5/2/2022

Sentiment Analysis in Drug Reviews using Deep learning, and Supervised Learning

APPLIED ARTIFICIAL INTELLIGENCE
UNIVERSITY OF HULL

Contents

Introduction

(Kaplan, 2019) defines artificial intelligence as a system's ability to translate outside data precisely, to acquire from such data, and use those learnings to achieve express goals and tasks through adaptable variation.

The foundation of this subject has been recorded by several authors and researchers over the years to be between the years the 1940s and 1950s. (Anyoha, 2017) documents that AI started in the early 1950s when scientists started coming up with theories of humans using the data available to them to make informed decisions. This took root when science fiction wrote about engineers Gregory Powell and Mike Donavan work on robots (Anyoha, 2017). The name artificial intelligence was formally invented in 1956 (Lewis, 2014).

Subjects such as image recognition, object identification, text classification, sentiment analysis, and speech recognition, today would not have been possible without acceptance and advancement in artificial intelligence over decades.

This project is a sentiment analysis problem, to classify customers' drug reviews. The motivation behind this paper is to foster a successful strategy for opinion examination of medication surveys on wellbeing data administration sites. Customers have a lot to share about their experience and rightly so. Good or bad, the problem is defining this 'warehouse of words' on a general sense of positive or negative experience.

There are two related works on drug review that will be discussed in this study. The first is (Basu, 2020) work on "Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms". In their approach, the authors experimented with supervised learning algorithms such as Support Vector Machine that uses margin to recognize and identify data points that look very much like in hindsight the other feature but it is able to classify correctly where it actually lies. They also built neural networks including LSTM, RNN etc. They narrowed their scope to the three most popular conditions from the drug review dataset, birth control, depression, and Pain; experimenting with 12 models' algorithms.

The second piece of literature by (Kyaing, 2015) on "Sentiment Analysis of User-Generated Content on Drug Reviews" focused more on using the linguistics approach by breaking down sentences into clauses, phrases, parts of speech, using bigrams, di-grams negation handling, linguistic decision trees to point to aspects of sentences that recognizes nuances, context and semantics before decisions on classification.

Background

Natural language processing is simply computer programming built to understand verbal or written language. Examples of Languages in their natural forms are English, German, Yoruba Language etc. (Zhao, 2018).

Processing: This is the way the computer processes commands and instructions to understand the Natural Languages.

There three common examples:

- Sentiment Analysis
- Topic Modeling
- Text Generation

Sentiment analysis is the process of computationally identifying and categorizing opinions in a piece of text, in order to determine whether the writer's view is negative or positive (Basu, 2020).

Topic modelling: Subjects such as fraud can be easily investigated in an organization using topic modelling. For example, fraud can be traced with document classification to highlight fraudulent schemes from documents and emails; prioritizing words relating to embezzlement.

Text Generation: for example, one can develop new quotes or articles by merging previously written texts and passing them through a text generating algorithm.

Data was sourced from the UCI repository. According to the resource on the data, the data was generated by crawling various pharmaceuticals websites. The dataset has 7 attributes, the drug name, condition, review, and rating were filtered for the project.

The rating is 1-10 according to customers' choices. To determine the target, the sentiment intensity analyzer library was used to provide polarity scores in three classes, negative, positive, and neutral. After careful review and comparison with ratings, ratings were preferred because the distribution of the sentiment analyzer was astoundingly skewed towards neutral. The rating was mapped to binary 1 and 0, text was vectorized and split into dependent and independent matrices of features, lastly, we trained the data using LSTM, ANN, Pre-trained BERT, Ktrain BERT and also Naïve Bayes, Linear Discriminant Analysis, Decision Tree, and Random Forest. Finally, we used our best model to make independent predictions.

Literature Review

(Basu, 2020) journal on Sentiment Analysis in Drug Reviews revealed that the authors trained their network with algorithms such as Random Forests, Support Vector Machines, Logistic Regression, Recurrent Neural Networks, and Artificial Neural Networks. The authors compared results to choose the best one.

They centred their study on the three most frequent conditions; Birth Control, Depression, and Pain. Twelve Models were trained over each condition for the purpose of this analysis with changing hyperparameters. They also performed a 10-Fold Grid Search on all models and then selected the one with the best result. In their experimentation, the authors tried the word2vec model to encode the text into numeric data but after training, they arrived at a very poor accuracy. They concluded that Word2Vec could not capture the importance of words within the context and also there were not enough data fed into the system.

This whole experiment arrived at the following conclusions:

1. Count Vector did better than TFIDF because it openly represents the survey words; it addresses the rate of recurrence of words of while TFIDF (frequency-inverse document frequency) represented important words which take away the nuances when classification is done.
2. Deep learning algorithms had better results because they captured more significant features (Basu, 2020). Neural networks utilize every one of the elements.

Similarities and differences: Their study is closer to the approach taken in this project by experimenting with various supervised learning algorithms and deep learning algorithms, benchmarking accuracies across models. Our study used all conditions as a feature as opposed to narrowing it down to three as used in (Basu, 2020). This report also included scrapping commercial health websites for independent drug reviews and then classifying them using the pre-trained hugging face transformers model which takes reviews and creates 1–5-star ratings. In extended research in the future, deploying this research for public use as a web application on Heroku using python flask can be explored.

(Kyaing, 2015)'s paper on "Sentiment Analysis of User-Generated Content on Drug Reviews".

