Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models

Akhil Shiju
Florida State University
Tallahassee, Florida, USA
aks19m@my.fsu.edu

Zhe He
School of Information
Florida State University
Tallahassee, Florida, USA
Zhe.He@cci.fsu.edu

ABSTRACT

Social web contains a large amount of information with user sentiment and opinions across different fields. For example, drugs.com provides users' textual review and numeric ratings of drugs. However, text reviews may not always be consistent with the numeric ratings. In this project, we built different classification models to classify user ratings of drugs with their textual review. Multiple supervised machine learning models including Random Forest and Naive Bayesian classifiers were built with drug reviews using TF-IDF features as input. Also, transformer-based neural network models including BERT, BioBERT, RoBERTa, XLNet, ELECTRA, and ALBERT were built for classification using the raw text as input. Overall, BioBERT model outperformed the other models with the overall accuracy of 87%. This research demonstrated that transformer-based classification models can be used to classify drug reviews and identify reviews that are inconsistent with the ratings.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence; Machine learning; • Applied computing → Document management and text processing.

KEYWORDS

Drug reviews, Transformer models, Text classification

ACM Reference format:

Akhil Shiju and Zhe He. 2021. Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models. XXXX

1 Introduction

The evaluation of the efficacy and safety of drugs heavily relies on the randomized controlled trials with rigorous inclusion and exclusion criteria [1]. However, such processes are limited to a small number of individuals enrolled in the study and are often constrained to participants in the target population that meet possibly restrictive eligibility criteria, limiting the population representativeness and subsequent study generalizability [2] [3]. The ramifications of these acclimations could potentially have resulted in the overestimate of the efficacy of the product and misidentification of adverse events/side effects in the diverse population [4]. To counter such issues, approaches such as post-marketing drug surveillance have been introduced to optimize the safety of the drug after its regulatory approval and mass production [5].

There are two major forms of post-marketing drug surveillance. Some are formed by government regulators such as the Vaccine Adverse Event Reporting System (VAERS) by the United States Food and Drug Administration [6] or the Yellow Card Scheme by the United Kingdom Medicines and Healthcare Products Regulatory Agency [7]. Also, public/private organizations have a system to monitor drug side-effects such as the Research on Adverse Drug events And Reports and various websites [8].

Prior methodologies consisted of the use of statistical measures to determine adverse events such as side effects. The typical methodology involved looking at query logs or other categorical variables for the identification of these events but did not focus specifically on the sentiment of the user through actual reviews [9].

The application of post-market drug surveillance has been successfully applied in the identification of adverse events through safety reports by the introduction of deep learning-based methods including the extraction of temporal events, the procedure performed, and social circumstance [10]. Adverse events have also been utilized for the identification of drug compounds, but also molecular drug compositions have also been used to predict adverse side effects [11].

In the era of Web 2.0, the Internet has opened up new unique pathways to obtain information about consumers' drug reviews through websites on drug reviews. These reviews have contained a plethora of information regarding individual experiences associated with the drugs. Such reviews have contained an extensive amount of user sentiment related to a particular condition, which could be leveraged to detect the side effects and efficacy of drugs [12].

However, many barriers exist in the extraction of sentiment from these online medical reviews. For instance, user reviews of drugs in such online forms are typically unconventional and most reviewers lack medical knowledge which may pose barriers for meaningful extraction of information. In addition, many review websites have some form of numerical rating that has served the role of quantifying such a sentiment. Nevertheless, such

applications may have introduced biases as individual users may have different understandings as to what a high score means versus what would have constituted a low score. Users have tended to reduce the effort required in reporting values by rating all qualities as highly important, thus resulting in overtly positive ratings [13]. This could lead to an unintended positive view of the overrated drugs by the general public, albeit less effective for certain population subgroups. Prior research [14] has found that web-based reviews have the potential to be viewed as an applicable source of information for analysis, but the direct reliance on consumer ratings could be biased by the consumer experience. For example, addictive drugs have been observed to be typically highly rated in comparison to other drugs which have treated the same condition, even if these additive drugs underperformed [14]. Thus, the ratings of those drugs may be skewed, thus a potential solution could be focusing on the relationship between the review and the rating, to identify differentiation between the skewed rating and what was typed in by the user [15].

