Decision-Making Based on Information in Medical Media Resources

Masuma Mammadova
Department of Number 11
Institute of Information Technology
Baku, Azerbaijan
0000-0002-2205-1023

Zarifa Jabrayilova
Department of Number 11
Institute of Information Technology
Baku, Azerbaijan
djabrailova_z@mail.ru

Nargiz Shikhaliyeva
Department of Number 11
Institute of Information Technology
Baku, Azerbaijan
0000-0002-5427-0765

2023 5th International Conference on Problems of Cybernetics and Informatics (PCI) | 979-8-3503-1906-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCI60110.2023.10325933

***Abstract*—The article proposes an approach for decision-making based on the information gathered from the medical social media environment. It develops an algorithm based on sentiment analysis and machine learning methods through the Valence Aware Dictionary and Sentiment Reasoner approach to select the best clinic based on patient reviews collected in medical media resources. The results of the experiment on patient reviews are presented in the cms_hospital_satisfaction_2020 database by the Kaggle company.**

Keywords*—*medical social media resources, sentiment analysis, lexicon-based approach, machine learning

I. Introduction

Today, the information collected on social media tools that people use to build relationships and share information has become a valuable resource. Social media sites, professional social societies, online forums, and large amounts of information collected in personal blogs act as people's advisors in solving various problems, and information generated by the crowd supports making "wise decisions" [1]-[4].

The increase in studies devoted to the use of information collected in the social media for improving the quality of medical care and making more objective decisions confirms this trend [5]-[7]. Currently, the emergence of a large number of professional medical social communities in the Internet environment, as well as the expansion of communication opportunities among stakeholders of the medical social media environment, i.e., patients, doctors, medical institutions, nurses, etc., are upsurging the value of the information collected in this segment day by day. A study conducted within the framework of the Pew Internet & American Life Project1 found that almost 80 percent of Internet users in the United States search for health-related topics online, 63% of users seek information about a specific medical problem, and approximately 47% seek information about medical treatment or procedures in medical social networks [8].

Patients using medical social networks get in touch with doctors and clinics to ask questions they are interested in, get answers, establish relationships, and express their opinions. Thus, in the patient-doctor and patient-clinic segments of social media, patients' opinions about doctors and clinics are collected. Users accessing social networks to solve their problems get acquainted with these reviews, analyze them, and try to choose a better clinic or doctor. Solving this process with intelligent technologies requires the application of sentiment analysis methods [8]-[10].

This paper proposes a hybrid approach based on lexicon-based sentiment analysis and machine learning methods to determine the best clinic based on patient satisfaction in the feedbacks collected in the patient-clinic segment of medical social media.

II. Problem Statement

Within the framework of the research conducted in the field of the formation of digital medicine, the authors of [11, 12] examine the integration of social media into electronic health care, medical social networks, societies, social media activities of interested parties, and segment the medical social media environment according to user relations. [13] proposes the conceptual basis of the use of information collected in medical social media for health decisions.

[14] develops indicators for evaluating the medical social media activity of interested parties by referring to the statistical analysis of surveys collected in the doctor-patient segment of medical media and the demographic indicators of e-patients. It also formally describes the medical decision-making process based on these proposed indicators. [15] explores the possibilities of applying sentiment analysis for the use of information collected in the medical social media environment for health decisions. The possibilities created by sentiment analysis referring to lexicon-based approach and machine learning methods for the study and classification of general opinion according to the opinions mentioned about the medical media participants are explained in detail.

The goal of submitted paper is to evaluate patient satisfaction with the activity of clinics based on the information collected in the patient-medical institution (clinic) segment of medical social media, and to determine the rating of clinics according to patient satisfaction.

The problem solution requires tackling the following sub-problems:

- sentiment analysis of opinions;
- classification of opinions and determination of patient satisfaction regarding the activity of clinics;
- determination of clinics' rating according to patient satisfaction.

III. Problem Solution

Problem solution is performed in the following stages:

Stage 1

Pandas, Numpy, Matplotlib, Seaborn, NLTK libraries and Python environment are used to collect patients' opinions about clinics. An open database cms_hospital_satisfaction_2020 by the Kaggle company is selected for the opinions' analysis [16]. 442587 patient feedbacks about the clinics' activity are analyzed in the database (figure 1).

```
[ ] # Read in data
    df = pd.read_csv('cms_hospital_patient_satisfaction_2020.csv',low_memory=Fals
    print(df.shape)
```

Fig. 1. Access to open database cms_hospital_satisfaction_2020 by Kaggle company for sentiment analysis.

Stage 2

This stage is the Data Pre-Processing stage and performs the data cleaning (tokenization) process, removes spaces and special characters, and the remaining ones are called tokens.

Stage 3

This stage is called Extraction Opinions and prepares the processed opinions for sentiment analysis.

Stage 4

Applies Lexicon Based Sentiment Analysis algorithm. This approach refers to the sentiment lexicon, which consists of words and phrases commonly used to express positive and negative attitudes.