The researchers adopted an unadulterated linguistic approach, looking at it from a multiclass angle of positive, negative, and neutral emotions; taking into cognizance, grammar, semantics, and parts of speech other than a reference line approach used in machine learning. This was achieved by highlighting all varying challenges through error analysis. A domain approach was utilized to pass different well-being and clinical terms in users' opinions to semantic types in a Unified Medical Language System called MetaMap. The labelled semantic data was used for opinion examination (Kyaing, 2015). Dr Alan (Lan) Aronson developed the program at the National Library of Medicine to map biomedical text to the medical dictionary to discover medical term concepts denoted in the text (National Library of Medicine - Lister Hill National Center for Biomedical Communications, 2020). They steered a primer report where a provisional level opinion order classification was created. They argued in line with (Wilson et al., 2009) that many pieces of research have been done on a phrase-level logical feeling examination, yet phrases are frequently not long enough to comprise both opinion and true emotion together for **viewpoint-based investigation**. The authors stated that there is two approaches to sentiment analysis and these are linguistic and machine learning approaches. They argued that since clauses are very short and don't contain numerous emotional words, the ML approach alone would for the most part experience information sparseness issues.

The authors opined that so as to produce a more exact and proficient sentiment analysis, sentences with two or more clauses such as “this is some drug and it did a great job” will be broken down into several clauses. “This is some drug” and “it did a great job”.

Results: To benchmark the linguistics approach, the authors conducted a supervised learning classification using a Support Vector Machine (SVM). The ML algorithm performed at 66% accuracy while the linguistic approach made better at 69% accuracy.

Similarities and differences: This research argued that clause, phrasal, context level classification would be better for reviews such as this while our study focused more on vectorizing the words, identifying patterns, and testing against customers’ ratings. The difference with this study is that we took a machine learning approach but their study took the linguistic approach, benchmarking against the ML approach in comparison. Overall, both need the other. The linguistics and machine learning approach can be incorporated to have even better classification results. The linguistic approach will be a little difficult to deploy because of the intrinsic linings that follow; trees, plus and minus negation handling, error analysis, inference error analysis, lexicon error analysis etc.

Objectives

Aim

This research looks to build machine learning models to categorize reviews into positive or negative. To carry this out, the following specific objectives were set out:

Objectives

1. Develop deep learning model with Ktrain BERT, Pretrained -BERT, LSTM, ANN, and supervised learning models Naïve Bayes, Linear Discriminant Analysis, Decision Tree, and Random Forest to classify data sourced from the UCI machine learning repository online.

2. To develop models able to compare levels against other related study results.

3. To analyse and visualize the current predictions with data visualization libraries such as seaborn and matplotlib etc.

Methodology

This study is designed to define features that describes the dynamic information in drug reviews for the easy interpretation of customers' experience from on-shelf drugs, for manufacturers to reconcile feedbacks and make needed improvements.

Data collection

Predictors: Drug name, conditions, rating, and review are the variables extracted from the dataset sourced from "University of California, Irvine, School of Information and Computer Science, Center for Machine Learning and Intelligent Systems online dataset archive" (University of Carlifornia Irvine, 2018) and (Felix Gräßer, 2018).

Procedures: Data Mining, Data pre-processing, text cleaning using stop words, regex, stemming, vectorizing (Bag of Words), the building of machine learning algorithms and deep learning architectures, and making predictions.

Outcomes: The framework will deliver algorithms to categorize drug reviews.

Experiment

- Data Pre-processing
- Text Cleaning
- Exploratory Data Analysis
- Sentiment Intensity Analysis (Sentiment Analyzer)
- Word Cloud
- Bag of Words
- Splitting the dataset into the Training set and Test set
- Supervised Learning Models
- Neural Networks Architectures
- Model Accuracy Comparison
- Visualization

Data Preprocessing

To prepare data for training, the data was imported, and 20,000 samples extracted using pandas.

Text Cleaning

The review attribute was cleaned using the Natural Language Toolkit (NLTK) module. Articles like a, the, an etc. were all removed during this process, characters such as @&^ etc. were removed with regex.

The data was stemmed using Porter Stemmer to adjust words to their natural form for better interpretation. E.g. ‘Bringing’ will be cut down to ‘bring’.

Exploratory Data Analysis

Some exploratory analysis was made to have a better look into the data distribution in order to decide the direction to go with predictors.