Publicly available information on the Internet offers an easily attainable resource that could be potentially leveraged to gain a deep understanding of the drug reviews by the users. Entire user reviews are fully available on drug review websites, on which users can comment on their personal experiences of the drugs they have taken for a specific condition. Unlike many other forms of medical data, this information is not filtered through medical professionals. Since these reviews are given by anonymous users, there is no risk of patient health record violation for confidentiality.

The application of machine learning, especially through transformer-based language models pretrained with enormous amount of data, offers a unique approach to classify textual information through natural language processing [16]. In this project, we evaluated the feasibility of leveraging machine learning and natural language processing to classify user ratings based on their textual review to identify the locations of contingency. In addition, the model can be used to identify overtly positive or negative scores. Overtly positive or negative scores are user rating which was incorrectly classified by the model in order to identify some generalizations. We also used a computational technique to highlight phrases in the text that have positive or negative impact on the classification results.

2 Methods

2.1 Dataset Preparation

We obtained the dataset from the UCI Machine Learning Repository [17]. These instances were collected from Drugs.com using Beautiful Soup. As can be seen in Table 1, the dataset consists of user drug reviews, drug names, related medical conditions, and a 10-point rating. Figure 1 shows the distribution of reviews by ratings. The ratings were shown to be skewed to the left to suggest that most drugs received a relatively high score. Prior analysis of this dataset focused primarily on the sentiment analysis [1]. They did not focus on classification of reviews [18]. Neither was there an emphasis on the error analysis of the models.

In total, the dataset consists of 215,063 instances [19]. The dataset ratings had a mean of 7.00 with a standard deviation of 3.27. There are 836 classified medical conditions in the dataset.

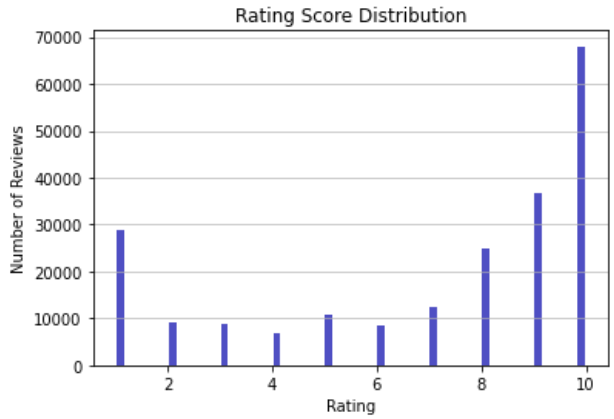

Figure 1: Total number of reviews in the dataset. The rating were integer values ranging from 1 to 10 with 10 being the highest possible rating.

Drug Name	Condition	Review	Rating
Chantix	Smoking Cessation	I smoked for 50+ years. Took it for one week and that was it. I didn't think it was possible for me to quit. It has been 6 years now. Great product.	10.0
Excedrin	Migraine	Does not work for people sensitive to caffeine. I was jittery and nervous and queasy after using a single dose.	2.0

Table 1: Two examples of a high review versus a low review along with the information including condition, drug name, and rating.

Since the primary focus of this study was to classify textual reviews, the data was broken down using a median of the ratings: ratings 8 or above were considered **above average** and below 8 were considered as **below average**. Binary classification was chosen over multiple classes since the other objective of this project was to identify overtly positive and overtly negative scores which would be identified by misclassification of the system versus the actual score. Thus, misclassification identification in its simplest sense would only be possible with a binary system. Instances in which the reviews contained more than 514 tokens were removed from the study due to the input size limit of the transformer-based language models.

