YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.gov.tr/ybs

Yayın Geliş Tarihi: 06.02.2020
Yayına Kabul Tarihi: 30.04.2020
Online Yayın Tarihi: 15.05.2020

Cilt:6, Sayı:1, Yıl:2020, Sayfa:42-50
e-ISSN: 2360-550X

MEDICAL SENTIMENT ANALYSIS BASED ON SOFT VOTING ENSEMBLE ALGORITHM[1]

Akın Özçift[1]

[1]Software Engineering Department, Hasan Ferdi Turgutlu Technology Faculty, Manisa Celal Bayar University, Manisa Turkey

ABSTRACT

Digital information is continuously generated from various sources such as social media, user reviews for services. The processing of this written information to extract user opinions is critical for developing customer satisfaction. In particular, medical services may be improved with customer feedbacks if the user opinions or sentiments are inferred from user reviews. There is an ongoing effort to develop automated software systems to evaluate these customer reviews. Machine Learning (ML) algorithms combined with Natural Language Processing (NLP) techniques are used to assess customer feedbacks. There are many studies related to English language in the literature to evaluate sentiments of user reviews. However, Turkish language needs research and it has abundant search opportunities in terms of sentiment classification. This work develops a soft voting ensemble (SVE) algorithm that combines predictions of Logistic Regression (LR), Random Forest (RF) and Decision Tree (DT) to analyze a newly collected medical review data. The accuracies of sentiment classifications of LR, RF and DT are 90.68%, 89.03% and 85.41%. The sentiment classification accuracy of SVE, combination of three algorithms, is 91.12%. The obtained results are promising for an automated Turkish medical sentiment identification algorithm.

Keywords: Medical Information, Sentiment Classification, Ensemble Learning, Soft Voting

INTRODUCTION

The huge information generated through popular social media channels gave rise to many fields for the analysis of data. In particular, sentiment analysis (SA) concerns to extract user opinions about a product or service from the corresponding piece of text. Since the amount of information generated is massive, the raw text needs to be analyzed with the use of ML algorithms combined with NLP techniques. The main goal of the ongoing SA research is to infer opinions of users or customers about a service or a product on behalf of user satisfaction (Ayyoub et al.,2019).

SA methods has been applied to many fields such as hotels, airlines and e-business web sites. Availability of datasets extracted from user reviews have increased implementation of automated NLP techniques to analyze opinions. In particular, sentiment classification task attempts to identify subjective polarity (i.e. positive or negative) of a piece of text. The overall SA is then used to improve quality of the services, marketing strategies and product recommendation systems. Medical informatics data related to patient opinions about hospitals or doctors is another domain for the application of SA methods. In this context patients can make use of user opinions about a physician when they need to choose one. Furthermore, hospitals may use the patient opinions to improve the quality of their whole services from doctors to health-care facilities (Zafra et al.,2019).

There are many techniques in the literature that is used SA in some way. From literature survey point of view, we particularly focus on Turkish SA studies in general. Then we will emphasis on health-care domain SA studies in recent medical studies in any language.

SA studies may be classified as ML based methods, lexicon based techniques and the hybrid approaches makes use of ML and lexicon combinations. A few recent studies from Turkish SA literature that makes use of one of the mentioned methods are as follows: In their study, Ersahin et al. used a combination of sentiment dictionary and ML algorithms to develop a Turkish sentiment analysis algorithm for hotel customer reviews (Ersahin et al., 2019). Another study that collects a Turkish dataset that can be used for SA tasks (Makinist et al., 2017). The authors developed an effective spelling correction algorithm based on Hadoop platform. SA for Turkish movie reviews was conducted by Dehkharghani et al. with the use of ML algorithm and their sentiment dictionary (Dehkharghani et al., 2016). They have made a detailed research at various granularity levels of Turkish from SA point of view. In another recent study by Karcioglu and Aydin was compared bag of word (BOW) model with word embedding approach to extract opinions from Turkish tweets (Karcioglu and Aydin, 2019). They have used LR and Support Vector Machine (SVM) algorithms to evaluate the collected Turkish tweets. They have found word embedding model is better than BOW model in terms of sentiment extraction task. In their study, Shehu et al. SVM and RF algorithms on top of a polarity lexicon to infer sentiments of tweets (Shehu et al., 2019). They have empirically showed that the use of polarity lexicon has improved the accuracy of SA extraction process. A recent SA study using novel Extreme Learning Machine (ELM) was conducted by Coban et al. to extract opinions of Turkish tweets (Coban et al., 2018). They compared ELM and SVM algorithms on two different Turkish tweets datasets and they found that SVM was slightly better in terms of sentiment extraction quality. Though not frequent as for English language, there can be found other SA studies for Turkish. For the sake of convenience, we had only focused on recent studies from the literature about various sentiment tasks.

