LEXICON-BASED SENTIMENT ANALYSIS OF MEDICAL DATA

Masuma Mammadova
Corresponding Member of Azerbaijan National Academy of Sciences, Doctor of Technical Sciences, Professor[1]
mmg51@mail.ru

Zarifa Jabrayilova
Doctor of Technical Sciences, Associate Professor
Chief Researcher[1]

Nargiz Shikhaliyeva
Engineer-Programmer[1]

[1]*Department of Number 11*
Institute of Information Technology
MInstry of Science and Education Republic of Azerbaijan
9 B Vahabzada str., Baku, Azerbaijan, AZ1141

Abstract: The article explores the possibilities of applying sentiment analysis for the use of information collected in the medical social media environment in medical decision-making. Opinions and feedbacks of medical social media subjects (physician, patient, health institution, etc.) make media resources an important source of information. The information collected in these sources can be used to improve the quality of health care and make decisions, taking into account the public opinion. Researches in this field have actualized the application of artificial intelligence methods, i.e., sentiment analysis methods. In this regard, it segments the medical social media environment in accordance with user relationships, and shows the nature of the information collected on each segment and its importance in decision-making to improve the quality of medical services. The possibilities of applying the lexicon-based sentiment analysis method for studying and classifying the collected data are explained in detail. The open database cms_hospital_satisfaction_2019 by the Kaggle company is used, and the opinions collected from patients about the services provided by a specific medical center are analyzed. This study analyzes opinions using the Valence Aware Dictionary and Sentiment Reasoner lexicon and classifies them as neutral, positive and negative and the implementation of this process is described in stages. The importance of the obtained results in decision-making regarding the better organization, evaluation and improvement of the activity of the medical institution is shown.

Keywords: medical social media resources, sentiment analysis, lexicon-based approach, machine learning.

1. Introduction

Social media emerged as a result of the rapid development of technology are virtual space resources for expressing people's thoughts, opinions and attitudes about certain topics, products and services. The information collected in this environment becomes a valuable resource for decision-making in relevant fields, and intelligent technologies are the utmost required tool for achieving successful results in this field [1, 2]. In this regard, medical social media tools, which are one of the main indicators of the formation of e-medicine, have led to the transformation of doctor-patient-medical institution relations, to the change of treatment-diagnosis and prevention methods, to the improvement of health monitoring processes. These tools have become an important resource for decision-making by considering public opinion in medical decisions. Today, physicians and patients turn to online platforms such as blogs, social media, and websites to share their thoughts on health issues. The large amount of medical information collected on social media sites, online forums, and personal blogs is becoming a source of better outcomes for physicians, patients, and medical institutions.

In [2, 3], the authors propose a conceptual approach for solving a number of medical decision-making issues based on the statistical analysis of information collected in medical social media, and in [4] they presented a methodology for evaluating the media activity of users. One of the important points here is the content analysis of the information related to the physicians-patient-medical institution, which are media subjects, and determining the opinion about media subjects in applications. Determining such an opinion may provide decisions support related to the evaluation of the activities of media subjects, better organization of work, improvement of the quality of health care provided to patients, etc.

In this study, let's show the possibilities and methods of applying sentiment analysis for determining mass opinion about medical media subjects in the data collected on medical social media resources. Let's also review the issue of sentiment analysis of patient opinions collected about the medical institution and interpret the results.

2. Methods

Currently, a large number of professional medical social communities have emerged in the Internet environment. *Sermo, Doximity, QuantiaMD, Among Doctors, Figure1, DoctorsHangout, MomMD, DailyRounds* are medical social networks where only doctors can register [5, 6]. *DailyRounds.org* reports about 1.3 million doctors being registered in more than 16 countries. Medical social networks *Ozmosis, Doc2Doc, Healtheva* are the best developed platforms that enable communication between patients and doctors. *Medihost.ru, adam.com, DoctorSpring, likar.info, health.mail.ru* are medical social networks designed for patient-patient communication.