2.2 Model Construction

2.2.1 Transformer-Based Models. The common methodology for transfer learning has been through the application of pre-training

on a large unannotated corpus that was capable of understanding the composition of the data type such as patterns in the language. This process could be considered as self-supervised learning. This pre-trained model is then followed by the fine-tuning process which focused on the training on an application-specific dataset.

BERT: Some common language models are pre-trained by predicting the next word in a sequence, but Bidirectional encoder representation from transformer (BERT) looked at bidirectional predicting context masked intermediate text tokens in the pretraining from Wikipedia and BookCorpus and next sentence prediction. Bert-base-uncased was used for this project[20].

In addition, BERT models such as clinical **BioBERT** have been pre-trained with a medical corpus from publicly available data from Pubmed and PMC [21]. The model which was used was from Huggingface labeled Bio_ClinicalBERT.

ALBERT: A Lite BERT (ALBERT) is a model which focused on being a less memory-heavy and faster version of BERT through the separation of the word embedding into two matrixes and by cross-layer parameter sharing[22]. Albert-base-v2 was used for this model.

RoBERTa: Robustly Optimized BERT Approach (RoBERTa) has been considered a pretraining model that eliminates the next sentence prediction task and adapts a novel approach of dynamic masking which randomized the masked token between training epochs[23]. RoBERTa out-performed BERT of multiple results such as GLUE, RACE, and SQuAD. Roberta-base was the model selected for this project.

XLNet: As a more computationally expensive model, the Generalized Auto-Regressive model (XLNet) implemented a system where all masked tokens were predicted, but in a randomized fashion [24]. Contrary, to BERT where only 15% of the masked token were predicted. XLNet-base-cased was the model selected for this project.

ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) replaced the masked language task with a generator and pre-trains the model to identify which token has been replaced[25]. The Electra-base-discriminator was used for this project.

2.2.2 Classification Model. We split the dataset into a training set (60%), a validation set (20%), and the test set (20%). These datasets were further classified into lists which were then converted into Transformer dataset that could be trained by a neural network to generate a model.

We constructed these transformer-based text classification models utilizing the Huggingface transformers using the Python k-train pipeline wrapper class for text classification. The models used for this project consisted of BioBERT, ELECTRA, RoBERTa, XLNet, ALBERT, and BERT. The parameter included a 514 max token length, a 5e^-5 learning rate, and a batch size of 6. The train test dataset was fed into the neural network trained to minimize validation data loss. After the training was completed, a confusion matrix of the test data was generated to determine F1 scores for the classes and the accuracy in comparison to the user ratings.

As a baseline measure for the comparison of these models with the transformer-based models, bag-of-words (BOW) models were constructed based on term frequency and inverse document frequency (TF-IDF). The textual reviews were converted into a bag of word representation. Afterward, a term frequency-inverse document frequency score matrix was computed for the bag of words representation. A Random Forest classifier and a Naïve-Bayes classifier were evaluated with the BOW features.

2.2.3 Classification Modeling Interpretability. After the best performing transformer model was selected, to provide some expandability for the model, Eli5 metrics were applied to the model. Eli5 has been used to understand why a certain classification through the identification of important features such as highlighting significant text features [26]. This is accomplished by inspecting the model parameters to discover the global implications. This was performed for many reviews to establish some sense of how the model performed these classifications. Top scores were also computed through the Eli5 metrics.

2.2.4 Condition Specific Classification Modeling The test data was stratified for the top 10 ten conditions based on the test data user reviews as seen in Table 2. The transformer models were then used to classify each of the different conditions to determine condition-specific f1 score and accuracy. The overall workflow is outlined in Figure 2.

Top 10 Conditions	**Instances in the Dataset**	**Categories Mean Specific Rating**	**Categories Median Specific Rating**
Birth Control	38345	6.08	7
Depression	12129	7.10	8
Pain	8241	7.62	9
Anxiety	7795	7.69	9
Acne	7411	7.37	9
Bipolar Disorder	5598	7.12	8
Insomnia	4895	6.69	8
Weight Loss	4852	8.03	9
Obesity	4748	7.73	9
ADHD	4497	7.31	8

Table 2: Condition-specific statistics. Birth Control was the most common condition which users reviewed followed by depression, pain, anxiety, and acne.