Medical or health-care information studies are also drawn attention of SA researches. In particular, patients continuously make research about health information and they produce large amount of opinions about hospital services, treatments or doctors. Particularly, there are SA studies in health domain that try to identify user opinions with the use of various strategies. In their study, Rodrigues et al. developed a tool named as sentihealth-cancer to extract mood of the cancer patients from social media (Rodrigues et al., 2016). They analyzed Facebook groups of cancer patients for Portuguese language and they used a special Portuguese lexicon to infer patients' opinions. A study conducted by Kumar et al. made use of SA methods to determine patient satisfaction for an electroencephalogram (EEG) brand (Kumar et al., 2018). The ML methods used in the study are RF optimized with Artificial Bee Colony (ABC). Korkontzelos et al. identified adverse drug reactions from patient blogs and tweets with various NLP techniques and they obtained 80.14% success in terms of F-measure metric (Korkontzelos et al., 2016). In his research, Rajput looked at the applicability of SA techniques to mental health (Rajput, 2020). A last work from literature was conducted by Crannel et al. to identify opinions or moods of various types of cancer patients from Twitter. They concluded that the obtained results could have been used as support for clinical surveys (Crannel et al., 2016).

ML algorithms are widely used in SA literature to infer user opinions from written documents. However there is no universal ML algorithm that guarantees a prediction accuracy above a predefined baseline. It is a continuous effort to develop high accurate algorithms in any information processing task (Obulesu et al., 2018). Ensemble algorithms emerge as solutions to this research to some extent. In particular, ensemble methods may improve the predictive performance of a single predictor with state-of-the-art techniques developed. More precisely, ensemble methods train multiple ML models to combine predictions of each single algorithm in some way to obtain an improved prediction accuracy compared to constituents' algorithms (Sagi end Rokach, 2018). In the literature, many ensemble techniques have been developed such as Bagging, Boosting, Stacking and Voting (Ribeiro and Coelho, 2020). The details of the mentioned methods will be explained in Method section.

The main motivation of this work is the lack of SA research in health-care domain for Turkish language. To the best of our knowledge, this is one of the first SA studies focusing on Turkish medical literature. The second contribution of this study is the generation of a Turkish dataset that can be used for health-care opinion mining domain. Though Turkish is a widely spoken language it has limited data sources for researchers. Because of lack of data sources, it is a low-resource language compared to the English. Medical review domain is obviously important for evaluation of services. Our dataset was collected with mentioned limitation in mind. It has 2000 positive and 2500 negative physician reviews in various health problems. One of the problems of such data collections is that natively it is easy to get negative comments on oppose to positive reviews. We were able to collect negative reviews using just a single hospital web site. On the other hand, we collected positive reviews using a few internet resources.

METHOD

In this study, we used supervised ML algorithms to predict sentiment of newly collected medical dataset in Turkish. Having obtained sentiment classification accuracies of LR, RF and DT algorithms, we then developed a SVE algorithm to improve overall sentiment classification accuracy further. The basic pipeline of the proposed system has mainly six steps as: (i) collection of dataset, (ii) pre-processing of data, (iii) generation of term vector model, (iv) splitting data into 80% train and 20% test sets, (v) obtaining sentiment identification accuracies with LR, RF and DT and (vi) calculating overall accuracy of SVE of three algorithms. The corresponding pipeline is summarized in Figure 1.

Data Collection and Labeling

Extraction of data from web sites is an important step for data based mining projects such as customer review evaluations to enhance the services. However, web pages involves many types of data and there are a few preliminary steps to be taken while collecting data for a ML analysis pipeline such as sentiment identification. The structure of web data is mostly unordered. In other words, the data should be processed and cleaned before a data analysis step.

The data collection from web pages is known as web scrapping and this practice requires some flexible tools. Particularly, Python one of the most used programming language for web scrapping and we used the following steps while we collect data from web services of hospitals: (i) Extraction of html and parsing the related content, (ii) locating the review sections from the pages with unique identifiers, and (iii) cleaning html tags to obtain texts for each review. The final data used in the study was collected from various open-source hospital web pages using the mentioned steps and python web scrapping tool.