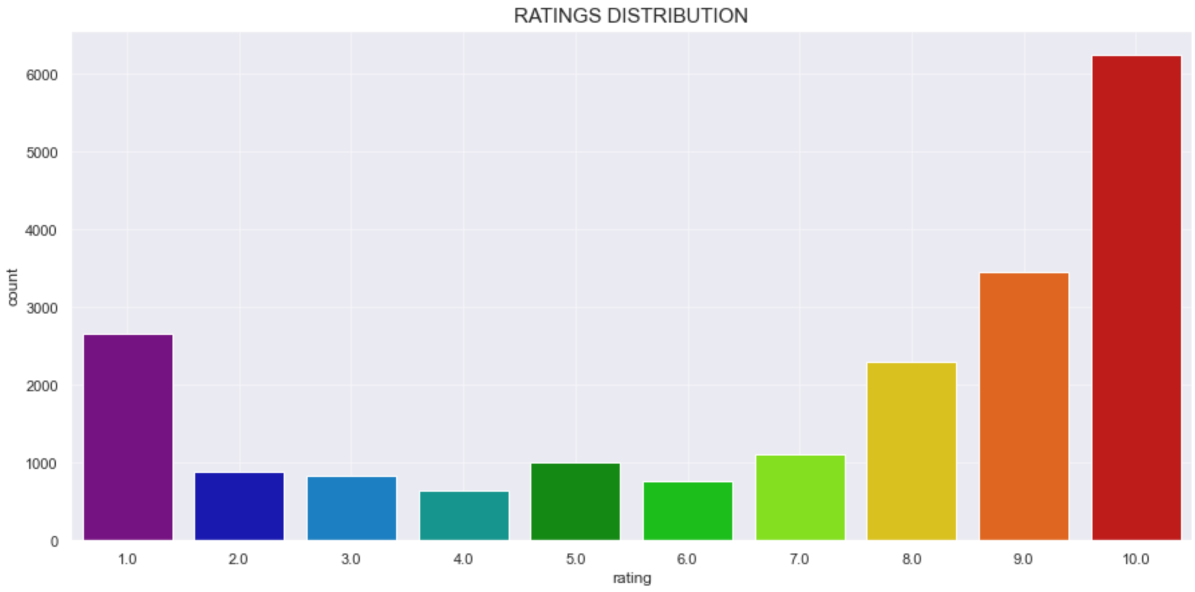

Figure 1: Ratings

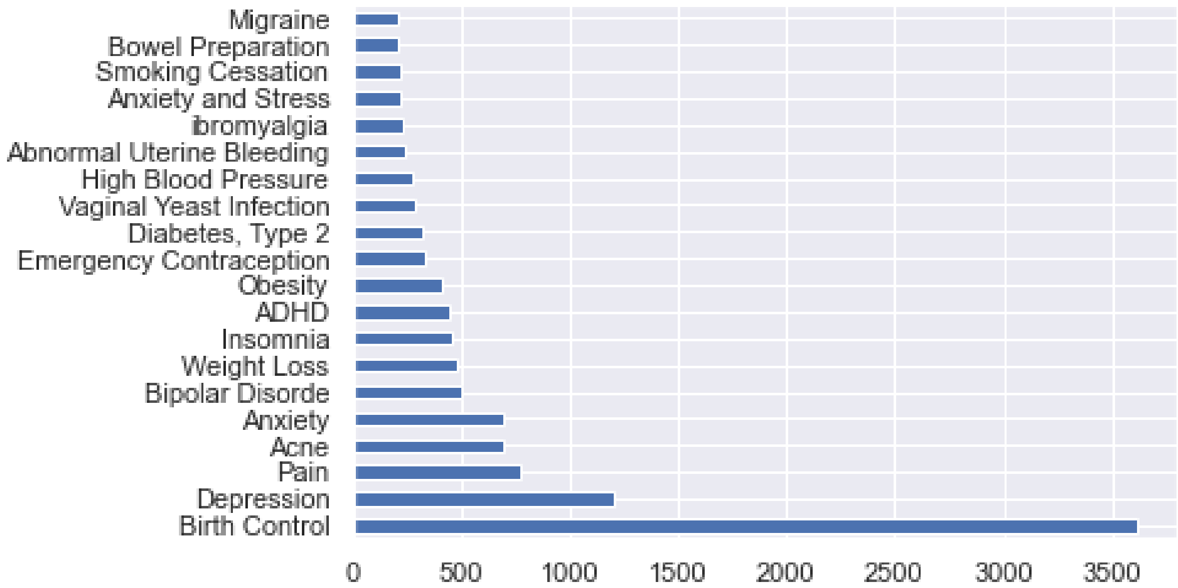

Figure 2: Conditions

Sentiment intensity analyzer (L) distribution below revealed the neutrality of the reviews was quite high. After careful review of the distribution of both ratings (R) and sentiments analyzer, rating is preferred as target because it is fairly evenly spread. The rating was mapped (<=5, negative, > 5, positive) to binary 1 and 0.

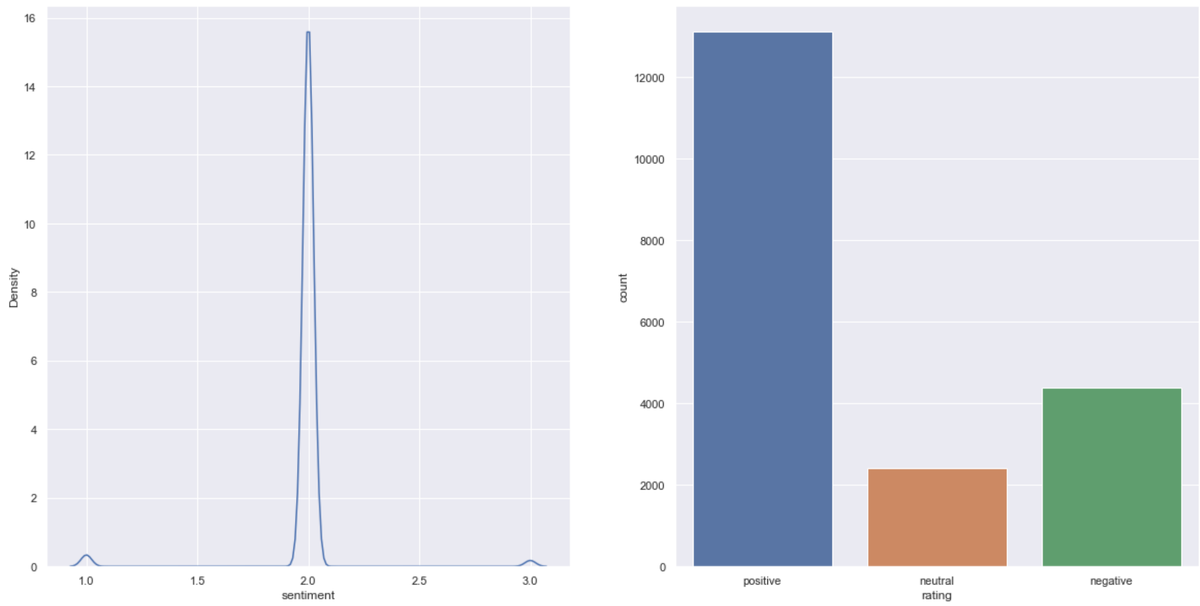

Figure 3: Sentiment Analyzer & Rating

Word Cloud

This is the graphical representation of word recurrence that give more prominent noticeable quality in the source text (BetterEvalaution, 2020).