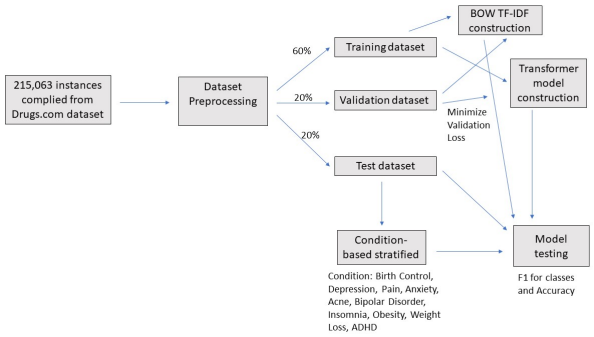

Figure 2. The workflow of the project.

3 Results

Overall, the model generated by the BioBERT and ELECTRA outperformed the other models on a variety of metrics as displayed in Table 3. The BOW models showed lower accuracy compared to the other constructions. XLNet had the longest training time compared to the other models.

Model	Above Average F1	Below Average F1	Accuracy
BERT	0.84	0.84	0.84
RoBERTa	0.83	0.83	0.84
XLNet	0.84	0.84	0.84
BioBERT	**0.87**	**0.87**	**0.87**
ELECTRA	0.85	0.87	0.86
ALBERT	0.75	0.81	0.78
Random Forest (BOW)	0.77	0.45	0.68
Naïve Bayes (BOW)	0.76	0.03	0.61

Table 3: Overall condition validation from the test dataset for the minimized loss for the top preforming models

Condition	Model	Above Average F1	Below Average F1	Accuracy
Birth Control	**ELECTRA**	**0.89**	**0.94**	**0.92**
	BioBERT	0.86	0.92	0.90
Depression	**ELECTRA**	**0.88**	**0.88**	**0.88**
	BioBERT	0.87	0.87	0.87
Pain	ELECTRA	0.83	0.81	0.82
	BioBERT	**0.85**	**0.83**	**0.84**
Anxiety	ELECTRA	**0.87**	**0.82**	**0.85**
	BioBERT	0.87	0.81	0.85
Acne	**ELECTRA**	**0.90**	**0.88**	**0.89**
	BioBERT	0.86	0.84	0.85
Bipolar Disorder	**ELECTRA**	**0.88**	**0.84**	**0.86**
	BioBERT	0.84	0.87	0.86
Insomnia	**ELECTRA**	**0.82**	**0.86**	**0.84**
	BioBERT	0.82	0.84	0.83
Weight Loss	ELECTRA	0.87	0.80	0.85
	BioBERT	**0.89**	**0.79**	**0.86**
Obesity	ELECTRA	0.82	0.77	0.80
	BioBERT	**0.85**	**0.79**	**0.83**
ADHD	ELECTRA	0.86	0.88	0.87
	BioBERT	**0.88**	**0.87**	**0.87**

Table 4: Condition-specific statistics for the top 10 conditions.

