Journal of Research in Medical and Dental Science
2020, Volume 8, Issue 5, Page No: 6-12
Copyright CC BY-NC 4.0
Available Online at: www.jrmds.in
eISSN No.2347-2367: pISSN No.2347-2545

Sentiment Analysis of Hospital Service Satisfaction

Migena Ceyhan[1*], Zeynep Orhan[2], Dimitrios A. Karras[3] and Senol Dane[4]

[1]Department of Mathematics and Informatics, University of Shkodra "Luigj Gurakuqi" Shkodër, Albania

[2]Department of Computer Science, Union College, Schenectady, New York, USA

[3]Department of Computer Engineering, Epoka University Tirana, Albania

[4]Department of Physiology, Faculty of Basic Medical Sciences, College of Health Sciences, Nile University of Nigeria, Abuja, Nigeria

ABSTRACT

Introduction: Measuring customer satisfaction is one of the most important aspects of every successful enterprise trying to improve its service quality, so accumulating reviews is highly encouraged. But, just collecting this data is not sufficient, without possessing an efficient and reliable automatized system able to analyze this data and take out the priceless information for further enhancement. With the scarcity of similar works in the health area, and especially in Turkish, this study tries to fill this gap by analyzing health service satisfaction.

Methods: 2018 positive and 1394 negative comments collected from patients. Binary List, Frequency List, Binary Words and Words Frequencies feature selection methods were used to train and test a classification system by using machine learning methods such as Naïve Bayes, Support Vector Machine (SMO) and J48 tree algorithms. More compact feature subsets are used after eliminating mostly irrelevant common features from both or just one of the positive and negative feature lists. This data elimination may increase the negative miss ratio, being an important measure especially for health reviews domain.

Results: The results obtained are very efficient and have high average prediction rates.

Discussion: Binary Words feature selection methods outperform the others with the best average accuracy for Naïve Bayes as 98%, while the poorest results are obtained from the Binary List feature selection method and NB classifier. True and False Negative Rates (specificity and miss rates, respectively) are also evaluated to measure the best matching results.

Conclusion: Generally speaking, Words (both Binary and Frequency) feature selection methods are superior to Lists ones, providing more detailed information for each comment. Frequency methods in some cases slightly outperform Binary methods, but the shortness of the texts makes this change not very significant. NB, which is a very efficient algorithm in terms of time forms better classification models than SMO. J48, however, is generally better at Frequency Lists compared to the other ML algorithms, acquiring the highest rate of 99% for TNR in Binary Lists when all the features are used.

Key words: Text classification, Turkish, sentiment analysis, opinion mining, health care reviews classification

HOW TO CITE THIS ARTICLE: Migena Ceyhan, Zeynep Orhan, Dimitrios Karras, Senol Dane, Sentiment Analysis of Hospital Service Satisfaction, J Res Med Dent Sci, 2020, 8(5): 6-12

Corresponding author: Migena Ceyhan
e-mail✉: migena.ceyhan@unishk.edu.al
Received: 03/07/2020
Accepted: 06/08/2020

INTRODUCTION

The web is now a rich data ecosystem created cooperatively by all Internet users without any cultural or physical borders offering a diverse information ocean. Being heavily populated with user-generated content, commenting mainly on politics, sports, news in general, social media posts; reviews of goods , services, entertainment events and much more, it offers a decent source for their interpretation and classification, which is at the same time too much work and time consuming to manually analyze. Recently, this ocean of data has become a valuable tool for extracting consumer opinions for a range of purposes ranging from customer experience management to monitoring public opinion. The huge volume of data guarantees reliability and comprehensibility for most users. These two facts on social media feedback makes them preferable in the decision-making process as regards brand analysis, business intelligence, stock market forecasting and image monitoring [1].

There are several fields where technology helps us evaluate a possible solution or discover a possible one. In general, in the technical applications of these fields, natural language processing (NLP) and sentiment analysis (SA) are used. NLP is a branch of Artificial Intelligence (AI) which deals with human language analysis and understanding. NLP's goal is to develop and build a framework that will analyze and generate human languages. NLP techniques on derived data use Machine Learning (ML) algorithms.