In Azerbaijan, social societies such as "*Həkim.tap*", "*həkimsənaz.az*", "*sağlamolun.az*", "*doctormap.az*", "*Tibb.az*" bring together the doctors specializing in various fields of medicine on one portal [3, 4, 7]. Patients may get detailed information about doctors by using the "doctor-search" section in these networks, they can contact those doctors and ask them questions.

The expansion of medical social media, the activity of medical specialists, doctors, patients, and medical clinics on social media has led to the formation of various stakeholders and the emergence of virtual medical relations between them. In [3, 4], user relationships shaped in the medical social media environment are divided into the following segments:

– physician – physician. Through social media, physicians discuss effective treatment methods for various diseases with their colleagues;

– physician – patient. Physicians interact with patients and monitor the health of their patients, give advice, and at the same time, patients ask doctors about various problems;

– patient – patient. Patients share their thoughts and opinions about disease diagnosis, symptoms, treatment methods, medications, treatment methods prescribed by doctors for the same disease, treatment results, etc.;

– physician – patient – nurse. Physician and nurse are in contact with the patient. The nurse performs the tasks of the physician regarding the patient, conveys the information received about the patient's health condition to the physician;

– physician – clinic. Medical institutions access the personal social media resources of doctors and collects information about their treatment methods, treatment results, activity, experience, etc.;

– patient – medical institution (clinic). Patients can use the social media site to get information about the clinic's work, medical staff, contacts, address and working hours, appointment rules, prices, etc.

Depending on the nature of the information collected in these relationship segments, scenarios are formed for making different

types of decisions related to improving the quality of medical care. For example, based on the patient surveys collected in the segment of doctor-patient relations it can be defined which doctor is the most consulted and in which field of medicine, the activity of women and men among e-patients, social media activity of e-patients by different age groups, regions, etc. However, based on these applications, determining "patient satisfaction" and "a doctor who has won mass sympathy" for various diseases requires the analysis of the content of the requests, and therefore, the use of sentiment analysis methods become essential.

3. Results

Sentiment analysis (SA), also known as opinion extraction, is a natural language processing that enables automatic classification of the content (opinion) expressed in a text [8, 9]. Although the history of SA goes back to the 1990s, the studies on its application has expanded since the emergence of Web 2.0, increased access to information generated by network users, and the proliferation of social media platforms. SA is currently involved in industry, economy, healthcare, etc. and used as a valuable tool in various fields.

Machine learning-based, lexicon-based and hybrid methods of SA are also available (**Fig. 1**) [10].

Lexicon-based SA approach is considered to be a simple approach. This approach refers to the sentiment lexicon, which consists of words and phrases commonly used to express positive and negative attitudes. Unlike machine learning-based and hybrid methods, this method does not require any training data [11].

The Valence Aware Dictionary and Sentiment Reasoner (VADER) approach is a lexicon and rule-based system. The VADER approach classifies the text by giving scores such as negative, positive, neutral and compound.

Manually developed sentiment lexicons are as follows [11]:

1. Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon is a subjective key-referencing lexicon of over 8,000 words, each classified as positive or negative. This lexicon includes 2718 positive and 4909 negative words, as well as a number of personal adjectives, adverbs, any POS (part of speech).

2. Bing Liu Lexicon is based on a list of English words (about 6800 words) expressing positive and negative opinions (feelings).

Auto-generated sentiment lexicons include:

1. NRC Hashtag Sentiment Lexicon classifies tweets by indicating the association with a positive emotion with a positive score, and the association with a negative emotion with a negative score.

2. Sentiment140 Lexicon.

3. SentiWordNet automatically classifies all WORDNET synsets according to their positive, negative and neutral degrees.

4. Discussion

Lexicon-based SA of medical data uses *Pandas, NumPy, Matplotlib, Seaborn, NLTK* libraries. SA with the *VADER* lexicon is performed in the Python programming environment [12] in the following stages [13, 14]:

1. Data Collection. An open database cms_hospital_satisfaction_2019 by Kaggle company, which expresses the attitudes (opinions) of patients about a certain medical center, is used [15], and a sentiment analysis of 442,401 patient opinions about the level of services provided by the medical staff and the center is performed using a lexicon-based approach.

