AJIT-e: Bilişim Teknolojileri Online Dergisi
2020 Winter/Kış – Cilt/Vol: 11 - Sayı/Issue:40
DOI: 10.5824/ajite.2020.01.001.x

Date Received: 26.02.2020
Publishing Date: 3.05.2020

Application of Natural Language Processing with Supervised Machine Learning Techniques to Predict the Overall Drugs Performance

Pius MARTHIN, *Hacettepe University, Department of Statistics, PhD scholar, martinpius01@gmail.com, ORCID: 0000-0003-3529-0311*

Duygu İÇEN, *Hacettepe University, Department of Statistics, Associate Professor, duyguicn@hacettepe.edu.tr, ORCID: 0000-0002-7940-5064*

ABSTRACT

Online product reviews have become a valuable source of information which facilitate customer decision with respect to a particular product. With the wealthy information regarding user's satisfaction and experiences about a particular drug, pharmaceutical companies make the use of online drug reviews to improve the quality of their products. Machine learning has enabled scientists to train more efficient models which facilitate decision making in various fields. In this manuscript we applied a drug review dataset used by (Gräßer, Kallumadi, Malberg,& Zaunseder, 2018), available freely from machine learning repository website of the University of California Irvine (UCI) to identify best machine learning model which provide a better prediction of the overall drug performance with respect to users' reviews. Apart from several manipulations done to improve model accuracy, all necessary procedures required for text analysis were followed including text cleaning and transformation of texts to numeric format for easy training machine learning models. Prior to modeling, we obtained overall sentiment scores for the reviews. Customer's reviews were summarized and visualized using a bar plot and word cloud to explore the most frequent terms. Due to scalability issues, we were able to use only the sample of the dataset. We randomly sampled 15000 observations from the 161297 training dataset and 10000 observations were randomly sampled from the 53766 testing dataset. Several machine learning models were trained using 10 folds cross-validation performed under stratified random sampling. The trained models include Classification and Regression Trees (CART), classification tree by C5.0, logistic regression (GLM), Multivariate Adaptive Regression Spline (MARS), Support vector machine (SVM) with both radial and linear kernels and a classification tree using random forest (Random Forest). Model selection was done through a comparison of accuracies and computational efficiency. Support vector machine (SVM) with linear kernel was significantly best with an accuracy of 83% compared to the rest. Using only a small portion of the dataset, we managed to attain reasonable accuracy in our models by applying the TF-IDF transformation and Latent Semantic Analysis (LSA) technique to our TDM.

Keywords **:** ***Term Document Matrix (TDM), Machine Learning, Sentiment Analysis, Cross-Validation, Term Frequency-Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA)***

AJIT-e: Bilişim Teknolojileri Online Dergisi
2020 Winter/Kış – Cilt/Vol: 11 - Sayı/Issue: 40
DOI: 10.5824/ajite.2020.01.001.x

Genel İlaç Performansını Tahmin Etmek İçin denetimli Makine Öğrenme Yöntemleriyle Doğal Dil İşleme Uygulaması

ÖZ

Çevrimiçi ürün incelemeleri, belirli bir ürünle ilgili müşterilerin karar almasını kolaylaştıran değerli bir bilgi kaynağı haline gelmiştir. İlaç şirketleri, ürünlerinin kalitesini artırmak adına kullanıcının memnuniyeti ve belirli bir ilaçla ilgili deneyimleri hakkındaki zengin bilgilerle donatılmış olan çevrimiçi ilaç incelemelerini kullanır. Makine öğrenimi, bilim insanlarının çeşitli alanlarda karar vermeyi kolaylaştıran daha verimli modeller geliştirmelerini sağlamaktadır. Bu makalede UCI makine öğrenimi veri havuzu web sitesinden Gräßer, Kallumadi, Malberg ve Zaunseder (2018) tarafından kullanılan bir ilaç inceleme verisini ele aldık. Amacımız kullanıcıların yaptıkları incelemelerine göre genel ilaç performansının daha iyi tahmin edilmesini sağlayan en iyi makine öğrenme modelini belirlemektir. Model doğruluğunu artırmak için yapılan çeşitli manipülasyonların yanı sıra, metin temizliği ve makine öğrenme modelleri uygulamak için metinlerin sayısal formata dönüştürülmesi dahil olmak üzere metin analizi için gerekli tüm prosedürler izlenmiştir. Modellemeye geçilmeden önce, müşterilerin ilaçlar hakkında yaptıkları incelemeler için genel duygu puanları elde ettik. Müşterilerin yorumları, en sık kullanılan terimleri keşfetmek için bir çubuk grafiği ve kelime bulutu grafiği kullanılarak özetlendi ve görselleştirildi. 161297 gözlemli eğitim verisinden rastgele 15000 gözlem seçtik ve 53766 gözlemli test verisinden 10000 gözlem rastgele seçildi. Çeşitli makine öğrenme modelleri, tabakalı rastgele örnekleme altında gerçekleştirilen 10 kat çapraz doğrulama kullanılarak eğitildi. Eğitim için kullanılan modeller: Sınıflandırma ve Regresyon Ağaçları (CART), C5.0 algoritması, lojistik regresyon (GLM), Çok Değişkenli Uyarlanabilir Regresyon Eğrileri (MARS), Destek vektör makinesinin (SVM) hem radyal hem de doğrusal çekirdekleri ve Rastgele Orman (Random Forest) algoritmalarıdır. Model seçimi doğruluk ve hesaplama verimliliğinin karşılaştırılması yoluyla yapılmıştır. Lineer çekirdekli destek vektör makinesi (SVM), diğerlerine kıyasla% 83 doğrulukla önemli ölçüde en iyi tahmin sonuçlarını vermiştir. Veri kümesinin sadece küçük bir kısmını kullanarak, TF-IDF dönüşümünü ve Latent Semantik Analiz (LSA) ile TDM'imize uygulayarak modellerimizde makul doğruluk elde etmeyi başardık.