SA seeks to classify and extract subjective knowledge in source materials through the use of NLP, text analysis, computational linguistics and the automated classification of texts attempting to assess the author's attitude according to a subject or the entire document. This finds its position widely in the study of feedback and social media, from

marketing to customer service. Public preferences play a major role in developing new products, preparing future plans and delivering tailored goods according to the profiles of customers. Hence, discovering reliable and cost-effective ways of using the views, desires, behaviors and thoughts of their consumers in real time is a critical and beneficial problem for businesses. In addition to its difficulties, SA remains an insurrection phenomenon worth pursuing offering several domains, with comments being one of the easiest to access and featuring a lot of opinions [2].

The SA region is newly being studied in morphologically rich languages and particularly in Turkish, and not much research has been published in the field. New words can be extracted in Turkish by adding suffices which makes SA a challenging task. Working with the Turkish language in itself is difficult, with only few research carried out in this field.

Engineering and computational areas such as AI and NLP help the analysis of data from various fields. Reviews can be of great help in analyzing customer feedback and one of the areas that can benefit a lot from similar research is the health system. People are very concerned about health and well-being issues, so nearly all patients or potential patients want to consult and learn the experiences of other people before they receive any health service. The Internet became like the first contact of people with medical knowledge, even before consultation of medical experts.

A research on health care in Turkey showed that the first combined selection criteria in hospitals in Turkey were based on a specialist doctor (43.7%), the second was based on confidence (40.2%), while the next came access facilities (33.9%) and overall satisfaction (23.1). Those are accompanied by preferences, accessibility of rates, advice from relatives and acquaintances, and institutional study. Since trust and doctor preference is important in the health care system, people make sure they get accurate details from family or friends, or even unknown individuals willing to share their experiences. Disliking a hospital was primarily measured by inadequate exams (28.1%), incompetence of the doctor (24%), high rates (19.1%), accompanied by unethical behavior, lack of empathy, inadequate cleanliness, medical supplies, insufficient physical conditions, etc.

For this reason, it is not enough to only learn from good experiences, people want to learn more from adverse experiences in order to take steps before anything unpleasant can happen. For this purpose detecting all negative feedback is a priority, and it is important to reach all potential negative comments as this method is automated, besides the consistency of the tests. [3]
According to Turkey's 2017 annual report, 22 percent of total health spending (almost $7.9 billion, recognizing that the overall per capita expenditure is $445 and the population was $80,745,000) was invested in the private health sector, which is a substantial amount[4]. Hospitals are very interested in holding and growing their customer share and they are mindful that this way passes by assessing the satisfaction and reducing the dissatisfaction of their patients and their families.

According to the facts of the Turkish Statistical Institute, health sector spending has risen significantly, particularly during the last 15 years. While the national healthcare expenditure amounted to 8,248 million Turkish Liras, this amount increased to 94,750 million Turkish Liras in 2014 [3].

Very few interdisciplinary studies are available that apply statistical methodologies to automatically detect opinions in data collected from the health system. In a very interesting study, diagnoses of mental disorder were targeted. The study was conducted on depression, mania and healthy adults diagnosed beforehand. With Naïve Bayes, Bayesian Logistic Regression and Support Vector Classifier machine learning algorithms, the semantic categories found in the Turkish version of the General Inquirer Harward III dictionary were used in addition to syntactic characteristics. A mobile platform that detects disturbed psychological conditions in patients exchanging messages was designed to build specific dictionaries for each psychological disorder [6].

Another remarkable research is investigating the challenge of determining the necessary tests that the person should go through differing from the symptoms and anamnesis of the patient. Six different ML techniques were used to assess the accuracy of the required examinations up to 90 percent [7].

The paper is based on the patient comments rating on the health care they were provided in a private hospital. The following sections will cover experimental setup to clarify all the steps involved in the analysis, such as data collection, pre-processing techniques, methodology, assessment and conclusion.

Experiment setup

In this paper, a novel system for detecting positivity or negativity of comments on health care services is proposed. The following sections will briefly describe the details related to data collection, data pre-processing, feature selection methodology, training and test of the system by using ML classification algorithms and lastly the evaluation of the results.