2. Data Pre-Processing. This stage performs the process of data cleaning (tokenization), i.e., spaces, special characters, symbols are deleted and the remaining ones are called tokens.

3. The stage of Extraction Opinions prepares processed opinions for analysis.

4. Application of Lexicon Based Sentiment Analysis algorithm.

5. The stage of Classification of Opinions applies the VADER approach for sentiment analysis of opinions and classifies opinions.

6. Result stage interprets the classification results.

Fig. 2 shows a fragment from the *Kaggle cms_hospital_satisfaction_2019* database following the classification based on sentiment analysis. For database classification with VADER approach, columns "neg", "neu", "pos", "compound" are added to the dataset, the final result of these columns is represented in the column "comp_score".

The next step is the process of visualization of the obtained results using the *Bar plot*, a visualization tool of the *Python* programming environment (**Fig. 3**).

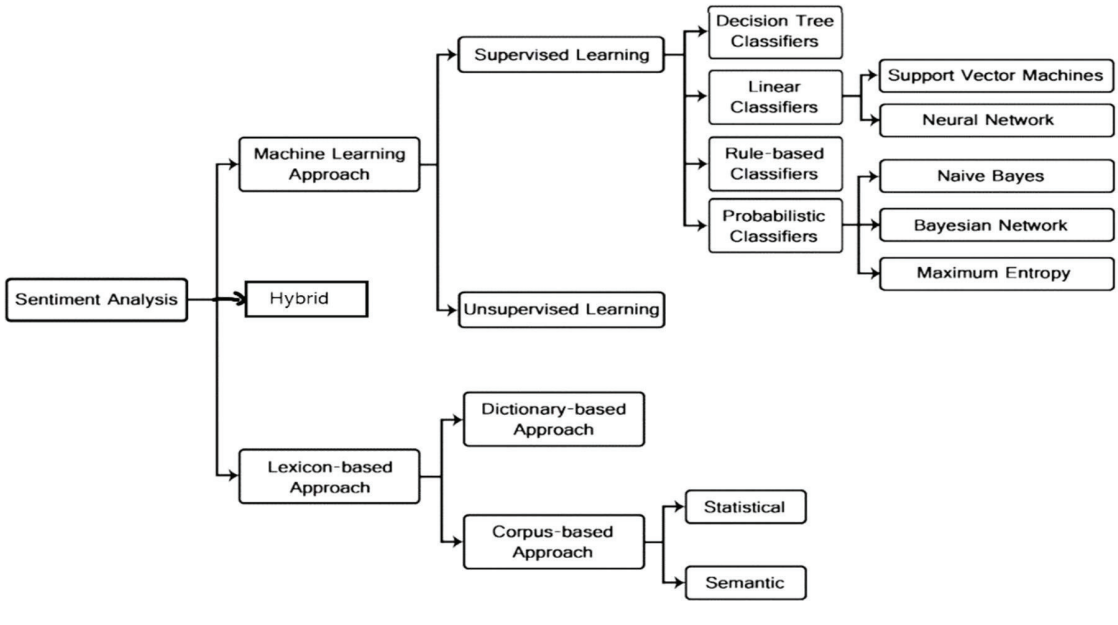

Fig. 1. Sentiment analysis methods [10]

As a result of the conducted research, it was determined that 218822 positive, 190280 neutral, and 33299 negative scores were received from 442401 pieces of information expressing the patient's opinion and included in the column "*HCAHP Answer Description*" in the database**. Fig. 4** shows a visual image of opinions classified by positive, neutral, negative scores.

```
vaders[["HCAHPS Measure ID","HCAHPS Answer Description", "comp_score"]].tail(-1)
```

	HCAHPS Measure ID	HCAHPS Answer Description	comp_score
1	H_COMP_1_A_P	Nurses "always" communicated well	pos
2	H_COMP_1_A_P	Nurses "always" communicated well	pos
3	H_COMP_1_A_P	Nurses "always" communicated well	pos
4	H_COMP_1_A_P	Nurses "always" communicated well	pos
5	H_COMP_1_A_P	Nurses "always" communicated well	pos
...	...	...	...
442396	H_STAR_RATING	Summary star rating	neu
442397	H_STAR_RATING	Summary star rating	neu
442398	H_STAR_RATING	Summary star rating	neu
442399	H_STAR_RATING	Summary star rating	neu
442400	H_STAR_RATING	Summary star rating	neu