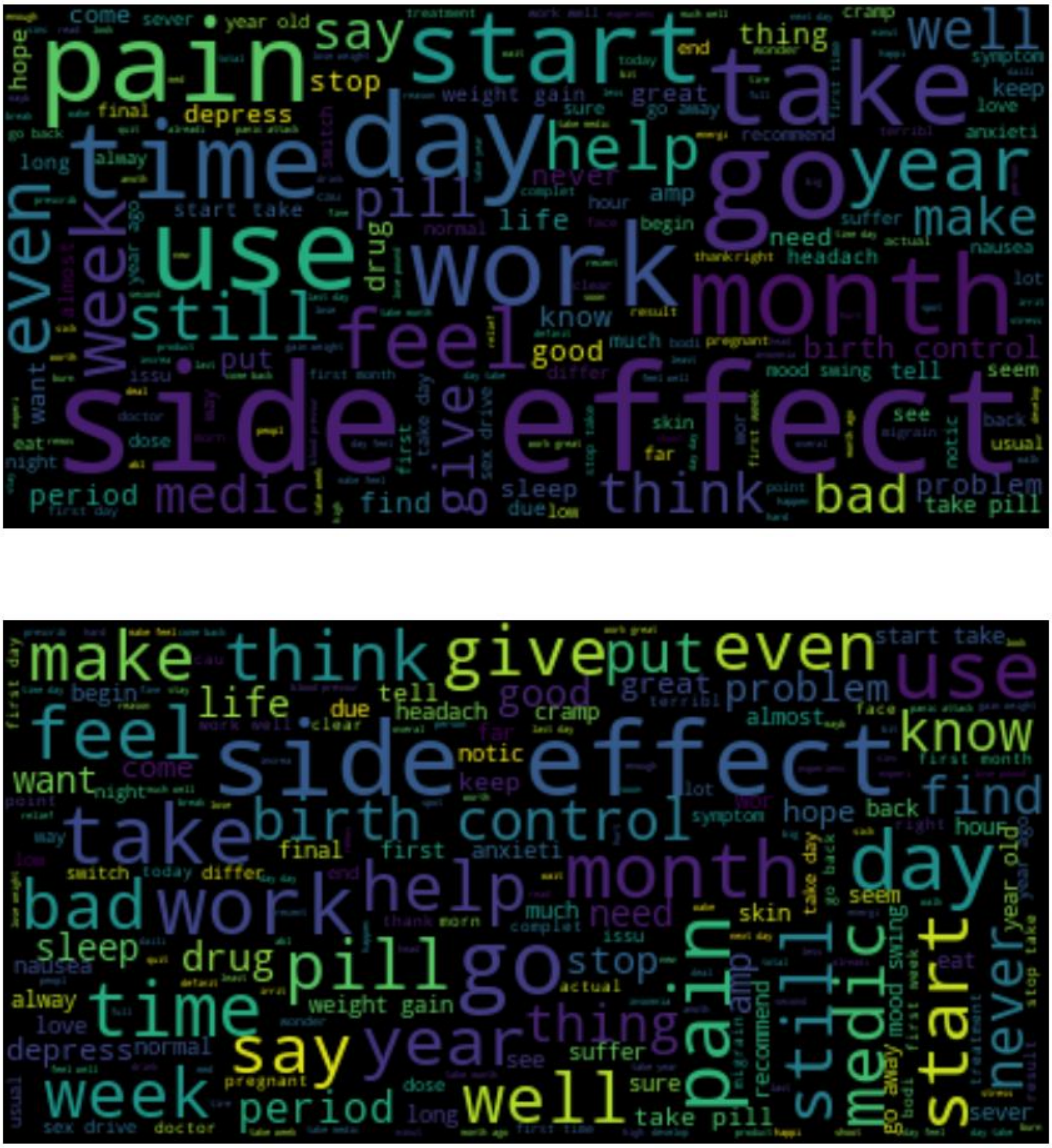

Figure 4: Word Cloud

Bag of Words

A text document is addressed as though it were a bag-of-words, that is, an unordered arrangement of words with their position disregarded, keeping just their frequency (Jurafsky, 2020). The cleaned texts were passed through a count vectorizer which counts the occurrence of the tokens into vectors. See the illustration below:

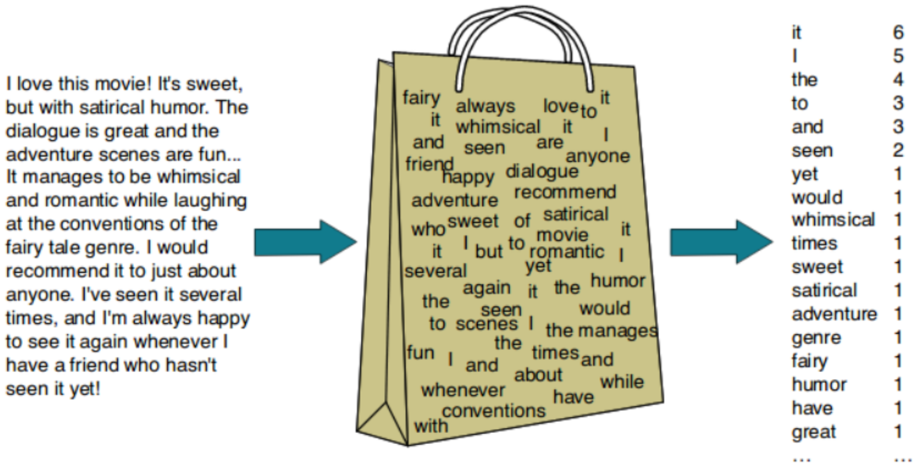

Figure 5: Bag of Words

[Source](): Bag of Words

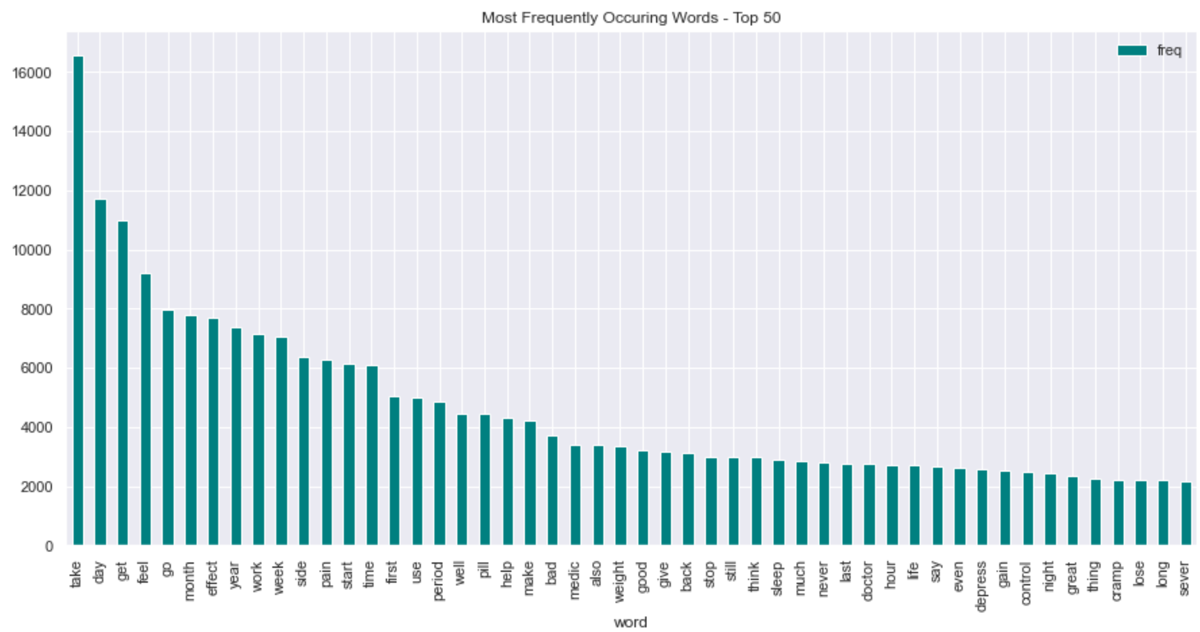

Figure 6: Most frequent Words

Splitting the dataset into the Training set and Test set

Data was split with train_test_split library – 20% test.

Supervised Learning

Trained the set with Gaussian Naves Bayes but the training accuracy was very low 58% and 41% with a feature scaled data. From Scikit learn documentation, Multinomial Naive Bayes is better used on a text classification problem. Find below:

Model Name	Training Accuracy	Validation Accuracy
Naive Bayes	0.7939	0.7744
Decision Tree	0.7102	0.7197
Random Forest Classifier	0.9997	0.7919
Linear Discriminant Analysis	0.8455	0.7895

Table 1: Supervised Learning Models

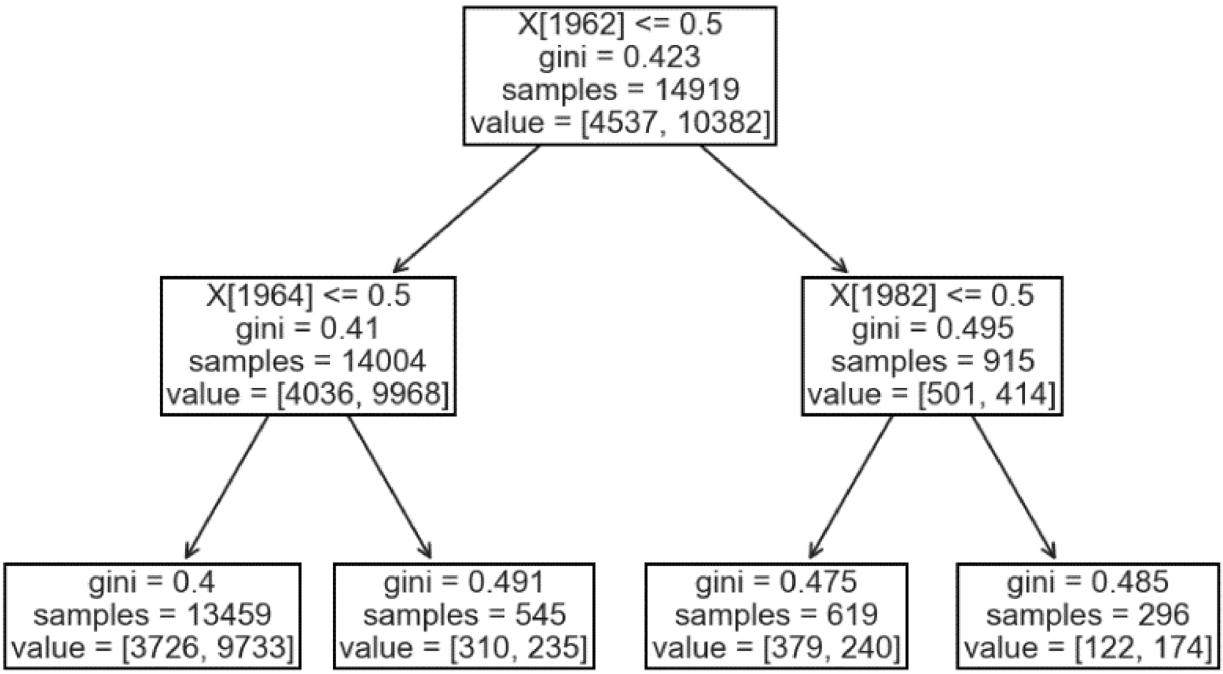

Figure 7: Decision Tree

Neural Network

It should be noted that LSTMs and Recurrent networks are best for text analysis because they can be used repeatedly to predict text as tokens are ingested (Datarobot, 2019). Artificial neural network was only used here for experimental purpose.

Long Short-Term Memory (LSTM) 1

We experimented with binary as a loss function, it performed poorly because our target is categorical variables. Categorical cross-entropy is best used for the multi-class classification model with categorical variables (Kumar, 2020).