Data collection

Numerous comments of satisfaction or complaints toward doctors, nurses or other hospital personnel, as well as health service quality, personnel attitude towards them, hygiene, price, etc. were gathered during the last 7 years in a private hospital.

The data were selected out of 2018 positive comments and 1394 negative comments. Since the number of data is large enough, in order not to create huge datasets, which could be too much time consuming for the system, they were divided into four datasets of 500 positive and 500 negative comments, then the averages of all of the datasets were evaluated. Out of each dataset 450 of the positive and the same amount of negative comments were used for training, while 50 positive and 50 negative comments were randomly divided to be used for testing. Information about data is shown in Table 1. Each comment was stored into a separate text file which was further processed before

applying the feature selection methods and ML algorithms, as will be shown in detail in the following sections. Another dataset, Dataset 5, was built again of 500 reviews for each class, but the number of the average words per comments were chosen to be similar, given that the negative comments tend to be much longer and detailed than the positive ones. Information on datasets is explained further in Table 1.

Table 1: Statistical information about datasets.

Classes	Total comments	Train comments per dataset		Test comments per dataset	
	Datasets	D1 – D4	D5	D1 – D4	D5
Positive	2018	450	335	50	165
Negative	1394	450	335	50	165

Data pre-processing

Morphological analysis of data

Since Turkish is a very rich suffix agglutinative language (using consecutive suffixes to obtain different word derivations, negation, personal or tense conjugations, plural forms, etc.), in order to detect specific words in their nominative or infinitive case, regardless of the affixes or suffixes, word roots were extracted by using a Morphological Analyzer (MA) [7]. Morphological analysis is a simple creative method of forced association of attributes that simply divides the words into root and suffixes in possible formats. Before morphologically analyzing the data, each sentence was identified according to the punctuation marks and converted into suitable input format to the morphological analyzer, so that each sentence was included between <S><S>and <\\S><\\S>tags, and each word was displayed in a separate line, as shown in Figure 1A and 1B

Disambiguating morphologically analyzed data

Generally, the morphological analysis result yields multiple and ambiguous results about each word, which brings the necessity of morphological disambiguation (MD) to select the correct one [7]. Disambiguation is a natural language processing application that tries to determine the intended meaning of a word or phrase by examining the linguistic context. The output after the disambiguation process is displayed in Figure 1B and 1D.