Classification result	Review	Probability	Green Indicates
Correctly Classified Above Average	"bactrim worked miracles when i first started taking it. i was on it for almost 5 years, but it began to lose its potency so i had to change. overall, it helped a lot in clearing my acne. i did experience less of an appetite, but other than that it was a positive experience." (Original rating: 8)	0.995	Above Average features
Correctly Classified Below Average	i have been on bydureon for just under a week. i have not been able to eat very much. i am nauseous every day and no very little energy. i've been considering going to see my doctor to see whats wrong with me but i see from these messages that it's probably from the injection. of course, i've lost a few pounds because i cannot eat anything. i hate how i feel! (Original rating: 6)	0.995	Below Average features
Incorrectly Classified Below Average	"i started taking saxenda the beginning of september and i have lost 28 lbs in a month... i am very pleased with this medicine being that it is very hard for me to lose weight due to my pcos. i exercise at least 3 to 4 times a week and I'm eating a little more healthy. the only thing about this drug is i have been experiencing eggy burping, nausea and diarrhea since i have been on the 3.0 dosage. hoping it will subside so on. I'm hoping to lose at least 60 more lbs and i am very confidant that saxenda will help me reach my goal! i went 0nline and signed up for the discount card to help with the cost I was paying 115.00 now i only pay 30." (Original rating: 10)	0.524	Below Average features
Incorrectly Classified Above Average	"this gel has completely changed my life. its been only 10 days that i am using it and i can see a huge improvement on my skin. scars have started fading but don't know why they still hurt when i touch them. i'm hoping for the best. i would suggest people who have acne like me that pls go and get it from your dermatologist. i use a homemade face pack on your skin after every 2 days to have a healthy peeling. i guess it works by peeling the surface of the acne prone area. best of luck guys this duac gel is a lifesaver and it's worth the money" (Original rating: 5)	0.714	Above Average features

Table 5: Example classification and interpretation generated by the Eli5 score metrics.

Table 4 shows the condition-specific classification for the top 10 conditions. The classification accuracy of instances of drugs used for pain, obesity, and insomnia are slightly lower than the other conditions.

The results generated in Table 5 are produced using the Eli5 metrics using the BioBERT trained model. There is a clear relationship

between the words highlighted and the classification that was made by the model. Terms highlighted in green supports the classification generated by the model, while terms generated in red opposes the predictions. Phrases related to side-effects were typically highlighted as below average features. An example for a incorrectly classified below average review revealed that terms such as burping, nausea, and diarrhea had a significant impact on the model classification for a below average review. Terms for different medications often had a strong impact on the classification. One factor to note was that "Bactrim", the drug, was highlighted as an above-average feature even though it is a drug. This suggested that the model considered this term as important which could be classified as an example of overfitting that occurred in this model. It also indicates that the overall experience of this drug is positive.

4 Discussion

We built multiple classification models through transformer-based architecture. These models were then to classify test data including condition-specific data. Afterward, the Eli5 toolkit was applied to help explain the models classification.

From these results, it was clear that the consumers' online drug reviews contain a vast quantity of information regarding the sentiment expressed by the user. Transformer-based models have the potential to serve as a methodology to discriminate between overtly positive and truly positive scores. Overall, this research outlined a potential process to identify consumer drug review bias. This was concurrent with other studies which have found that subjective effects are often distorted in rating systems [27].

On a comparison of the transformer models, the BioBERT and ELECTRA model outperformed the other models on a multitude of metrics when the same amount of information was present, which was consistent with other research papers that involved the application of transformer-based models [28]. Surprisingly, BioBERT outperformed many other models such as RoBERTa, XLNet, and BERT. One potential reason for this occurrence was the fact that BioBERT model was pretrained through medical texts which were topically related to the drug reviews [29]. Since these drug reviews had reference to different condition and medication, this reviled that pretraining with text discussing these ideas is beneficial.

Overtly positive and negative scores could be identified by this model through the comparison of the model predicted score versus the score which the user rating (More examples are available on https://github.com/akhilfsu/BCBACM) Overtly positive tend to suggest that the review given by the user does not reflect the rating that the user gave and conversely for overtly negative scores. Disparities between user reviews and rating may signify that there may be a knowledge gap for the score criteria in the users or might indicate bias. The low classification accuracy of the model for pain-specific ELECTRA model indicated that many pain medications are addictive [30] which may result in overtly high classification even though the review may indicate that there have been issues with the medication such as side-effects or other adverse events.

4.1 Future Work

The binary classification application of ELECTRA and other transformer models could serve the potential of finding user reviews that contain complaints about medication. Another area of focus could be expanding this model for multi-class identification as this may be much more beneficial in the identification of highly negative reviews. In addition, this dataset could be applied to identify adverse effects for medication based solely on online social media data.