Stage 5

The stage of Classification of Opinions applies the Valence Aware Dictionary and Sentiment Reasoner (VADER) approach for sentiment analysis of opinions and classifies opinions. The VADER approach classifies texts expressing patient opinions into 3 classes as negative, positive, and neutral.

In order to classify the base with the VADER approach, the "sentiment_type" column, which receives values such as "NEGATIVE", "NEUTRAL", "POSITIVE", is added to the dataset.

Figure 2 illutrates a fragment taken from Kaggle's cms_hospital_satisfaction_2020 database following the classification based on sentiment analysis.

```
df[["Facility ID","Facility Name",  "sentiment_type"]].tail(-1)
```

	Facility ID	Facility Name	sentiment_type
1	010001	SOUTHEAST ALABAMA MEDICAL CENTER	POSITIVE
2	010001	SOUTHEAST ALABAMA MEDICAL CENTER	POSITIVE
3	010001	SOUTHEAST ALABAMA MEDICAL CENTER	NEUTRAL
4	010001	SOUTHEAST ALABAMA MEDICAL CENTER	NEUTRAL
5	010001	SOUTHEAST ALABAMA MEDICAL CENTER	POSITIVE
...	...	...	...
442582	670130	SOUTHCROSS HOSPITAL	POSITIVE
442583	670130	SOUTHCROSS HOSPITAL	POSITIVE
442584	670130	SOUTHCROSS HOSPITAL	POSITIVE
442585	670130	SOUTHCROSS HOSPITAL	POSITIVE
442586	670130	SOUTHCROSS HOSPITAL	NEUTRAL

442586 rows × 3 columns

Fig. 2. Post-classification representation of Kaggle's open database cms_hospital_satisfaction_2020.

Stage 6

MultinominalNB and SVM machine learning models are built based on the received indicators 80% of the dataset is allocated to machine training data and 20% to test data (figure 3).

```
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X_fit,y, test_size = 0.2, random_state=0)
```

Fig. 3. Allocating 20% of the dataset to test data

Stage 7

To evaluate the detection performance of classifiers in machine learning, precision, recall, false positive rate (FPR), true positive rate (TP), f-measure and accuracy criteria are used [17].

Precision (P) denotes the proportion of the number of true positives to whole predicted positives and is defined as follows:

$$P = \frac{T_p}{T_p + F_p}$$

Here: T_p denotes the number of correctly classified, prediction-related data; F_p is the number of misclassified data non-related to the prediction.

Recall (R) is defined as the proportion of the number of true positives to all actual positives is calculated using the following formula:

$$R = \frac{T_p}{T_p + F_n}$$

Where: F_n is the number of data unrelated to the prediction classified as errors.

F1-Score is defined as the harmonic mean of the recall and precision, and is calculated by the following formula:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Accuracy is defined as follows:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

Stage 8

This stage analyzes the results obtained from the application of machine learning algorithms.

Figure 4 illustrates the text classification results according to the MultinominalNB machine learning model.

```
model.fit(X_train,Y_train)

▾ MultinomialNB
MultinomialNB()

y_pred = model.predict(X_test)
from sklearn.metrics import classification_report
classification = classification_report(Y_test,y_pred)
print(classification)
```

	precision	recall	f1-score	support
NEGATIVE	0.79	1.00	0.88	6715
NEUTRAL	1.00	0.95	0.97	38201
POSITIVE	0.96	0.96	0.96	43602
accuracy			0.96	88518
macro avg	0.91	0.97	0.94	88518
weighted avg	0.96	0.96	0.96	88518

Fig. 4. Accuracy of text sentiment analysis by MultinominalNB machine learning model

Figure 5 illustrates the text classification results according to the SVM machine learning model.

```
y_pred = top_model.predict(X_test)
from sklearn.metrics import classification_report
classification = classification_report(Y_test,y_pred)
print(classification)

              precision    recall  f1-score   support

    NEGATIVE       0.83      0.72      0.77      6715
     NEUTRAL       1.00      1.00      1.00     38201
    POSITIVE       0.96      0.98      0.97     43602

    accuracy                           0.97     88518
   macro avg       0.93      0.90      0.91     88518
weighted avg       0.97      0.97      0.97     88518
```

Fig. 5. . Accuracy of text sentiment analysis by SVM machine learning model

Stage 9

Clinics are rated according to the patient satisfaction (more precisely, according to the number of positive opinions). In this case, the column "Hospitals" of the referenced database and the columns expressing patient satisfaction according to the opinion are analyzed in the Excel program. Figure 6 presents the results.

