Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 28, S. 690-693, Kasım 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 28, pp. 690-693, November 2021
Copyright © 2021 EJOSAT
Research Article

Evaluation of Ensemble Algorithms and Deep Learning Transformers in Medical Sentiment Prediction

Akın Özçift[1*], Mehmet Bozuyla[2]

[1*] Manisa Celal Bayar University, Hasan Ferdi Turgutlu Technology Faculty, Department of Software Engineering, Manisa, Turkey, (ORCID: 0000-0003-2840-1917), akin.ozcift@mcbu.edu.tr

[2] Pamukkale University, Faculty of Enginnering, Departmant of Electrical and Electronics Engineering, Denizli, Turkey, (ORCID: 0000-0002-7485-6106), mbozuyla05@posta.pau.edu.tr

(1st International Conference on Applied Engineering and Natural Sciences ICAENS 2021, November 1-3, 2021)

(**DOI:** 10.31590/ejosat.1010241)

ATIF/REFERENCE: Özçift, A. & Bozuyla, M., (2021). Evaluation of Ensemble Algorithms and Deep learning Transformers in Medical Sentiment Prediction. *European Journal of Science and Technology*, (28), 690-693.

Abstract

Social media continuously produces digital information that can be used to improve service quality. In this aspect sentiment prediction, automated analysis of written user reviews, is an important research area from service quality point of view. Online sentiment prediction is a rich research area from e-business perspective. However, identification of sentiment from medical service user reviews is particularly researched less frequently. From Turkish language point of view, the medical informatics literature needs more research to design automated medical sentiment systems. Automated sentiment analysis systems particularly make use of Machine Learning (ML) algorithm in tandem with Natural Language Processing (NLP) methods to address written user reviews. In this work, ensemble learning approaches are compared with newly developed deep learning variations, Bidirectional Encoder Representations from Transformers (BERT), to investigate medical sentiments. As the obtained results are evaluated, it is observed that newly proposed transformer models are perfectly successful to identify sentiment of Turkish medical reviews.

Keywords: Ensemble Learning, Bidirectional Encoder Representations from Transformers, Medical Review, Sentiment Identification

Tıbbi Duyarlılık Tahmininde Topluluk Algoritmalarının ve Derin Öğrenme Transformatörlerinin Değerlendirilmesi

Öz

Sosyal medya, hizmet kalitesini artırmak için kullanılabilecek dijital bilgileri sürekli olarak üretmektedir. Bu yönüyle duygu tahmini, yazılı kullanıcı yorumlarının otomatik analizi, hizmet kalitesi açısından önemli bir araştırma alanıdır. Çevrimiçi duygu tahmini, e-iş perspektifinden zengin bir araştırma alanıdır. Bununla birlikte, tıbbi servislere ait kullanıcı incelemelerinden duyguların belirlenmesi özellikle daha az sıklıkla araştırılmaktadır. Türk dili açısından bakıldığında, tıbbi bilişim literatürünün otomatikleştirilmiş tıbbi duyarlılık sistemleri tasarlamak için daha fazla araştırmaya ihtiyacı vardır. Otomatik duygu analizi sistemleri, yazılı kullanıcı incelemelerini ele almak için özellikle Doğal Dil İşleme (DDİ) yöntemleriyle birlikte Makine Öğrenimi (MÖ) algoritmalarını kullanır. Bu çalışmada, tıbbi yorum duygularını araştırmak için topluluk öğrenme yaklaşımları, yeni geliştirilen derin öğrenme varyasyonları olan Transformers'dan Çift Yönlü Kodlayıcı Gösterimleri (TÇYK) ile karşılaştırılmıştır. Elde edilen sonuçlar değerlendirildiğinde, yeni önerilen transfers modellerinin Türkçe tıbbi incelemelerinin duyarlılığını belirlemede mükemmel derecede başarılı olduğu görülmektedir.

