2023 International Conference on Smart Applications, Communications and Networking (SmartNets) | 979-8-3503-0252-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/SmartNets58706.2023.10215681

# A Transfer Learning Application on the Reliability of Psychological Drugs' Comments

Tarık Üveys Şen
*Department of Computer Engineer*
*Abdullah Gul University*
Kayseri, Turkey
Email: tarikuveys.sen@agu.edu.tr

Gokhan Bakal
*Department of Computer Engineer*
*Abdullah Gul University*
Kayseri, Turkey
Email: gokhan.bakal@agu.edu.tr

***Abstract*—As digitalization and the Internet stay emerging concepts by gaining popularity, the accuracy of personal reviews/opinions will be a critical issue. This circumstance also particularly applies to patients taking psychological drugs, where accurate information is crucial for other patients and medical professionals. In this study, we analyze drug reviews from drugs.com to determine the effectiveness of reviews for psychological drugs. Our dataset includes over 200,000 drug reviews, which we labeled as positive, negative, or neutral according to their rating scores. We apply machine learning (ML) models, including Logistic Regression, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) algorithms, to predict the sentiment class of each review. Our results demonstrate an F1-Weighted score of 85.3% for the LSTM model. However, by applying the transfer learning technique, we further improved the F1 score (nearly 3% increase) obtained by the LSTM model. Our findings proved that there is no contextual difference between the comments made by the patients suffering from psychological or other diseases.**

***Index Terms*—Machine Learning, Deep Learning, Transfer Learning, Natural Language Processing**

## I. Introduction

In the digital era, online shopping over web retailer stores has become ubiquitous in daily activities. With the convenience of the internet, people can easily purchase products from anywhere in the world with just a few clicks needed. However, one of the biggest challenges of online shopping is determining the quality of a product based on the reviews provided by other customers. While the internet provides us with a platform to share our opinions and experiences, the accuracy of these reviews is often questionable and needs to be validated [1]. This situation is also applied to the medical domain, specifically psychological drug reviews/comments, where accurate information is even more critical for patients and medical professionals.

In this study, we sought to address this issue by applying transfer learning to the experiments analyzing drug reviews. By leveraging the power of machine learning, we aimed to determine the accuracy of reviews for psychological drugs and improve our understanding of their effectiveness. Our dataset consisted of over 200,000 drug reviews from drugs.com, which included information on the drug name, condition, rating, and useful count given by other users. During the experimental studies, we employed various machine learning models, including Logistic Regression, RNN, and LSTM algorithms. To utilize supervised learning methodologies, we automatically labeled the review entries as positive, negative, or neutral classes by the rating scores and trained our models to predict the sentiment class of each review. Our results showed that we achieved an F1-Weighted score of 85.3% for the LSTM model, while we further improved the F1 score by the LSTM model boosted by applying transfer learning.

The following sections will discuss the background & related work, the dataset used in the experiments, the methodology applied, and finally, the results of the experiments. Through this research, we aim to shed light on the accuracy of drug reviews. Besides, as a principal motivation, we wanted to prove that ***psychological drug reviews are reliable enough to be employed in research studies*** and should not be differentiated from other drug reviews of non-psychological disorders. While patient reviews may not provide infallible information, they can still serve as valuable indicators for research purposes. Thus, we utilized the patients' reviews as the primary data elements.

## II. Background and Related Work

In this section, we describe the technical details of the model architectures used in this study. Besides, we also mentioned the key related works in the corresponding field.

### A. Logistic Regression Classifier

Logistic regression is a linear model that uses a sigmoid function to map the input features to a probability of belonging to a particular class [2]. The model is trained by minimizing the negative log-likelihood of the data, and the coefficients learned during training are used to make predictions on new/unseen instances. Even though it is commonly employed for binary classification problems, it is also possible to build a multi-class classification in a one-vs-rest manner. It is a simple yet powerful algorithm that is widely used in many fields, such as medicine, finance, and social sciences [3].

979-8-3503-0252-3/23/$31.00 ©2023 IEEE

### B. RNN Architecture

Recurrent Neural Networks (RNNs) are a subtype of neural network architecture that is widely used in various fields involving sequential data elements, such as natural language processing and speech recognition [4] Considering the technical aspect, RNNs are designed to handle sequential data by processing one input at a time while maintaining an internal state that captures the context of previous input data. Nevertheless, despite their popularity, we admit that RNN models have critical problems that can affect their performance and usability compared to other modern models.

Here, the major problem is capturing long-term dependencies among the input data. Since the RNN's internal state is updated based on the previous input and current input, the information from the prior inputs can quickly decay as the sequence grows longer. This problem, known as the vanishing gradient problem, can make it difficult for RNNs to capture long-term dependencies in sequential data.