Any supervised ML analysis application require labeled data for a train/test scheme. Having collected and cleaned the data, we prepared a labeling setup with the help of three domain experts. The labeling is evaluated as follows: If a user comment is voted as positive at least the two out of three experts then it is labeled as positive. The same scheme is also applied to negative reviews and all collected the data is labeled in this approach. The brief properties of data is given in Table 1.

Table 1. Properties of Medical Sentiment Data

SENTIMENT	NUMBER of TEXTS
positive	2000
negative	2500

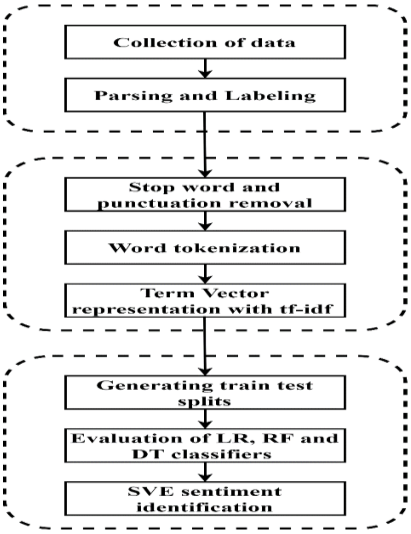
Data Pre-processing and Term Vector Generation

Any text processing task including SA generally starts with a pre-processing step in order to make the data to be ready for ingress to ML algorithms. In more clear terms, NLP processing steps such as n-gram term frequencies, stemming, stop word and punctuation marks removal etc. may be applied before the language analysis task (Etaiwi and Naymat, 2017). It is empirically shown that language pre-processing may positively affect the performance of ML algorithms (Uysal and Gunal, 2014).

In this work, we preferred basic language processing tasks of stop-word removal and punctuation filtering. While removing stop-words we used python nltk module with Turkish stop-word list. Stop-words are defined as the words that has negligible or no semantic importance from specific language task perspective (Rani and Lobiyal, 2018). For example, in Turkish "bu-this", "belki-maybe", "benim-mine" are examples of stop-words that can be removed without affecting the opinions of users semantically.

Having pre-processed the data, the sentences are tokenized to obtain single words. An important step for any language classification task such, data requires to be represented as a term vector model to be evaluated by classifiers. In other words, a text should be represented in numeric vectors with a model such as BOW. BOW represents a text in terms of word histograms or in terms of word counts. However, one of the problems of BOW model is that this approach assigns identical weights to each word in the corpus and it neglects respective weights of words occurrences. This problem was overcame with the use of term frequency-inverse document frequency (tf-idf) approach. Being a word weighting factor, a tf-idf value is obtained with the multiplication of *tf* and *idf* functions. In mathematical terms, weight of a word is calculated with Equation 1.

Figure 1. The Generation of SVE Algorithm for Sentiment Identification


Collection of data
Parsing and Labeling
Stop word and punctuation removal
Word tokenization
Term Vector representation with tf-idf
Generating train test splits
Evaluation of LR, RF and DT classifiers
SVE sentiment identification

$$TFIDF(word, doc) = TF(word, doc) * IDF(word) \qquad (1)$$

Where TF and IDF in Equation 1 is calculated with Equation 2 and Equation 3 below.

$$TF(word, doc) = \frac{Frequency of word \in the doc}{No. of words \in the doc} \qquad (2)$$

$$IDF(word) = \log_e\left(1 + \frac{No. of docs}{No.\ of docs with word}\right) \qquad (3)$$

TF is defined the inverse document frequency of a word in the whole document corpus and IDF is term frequency of each word in each document (Das and Chakraborty, 2018). We then obtained tf-idf weights of the sentiments with the mentioned protocol and we finally obtained term vector model suitable for ML algorithms.

Sentiment Identification with ML Algorithms and SVE

Having pre-processed data to obtain term vector model on top of ti-idf representation, we divide the dataset into 80% train and 20% test splits. In this study, since the dataset is not skewed to any class, we used accuracy for evaluation of sentiment classification results.

Accuracy (ACC) is a metric that measures the percentage correctly classified instances and it is derived from confusion matrix given in Table 2.