Anahtar Kelimeler: Topluluk Öğrenimi, Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri, Tıbbi Yorum, Duygu Tanımlama

[*] Corresponding Author: akin.ozcift@mcbu.edu.tr

1. Introduction

Digital information in terms of user reviews are generated constantly. In particular, Sentiment Analysis (SA) is used to extract user opinion from written reviews to enhance the quality of all kinds of services such as products, hotels, airlines etc. analysis of valuable information from these reviews are principally important. From this point of view, medical sentiment analysis is a research field that is effectively used to develop service quality of medical domain (Rahim et al., 2021). More precisely, this anaylsis focus on patient opinions about hospitals or doctors. In other words, patients may use this type of analysis to choose a physician or hospitals use these opinions to improve the quality of health-care facilities (Jiménez-Zafra et al., 2019).

Automation of any SA task is important from data abundance point of view. Since, it is not possible to extract information from social media, automated analysis models are strictly becomes vital (Lin et al., 2020). Most of the automated analysis tasks make use of ML and NLP methods in tandem to extract user opinions or to obtain sentiment of users for any service (Ullah et al., 2020).

In the literature, there are many SA studies using ML/NLP combinations in some way. In this work, we focus on recent Turkish SA research and particularly Turkish medical sentiment analysis.

Alqaraleh, in his Turkish movie sentiment analysis study, made use of Random Forests (RF) and AdaBoost (ADB) esemble learners and he obtained promising results (Alqaraleh, 2020). Catal et al. combined Bagging (BG), Support Vector Machine and Naïve Bayes in a majority voting ensemble strategy in their work to detect Turkish sentiments (Catal & Nangir, 2017). In his recent work, Onan proposed an ensemble architecture to evaluate Turkish sentiments with the use of BAG and ADB algorithms (Onan, 2021). In another work, Tocoglu used AdaBoost, Bagging and Voting ensembles to analyse sentiments in software domain (Toçoğlu, 2020). A recent study focusing on Turkish movie and SemEval-2017 datasets made use of Stacking ensemble strategy to detect sentiments (Görmez et al., 2020).

For Turkish medical sentiment domain, a recent study by Ozcift used a majority Voting strategy to analyse medical reviews (Özçift, 2020). Another study that analyses medical records with the use of Naïve Bayes (NB), J48 tree and Support Vector Machine (SVM) (Ceyhan et al., 2017). Twitter based reviews for physical activities was studied in (Şahin et al., 2021) to analyse sentiments.

As it is observed from literature survey, Turkish medical sentiment analysis domain is relatively insufficient. This work in this aspect is a contribution to the medical sentiment literature with the use of newly developed deep learning transformer algorithms. In this context, we first tested performance of various ensemble algorithms and then we compared them with the newly proposed transformer algorithms.

The rest of the paper is as follows: We explain the related framework and the experimental setup in Section 2. The results of the conducted experiments are given in Section 3. Our research ends with conclusion in Section 4.

2. Material and Method

In this section, we explain the medical sentiment evaluation pipeline. Our framework is composed of sections such as data, experimental setup, results of experiments and evaluation-validation metrics.

2.1. Medical Sentiment Data

The medical sentiment data is obtained from (Özçift, 2020) and after the data processed it consists of 1843 positive and 2319 negative instances. In this aspect, the data is balanced and we may use Accuracy (ACC) as performance evaluation metric.

2.2. Experimental Layout

Traditional ML algorithms need a text pre-processing step to be able to make a prediction such as sentiment classification. In more clear terms, extraction of features from written text requires a proper encoding such as term frequency-inverse document frequency (TF-IDF). We therefore used this encoding scheme to represent medical reviews to be able to use in ensemble models. For the sake of reliability of comparision among all algorithms, we pereferred to use an 80/20 (train/test) split in all the experiments.

Since the goal of this study was to compare new transformer models with ensemble algorithms, we first selected widely used ensemble algorithms and multi-lingual transformers from literature. As a second step, we evaluated performance of the selected algorithms in terms of Accuracy (Acc) and we tested their confidence in terms of Matthews Correlation Coefficient (MCC)

As ensemble algorithms, we selected ADB, BG, LogiBoost (LGB), Random Subspace (RS), Rotation Forests (ROTF), Random Committee (RCM) and Random Forests (RANF) (Dong et al., 2020) from WEKA suit. Having obtained features from medical texts, we then used 80/20 split to get corresponding Acc and MCC values.