Another limitation of RNNs is their susceptibility to overfitting. Because RNNs are capable of modeling complex patterns in sequential data, they can easily memorize the training data, leading to a poor generalization of unseen data. This problem can be mitigated by using regularization techniques such as dropout and weight decay. Finally, memory capacity is also a significant limitation for both traditional RNNs and LSTM networks due to the limited number of memory blocks in the network. Simply increasing network size may not be a feasible solution, as modularization of the network topology is necessary to maintain an effective learning process. However, the exact methods for implementing modularization and how network modules should be interconnected are not yet fully understood [5].

### C. LSTM Architecture

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that has gained popularity recently due to its ability to handle long-term dependencies in sequential data [6]. LSTM networks are practical and powerful for language modeling, speech recognition, and text classification tasks. However, despite their success, LSTM networks have a few limitations that can affect their performance and usability.

One of the main limitations of LSTM networks is their higher computational cost and memory requirements compared to traditional RNNs. This is due in part to their increased complexity, including the use of memory cells and gates that allow for greater memory capacity but also require more computational resources to train and deploy. As a result, designing and optimizing LSTM networks often involves a trade-off between memory capacity and computational cost, with larger networks that can store more information over time typically requiring more computational resources to operate effectively. Consequently, they may not be suitable for some applications, particularly those with limited computing resources or real-time constraints. Another limitation is the vulnerability to over-fitting. Since the LSTM networks are capable of modeling complex patterns in sequential data, they can easily tend to memorize the training data, leading to poor generalization performance on unseen data. This problem can be reduced by regularization techniques such as dropout and weight decay approaches.Lastly, the LSTM networks can also suffer from vanishing and exploding gradients, although they are designed to suppress these issues occurring in plain RNN architectures [7]–[9].

## III. Dataset Curation and Stats

When analyzing textual data, the quality and relevance of the dataset play a critical role in terms of the accuracy and interpretability of the results. In this section, we describe the curation and key statistics of a dataset obtained from drugs.com. The original dataset comprises 215,063 records; however, 1,194 records were removed since they did not contain a "condition" column, leaving 213,869 records. Using the remaining records, we divided the dataset into three groups based on the "**rating**" column, yielding 141,560 records as annotated positive-class, 63,906 instances as annotated negative-class, and 8,403 examples as labeled neutral-class, respectively. The graphical representation of the data distribution is shown in Figure 1.

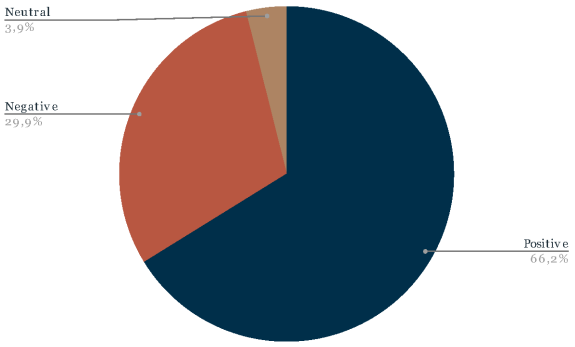


Fig. 1: Class distribution of data instances.

By analyzing this dataset, we may identify patterns and trends in drug efficacy, safety, and potential factors influencing patient perceptions and experiences. Our findings showing the meaningfulness of the drug comments may have important implications for healthcare professionals and policy-makers to optimize treatment plans and improve patient outcomes.

To ensure the accuracy and relevance of our findings, we divided our dataset into two subsets: the first set (named psychological reviews set) containing only reviews of medications used to treat psychological conditions and the second set (called all-condition reviews set) containing reviews of all other drugs. The psychological review set comprises 42,219 data instances, which is a decent amount of data for our analysis. By focusing on this subset, we aim to gain deeper insights into the experiences and attitudes of patients seeking treatment for mental health conditions. Besides, we also target to identify any patterns or trends in which they express their opinions and emotions about these medications.

## IV. Methodology

In this section, we explain the methodological details, including data preprocessing, machine learning & deep learning model building, and proposed transfer learning model to reach the goal we mentioned in Section I.

### A. Textual Feature Extraction

In this study, we have not performed any text cleaning or n-gram separation. Instead, we converted the "review" column into an array using the `Tokenizer` function from the Keras library [10]. We then determined the mean, maximum, and minimum length values, which were 89, 1992, and 0, respectively. Based on the length values, we set the max length of the padding sequence to 1000. Also, we used the `TFIDFvectorizer` function from the sklearn library [11] for the traditional machine learning model built by the logistic regression algorithm.