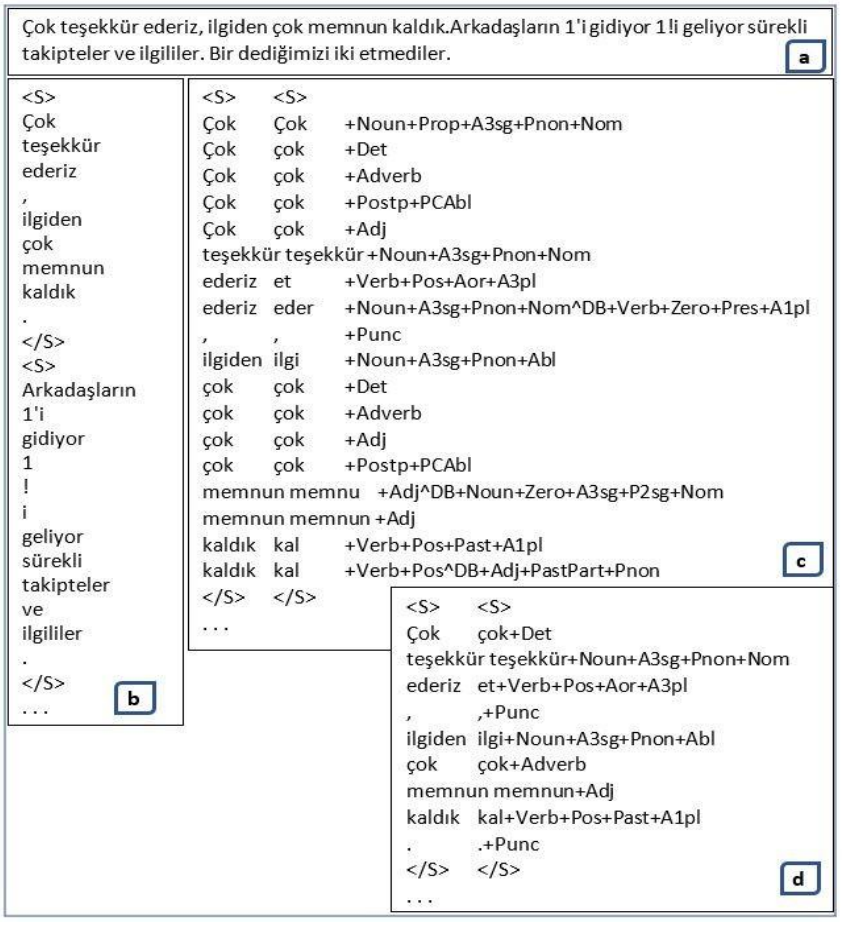

Figure 1: The conversion of data from (A) The row text. (B) The input to morphological analyzer (C) The output of morphological analyzer/ input to morphological disambiguator to (D) The output of morphological disambiguator.

Methods

The data were grouped into two categories, namely positive and negative comments. In four of the datasets, data were fully used, while in the fifth dataset reviews with similar text average lengths were selected, in order to equilibrate the data after normalization was performed. Normalization takes into account the length of the texts because each word's frequency is computed in the reverse ratio with the text length. Each dataset contains 500 positive and 500 negative comments. While in the first four sets 90% of the comments in each group were randomly assigned to train the system, while the remaining 10% to test it, in the fifth dataset 33% of the comments were used to test the system. ML techniques were used to analyze the polarity of documents in this study. ML is a scientific discipline that explores the computational approaches to learning. Feature selection is the first and most important step in ML methods. In this study, the root of each word used in the comments was considered as a feature.

Four main feature selection methods were employed according to the words (binary - existence) and frequency information. The first method used the binary information of the existence or nonexistence of the root of a word in the prepared set of features for each subject. The second method also provided information about the frequency of the usage of each word.

Binary List All the distinct roots of the words used by each class of persons were counted and used to create category-based lists (namely Positive List and Negative List). Then the words appearing in these lists were counted and the values were added up for each person. Finally, the largest value determined the class of the comment as positive or negative.

Frequency List The number of words used by each person was multiplied by the respective frequencies of the words appearing in Positive List and Negative List, and the values were added up. Finally, the largest value determined the class of the comment as positive or negative.

Binary Words All the distinct roots of the words used by the persons in the train sets were counted and used as features and their existence or nonexistence information was used, as shown in Figure 2a.

Frequencies Words All the distinct roots of the words used by the persons in the training set were counted and used as features and their frequency information was used. When the frequency was used, normalization was also applied by dividing the calculated frequency by the text size of the comment, as shown in Figure 2 b.

Id	W1	W2	W3	. . .	Wn	Class (N:Neg P:Pos)
1	0	1	0	. . .	1	N
2	1	0	1	. . .	1	P
3	1	0	0	. . .	0	P

(a) Binary

Id	W1	W2	W3	. . .	Wn	Class (N:Neg P:Pos)
1	0	3	0	. . .	2	N
2	4	0	1	. . .	6	P
3	3	0	0	. . .	0	P

(b) Frequency

Figure. 2: Training data including all distinct words as columns and comments as rows, where: (A) Binary and (B) Frequency values indicate word existence and frequencies in comments.

Additional reduced features sets in number for both Lists and Words methods were obtained by erasing similar words appearing in both positive and negative word lists according to a threshold value. This is done with the assumption that those words with close values are not determinant to identify the sentiment of the document. According to the erase type, either similar frequency words were deleted from both lists, or from only the list where they appeared less frequently, updating the frequency value in the opposite list with the difference of their frequency values.

CLASSIFICATION

Classification is the process of identifying the new observations whose classes are unknown by using the pre-classified data. Classification is achieved using ML methods provided by a data mining tool named Weka [8]. Weka provides a collection of many well-known ML algorithms to test and train the classifiers. In this study for the classification of subjects, Naïve Bayes (NB), Sequential Minimal Optimization (SMO), and a decision tree algorithm J48 are used.