Anahtar Kelimeler **:** ***Terim Belge Matrisi, Makine Öğrenme, Duygu Analizi, Çapraz Doğrulama, Terim Frekansı-Ters belge Frekansı, Gizli Semantik Analiz***

INTRODUCTION

Pharmaceutical companies ensure the safety of their products depending mostly on clinical trials and specific test protocols used to test drug effectiveness. Due to a limited number of test subjects and time span, high variations and biases in patient selection may be inevitable for such kinds of studies (Gräβer *et al*, 2018). Consequently, a significant impact on the effectiveness of the drug and unexpected adverse drug reactions may occur.

According to the study conducted by (Pirmohamed, James, Meakin, Green, Scott, Walley, Farrar, Park, & Breckenridge), adverse Drug Reactions (ADRs) is one of the major public health issues and one of the leading causes of morbidity and mortality. Korkontzelos, Ioannis, Nikfarjam, Azadeh, Shardlow, Matthew, Sarker, Abeed, Ananiadou, Sophia, Gonzalez &, Graciela, (2016) found that although the efficiency and safety of drugs are tested during clinical trials, many ADRs remain latent and may only be revealed under specific cases such as: after long-term use, when used in combination with other drugs, or when used by patients who were excluded from the trials such as adults with other morbidities, children, the elderly or pregnant women. Therefore, the use of systematic drug reviews that aggregate the available information in a neutral manner is very essential in order to uplift customer satisfaction, achieve business objectives and improve community health in general. Procedures that lead to ideal personalized treatment options for a given patient and time specifically depend on structured data (Gräβer *et al*, 2018). The amount of such data often appears to be limited as it requires intense preparation which is not usual in clinical routine and therefore other targets of information such as user reviews are of great demand (Gräβer *et al*, 2018). With the rapid growth of social media on the Web, individuals and organizations are increasingly using public opinions in these media for their decision making (Liu and Zhang, 2012). Although the Accessibility of all-important data from an unstructured source is a challenge, it can significantly increase the healthcare practitioners' knowledge of the patient if the information embedded in these sources can be exposed (IBM Corporation, 2013).

Sentiments analysis for opinions presented via medical platforms provides significant usefulness in decision making concerning public health (Gräβer *et al*, 2018). Positive and negative effects of a treatment can be assessed for clinical evidence; relations between symptoms, lifestyle and effectiveness can also be studied (Gräβer *et al*, 2018). Information on the health status and psychological status of a patient can be collected for example by analyzing information generated within a patient-doctor social network (Kerstin, 2015).

Texts obtained from other social networks; opinions may be conveyed through facts that are interpretable by emotions they convey (Denecke and Deng 2015). In a similar manner we can compare to sentiment analysis in the healthcare domain where sentiment are fetched through diseases, treatments or medical conditions and their impact on a patient's life quality and health status (Denecke and Deng 2015).

DOI: 10.5824/ajite.2020.01.001.x

However, users of online medical platforms express their views in a unique manner as the language used to comment on a particular drug or medication differs much from other usual platforms such as sports, business, etc. This imposes a limitation in applying sentiment analysis using typical lexicons.

In most of the existing literature, the sentiment is often taken as polarity, i.e. positive, negative or neutral polarity towards some subject (Denecke and Deng 2015). In contrast to products or persons where sentiment mainly comprises of like or dislike towards a person or product, opinions or sentiments towards medications, treatments or even diagnoses sentiments have even more facets and are expressed in different words (Denecke and Deng 2015).

Accordingly, alternative procedures that consider the problem as either classification or regression may be carried out where machine learning can be applied to provide possible solutions. Machine learning techniques can appropriately used to train classifiers on domain-specific data sets to detect the polarity at sentence or document level and performing sentiment analysis over multiple facets of issues (Gräβer *et al*, 2018).

Therefore, using machine learning as an alternative remedy, several studies on analyzing online drug reviews from different medical platforms have been conducted including but not limited to (Jimene, Martín, & Urena, 2019), who applied supervised learning and lexicon-based sentiment analysis approach over two different corpora extracted from social web specifically focused on drugs and doctors, (Kho, Padhee, Bajaj, Thirunarayan, & Sheth, 2019) discussed the need to go beyond data-driven machine learning and natural language processing and incorporate deep domain knowledge, (Bhargava, 2019), applied the k-means clustering algorithm on a textual dataset of unlabeled reviews of medicinal drugs in order to group the drugs with similar usage and benefits, (Gräβer *et al*, 2018), performed multiple tasks over drug reviews with data obtained by crawling online pharmaceutical review sites (same dataset applied in this paper) to perform sentiment analysis to predict the sentiments concerning overall satisfaction, side effects and effectiveness of user reviews on the specific drug.