Accuracy	Value
Training Accuracy	0.9530
Test Accuracy	0.8002
Training Loss	0.1407
Test Loss	0.5612

Table 2: LSTM

Parameter	Value
Activation function	softmax
SpatialDropout1D	0.2
Dropout	0.2
Epoch	10
Batch Size	32
Optimizer	Adam
loss function	Categorical cross-entropy
embed_dim	128
lstm_out	100
max_features	2000

Result: The above table revealed that the model trained well but overfit a little as seen from the validation accuracy.

Long Short-Term Memory (LSTM) 2

The model performance is not tangible looking at the accuracy but it looks like a good fit as shown by the little distance between train and test.

Training Accuracy	0.8769
Test Accuracy	0.8217
Training Loss	0.3194
Test Loss	0.4035

Table 3: LSTM 2

Parameter	**Value**
Activation function	sigmoid
SpatialDropout1D	0.2
Dropout	0.2
Epoch	5
Batch Size	32
Optimizer	RMSProp
loss function	Categorical cross-entropy
embed_dim	128
lstm_out	200
max_features	2000

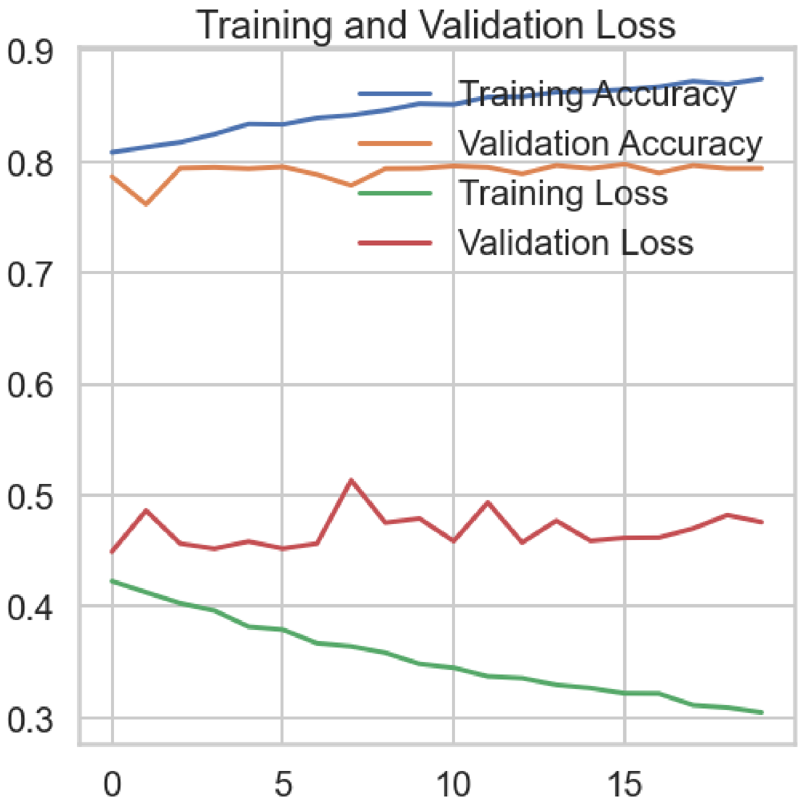

Figure 8: LSTM 2

Bidirectional LSTM

Accuracy	Value
Training Accuracy	0.6968
Test Accuracy	0.6973
Training Loss	0.6096
Test Loss	0.6099

Table 4: Bidirectional LSTM

Parameter	Value
Activation function	Relu, softmax
Padding	same
Dropout	0.4
Epoch	20
Batch Size	32
Optimizer	SGD
loss function	Categorical cross-entropy
embed_size	32
Learning rate	0.1
decay_rate	learning_rate / epochs
filter	32
Kernel size	3
Vocab_size	10000

Note: This step will duplicate the first recurrent layer, placing it side by side first layer, and further making a copy of the input sequence as the second recurrent network (Brownlee, 2021).

Result: The model ran at a very fast rate but failed to learn.

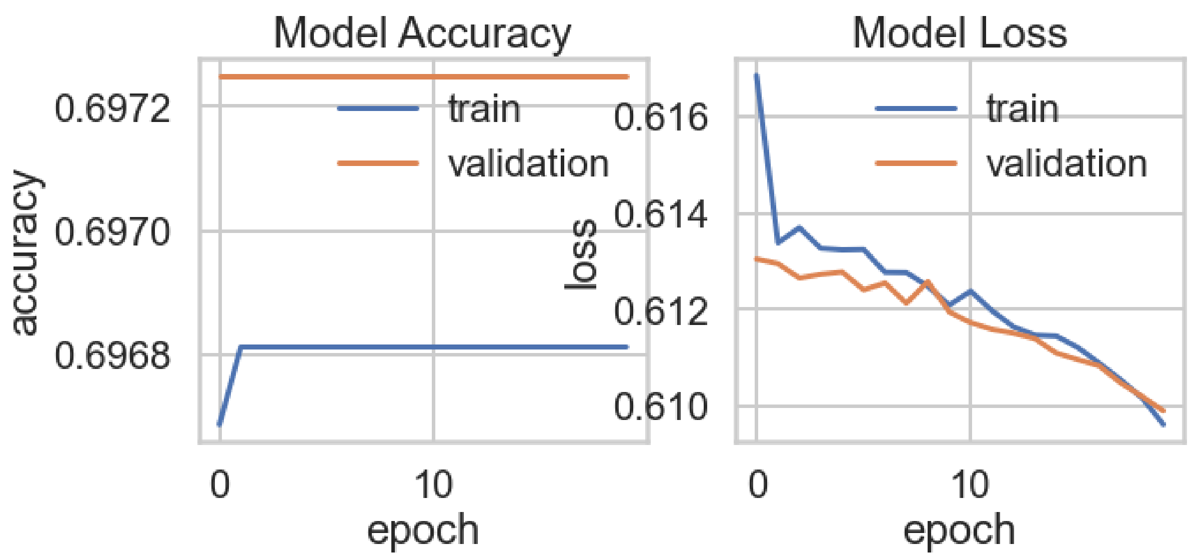

Figure 9: Bidirectional LSTM

Artificial Neural Network

Parameter	Value
Activation function	Relu, sigmoid
Dropout	0.2,
Epoch	10
Batch Size	64
Optimizer	Adam
loss function	Binary cross-entropy
Dense	2000, 1

Training Accuracy	0.9997
Test Accuracy	0.8609
Training Loss	0.0024
Test Loss	0.8162

Table 5: Artificial Neural Network

Result: The model learned well and it is slightly a good fit.