Naïve Bayes uses estimator classes which are based on Bayes theorem with strong and Naïve independence assumptions. It is not computationally intensive, and it requires a small amount of training data, resulting in training time significantly smaller as opposed to alternative methods [9].

SMO algorithm trains a Support Vector Classifier (SVC) with John Platt's SMO. An SVC is a classifier that takes a set of data and for possible two classes predicts the membership of each data according to those classes.

J48 is a decision tree algorithm for generating a pruned or un-pruned C4.5 decision tree. It uses formulas based on information theory to evaluate how much good a test extracting the maximum amount of information from a set of cases, given the constraint that only one attribute is tested [10].

Together with the accuracies, the True Positive (TPR) and False Positive Rates (FPR) in percentage for the negative set of data were evaluated. While TRP is aimed to have the highest values as close as possible to 100, the opposite is true for FPR, ideally close enough to 0. This was done to identify the Feature method – ML algorithm pair with the highest precision in not classifying negative data wrongly because health issues are so delicate that nobody would

like to risk not to identify problems already happened to other reviewers, so that they can possibly avoid them The test results of the accuracy, TNR, and FNR of texts by using each of the methods when applied to the ML techniques, are given in Table 2. Tables 2-3 and Figures 2-5 show the results of the study.

Table 2: The average accuracy percentages obtained from three algorithms for four different datasets (DS1 to DS4)

Datasets	Average Accuracies of Datasets 1-4				Accuracies of Dataset 5			
Methods	**Binary List (BL EN)**	**Frequency List (FL EN)**	**Binary Words (BW EN)**	**Frequency Words (FW EN)**	**Binary List (BL EN)**	**Frequency List (FL EN)**	**Binary Words (BW EN)**	**Frequency Words (FW EN)**
Naïve Bayes	82.6	90.4	91	89.4	68	92	**98**	96
SMO	90.8	90.4	95.4	**95.8**	86	93	96	97
J48	84.8	91	93.4	91.2	73	92	92	92

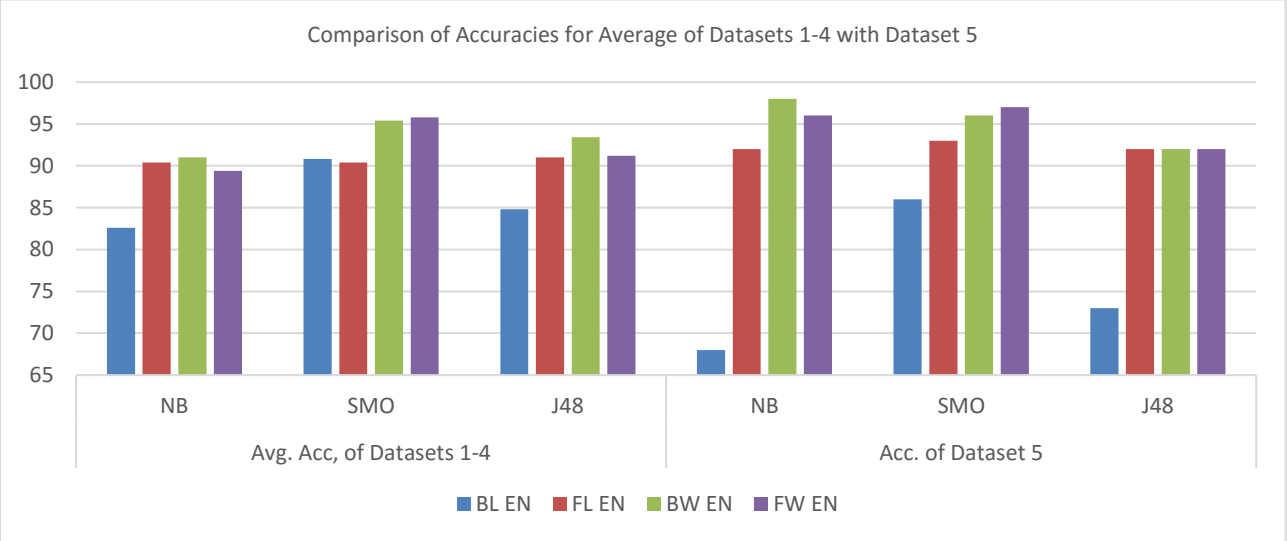

Figure. 3: Comparison of the results between the average accuracy of Datasets 1-4 with Dataset 5, where: Binary and Frequency values indicate word existence and frequencies in comments.