In this manuscript, we apply the drug review dataset used by (Gräβer *et al*, 2018) available freely from from machine learning repository website of the University of California Irvine (UCI) to perform sentiment analysis on drug reviews in order to identify the best machine learning model which provides a better prediction of the overall drug performance with respect to users' reviews. In this study, we apply Latent Semantic Analysis (LSA) to select few most important predictive features and penalize mostly frequent terms using TF-IDF transformation to attain reasonable accuracy for our models in the most efficient manner using only a portion of the dataset.

The rest of the manuscript is organized as follows: in section 2 we discuss the dataset used to train our machine learning models, section 3 covers material and methods, section 4 includes

the results and discussion. Concluding remarks and some possible future perspectives are addressed in section 5.

DATASET

The drug data set was created by (Gräβer *et al*, 2018) and is available freely from the machine learning repository website of the University of California Irvine (UCI). The texts files are downloaded containing both training and testing datasets. The datasets consisted of 6 features which defines drug name, patient condition, patient review (text), ratings (10-star patient rating), review date and number of users who found the review useful (For more information about the data set please see machine learning repository website of the University of California Irvine (UCI) with the link provided in the reference list).

Due to scalability issues we are able to use only the sample of the dataset. We randomly sample 15000 observations from 161297 training dataset and 10000 observations are randomly sampled from 53766 testing dataset. For the purpose of this study, we select two features including reviews (text) and ratings. We create our target variable which represents overall drug performance (binary) by converting ratings into a factor and redefining its levels as high if it has 6 or higher star-patient rating score and low otherwise.

MATERIAL and METHODS

All necessary procedures for text analytics are applied to clean the corpus (reviews collections) includes removal of (URL, stop-words, punctuations, white space), converting to lowercase, steaming the document and finally converting to the term-document matrix (TDM). We also obtain a data frame consisting of terms (words) with their respective frequencies to be used for word clouds and bar plot of most frequent terms in the corpus together with other necessary computations such as term frequency-inverse document frequency (TFIDF) transformation.

For feature space, both unigram and bi-gram cases are considered but no improvement in the model accuracy through bi-gram is achieved and therefore we rely completely on unigram models. Afterward, we also engineer two new features by utilizing review length and cosine similarities respectively. The new feature due to review length does not produce any improvement to our model's accuracy and hence it will not be used.

A new feature with respect to cosine similarity is computed under the hypothesis that low-rated drugs have low cosine similarities with highly rated drugs or vice versa. Cosine similarity calculates similarity by measuring the cosine of the angle between two vectors. Given the two vectors A and B, cosine similarity is given with Equation 1.

DOI: 10.5824/ajite.2020.01.001.x

$$cos\ cos\ (\theta) = \frac{A.B}{||A||*||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\left(\sqrt{\sum_{i=1}^{n} {A_i}^2}\right)*\left(\sqrt{\sum_{i=1}^{n} {B_i}^2}\right)} \quad (1)$$

In the equation above the numerator is the usual dot product and the denominator is the Euclidean distances or magnitude for the two vectors as suggested in the study conducted by (Luo, Zhan, Xue , Wang , Ren, & Yang , 2018). This formula is applied to our term frequency matrix (TF) to compute the new predictor.

We then apply the Latent Semantic Analysis (LSA) to extract 300 most influential features. LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. It is an information retrieval technique that analyzes and identifies the pattern in an unstructured collection of text and the relationship between them through singular value decomposition (SVD) (Landauer, Foltz, & Laham, 1998).

Further, we standardize our term-document matrix (TDM) through term frequency-inverse document frequency (TFIDF) transformation in order to penalize most frequent terms. TF-IDF scores are computed as a product of term frequency (TF) and inverse document frequency (IDF) for specific words. Therefore, the score of any word in any document (review) can be obtained as TF-IDF (word, doc) = TF ($word, doc$) $*$ IDF ($word$) . TF and IDF matrices are computed using Equation 2 and Equation 3 provided by Liu, Sheng, Wei, & Yang, (2018).

$$\text{TF } (word, doc) = \frac{Frequency\ of\ words\ belong\ to\ the\ document}{Number\ of\ words\ belong\ to\ a\ document} \quad (2)$$

$$\text{IDF } (word) = log\left(1 + \frac{number\ of\ documents}{Number\ of\ documents\ with\ word}\right) \quad (3)$$

RESULTS

Prior to modeling, an unsupervised machine learning approach using a high-quality, moderate-sized emotion lexicon developed by Saif Mohammad and Peter Turney (2010) is conducted to obtain overall sentiment scores for the reviews of the drug as summarized in Figure 1 below.