Table 2. Confusion Matrix for Binary Classification

		Actual Values	
Predicted Values		Positive	Negative
	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

From Table 2, the ACC is defined with Equation 4.

$$ACC = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad (4)$$

The ML algorithms used in this work are LR, RF and DT respectively. We give brief explanations of the algorithms and then we explain the proposed SVE approach.

LR given in Equation 5 use a Logistic function in estimation of positive or negative class labels denoted by y for features of data w.

$$p(y = \pm 1 | x, w) = \frac{1}{1+e^{-w^T h(x)}} \quad (5)$$

From Equation 5, class of a sentiment text is calculated based on the learnt coefficient (w) and each feature (h) for given input (x) (Omari et al., 2019).

RF grows multiple classifier trees with bootstrap sampling with replacement to train each tree with a different part of the dataset. For a training set with N samples, the probability that each tree is trained randomly with different sample is calculated with $(1 - \frac{1}{N})^N$ (Peng et al., 2019). As N increases the sample becomes more diverse. While RF is trained with multiple trees, the node splitting for a tree is evaluated with randomly selected features. The importance of a feature may be decided with a measure such as Gini index and overall class identification is evaluated from predictions of trees cumulatively.

DT is a widely used classifier algorithm and it is based on entropy theorem to select the attribute (feature) for node split. In other words, DT requires to measure importance of attributes for the splitting nodes at each level. When a node is used in DT for partition of training instances, the entropy changes. Information Gain (IG) given in Equation 6 is the measure of this entropy and hence it is used in the decision of node splitting process (Kaewrod and Jearanaitanakij, 2018).

$$IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} . Entropy(S_v) \quad (6)$$

In Equation 6, S denotes set of instances with attribute set A and S_v is the subset of S for $A = v$.

Ensemble Learning (EL) algorithms makes use of methods to provide cooperation between a set of ML algorithms. The goal of EL is to organize predictions of the ML algorithms for a specific task (e.g. sentiment classification) to outperform the accuracy of the constituents. In the literature, EL approaches are basically bagging, boosting, stacking and voting. Bagging and boosting ensemble methods with N learners are generated (trained) with a random sampling with replacement strategy. Because of *replacement* notion, some instances may be repeated while training the ensemble members. The main difference between bagging and boosting is that while in bagging any instance from train data has equal probability for replacement, boosting assigns different weights for instances. Briefly, boosting samples may be included more frequently than others. RF is a well-known ML algorithm using bagging and Adaboost is a widely-used ensemble approach based on boosting (Tabassum and Ahmed, 2016). In stacking the outputs (predictions) of classifiers are used as attributes of a generalization algorithm such as Naïve Bayes and this algorithm is used to make predictions based on the outputs (predictions) of the classifiers.

In case of Voting Ensemble, there are mainly two approaches as hard (majority) voting and soft (probabilistic) voting. Hard voting obtains an overall class prediction based on majority of predictions of ML algorithms in the ensemble. This is shown mathematically in Equation 7.

$$\tilde{y} = argmax(N_c(y_t^1),\ \ N_c(y_t^2), \dots,\ N_c(y_t^n) \qquad (7)$$

On the other hand, soft voting uses a different approach rather than a binary voting mechanism. Soft voting obtains final class prediction with the use of confidence of a ML classifier in terms of class prediction probability. Clearly, soft voting is applicable for the ML algorithms that can generate class predictions based on probabilities. Then soft voting is average of the sum of the probability vectors obtained from each classifier in the ensemble (Bonaccorso, 2018) and this is given in Equation 8.

$$\tilde{y} = argmax = \frac{1}{N_{Classifiers}}\ \sum_{classifier}(p_1,\ p_2, \dots, p_N\ ) \qquad (8)$$

In this study, we used a soft voting approach, i.e. SVE, to obtain an overall sentiment prediction based on the predictions of LR, RF and DT. The hard voting and soft voting approaches for classifiers LR, RF and DT are compared in Figures 2 and Figure 3 side by side.