Table 3: ML results (Accuracy, TNR and FNR) for Dataset 5 for different Feature selection methods, with or without erasing from positive and negative lists of words according to thresh 75%.

Method	Accuracy			TNR			FNR		
Binary Lists (BL)	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**
No Erase (EraseNone)	68	86	73	55	93	**99**	19	22	52
Erase from both lists with threshing 75% (EraseBoth)	84	89	84	86	86	78	19	7	11
Erase from the smallest list with threshing 75% (EraseOne)	93	93	89	91	91	85	6	**5**	6
Frequency Lists (FL)	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**
No Erase (EraseNone)	92	93	92	93	92	91	10	6	6
Erase from both lists with threshing 75% (EraseBoth)	94	94	93	**98**	**98**	90	9	9	**4**
Erase from the smallest list with threshing 75% (EraseOne)	93	93	90	91	92	87	6	6	7
Binary Words (BW)	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**
No Erase (EraseNone)	**98**	96	92	**98**	94	92	**2**	**2**	8
Erase from both lists with threshing 75% (EraseBoth)	95	94	92	96	94	92	5	6	8
Erase from the smallest list with threshing 75% (EraseOne)	**98**	95	94	**98**	96	95	**2**	5	7
Frequency Word (FW)	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**	**NB**	**SMO**	**J48**
No Erase (EraseNone)	96	**97**	92	**98**	**97**	94	6	**3**	10
Erase from both lists with threshing 75% (EraseBoth)	91	91	92	89	90	93	7	8	10
Erase from the smallest list with threshing 75% (EraseOne)	94	55	92	96	87	96	9	**77**	12