1	HOSPITALS	POSITIVE	SUM OF HOSPITAL OVERALL RATING
2	MEMORIAL HOSPITAL	184	13
3	GOOD SAMARITAN HOSPITAL	184	11
4	ST JOSEPHS HOSPITAL	138	83
5	NORTHWEST MEDICAL CENTER	138	74
6	COMMUNITY HOSPITAL	92	74
7	ST MARY MEDICAL CENTER	138	65
8	ST MARY'S MEDICAL CENTER	92	65
9	JOHNSON MEMORIAL HOSPITAL	92	55
10	GREAT RIVER MEDICAL CENTER	92	46

Fig. 6. Rating of (positive) clinics by patient satisfaction

IV. CONCLUSION

This study was devoted to the assessment of public satisfaction and the determination of the rating of clinics according to the opinions collected in the clinic-patient segment of the medical media. Lexicon-based sentiment analysis of texts expressing mass opinion about clinics in medical social media resources was performed for this, and texts were classified as "neg", "neu", "pos" using the VADER approach. The classification accuracy was determined by MultinominalNB and SVM machine learning models, and the activity of the clinics was evaluated according to patient satisfaction ("neg", "neu", "pos"). The clinics involved in the research conducted on the cms_hospital_satisfaction_2020 database of the Kaggle company were rated according to patient (positive) satisfaction. It should be noted that with an analogous approach, it is possible to solve the problem of determining the best doctor and their rating according to the opinions collected in the doctor-patient segment.

As a continuation of the proposed approach, it is possible to obtain the indications that patients are interested in, as well as their importance, from the opinions about the clinics with better (more "positive" feedback) or worse (more "negative" feedback) results. This is of great importance for improving and updating the functioning of clinics. On the other hand, the comprehensive evaluation and rating of clinics according to "positive", "negative" and "neutral" opinions is also relevant, and these issues will be included in our further studies.

REFERENCES

[1] Rasim M. Alguliyev, Ramiz M. Aliguliyev, and Farhad F. Yusifov, "Role of Social Networks in E-government: Risks and Security Threats. Online Journal of Communication and Media Technologies, 8(4), 2018, vol. 8, no. 4., pp. 363-376 p. https://doi.org/10.12973/ojcmt/3957.

[2] M. E. Aksoy, "A Qualitative Study on the Reasons for Social Media Addiction", European Journal of Educational Research, 2018, vol. 7, no. 4, pp. 861-865. doi:10.12973/eu-jer.7.4.861

[3] A. Simsek, K. Elciyar, and T. Kizilhan, "A Comparative Study on Social Media Addiction of High School and University Students", Contemporary Educational Technology, 2019, vol. 10, no. 2, pp. 106-119. https://doi.org/10.30935/cet.554452

[4] A. Tunc-Aksan, A., and S. E. Akbay, "Smartphone Addiction, Fear of Missing Out, and Perceived Competence as Predictors of Social Media Addiction of Adolescents", European Journal of Educational Research, 2019, vol. 8, no. 2, pp. 559-569. doi:10.12973/eu-jer.8.2.559

[5] S. Campanini, "Outstanding Statistics & Figures on How Social Media has Impacted the Health Care Industry", 2016. Mashable, Linkedin, Avaliable at: www.linkedin.com/pulse/24-outstanding-statistics-figures-how-social-media-has-campanini,

[6] N. Cesare., C. Grant., J. B. Hawkins, "Demographics in Social Media Data for Public Health Research: Does it matter?", Bloomberg Data for Good Exchange Conference, 2017, New York City, NY, USA, http://www.arxiv.org/ftp/arxiv/papers/1710/1710.11048.pdf

[7] R. D. Krithika, and J. B. Rosiline, Dynamic and Reliable Intelligent Data Mining Technique on Social Media Drug Related Posts", IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (IEEE ICPCSI), 2017, pp. 1788-1794.

[8] Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya, "Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System", Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki,Japan, 2018. Pp. 2790-2797.

[9] S. Nadali., M. A. A. Murad, and R. A. Kadir, "Sentiment classification of customer reviews based on fuzzy logic", Proceedings of the International Symposium on Information Technology (ITSim' 10), 2010, pp. 1037-1044.

[10] A.-D. Vo, and C.-Y. Ock, "Sentiment classification: a combination of PMI, sentiWordNet and fuzzy function", Proceedings of the 4th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI '12), 7654 (2) of Lecture Notes in Computer Science, 2012, pp. 373–382.

[11] M. H . Mammadova, and A. M. Isayeva, "E-health activity in social media environment", Problems of information society, 2018, vol. 9, no. 1, pp. 52–62. DOI: 10.25045/jpis.v09.i1.05.

[12] M. H. Mammadova, and Z. G. Jabrayilova, "Electronic medicine: formation and scientific-theoretical problems", Baku: "Information Technologies" publishing house, 2019, 319 p.

[13] M. H. Mammadova, Z. G. Jabrayilova, and A. M. Isayeva, "Analysis of physician-patient relations segment of social media: opportunities and challenges", Problems information society, 2019, vol. 10, no. 2, pp. 41-50.

[14] M. H. Mammadova, Z. G. Jabrayilova, and A. Isayeva, "Conceptual Approach to the Use of Information Acquired in Social Media for Medial Decisions", Online Journal of Communication and Media Technologies,2020, vol. 10, no. 2, e202007. https://doi.org/10.29333/ojcmt/7877

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

[16] U. S. Hospital Customer Satisfaction 2020 . https://www.kaggle.com/datasets/abrambeyer/us-hospital-customer-satisfaction-20162020?select=cms_hospital_patient_satisfaction_2020.csv

[17] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over holdout validation on colossal datasets for quality classification," in 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78–83.