### B. Data Processing & Partitioning

To label the raw review records in the dataset, we utilized the "rating" column to generate the class distribution. We set a numerical limit for the rating column and chose six as the limit. We classified the reviews where the rating value is less than the limit as negative, the reviews having a rating value higher than the limit as positive, and the records containing a rating value equal to the limit as neutral. We employed lexicon-based models using the `Vader` tool from the NLTK library and the `Textblob` approach to validate the classification results obtained from our models. After cleaning our dataset, we only kept the classes and the review column. We used the label-encoder package from the Sklearn library to convert the class distributions to multi-class digital data.

Finally, the dataset was split into 70% as training, 20% as testing, and 10% as validation sets.

### C. Experimental Model Configurations

In this section, for the sentiment analysis experiments, we explain both the traditional machine learning models built by the logistic regression algorithm and deep learning models using RNN and LSTM architecture with and without transfer learning technique generated through the drug reviews overall the diseases.

We intensively exploited the sklearn machine learning library to build the logistic regression model. Since there was almost a negligible improvement in the performances and we wanted them aligned with the deep learning models, we did not perform any fine-tuning on the parameters and used the default parameter values.

*1) Recurrent Neural Network (RNN):* To build the deep-learning models, including RNN and LSTM algorithms, we utilized the Keras deep neural network library. In the RNN model, we set the input shape to 1000, as the input array is padded with the pad sequence arrangement. We then created a bidirectional RNN layer through the embedding layer containing 16 neurons, followed by the dropout layer. We called this model the vanilla-RNN model.

Afterward, we created the same Bidirectional simple RNN layer again and added the final dropout layer, followed by the dense layer used as the output layer. In the network, "*softmax*" is used as the activation function, and the "*adam*" algorithm is employed as the optimizer, while the `sparse-categorical-cross-entropy` is used as the loss function. We set 64 as the batch size and trained the model for 25 epochs with a dropout rate of 0.5.

*2) Long Short-Term Memory (LSTM):* As selected in the vanilla-RNN model, the input shape was also set to 1000, with the same padding-sequence approach. Then, we created a Bidirectional LSTM layer through the embedding layer. After adding the dropout layer for regularization purposes, we added another Bidirectional LSTM layer with eight neurons, followed by the final dropout layer, and a dense layer used as the output layer. Similarly, "*softmax*" is utilized as the model's activation function, the "*adam*" optimization algorithm is employed as the optimizer, and the `sparse-categorical-cross-entropy` is used as the loss function of the LSTM network model. During the training, we set the batch size as 64 and trained the whole network for 25 epochs with a dropout value of 0.5.

*3) Transfer Learning Architecture:* Transfer learning is a powerful machine learning technique that involves leveraging a pre-trained model to solve a related task [12], [13]. Thus, in our case, we deliberately exploited transfer learning to improve the accuracy of our sentiment analysis model [10]. Specifically, we used an in-house pre-trained LSTM model trained on a large dataset of drug reviews. We first loaded the pre-trained LSTM model using the Keras library to implement transfer learning. Then, we froze all the layers in the model to prevent them from being retrained. Next, we added a Bidirectional LSTM layer with 256 neural units to the model and connected it to the output of the second-to-last layer in the pre-trained model. Afterward, we integrated a dropout layer to reduce over-fitting and added another Bidirectional LSTM layer with 256 neurons with another following dropout layer. Finally, we added a dense layer with 16 neurons with a relu activation function for the nonlinearity concerns, and then the model network was finalized by the output layer. The model structure is illustrated in Figure 2.

The transfer learning model employed the "*softmax*" as the activation function, the "*adam*" algorithm as the optimizer, and `sparse-categorical-cross-entropy` as the loss function. We set the batch size to 64 and trained the model for 30 epochs with a dropout rate of 0.1 to reduce the over-fitting problem. Generally, transfer learning allowed us to achieve higher accuracy than traditional machine learning models. This outcome is because the pre-trained LSTM model already learned crucial features from a large dataset, which we could leverage for our sentiment analysis task.