442400 rows × 3 columns

Fig. 2. *Kaggle* open database *cms_hospital_satisfaction_2019* following the sentiment analysis classification

```
ax = sns.barplot(data=vaders, x='HCAHPS Answer Description', y='compound')
ax.set_title('Compund Score by answers')
plt.show()
```

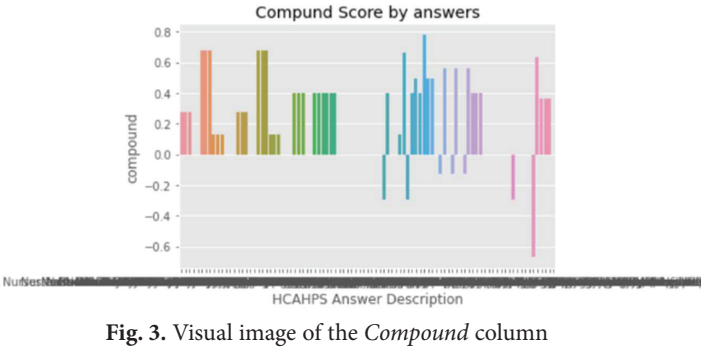

Fig. 3. Visual image of the *Compound* column

It turns out that although the vast majority of the opinions expressed by patients about the medical institution in the database used in the study are neutral, the positive opinions are much more than the negative ones.

5. Conclusions

Based on the data collected in medical social media resources, the necessity of studying public opinion for solving issues related to the quality of medical services was substantiated, and the possibilities of applying SA methods to solve this problem were explored. In order to represent the application possibilities of lexicon-based SA methods in the patient-medical institution segment of medical social media, the open database *cms_hospital_satisfaction_2019* by the *Kaggle* company and *Pandas, NumPy, Matplotlib, Seaborn, NLTK* libraries were used, and the collected data were analyzed in the *Python* environment with the *VADER* lexicon. The procedures for achieving the final opinion by classifying the patient opinions collected in the database as "neg", "neu", "pos" are described, and the ranking of the patient opinions about the medical institution by the specified classes was visually presented. The results of the conducted research can be used in solving some issues such as evaluating and improving the activity of medical institutions, determining patient satisfaction related to its activity, and in making relevant decisions.

Let's note that it is planned to conduct a similar study on other relational segments of medical social media resources, to examine the possibilities of applying other SA methods to solve the considered issues, and these are among the further research areas of the authors.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

```
[ ] fig, axs = plt.subplots(1, 3, figsize=(12, 3))
    sns.barplot(data=vaders, x='HCAHPS Answer Description', y='pos', ax=axs[0])
    sns.barplot(data=vaders, x='HCAHPS Answer Description', y='neu', ax=axs[1])
    sns.barplot(data=vaders, x='HCAHPS Answer Description', y='neg', ax=axs[2])
    axs[0].set_title('Positive')
    axs[1].set_title('Neutral')
    axs[2].set_title('Negative')
    plt.tight_layout()
    plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:8: UserWarning: Tight layout not applied. tight_layout cannot make a
```

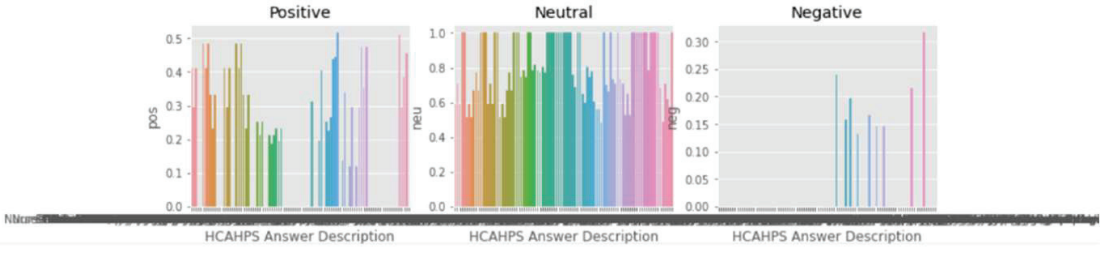

Fig. 4. Visualization of opinions classified by positive, neutral and negative scores

Financing
The study was performed without financial support.