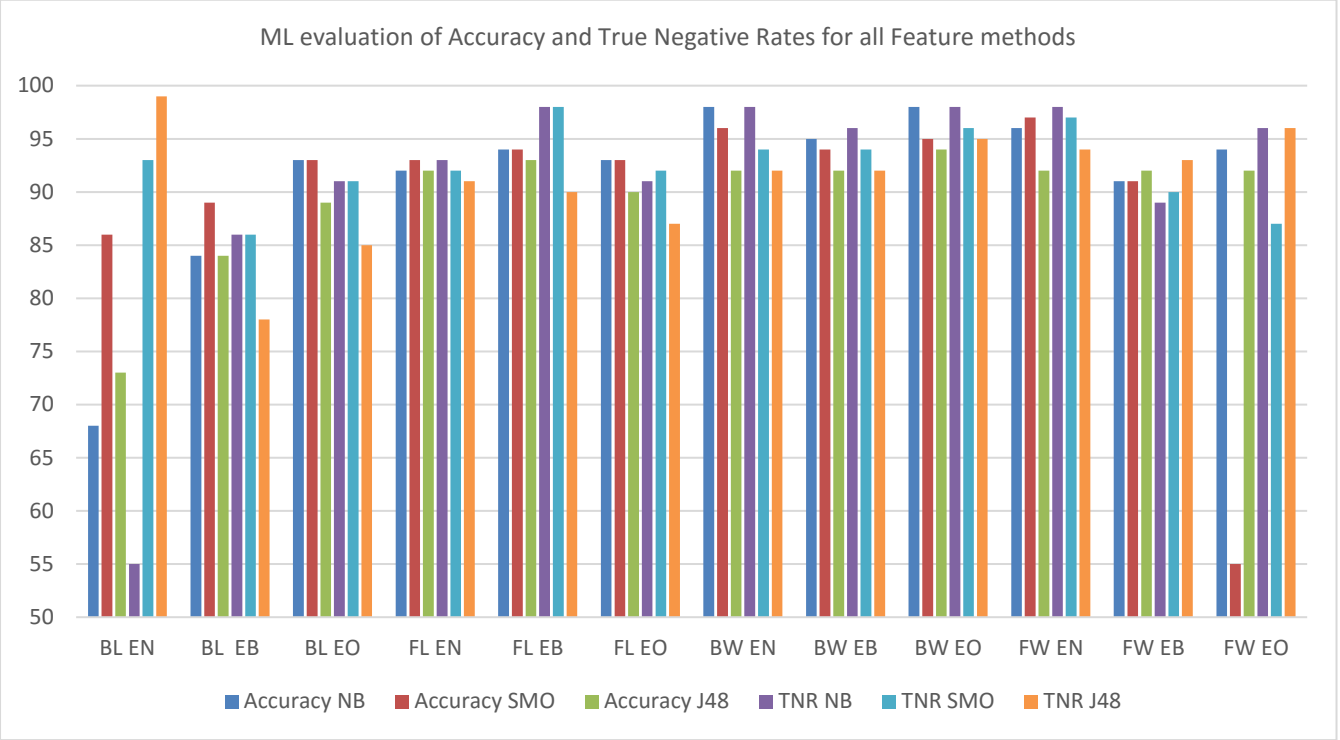

Figure. 4: ML evaluation of Accuracy and True Negative Rates for all Feature methods

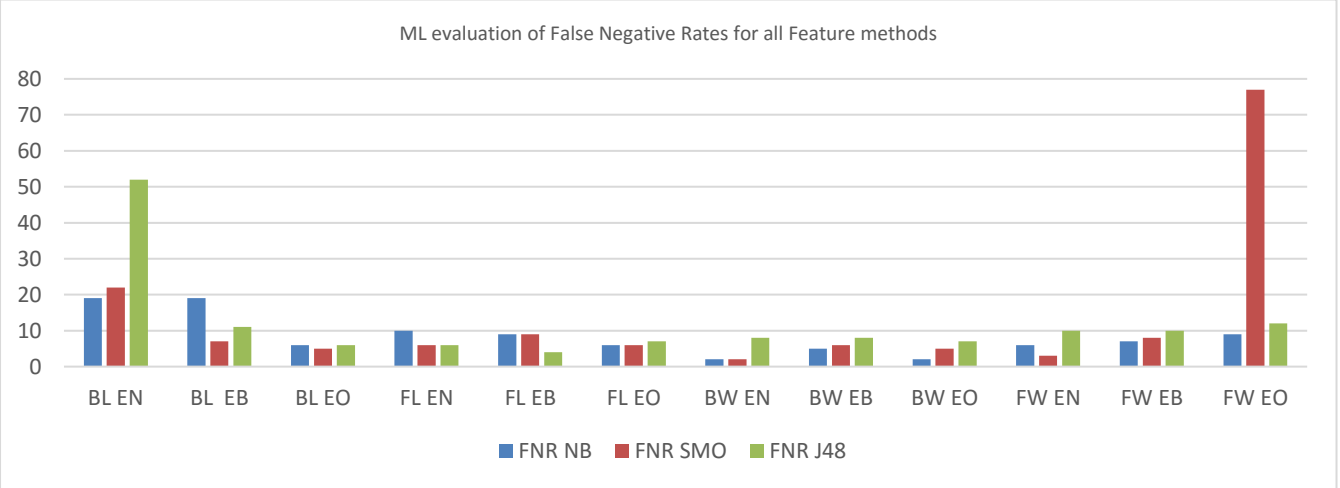

Figure. 5: ML evaluation of False Negative Rates for all Feature methods

INTERPRETATION OF RESULTS

As seen from the numerical and visual results, only Binary Lists (Erase None) does not profit from equilibrating the datasets according to the text length rather than only dataset size, for more accurate normalization, while all the other methods get obviously higher results.

In Dataset 5, for each ML classifier used, Binary Lists has the poorest performance, both accuracy-wise (Acc. of NB of 68%), and TNR-FNR-wise (TNR of NB 55%, FNR of J48 52%). These results are worse when no erasing is done, while erasing from one list is superior to erasing from both lists, especially for the FNR value (getting up to 5-6% for all ML algorithms). When erasing is performed, more relevant information is processed, thus especially when one entry having the largest value remains in the list, the information provided is more significant, since List methods process linear data. This is especially visible in the FNR result of J48.