Figure 2. Hard Voting Ensemble Sentiment Prediction for Three ML Algorithms

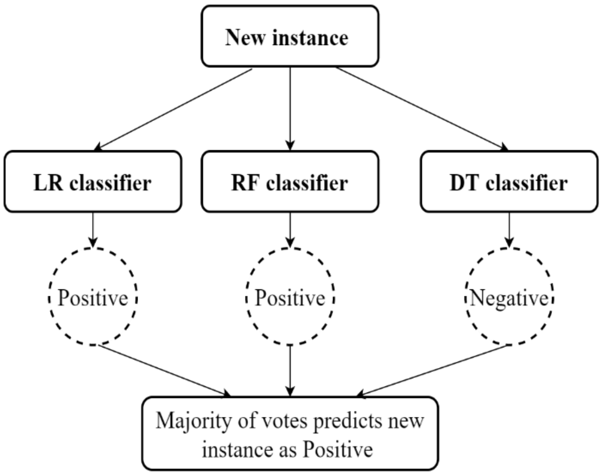
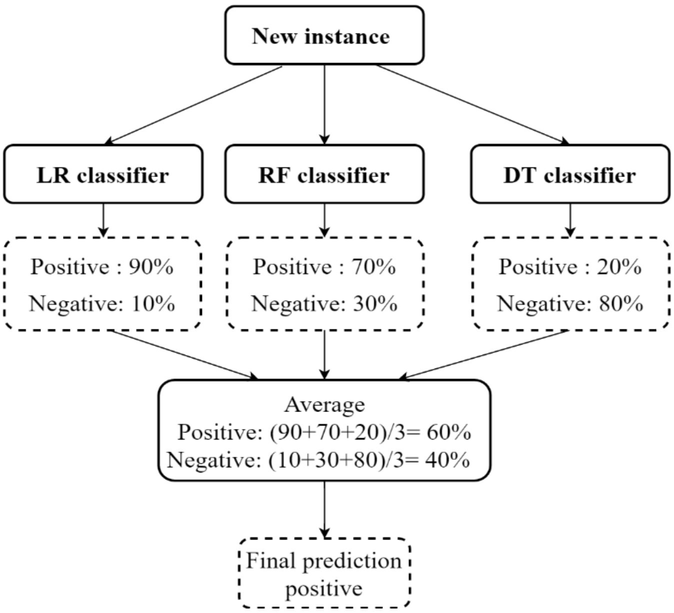


Figure 3. Soft Voting Ensemble Sentiment Prediction for Three ML Algorithms



Experimental Findings

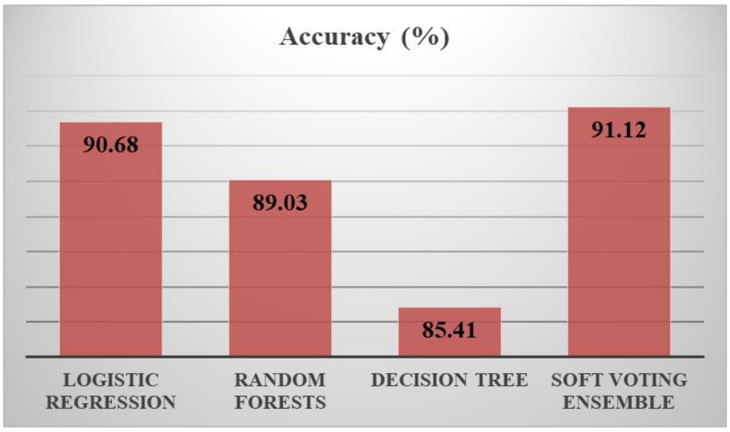
Having evaluated overall mentioned steps, we now present the experimental findings for sentiment prediction accuracies of LR, RF and DT algorithms and their SVE aggregation in Table 3.

Table 3. Experimental Findings for LR, RF, DT and SVE Algorithms in Sentiment Prediction

ML Algorithms	**Accuracy (%)**
Logistic Regression	90.68
Random Forests	89.03
Decision Tree	85.41
Soft Voting Ensemble	**91.12**

It is seen from Table 3 that the most accurate ML algorithm is LR and it has 90.68% sentiment prediction accuracy. And it is further observed from the table that the soft voting combination, SVE, increases this value to 91.12%. We may summarize the experimental findings in Figure 4.

Figure 2. Experimental Findings of Algorithms in Sentiment Prediction



CONCLUSION

There is a growing need for automated analysis of written text from digital sources. In particular, to improve quality of a service or to monitor customer satisfaction, sentiment analysis of customer reviews are valuable sources. In this context, development of accurate systems that extract opinions of customers with minimal human incidence gets more and more critical. Medical informatics or health-care services also make use of such opinion extraction systems. We proposed a system to extract sentiments of patients from a hospital online reviews. The ensemble system has three ML components, LR, RF and DT with the sentiment prediction accuracies of 90.68%, 89.03% and 85.41%. The prediction performance of mentioned ML algorithms with soft-voting aggregation is further increased to 91.12%.