4.2 Limitations

Although this model was able to successfully classify reviews in a binary system, the ability for large class identification is still unknown. One important issue with many transformer models was the issue of over-fitting. As seen in the misclassification there was a reliance on drug names in determining the classification. In addition, many transformer models such as XLNet are computationally expensive which may result in long training times.

5 Conclusion

This study covered the construction of a transformer-based model for the classification of drug reviews from drugs.com. The most successful model in this project was the BioBERT model with the high F1 scores. Overall, the transformer models outperformed the bag of words models. These binary transformer models tended to be effective at decerning highly optimistic reviews from reviews that contain a mixture of positive and negative feedback.

ACKNOWLEDGMENTS

This study was partially supported by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431; and in part by Florida State University-University of Florida Clinical and Translational Science Award funded by National Center for Advancing Translational Sciences under Award Number UL1TR001427. The first author would like to thank eHealth Lab at FSU and the Undergraduate Research Opportunity Program at Florida State University for the mentorship and guidance.

REFERENCES

[1] Robert M. Califf. 2012. Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007-2010. *JAMA* 307, 17 (May 2012), 1838. DOI:https://doi.org/10.1001/jama.2012.3424

[2] Kevin C. Farmer. 1999. Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. *Clinical Therapeutics* 21, 6 (June 1999), 1074–1090. DOI:https://doi.org/10.1016/S0149-2918(99)80026-5

[3] Zhe He, Xiang Tang, Xi Yang, Yi Guo, Thomas J. George, Neil Charness, Kelsa Bartley Quan Hem, William Hogan, and Jiang Bian. 2020. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. *Clin Transl Sci* 13, 4 (July 2020), 675–684. DOI:https://doi.org/10.1111/cts.12764

[4] Edward J Mills, Dugald Seely, Beth Rachlis, Lauren Griffith, Ping Wu, Kumanan Wilson, Peter Ellis, and James R Wright. 2006. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. *The Lancet Oncology* 7, 2 (February 2006), 141–148. DOI:https://doi.org/10.1016/S1470-2045(06)70576-9

[5] Ik Crombie. 1986. The role of record linkage in post-marketing drug surveillance. *British Journal of Clinical Pharmacology* 22, S1 (February 1986), 77S-82S. DOI:https://doi.org/10.1111/j.1365-2125.1986.tb02987.x

[6] Tom T. Shimabukuro, Michael Nguyen, David Martin, and Frank DeStefano. 2015. Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS). *Vaccine* 33, 36 (August 2015), 4398–4405. DOI:https://doi.org/10.1016/j.vaccine.2015.07.035

[7] B. O' Donovan, Ruth M. Rodgers, Anthony R. Cox, and Janet Krska. 2019. Making medicines safer: analysis of patient reports to the UK's Yellow Card Scheme. *Expert Opinion on Drug Safety* 18, 12 (December 2019), 1237–1243. DOI:https://doi.org/10.1080/14740338.2019.1669559

[8] Elad Yom-Tov and Evgeniy Gabrilovich. 2013. Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries. *J Med Internet Res* 15, 6 (June 2013), e124. DOI:https://doi.org/10.2196/jmir.2614

[9] Martin Kulldorff, Robert L. Davis, Margarette Kolczak†, Edwin Lewis, Tracy Lieu, and Richard Platt. 2011. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. *Sequential Analysis* 30, 1 (January 2011), 58–78. DOI:https://doi.org/10.1080/07474946.2011.539924

[10] Jingcheng Du, Yang Xiang, Madhuri Sankaranarayanapillai, Meng Zhang, Jingqi Wang, Yuqi Si, Huy Anh Pham, Hua Xu, Yong Chen, and Cui Tao. 2021. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. *Journal of the American Medical Informatics Association* (February 2021), ocab014. DOI:https://doi.org/10.1093/jamia/ocab014

[11] Sanjoy Dey, Heng Luo, Achille Fokoue, Jianying Hu, and Ping Zhang. 2018. Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinformatics* 19, S21 (December 2018), 476. DOI:https://doi.org/10.1186/s12859-018-2544-0

[12] Thu Dinh and Goutam Chakraborty. 2020. Detecting Side Effects and Evaluating the Effectiveness of Drugs from Customers' Online Reviews using Text Analytics, Sentiment Analysis, and Machine Learning Models. *sas-global-forum-proceedings* (2020), 1–23.