## V. Results and Discussion

To demonstrate the performance of the proposed transfer learning method, we evaluated the models using the f1-score as the primary performance metric. In addition, we

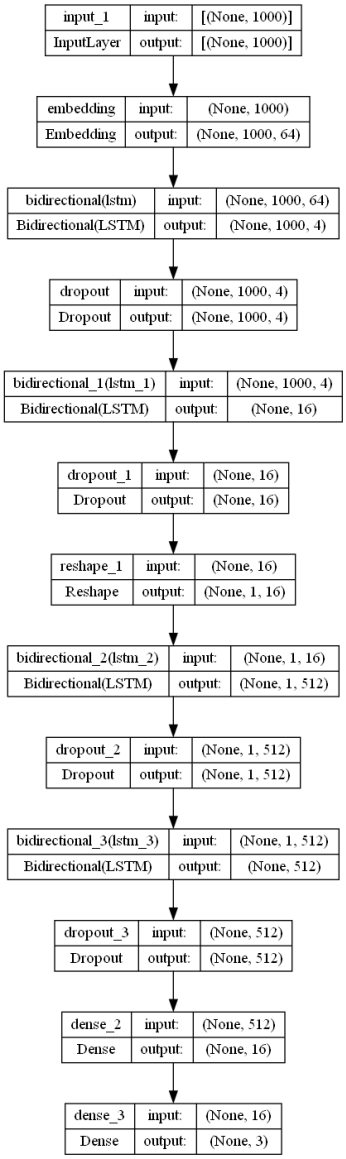


Fig. 2: The structure of the applied transfer learning model.

also reported other metrics, including precision and recall, to gain better evaluation insight. As can be obtained from Table I, the traditional machine learning model, the logistic regression, achieved f1 scores of 80%. When considering the deep learning models, the RNN and LSTM models yielded an F1 score of 85%. Here, the main possible reason for having better performances with the DL models is that both DL models could learn the contextual associations encoded in the sentences for each target label more than the n-gram feature space utilized in the logistic regression model.

As another evaluation indicator, the generation of the confusion matrices representing the numbers of actual and predicted instances for each class is also used in classification studies. Therefore, we presented the confusion matrix of the LSTM model with a color heatmap illustration using the green color as the core intensity level in Figure 3.

| | Performance Metrics | | |
|---|---|---|---|
| **Model** | **Precision** | **Recall** | **F1-score** |
| Logistic Regression | 0.8 | 0.82 | 0.8 |
| Vanilla RNN | 0.86 | 0.84 | **0.85** |
| LSTM | 0.86 | 0.84 | **0.85** |

TABLE I: Performance scores of traditional machine learning and deep learning models

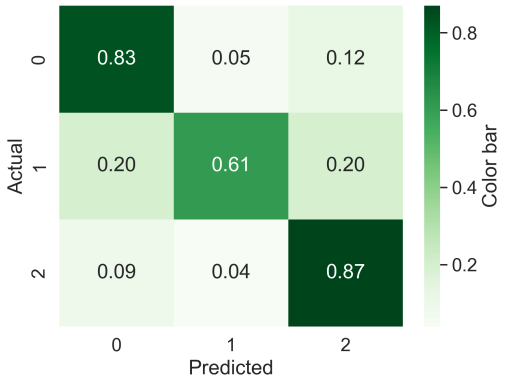


Fig. 3: The confusion matrix of the LSTM model.
**Note:** 0: Negative, 1: Neutral, 2: Positive

Considering the confusion heatmap matrix of the LSTM model, the set of positive test instances was classified better than other sentiment examples. As expected, the worst classification performance happened when separating the neutral examples from positive and negative samples. This outcome is predictable because it is relatively more challenging to identify a neutral state encoded in sentences than opposite polarities due to the natural language complexity.

To validate our dataset labeling based on the rate thresholding approach, we have built TextBlob [14] and NLTK-Vader [15] lexicon-based sentiment analysis models over the unlabeled raw instances for labeling the comments. The TextBlob model partitioned the dataset as 133,745 positives, 70,872 negatives, and 9,252 neutral comments, while the NLTK-Vader model classified 76,893 comments as positive, 70,376 examples as negative, and the rest of 66,600 instances as neutral. Following, we executed the same LR classification model using the dataset labeled by the lexicon-based approaches to justify that our labeling assumption is well enough for further experiments. Consequently, we obtained F1 scores of 73.8% and 91.1% from NLTK-Vader and TextBlob, respectively. As can be entailed from the results, the F1 scores are aligned with our original model's F1 score, and this outcome indicates that our thresholding approach is reasonable.