For transformer models, we selected multilingual Bidirectional Encoder Representations from Transformers (BERT) and its variation DistilBERT (*Web 1*, 2021). Furthermore, we made use of Turkish language dedicated transformer (BERTurk) as the third algorithm. We made use of 80/20 data split while we tune parameters of transformers and we also obtained Acc and MCC values for these experiments.

2.3. Evaluation Metrics

As we mentioned, our dataset is relatively balanced in terms of poisitive and negative number of samples. We therefore used Acc metric in the comparison of the ensemble and transformer algorithms. Acc is given in Equation 1 below.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In this equation True Positives (TP) and True Negatives (TN) are correct predictions. Incorrect predictions are denoted as False Negatives (FN) and False Positives (FP) in the same equation.

Any ML evaluation study needs to validate the experimental results statistically. One of the widely used statistical validation metric is MCC (Duysak et al., 2021) and it is calculated with Equation 2.

MCC metric for an experiment is statistically meaningful as the generated value is closer to 1.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{2}$$

3. Results

In this section, we present the experimental results obtained from previous section in Table 1 in terms of Acc and MCC values.

Table 1. Experimental results for Sentiment Identification

Algorithms	Acc	MCC
LGB	94.71	0.894
ADB	94.83	0.897
BG	96.39	0.927
RS	96.76	0.934
ROTF	97.12	0.942
RCM	97.60	0.951
RANF	97.96	0.958
DistilBERT	99.53	0.990
BERT	99.62	0.992
BERTTurk	99.97	0.999

Table 1 illustrates that the performance of transformer algorithms are better in terms of Acc compared to remaining models. We may observe that the best Acc among ensembles is achieved by RANF with 97.96%. This best performance is enhanced with all of the performances of transformers with Acc's of 99.53% to 99.97%. We compare Acc values of algorithms in Figure 1.

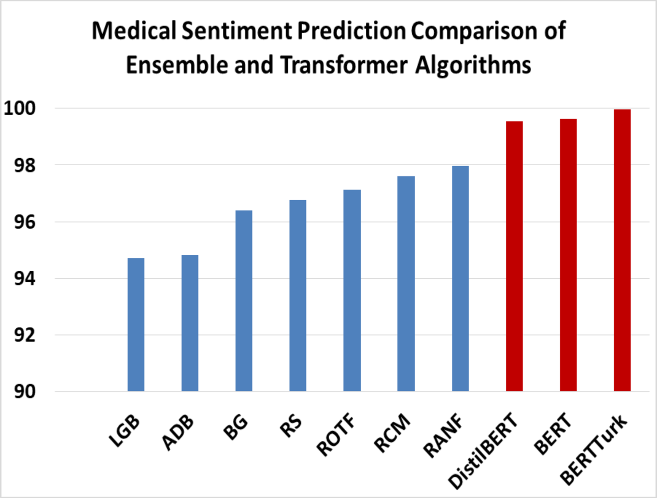

Figure 1. Accuracy of whole algorithms

The performance of the algorithms are also validated in terms of MCC values Figure 2.

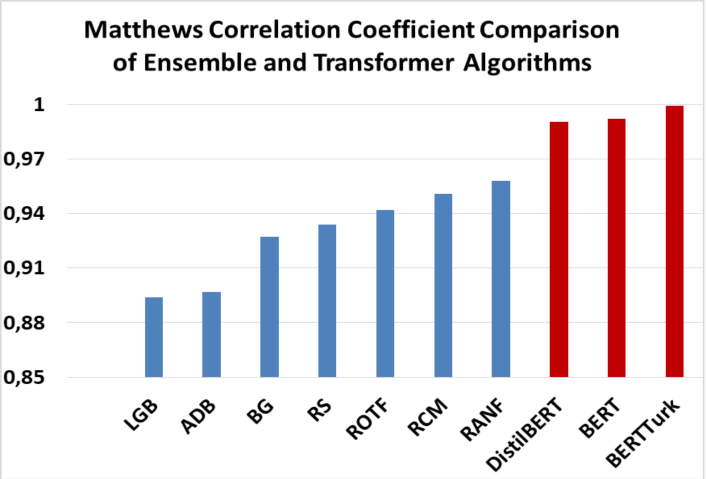

Figure 2. Kappa score of dataset

It is observed from Figure 2 that the performance of predictors are meaningfull with varying MCC values from 0.894 to 0.999.