[13] Airo Hino and Ryosuke Imai. 2019. Ranking and Rating: Neglected Biases in Factor Analysis of Postmaterialist Values. *International Journal of Public Opinion Research* 31, 2 (June 2019), 368–381. DOI:https://doi.org/10.1093/ijpor/edy007

[14] Paula Tanabe and MaryBeth Buschmann. 1999. A prospective study of ED pain management practices and the patient's perspective. *Journal of Emergency Nursing* 25, 3 (June 1999), 171–177. DOI:https://doi.org/10.1016/S0099-1767(99)70200-X

[15] Swarnaseetha Adusumalli, HueyTyng Lee, Qiangze Hoi, Si-Lin Koo, Iain Beehuat Tan, and Pauline Crystal Ng. 2015. Assessment of Web-Based Consumer Reviews as a Resource for Drug Performance. *J Med Internet Res* 17, 8 (August 2015), e211. DOI:https://doi.org/10.2196/jmir.4396

[16] David D. Lewis. 1996. Challenges in machine learning for text classification. In *Proceedings of the ninth annual conference on Computational learning theory - COLT '96*, ACM Press, Desenzano del Garda, Italy, 1-ff. DOI:https://doi.org/10.1145/238061.238062

[17] UCI. Drug Review Dataset (Drugs.com) Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

[18] Sairamvinay Vijayaraghavan and Debraj Basu. Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms. Retrieved from https://arxiv.org/abs/2003.11643

[19] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In *Proceedings of the 2018 International Conference on Digital Health*, ACM, Lyon France, 121–125. DOI:https://doi.org/10.1145/3194658.3194677

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). Retrieved March 5, 2021 from http://arxiv.org/abs/1810.04805

[21] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (September 2019), btz682. DOI:https://doi.org/10.1093/bioinformatics/btz682

[22] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]* (February 2020). Retrieved February 28, 2021 from http://arxiv.org/abs/1909.11942

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]* (July 2019). Retrieved March 5, 2021 from http://arxiv.org/abs/1907.11692

[24] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]* (January 2020). Retrieved February 28, 2021 from http://arxiv.org/abs/1906.08237

[25] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]* (March 2020). Retrieved February 28, 2021 from http://arxiv.org/abs/2003.10555

[26] Namita Agarwal and Saikat Das. 2020. Interpretable Machine Learning Tools: A Survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Canberra, ACT, Australia, 1528–1534. DOI:https://doi.org/10.1109/SSCI47803.2020.9308260

[27] M. Abou Taam, C. Rossard, L. Cantaloube, N. Bouscaren, G. Roche, L. Pochard, F. Montastruc, A. Herxheimer, J. L. Montastruc, and H. Bagheri. 2014. Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator ®) withdrawal in France. *J Clin Pharm Ther* 39, 1 (February 2014), 53–55. DOI:https://doi.org/10.1111/jcpt.12103

[28] Laiba Mehnaz. 2020. Automatic Classification of Tweets Mentioning a Medication Using Pre-trained Sentence Encoders. (2020).

[29] Keval Pipalia, Rahul Bhadja, and Madhu Shukla. 2020. Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, IEEE, Moradabad, India, 411–415. DOI:https://doi.org/10.1109/SMART50582.2020.9337081

[30] Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating Language Model Finetuning Techniques for Low-resource Languages. *arXiv:1907.00409 [cs]* (2019). DOI:https://doi.org/10.13140/RG.2.2.23028.40322