause negative consequences and complications for others. Therefore, as a continuation of the study, a more in-depth analysis of negative reviews and an assessment of the effectiveness of drugs based on the information in them is of particular importance.

References

Alexander, G., Bahja, M., Butt, G.F. (2022). Automating large-scale health care service feedback analysis: sentiment analysis and topic modeling study. JMIR Medical Informatics, 10(4): 1-15. https://doi.org/10.2196/29385

Basiri, M.E., Abdar, M., Cifci, M.A., Nemati, S., Acharya, U.R. (2020). A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. Knowledge-Based Systems 198(2): 1-19. https://doi.org/10.1016/j.knosys.2020.105949

Cavalcanti, D., Prudêncio, R. (2017). Aspect-based opinion mining in drug reviews. In: Oliveira, E., Gama, J., Vale, Z., Lopes-Cardoso, H., eds., Progress in Artificial Intelligence (EPIA), Lecture Notes in Computer Science, Springer Cham, 10423: 815–827. https://doi.org/10.1007/978-3-319-65340-2_66.

Colón-Ruiz, C., Segura-Bedmar, I. & Martínez, P. (2019). Sentiment analysis on health domain: Analyzing patient comments on drugs. Procesamiento de Lenguaje Natural. 63: 15–22. https://doi.org/.26342/2019-63-1.

Das, S., Kumar-Mahata, S., Das, A., Deb, K. (2021). Disease prediction from drug information using machine learning. American Journal of Electronics & Communication, 1(4): 16–21. https://doi.org/10.15864/ajec.1403.

Gräßer, F., Kallumadi, S., Malberg, H. and Zaunseder, S., (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In Proceedings of the 2018 International Conference on Digital Health 121-125. https://doi.org/10.1145/3194658.3194677

Issam, A., Saida, E., Fatna, E. (2021). Sentiment Analysis of Health Care: Review. E3S International Congress on Health Vigilance (VIGISAN 2021), Web of Conferences 319: 1-7. https://doi.org/10.1051/e3sconf/202131901064

Jiang, M.; Liang, Y., Feng, X.; Fan, X., Pei, Z., Xue, Y.; Guan, R., (2018). Text classification based on deep belief network and softmax regression. Neural Computing and Applications, 29: 61–70. https://doi.org/10.1007/s00521-016-2401-x

Jiménez-Zafra, S.M., Martín-Valdivia, M.T. & Molina-González, M.D., Ureña-López, L.A. (2019). How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. Artificial Intelligence in Medicine. 93: 50–57. https://doi.org/10.1016/j.artmed.2018.03.007

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, Donald E. Brown (2019). Text Classification Algorithms: A Survey Information, 10(4):1-68.

https://doi.org/10.3390/info10040150

Mammadova, M.H., & Jabrayilova, Z.G. (2019). Electronic medicine: formation and scientific-theoretical problems. Baku: "Information Technologies" publishing house, 1-318. https://ict.az/uploads/files/E-medicine-monograph-IIT-ANAS.pdf

Mammadova, M.H., Jabrayilova, Z.G., & Shikhaliyeva, N.R. (2022). Lexicon-based sentiment analysis of medical data. Technology transfer: fundamental principles and innovative technical solutions, 7–10. https://doi.org/10.21303/2585-6847.2022.002671

Mammadova, M., Jabrayilova, Z., Shikhaliyeva, N. (2023). Development of decision-making technique based on sentiment analysis of crowdsourcing data in medical social media resources. Eastern-European Journal of Enterprise Technologies, 5 (3 (125)): 75–85. https://doi.org/10.15587/1729-4061.2023.289989

Mammadova, M.H., Jabrayilova, Z.G., Isayeva, A.M. (2020). Conceptual approach to the use of information acquired in social media for medical decisions. Online Journal of Communication and Media Technologies, 10(2): 1-15. https://doi.org/10.29333/ojcmt/7877

Mishra, S. (2021). Drug review sentiment analysis using boosting algorithms. International Journal of Trend in Scientific Research and Development (IJTSRD), 5(4): 937–941. https://www.ijtsrd.com/papers/ijtsrd42429.pdf

Nasrullah Makhdom, H.N. Verma, Arun Kumar Yadav (2024). Sentiment Analysis of Patient Review in Drug. International Journal of Scientific Research in Computer Science Engineering and Information Technology. 11(5): 738-748. https://doi.org/10.32628/cseit241029

Panda, B., Panigrahi, C.R., Pati, B. (2022). Exploratory data analysis and sentiment analysis of drug reviews. Computación y Sistemas, 26(3): 1191-1199. https://doi.org/10.13053/CyS-26-3-4093.

Shiju, A., He, Z. (2021). Classifying drug ratings using user reviews with transformer-based language models. IEEE 10th International Conference on Healthcare Informatics (ICHI).163-169. https://doi.org/10.1101/2021.04.15.21255573

Tang, J. E., Arvind, V., Dominy, C.; White, C.A., Cho, S.K., Kim, J.S. (2023). How are patients reviewing spine surgeons online? A sentiment analysis of physician review website written comments. Global Spine Journal, 13(8): 2107-2114. https://doi.org/10.1177/21925682211069933.

Vijayaraghavan, S., Basu, D. (2020). Sentiment analysis in drug reviews using supervised machine learning algorithms. 1-9. https://doi.org/10.48550/arXiv.2003.11643

Yadav, S., Ekbal, A., Saha, S., & Bhattacharyya, P. (2018). Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2790-2797. https://aclanthology.org/L18-1442/.

Kaggle.com. Kaggle University. UCI ML Drug Review dataset. https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018