Frequency List methods have better outcomes compared to Binary List methods, understandably emphasizing the importance of using frequency information in linear computational processing. The highest accuracy of 94% is reached for this method when both similar values are erased from both positive and negative lists of words at a threshold of 75%, for NB and SMO. These results are supported by high values of 98% of TNR, too.

The best prediction is achieved with Binary Words method and Naïve Bayes (98% accuracy, 98% TNR and only 2% FNR), when all features are used (this is the case for no erasing or erasing from one list, being binary the information is not different). The next good prediction value is for SMO with no erasing (96% accuracy, 98% TNR, and 2% FNR). SMO, on the other hand, gets the highest results for Frequency Words method, with 97% accuracy,

97% TNR, and 3% FNR. The same ML algorithm, SMO, surprisingly gets a very poor result for Frequency Words method with erasing from one list with the threshing of 75%, with 55% accuracy, 87% PNR and a record of 77% of FNR.

According to the results, the best ones are obtained from Binary Words feature selection method, combined with Naïve Bayes feature selection method. In normal circumstances, Word Frequency feature selection is expected to give better results than any other one, but because of short reviews, Word Frequencies have no clear superiority.

CONCLUSION

In this paper, by considering the language features of the posts, we suggest a method for identifying the positivity or dissatisfaction of feedback on health care facilities. Negative and supportive feedback received from patients and analysis on grievances are automatically categorized using high quality ML methods. Such findings support the argument that language usage is a significant source in obtaining useful information concerning the views, attitudes and emotions of the people's health reviews.

This proposed system's approach uses all of the text's distinct terms as the attributes according to their nature, frequency or weighted frequency. Such features are provided to various Weka Library algorithms, such as Naïve Bayes, SMO, and a decision tree algorithm, J48. Among the methods used, the Binary Words process method and the Naïve Bayes classification algorithm achieve the highest scores.

The findings of this research effort, as one of the initiators of Turkish sentiment analysis and opinion mining in healthcare services, are very positive to the extent of our knowledge. A selection health service analysis dataset in Turkish taken from a private hospital was used to test the proposed model. It is a first step towards a model that does not involve complex and expensive data acquisition, which can be easily extended to other languages or other domains and contributes to several efficient, cost-effective designs of healthcare systems that will benefit patients, hospitals, medical experts and strategies developers.

REFERENCES

1. Balahur A, Mihalcea R, Montoyo A. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. Computer Speech Language 2014; 28:1–6.
2. Dave K, Lawrence S, Pennock DM. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on WWW '03, ACM, NY, USA, 2003; 519–528.
3. Tüfekci, N. & Asığbulmuş, H. (2016). The Factors that Effective in the Choice of Hospital and Patient Satisfaction: The Sample of Isparta. Journal of Current Researches on Health Sector, 2016, 6 (2), 71-92.
4. https://apps.who.int/nha/database, last accessed in June 2020.
5. Ceyhan M., Orhan Z., Domnori E. (2017) e-Medical Test Recommendation System Based on the Analysis of Patients' Symptoms and Anamneses. In: Badnjevic A. (eds) CMBEBIH 2017. IFMBE Proceedings, vol 62. Springer, Singapore. https://doi.org/10.1007/978-981-10-4166-2_98
6. Orhan Z., Mercan M., Gökgöl M.K. (2020) A New Digital Mental Health System Infrastructure for Diagnosis of Psychiatric Disorders and Patient Follow-Up by Text Analysis in Turkish. In: Badnjevic A., Škrbić R., Gurbeta Pokvić L. (eds) CMBEBIH 2019. CMBEBIH 2019. IFMBE Proceedings, vol 73. Springer, Cham. https://doi.org/10.1007/978-3-030-17971-7_59
7. http://www.denizyuret.com/2006/11/turkish-resources.html
8. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques with java implementations: Morgan Kaufmann 1999.
9. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. UAI ' 95, Morgan Kaufmann Publishers 1995; 338–345.
10. Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc 1993;235-240.