Data availability
Manuscript has no associated data.

References

1. Mammadova, M., Jabrayilova, Z. (2019). Electronic medicine: formation and scientific-theoretical problems. Baku: "Information Technologies" publishing house, 319. Available at: https://ict.az/uploads/files/E-medicine-monograph-IIT-ANAS.pdf
2. Mammadova, M., Isayeva, A. (2018). E-health activity in social media environment. Problems of Information Society, 09 (1), 52–62. doi: https://doi.org/10.25045/jpis.v09.i1.05
3. Mammadova, M., Jabrayilova, Z., Isayeva, A. (2020). Conceptual Approach to the Use of Information Acquired in Social Media for Medial Decisions. Online Journal of Communication and Media Technologies, 10 (2). doi: https://doi.org/10.29333/ojcmt/7877
4. Issue Brief: Social Networks in Health Care: Communication, collaboration and insights. Produced by the Deloitte Center for Health Solutions. Available at: http://healthinformationandcommunicationsystems.pbworks.com/w/file/fetch/93972338/SM%204b%20Full.pdf
5. Fogelson, N. S., Rubin, Z. A., Ault, K. A. (2013). Beyond likes and tweets: an in-depth look at the physician social media landscape. Clinical Obstet Gynecol, 2013, 56 (3), 495–508.
6. Tibb.az. Your virtual doctor. Available at: https://tibb.az/home
7. Aattouchi, I., Elmendili, S., Elmendili, F. (2021). Sentiment Analysis of Health Care: Review. E3S Web of Conferences, 319, 01064. doi: https://doi.org/10.1051/e3sconf/202131901064
8. Khan, M. T., Khalid, S. (2015). Sentiment Analysis for Health Care. International Journal of Privacy and Health Information Management, 3 (2), 78–91. doi: https://doi.org/10.4018/ijphim.2015070105
9. Kausar, S., Huahu, X., Ahmad, W., Shabir, M. Y., Ahmad, W. (2020). A Sentiment Polarity Categorization Technique for Online Product Reviews. IEEE Access, 8, 3594–3605. doi: https://doi.org/10.1109/access.2019.2963020
10. Yevseiev, S., Goloskokova, A., Shmatko, O. (2021). Researching a machine learning algorithm for a face recognition system. Technology Transfer: Fundamental Principles and Innovative Technical Solutions, 10–12. doi: https://doi.org/10.21303/2585-6847.2021.002222
11. Hamdan, H., Bellot, P., Bechet, F. (2015). Sentiment Lexicon-Based Features for Sentiment Analysis in Short Text. Research in Computing Science, 90 (1), 217–226. doi: https://doi.org/10.13053/rcs-90-1-17
12. Colab Research Google. Available at: https://colab.research.google.com/drive/17KLlgzCipalUtlID_ToGdoalKd-9A9O-B#scrollTo=5N3Vro5vYyap
13. Ramya Sri, V. I. S., Niharika, Ch., Maneesh, K., Ismail, M. (2019). Sentiment Analysis of Patients' Opinions in Healthcare using Lexicon-based Method. International Journal of Engineering and Advanced Technology, 9 (1), 6977–6981. doi: https://doi.org/10.35940/ijeat.a2141.109119
14. Rokade, P. P., D, A. K. (2019). Business intelligence analytics using sentiment analysis-a survey. International Journal of Electrical and Computer Engineering (IJECE), 9 (1), 613. doi: https://doi.org/10.11591/ijece.v9i1.pp613-620
15. U.S. Hospital Customer Satisfaction 2016-2020. Available at: https://www.kaggle.com/datasets/abrambeyer/us-hospital-customer-satisfaction-20162020

Received date 01.10.2022
Accepted date 07.11.2022
Published date 29.11.2022

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