As a future work, we plan to enrich the collected data with other sources from web and we plan to improve prediction performance using another ML algorithms.

REFERENCES

Al-Ayyoub, M., Khamaiseh, A.A., Jararweh & Y., Al-Kabi, M. (2019). A comprehensive survey of arabic sentiment analysis, Information Processing and Management: 320-342.

Bomaccorso, G. (2018). Machine Learning Algorithms: Popular algorithms for data science and machine learning, Packt Publishing, Birmingham, United Kingdom: 281-282.

Coban, O., & Ozel, S. A. (2018). An Empirical Study of the Extreme Learning Machine for Twitter Sentiment Analysis, International Journal of Intelligent Systems and Applications in Engineering: 178-184.

Crannell, C.W., Clark, E., Jones, J., James, T., & Moore, J. (2016). A pattern-matched Twitter analysis of US cancer-patient sentiments, Journal of Surgical Research: 536-542.

Das, B. & Chakraborty, S. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation, Computation and Language, arXiv:1806.06407.

Dekharghani, R., Yanikoglu, B., Saygin & Y., Oflazer, K. (2016).Sentiment analysis in Turkish at different granularity levels,Natural Language Engineering: 535-559.

Ersahin, B., Aktas, O., Kilinc, D. & Ersahin, M. (2019). A hybrid sentiment analysis method for Turkish,, Turkish Journal of Electrical Engineering & Computer Sciences: 1780-1793.

Kaewrod, N. & Kietikul, J. (2018). Improving ID3 Algorithm by Ignoring Minor Instances, International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand.

Karcioglu, A. A. & Aydin, T. (2019). Sentiment Analysis of Turkish and English Twitter Feeds Using Word2Vec Model, 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey.

Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S. & Gonzalez, G. (2019). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, Journal of Biomedical Informatics: 148-158.

Kumar, S., Yadava, M. & Roy, P.P. (2019). Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction, Information Fusion: 41-52.

Makinist, S., Hallac, R., Karakus, B. & Aydın, G. (2017). Preparation of Improved Turkish DataSet for Sentiment Analysis in Social Media, ITM Web of Conferences, 01-03 October 2017.

Obulesu, O., Mahendra, M. & ThrilokReddy, M. (2018). Machine Learning Techniques and Tools: A Survey, Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018).

Omari, M., Al-Hajj, M., Hammami, N. & Sabra, A. (2019). Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon, International Conference on Computer & Information Science (ICCIS), Sakaka, Saudi Arabia.

Peng, W., Silan, N. & Zhonghua, S. (2019). Random Forest Classification of Rice Planting Area Using Multi-Temporal Polarimetric Radarsat-2 Data, IEEE International Symposium Geoscience and Remote Sensing (IGARSS), Yokohama, Japan.

Rajput, A. (2020). Natural Language Processing, Sentiment Analysis, and Clinical Analytics, Innovation in Health Informatics, Cambridge, Massachusetts: 79-97.

Rani, R. & Lobiyal, D. K. (2018).Automatic Construction of Generic Stop Words List for Hindi Text, Procedia Computer Science: 362-370.

Ribeiro, M.H. & Coelho, L.S. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Applied Soft Computing Journal.

Rodrigues, R.G., Dores, R.M., Camilo-Junior, C. & Rosa, T. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks, International Journal of Medical Informatics: 80-95

Sagi, O. & Rokach, L. (2018). Ensemble learning: A survey, WIREs Data Mining Knowledge Discovery: 1-18.

Shehu, H.A., Tokat, S., Haidar, S. & Uyaver, S. (2019). Sentiment analysis of Turkish Twitter data, AIP Conference Proceedings.

Tabassum, N & Tanvir, A. (2016). A theoretical study on classifier ensemble methods and its applications, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India.

Uysal, A. & Gunal, S. (2014). The impact of preprocessing on text classification, Information Processing & Management: 104-112.

Wael, E. & Naymat, G. (2017). The Impact of applying Different Preprocessing Steps on Review Spam Detection, Procedia Computer Science: 273-279.

Zafra, S.M.,Valdivia, T., González, M. & Lopez, A. (2019). How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain, Artificial Intelligence in Medicine: 50-57.