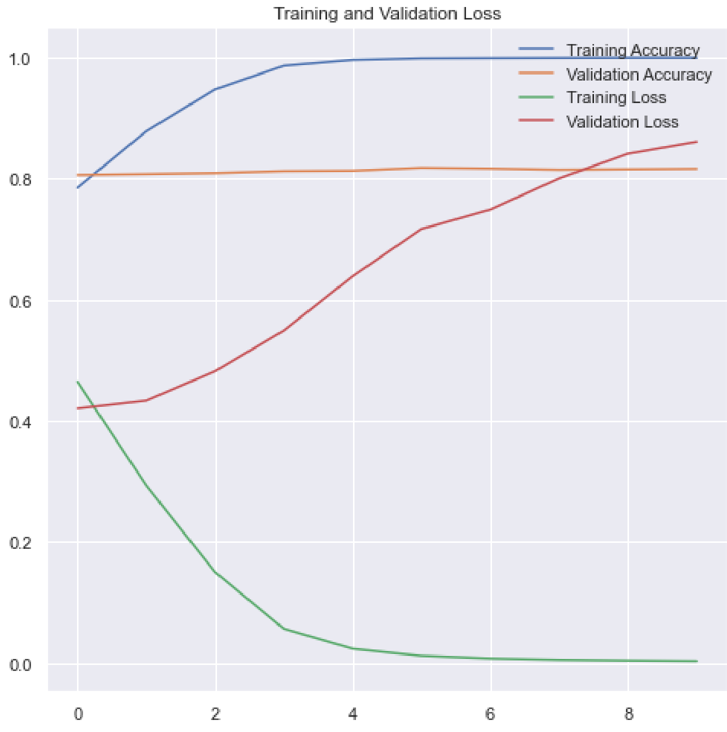

Figure 10: Artificial Neural Network

Pre- Trained BERT

We will use hugging face pre-trained transformer model to generate sentiments (ratings: 1 - 5 star) from our drug review dataset.

For experimentation, we will make an independent prediction by googling a random website(zavamed) to scrap for reviews on a birth control pill to make an independent prediction of the reviews, we will then further use the model on our dataset (Renotte, 2021).

Find in the plot below the distribution of ratings generated by the pre-trained transformers:

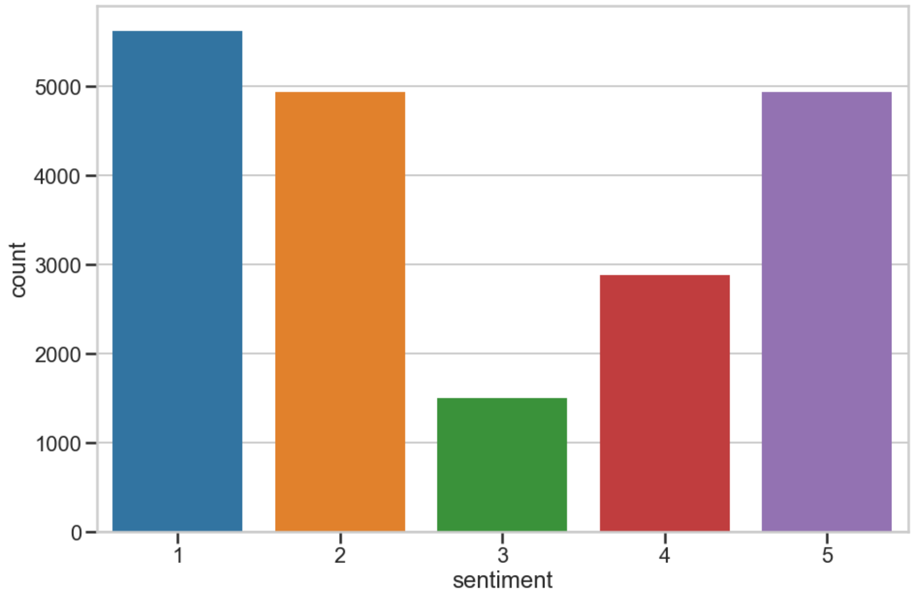

Figure 11: BERT (ratings)

BERT Ktrain

To determine the learning rate before training, we decided to simulate training to find the best learner. This is because the whole point of training is to minimize loss function (Maiya, 2019). The learner trained for an average 4 hours at one (1) epoch and this learner.lr_find() function will go through 1024 epochs, before showing the best range of learning rate on a plot. The process was interrupted was interrupted because of the slow rate of execution that could take days.

We decided to go for a 0.005 learning rate, 1 epoch to train our model. This trained for 4hours.

Result:

Training accuracy = 0.7840

Test Accuracy = 0.8625

A single prediction was made for positive and negative comments, it predicted accurately.

Note: Accuracy does not always determine the performance of a model (Brownlee, 2019) .