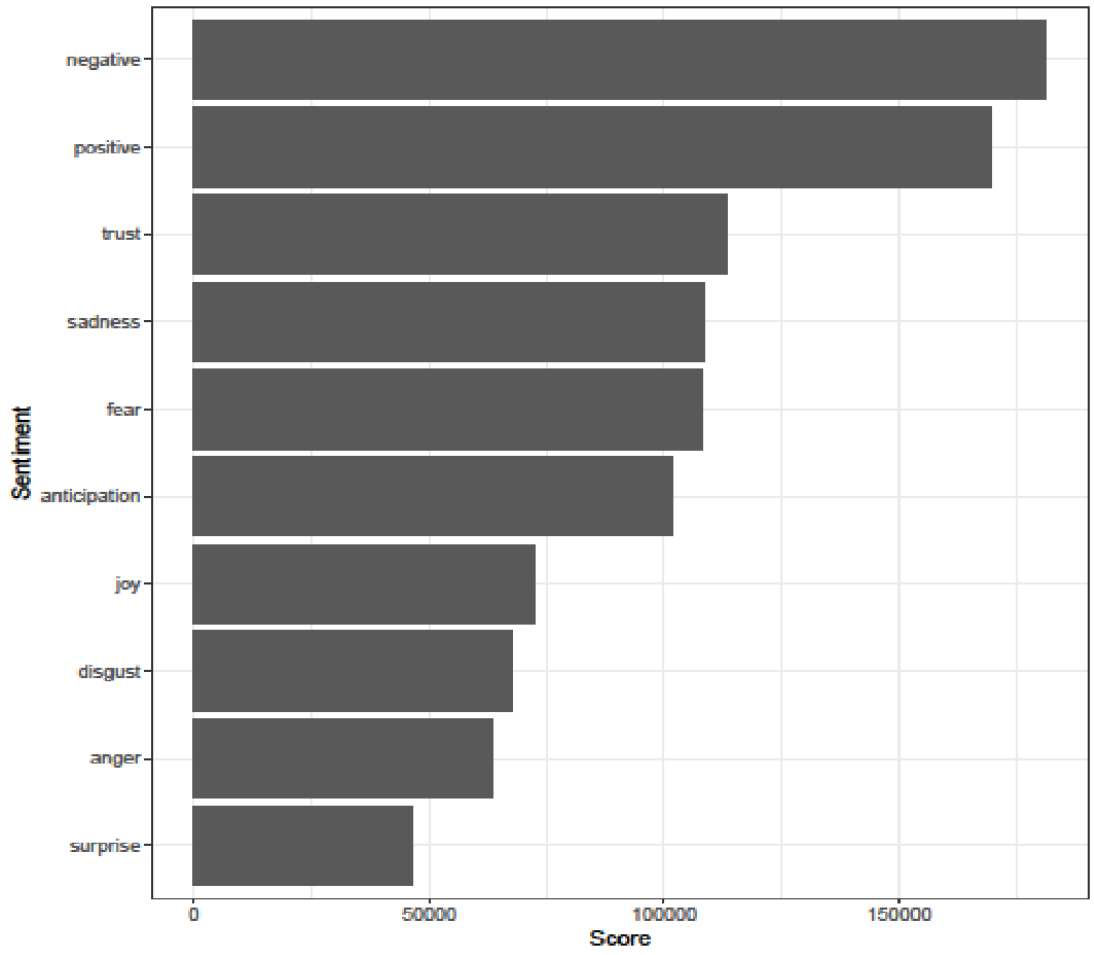

Figure 1. Overall sentiment score for the drug reviews

From Figure 1 we observe different emotions and polarity expressed by drug users. Apart from other emotions and feelings, we see the highest scores for negative and positive sentiments. Further, the score for negative sentiment is higher than that of positive sentiment. Due to limitations of sentiment analysis on medical reviews as discussed in section 1, we cannot generalize on drug performance strictly based on sentiment scores.

We also explored most frequently terms that occurred in our document matrix and summary results are shown in Figure 2 below.