4. Conclusions

Automated analysis of user reviews is important to improve quality of any service. In particular, advancing health-care services require to analyze medical reviews. In this manner, accurate systems that extract user opinions without or minimal human involvement becomes vital. From this point of view, we analyzed advanced ML algorithms from literature in medical sentiment prediction ability for Turkish language. It is deduced from experiments that newly developed transformer algorithms are more versatile and more performative in terms of prediction efficiency.

From Turkish language point of view, it can be deduced that use of tranformers is probably the new research direction.

References

Alqaraleh, S. (2020). Turkish Sentiment Analysis System via Ensemble Learning. *European Journal of Science and Technology*, 122–129. https://doi.org/10.31590/ejosat.779181

Catal, C., & Nangir, M. (2017). A sentiment classification model based on multiple classifiers. *Applied Soft Computing*, *50*, 135–141. https://doi.org/10.1016/j.asoc.2016.11.022

Ceyhan, M., Orhan, Z., & Domnori, E. (2017). Health service quality measurement from patient reviewsin Turkish by opinion mining. *Badnjevic A. (Eds) CMBEBIH 2017. IFMBE Proceedings*, *62*, 649–653. https://doi.org/10.1007/978-981-10-4166-2_97

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258.

Duysak, H., Ozkaya, U., & Yigit, E. (2021). Determination of the Amount of Grain in Silos with Deep Learning Methods Based on Radar Spectrogram Data. IEEE Transactions on Instrumentation and Measurement. tps://doi.org/10.1007/s11704-019-8208-z

Görmez, Y., Işık, Y. E., Temiz, M., & Aydın, Z. (2020). FBSEM: A Novel Feature-Based Stacked Ensemble Method for Sentiment Analysis. *International Journal of Information Technology and Computer Science*, *6*, 11–22. https://doi.org/10.5815/ijitcs.2020.06.02

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

Lin, H. C. K., Wang, T. H., Lin, G. C., Cheng, S. C., Chen, H. R., & Huang, Y. M. (2020). Applying sentiment analysis to automatically classify consumer comments concerning marketing 4Cs aspects. *Applied Soft Computing*, *97*, 106755. https://doi.org/10.1016/J.ASOC.2020.106755

Onan, A. (2021). Ensemble of Classifiers and Term Weighting Schemes for Sentiment Analysis in Turkish. *Scientific Research Communications*, *1*(1), 1–12. https://doi.org/10.52460/src.2021.004

Özçift, A. (2020). Medical Sentiment Analysis Based on Soft Votiıng. *Yönetim Bilişim Sistemleri Dergisi*, *6*(1), 42–50.

Rahim, A. I. A., Ibrahim, M. I., Musa, K. I., Chua, S. L., & Yaacob, N. M. (2021). Assessing Patient-Perceived Hospital Service Quality and Sentiment in Malaysian Public Hospitals using Machine Learning and Facebook Reviews. *International Journal of Environmental Research and Public Health*, *18*, 1–28. https://doi.org/10.3390/ijerph18189912

Şahin, T., Gümüş, H., & Gençoğlu, C. (2021). Analysis of Tweets Related with Physical Activity During COVID-19 Outbreak. *Journal of Basic and Clinical Health Sciences*, *1*, 42–48. https://doi.org/10.30621/jbachs.869506

Toçoğlu, M. A. (2020). Sentiment Analysis for Software Engineering Domain in Turkish. *Sakarya University Journal of Computer and Information Sciences*, *3*(3). https://doi.org/10.35377/saucis.03.03.769969

Ullah, M. A., Marium, S. M., Begum, S. A., & Dipa, N. S. (2020). An algorithm and method for sentiment analysis using the text and emoticon. *ICT Express*, *6*(4), 357–360. https://doi.org/10.1016/j.icte.2020.07.003

Web 1. (2021). https://huggingface.co/dbmdz