Model Comparison

MODEL NAME	ACCURACY (TRAIN)	ACCURACY (VALIDATION)
LSTM	0.9536	0.8002
LSTM 2	0.8769	0.8217
BIDIRECTIONAL LSTM	0.6968	0.6973
ANN	0.9997	0.8138
NAÏVE BAYES	0.7952	0.7744
LDA	0.8455	0.7895
CART	0.7102	0.7197
RANDOM FOREST	0.9997	0.7919
BERT(Ktrain)	0.7840	0.8625

Table 6: Model Comparison

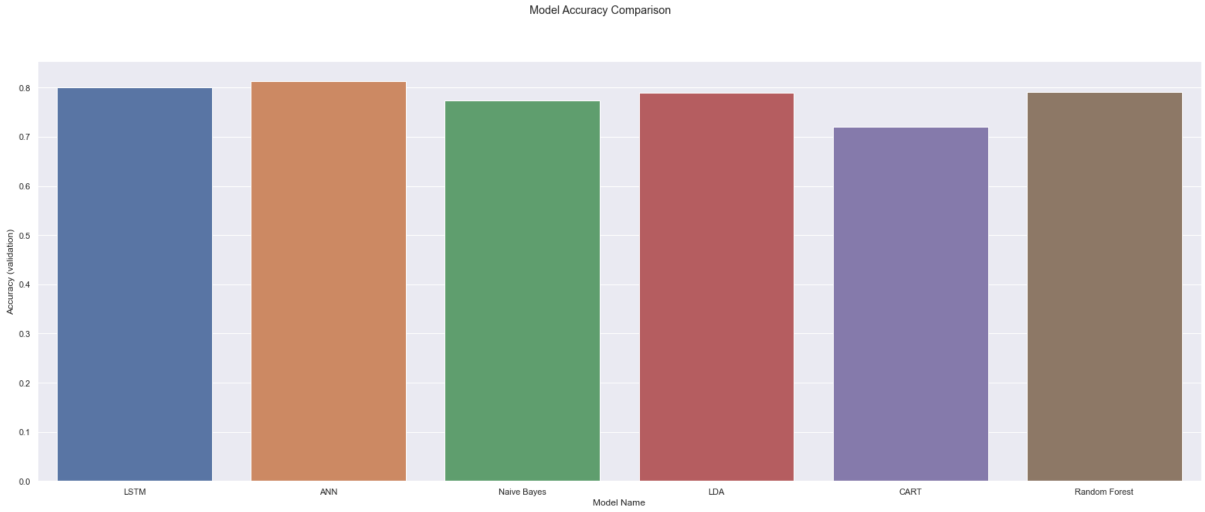

Figure 12: Model Comparison

Bibliography

Anyoha, Rockwell, (2017). *The History of Artificial Intelligence.*
Available at: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Basu, Debraj and Sairamvinay Vijayaraghavan, (2020). Sentiment Analysis in Drug Reviews using Supervised Machine Learning. *Sentiment Analysis in Drug Reviews,* Volume 1, 9.

BetterEvalaution, (2020). *Word Cloud.*
Available at: https://www.betterevaluation.org/en/evaluation-options/wordcloud#:~:text=Word%20clouds%20or%20tag%20clouds,in%20the%20document(s).

Brownlee, Jason, (2019). *Machine Learning Mastery.*
Available at: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/
[Accessed 2022 May 2019].

Brownlee, Jason, (2021). *Machine Learning Mastery.*
Available at: https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/
[Accessed 2022 May 2019].

Daniel Jurafsky James H. Martin, (2020). *Speech and Language Processing.*
Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
[Accessed 12 March 2022].

Datarobot, (2019). *Using Machine Learning for Sentiment Analysis: a Deep Dive.*
Available at: https://www.datarobot.com/blog/using-machine-learning-for-sentiment-analysis-a-deep-dive/

Felix Gräßer, Surya Kallumadi, Hagen Malberg, Sebastian Zaunseder, (2018), *Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning.* Lyon, France, Association for Computing Machinery - New York, NY, United States, 121-125.

HuggingFace, (2022). *nlptown/bert-base-multilingual-uncased-sentiment.*
Available at: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Michael, Haenlein. Andreas Kaplan (2019). *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence*. Berkeley Haas School Of Business. University of California Berkeley. California Management Review Vol. 61(4) 5–14, 1-10.

Ajitesh Kumar, (2020). *Keras – Categorical Cross Entropy Loss Function.*
Available at: https://vitalflux.com/keras-categorical-cross-entropy-loss-function/
[Accessed 18 February 2022].

Jin-Cheon Na, Wai Yan Min Kyaing, (2015). *Sentiment Analysis of User-Generated Content on Drug*. Journal of Information Science Theory and Practice, Volume 3(1), pp. Pages.6-23.

Tanya Lewis, (2014). *A Brief History of Artificial Intelligence.*
Available at: https://www.livescience.com/49007-history-of-artificial-intelligence.html
[Accessed 27 March 2022].

Arun Maiya, (2019). *Towards Data Science.*
Available at: https://towardsdatascience.com/ktrain-a-lightweight-wrapper-for-keras-to-help-train-neural-networks-82851ba889c
[Accessed 28 April 2022].

Daniel Jurafsky James H. Martin, (2020). *Speech and Language Processing - An Introduction to Natural Language Processing,.* Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

National Library of Medicine - Lister Hill National Center for Biomedical Communications, (2020). *MetaMap - A Tool For Recognizing UMLS Concepts in Text.*
Available at: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html
[Accessed 28 March 2022].

Pythonspot, (2022). *NLTK stop words.*Available at: https://pythonspot.com/nltk-stop-words/#:~:text=The%20stopwords%20in%20nltk%20are,the%20topic%20of%20your%20content.

Renotte, Nicholas, (2021). *Sentiment Analysis with BERT Neural Network and Python,* s.l.: s.n.

Theresa Wilson, Janyce Wiebe, Paul Hoffmann, (2009). *Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis*. Computational Linguistics, 35 (3), 399-433.

University of California Irvine, (2018). *Drug Review Dataset (Drug.Com) Data Set.*
Available at: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

Zhao, Alice, (2018). *Natural Language Processing in Python.*
Available at: https://www.youtube.com/watch?v=xvqsFTUsOmc