DOI: 10.5824/ajite.2020.01.001.x

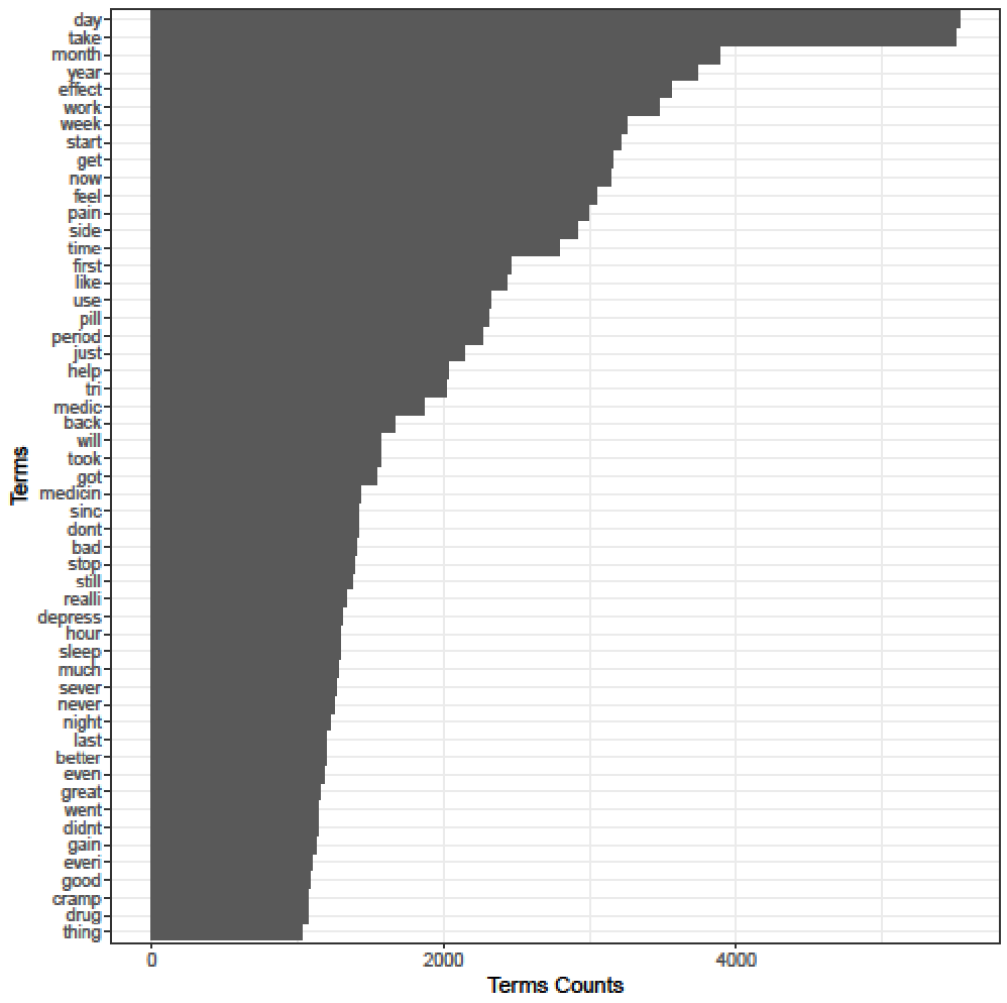

Figure 2. Most frequent terms in the corpus

Like as shown from Figure 2, most frequently words such as 'day', 'take', 'month' etc. will be penalized using TF-IDF transformations before training our machine learning models to avoid overfit problems since they are less informative on classifying newly incoming data.

We also visualize our term-document matrix (TDM) by constructing a word cloud. As shown in Figure 3 below, most frequently terms are much bigger in size as compared to less frequently terms. Like discussed above, large-sized words in the cloud represent the most frequent words which must be penalized to improve model accuracy.

Figure 3. Word cloud for the drug reviews

After completing all necessary manipulations on our text document as explained in section 3, five different machine learning models are trained to predict the overall drug's performance concerning users' reviews. The final data frame consists of 302 features of which 300 are the most important predictors obtained through LSA and two more features are engineered concerning review length and cosine similarity respectively. The new feature engineered concerning review length did not add any value and hence it was discarded.

Models are trained using 10 fold cross-validation through a stratified sampling approach to preserve the balance in the levels of our target variable. Our binary target variable which represents overall drug performance is modeled using 301 predictors. Models trained include Classification and Regression Trees (CART), classification tree by C5.0, logistic regression (GLM), Multivariate Adaptive Regression Spline (MARS), Support vector machine (SVM) with both radial and linear kernels and a classification tree using random forest (Random Forest). The tables below provide summary results for our model-fitting parameters.

DOI: 10.5824/ajite.2020.01.001.x

Table 1. Models Accuracy summary results

Model	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum	Rank
Random Forest	0.8289	0.8373	0.8420	0.8426	0.8482	0.8586	1
SVMRadial	0.8258	0.8340	0.8385	0.8384	0.8426	0.8493	2
GLM	0.8221	0.8307	0.8358	0.8351	0.8389	0.8511	3
SVMLinear	0.7675	0.8290	0.8335	0.8314	0.8377	0.8468	4
MARS	0.8061	0.8128	0.8178	0.8170	0.8215	0.8289	5
C5.0	0.7829	0.8018	0.8087	0.8071	0.8133	0.8189	6
CART	0.7906	0.8015	0.8071	0.8061	0.8104	0.8220	7

From tables 1.0 above, random forest model has higher accuracy 84% followed by SVM with radial kernel with an accuracy of 83%, logistic regression model (GLM) with an accuracy of 83%, SVM with linear kernel with an accuracy of 83%, MARS with an accuracy of 82%, C5.0 with an accuracy of 81%, and CART with an accuracy of 80% appeared to be the least performed model in predicting overall drugs performance.

Table 2. Kappa values summary results

Model	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
Random Forest	0.4199	0.4415	0.4581	0.4622	0.4850	0.5175
GLM	0.3889	0.4290	0.4404	0.4402	0.4547	0.4965
SVMRadial	0.3898	04172	0.4400	0.4348	0.4519	0.4760
SVMLinear	0.2841	0.4004	0.4152	0.4111	0.4320	0.4608
MARS	0.3025	0.3373	0.3507	0.3502	0.3666	0.4044
C5.0	0.2795	0.3189	0.3375	0.3383	0.3606	0.3864
CART	0.2464	0.2952	0.3217	0.3151	0.3371	0.3575

Table 2.0 above presents corresponding Kappa values for each model. Similar to Table 1.0 above, the random forest model has a better performance followed by SVM with the radial kernel, GLM, SVM with the linear kernel, MARS, C5.0, and CART which is the least performed model among the fitted models.