As a primary objective of this study, we built a transfer learning model using the comments made by drug takers for all diseases. By utilizing the transfer learning model, we wanted to prove that there are no contextual differences between the feedback comments generated by psychological drug takers

and other medical drug takers. Besides that, we targeted to achieve better sentiment analysis results through the transfer learning model. For this particular purpose, we trained a model using all disease comments and learned the appropriate weights in the neural units. Afterward, we transferred this model to classify the psychological disease comments based on the predefined sentiment classes. In Table II, we demonstrate the performance scores obtained by the transfer learning model.

| | Performance Metrics | | |
|---|---|---|---|
| **Transfer Learning Model** | **Precision** | **Recall** | **F1-score** |
| Macro average | 0.74 | 0.79 | 0.76 |
| Micro average | 0.88 | 0.88 | 0.88 |
| Weighted average | 0.88 | 0.88 | 0.88 |

TABLE II: Performance scores of the proposed transfer learning model

The first noticeable outcome from the transfer learning results is that the basic arithmetic average scores (by Macro averages) of the performance metrics are even lower than the lowest-performing models built without transfer learning. The potential reason for this case is that all classes were considered equally contributed. The other interesting point is that the weighted and micro averages of the metrics are almost 3% higher than the best-performing model without using the transfer learning approach. This outcome directly proved that transferring knowledge of all drug-disease experience comments to classify psychological experience comments helps to improve the classification performances. The most remarkable result is we can entail that the psychological drug-disease experience comments should not be treated as specific instances due to the nature of the psychological diseases themselves.

As can be seen in Figure 4, the rate of the correctly predicted test instances from the psychological drug-disease comments was also improved in general compared to the confusion matrix generated over the regular LSTM model without transferring knowledge from all other diseases.

The original project concept was conceived in response to insights provided by a psychology expert. The expert's viewpoint suggested that comments made by psychology patients regarding their experiences with drugs and disease may not be as useful as those made by patients with other medical conditions. This is because patients suffering from psychological diseases may not convey reliable experience information through their comments. However, as a core motivation of this study, we wanted to show that this opinion could be disproved by revealing at least equal or better classification performances when extra knowledge of the other diseases is transferred into the target model classifying the psychological test comments.

## VI. Conclusion

In this project, we developed several models for sentiment analysis using the dataset containing drug reviews made by multiple patients. By considering distinct approaches, we built logistic regression and deep learning models, namely RNN and LSTM. Besides, we employed the transfer learning technique to achieve better classification performances for the target psychology patients' comments. To do that, we leveraged a pre-trained LSTM model transferring knowledge from other feedback comments made by patients having other diseases than psychological ones.

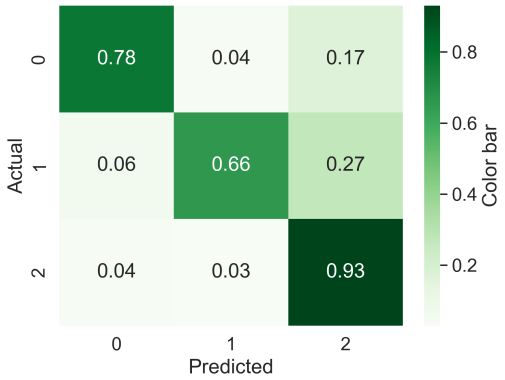


Fig. 4: Confusion Matrix of LSTM Transfer Learning model. **Note:** 0: Negative, 1: Neutral, 2: Positive

As can be seen in the experimental results, we showed that transfer learning was a powerful approach, which achieved a weighted-F1 score of 88% (yielding 3% improvement) when classifying the feedback comments about the psychological drug-disease pairs. By obtaining better performances, we disproved the expert's opinion regarding the reliability of the feedback comments made by patients suffering from psychological diseases. This outcome also highlights the potential of transfer learning for improving the performance of machine learning models, particularly in cases where large amounts of labeled data do not exist. Overall, our work demonstrates the importance of selecting appropriate models and techniques for sentiment analysis and highlights the benefits of using deep learning and transfer learning for a particular task involving the target set of medical feedback comments.

## Acknowledgment

We are thankful to Google Cloud Services for providing us with academic credit support to do this work. Plus, this study is partially supported by TUBITAK 3501 Career Development Program through grant 122E103.

## References

[1] G. Bakal and O. Abar, "On comparative classification of relevant covid-19 tweets," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2021, pp. 287–291.

[2] L. Regression, "A self-learning text," *Statistics for Biology and Health, Third Edition, David Kleinbaum, Mitchel Klein*, 1994.

[3] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[4] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[5] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.
[6] L. Yao and Y. Guan, "An improved lstm structure for natural language processing," in *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*. IEEE, 2018, pp. 565–569.
[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
[9] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. Pmlr, 2013, pp. 1310–1318.
[10] F. Chollet *et al.*, "Keras," https://keras.io, 2015.
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[12] D. R. Neog, P. Banerjee, and R. Jain, "Transfer learning in deep neural networks," 2020.
[13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
[14] S. Loria *et al.*, "textblob documentation," *Release 0.15*, vol. 2, no. 8, 2018.
[15] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.