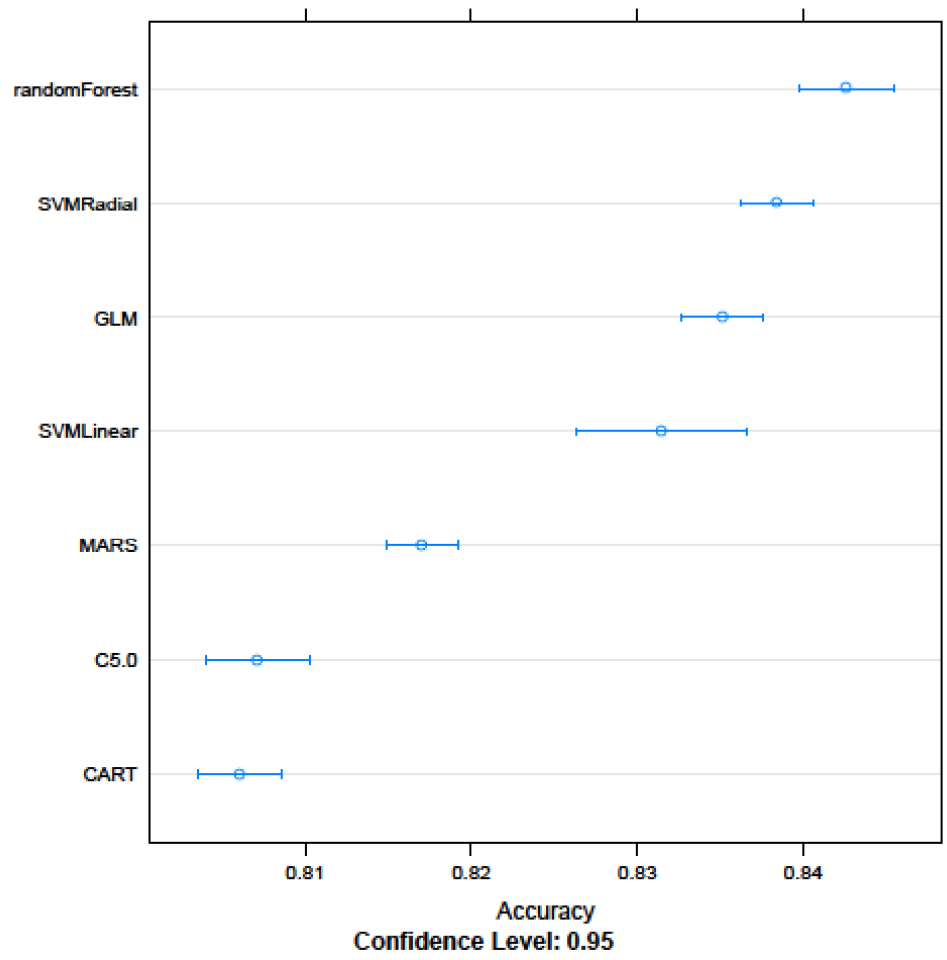

Figure 4. Model's Accuracy plot

Similar to table 1.0, figure 4.0 above displays the summary results of our model's performance. From the plot, the random forest model has higher accuracy followed by SVM with the radial kernel, GLM, SVM with linear kernel, MARS, C5.0, and CART.

DOI: 10.5824/ajite.2020.01.001.x

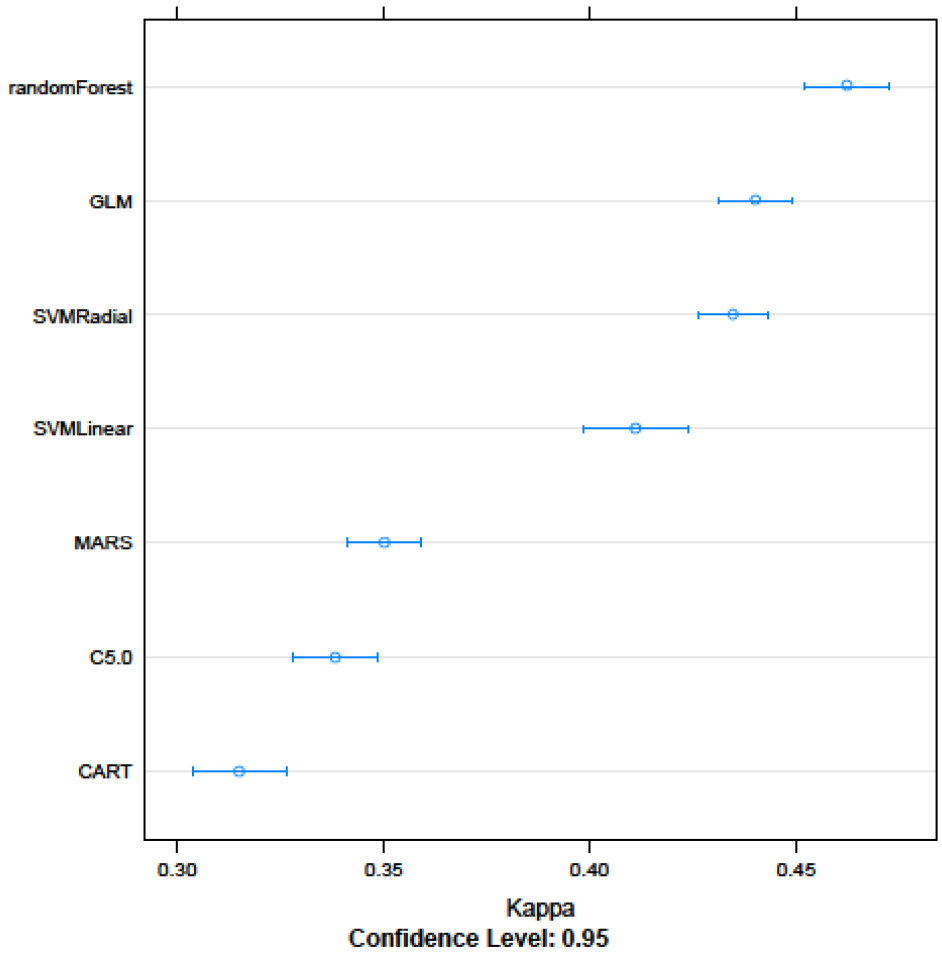

Figure 5. Model's Kappa values plot

Like in table 2.0 above, the same information is displayed in figure 5.0 where Kappa values are plotted. Similar to the above explanations, the random forest has the best performance accuracy followed by SVM with the radial kernel, GLM, SVM with linear kernel, MARS, C5.0, and CART.

Results from the above tables and figures indicate high competition across different models. Therefore, we performed a Bonferroni test to analyze the significant differences among the fitted models. The table below provides summary results of the test.

Table 3. Estimates of the differences and p-values

Accuracy	CART	MARS	GLM	RF	C5.0	SVMLinear	SVMRadial
CART		-0.01097	-0.0290	-0.03649	-0.0011	-0.025376	-0.032319
MARS	4.01e-09		-0.0180	-0.02552	0.0099	-0.014403	-0.021349
GLM	<2.2e-16	2.324e-24		-0.00746	0.02798	0.003658	-0.003285
RF	<2.2e-16	<2.2e-16	4.863e-06		0.03544	0.01111	0.004175
C5.0	1.0000	6.420e-06	4.495e-16	2.2e-16		-0.02472	-0.031265
SVMLinear	1.87e-09	0.0005	1.00000	0.008853	1.02e-07		-0.006943
SVMRadial	<2.2e-16	2.231e-16	0.13355	0.05982	<2.2e-16	0.0955111	

The upper diagonal values of table 3.0 represent estimated differences between the models while the lower diagonal shows the respective p-values to test if the difference is significant or not. Most p-values are below 0.05 except for some cases. We see that there is no significance difference between CART and C5.0 (p-value=1.000), GLM and SVMLinear (p-value=1.000), GLM and SVMRadial (p-value=0.13355), Random forest and SVMRadial (p-value=0.05982), and SVMLinear and SVMRadial (p-value=0.09551). Since the random forest has the best accuracy followed by SVM with radial kernel we can choose the SVM model due to its simplicity and computational efficiency. Also, since there is no significant difference between SVM with radial kernel and SVM with a linear kernel, we choose SVM with a linear kernel for simplicity.

We again train the SVM with the linear kernel by tuning the classifier with different values of costs. Results are provided in the below summary table and figure.

Table 4. Resampling results across turning parameters for SVM with linear kernel

Cost	Accuracy	Kappa
0.00	NaN	NaN
0.01	0.8271	0.3691
0.05	0.8288	0.3840
0.10	0.8295	0.3940
0.25	0.8293	0.3951
0.50	0.8293	0.3958
0.75	0.8292	0.3952
1.0	0.8289	0.3940
1.25	0.8294	0.3950
1.50	0.8291	0.3950
1.75	0.8288	0.3954
2.00	0.8299	0.4000
5.00	0.7526	0.2432

DOI: 10.5824/ajite.2020.01.001.x

From the above table, we see that SVM with a linear kernel attains the highest accuracy of (83%) when the cost value is C=2.

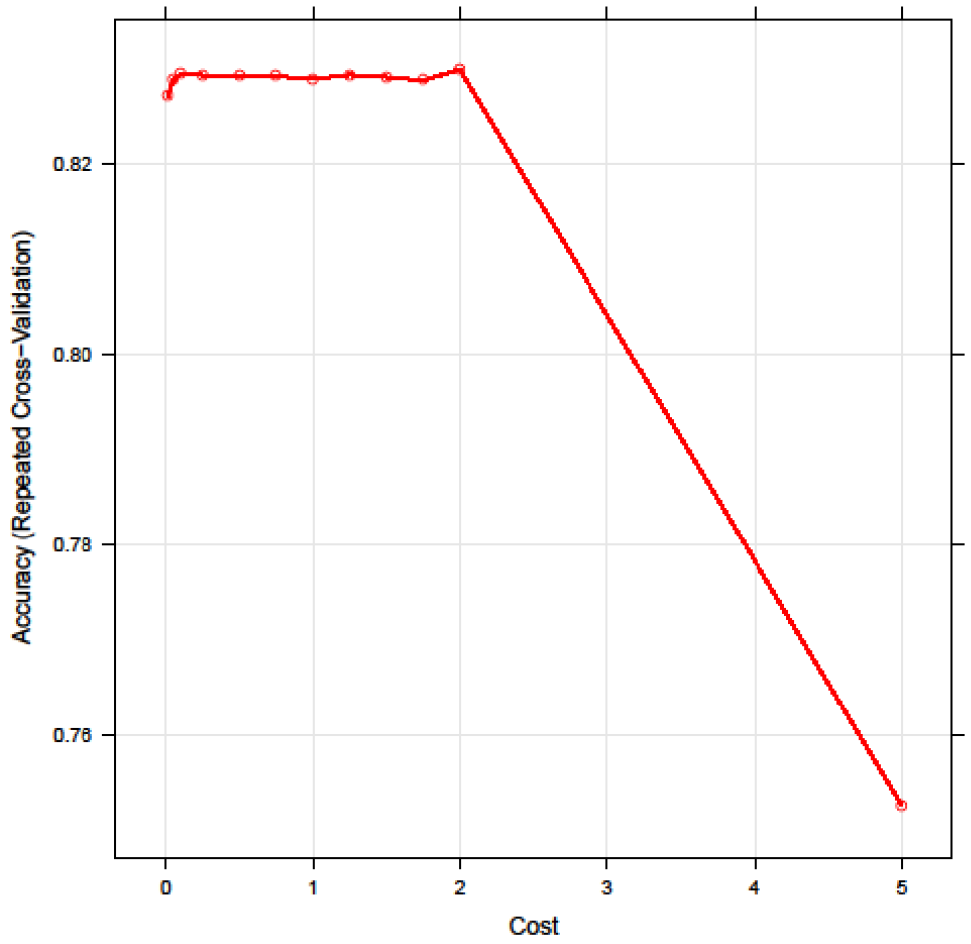

Figure 6. Resampling results across tuning parameters for SVM with linear kernel

Like table 4 above, figure 6 displays the accuracy of the SVM model with radial kernel across different cost values. The highest accuracy (83%) is attained when the cost value is C=2.

Therefore, results from machine learning models show the benefit of applying unstructured data (user reviews) to predict overall drug performance. Although we utilized only a sample of a dataset due to scalability issues, through Latent Semantic Analysis (LSA) and TF-IDF transformation we were able to train machine learning models with reasonable accuracies. Besides, the random forest has achieved the best accuracy of 84% to predict new drugs as either low-rated or highly-rated based on the reviews provided by users nevertheless SVM with the

linear kernel which attained maximum accuracy of 83% has been selected due to its simplicity and computational efficiency.

CONCLUSION

From the above discussion, supervised machine learning models provide a great remedy in predicting overall drug performance using unstructured textual data instead of completely relying on sentiment scores. Using only a small portion of the dataset, we managed to attain reasonable accuracy in our models by applying TF-IDF transformation to penalizes most frequent terms and Latent Semantic Analysis (LSA) technique to select few powerful predictive features. Further, the classification model by random forest appeared to be superior compared to all models considered in this study with an accuracy of 84% yet the SVM model with linear kernel was selected due to its simplicity and computational efficiency. Finally, we propose a future similar study to compare various features selection techniques such as Latent Semantic Analysis (LSA), Principle Components Analysis (PCA), Partial Least Square (PLS), Chi-Square method, Information Gain Ratio technique, and other methods found in the literature to analyze texts from the medical field domain using supervised machine learning approach.

REFERENCES

Bhargava, Apurva, (2019). *Grouping of Medicinal Drugs Used for Similar Symptoms by Mining Clusters from Drug Benefits Reviews*. Available at SSRN: https://ssrn.com or http://dx.doi.org/10.2139/ssrn.3356314

Denecke, K., Deng, Y, (2015). Sentiment analysis in medical settings: new opportunities and challenges. *Artif. Intell. Med.* ***64****(1),* 17–27.

Gräβer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. *Proceedings of the 2018 International Conference on Digital Health* - DH '18. doi:10.1145/3194658.3194677 https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

IBM Corporation, (2013). *Data-driven healthcare organizations use big data analytics for big gains. Somers,* NY: IBM Corporation.

Jimene-Zafra, S.M., Martín-Valdivia, M.T, Urena-Lopez, L.A., (2019). How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for the medical domain. *Artificial Intelligence in Medicine 93,* 50–57. doi: 10.1016/j.artmed.2018.03.007

Kerstin Denecke, (2015). Sentiment Analysis from Medical Texts. *Springer International Publishing*, Cham, 83–98. https://doi.org/10.1007/978-3-319-20582-3_10

Kho S.J., Padhee S., Bajaj G., Thirunarayan K., Sheth A. (2019). Domain-Specific Use Cases for Knowledge-Enabled Social Media Analysis. In: Agarwal N., Dokoohaki N., Tokdemir S. (eds) *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. Lecture Notes in Social Networks. Springer,* Cham.

Korkontzelos, Ioannis & Nikfarjam, Azadeh & Shardlow, Matthew & Sarker, Abeed & Ananiadou, Sophia & Gonzalez, Graciela. (2016). *Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of Biomedical Informatics*. 62. 10.1016/j.jbi.2016.06.007

Landauer, T. K., Foltz, P. W., & Laham, D. (1998*). Introduction to Latent Semantic Analysis. Discourse Processes*, **25**, 259-284

Liu, C., Sheng, Y., Wei, Z., & Yang, Y.-Q. (2018). Research of Text Classification Based on Improved TF-IDF Algorithm. *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE).* doi:10.1109/irce.2018.8492945

Liu B., Zhang L. (2012) *A Survey of Opinion Mining and Sentiment Analysis.* In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA.

Luo C., Zhan J., Xue X., Wang L., Ren R., Yang Q. (2018). *Cosine Normalization: Using Cosine Similarity Instead of Dot Product in Neural Networks*. In: Kůrková V., Manolopoulos Y., Hammer B., Iliadis L., Maglogiannis I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018. Lecture Notes in Computer Science, vol 11139. Springer, Cham.

M. Pirmohamed, S. James, Meakin, C. Green, A.K Scott, T.J Walley, K. Farrar, B.K. Park, A.M. Breckenridge, (2004). Adverse drug reactions as a cause of admission to hospital: prospective analysis of 18820 patients. *BMJ*, 329(7456)15-19. Doi:10.1136/bmj.329.7456.15

Saif Mohammad and Peter Turney, 2010. *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, LA, California.

T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, (2017). *Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review*, IEEE Access, vol. 5, pp. 16173-16192. doi: 10.1109/ACCESS.2017.2690342