UNIVERSITÄT BONN

Machine learning hands-on

LAB REPORT

Author: Danqi Wang

3289105 | WDANQI@LIVE.COM

Tutor: Sumit Madan, Manuel Lentzen

Professor: Prof. Dr. Holger Fröhlich

Contents

1. Introduction

Machine leaning (ML) is the science of getting computers to learn, to identify pattern and to make a decision with minimal human intervention [1]. It is a subset of artificial intelligence (AI) and has been widely used in the past years, including image recognition, traffic prediction, product recommendations, natural language processing (NLP) and so on [2]. In our lab course, we focus on the application of ML in the field of NLP for life sciences and medical healthcare.

NLP focuses on the procedure that enables the computers to understand and process human languages in text. Sequence classification is a typical NLP task which assigns labels to documents or sequences. It has wide range of applications, including spam filtering, DNA sequence classification, article screening and so on [2,3]. Nowadays, there is a great need for researchers to find better tools to efficiently skim through the literature. For instance, in PubMed (a repository for abstracts of biomedical literature and life science journals), even though Medical Subject Headings (MeSH) terms can be used to reduce the number of documents, but there is no guarantee for the accurate classification of different elements (objective, methods, results and conclusion) since the abstracts are unstructured. So sequence classification is proposed to greatly facilitate the process of literature querying and searching.

Different from sequence classification, sentiment analysis (SA), also known as opinion mining or emotion AI, is a machine learning tool which detects positive, negative or neutral sentiment in text [4]. For example, word ‘fantastic’ and ‘useless’ in Figure 1 express positive and negative sentiment, respectively, and the sentence ‘ok, I guess’ expresses neutral sentiment, since it presents an uncertainty. One of its applications is retrieving medical healthcare information. For example, customers utilize the online review sites of drug reviews to express their sentiments and experienced drugs. But it is very difficult to review all online comments about the drug before making a purchase decision. And it is hard to classify the comments into useful insights as well [6]. The main purpose of SA of drug review is to predict particular side effects of drug, overall experience and effectiveness of patients as we have seen in the lab course.

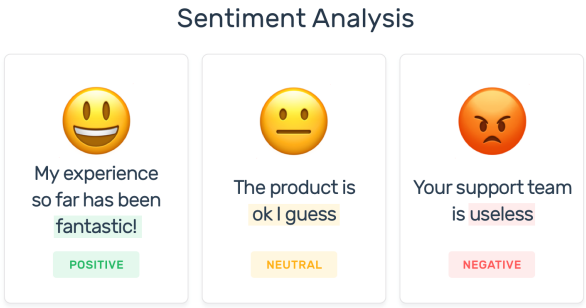

Figure 1. Sentiment analysis Go-To guide from MonkeyLearn [5]

2. Theoretical background

2.1. Text Mining

In order to analyze unstructured texts, the first thing we need to do is to discovery and explore the new information and patterns of them. Text mining (TM) employs a series of methods for converting the text into structured data, which is accessible to data mining algorithms. The whole procedure includes preprocessing of unstructured text, feature generation and selection, Data mining and analyzing the results [7]. The goal is to prepare the text for NLP.

2.2. Neural Networks and Transfer Learning

Neural networks (NNs), also known as artifical neural networks (ANNs), are compused of artificial neurons or nodes. The inspiration of the structure of NN is from

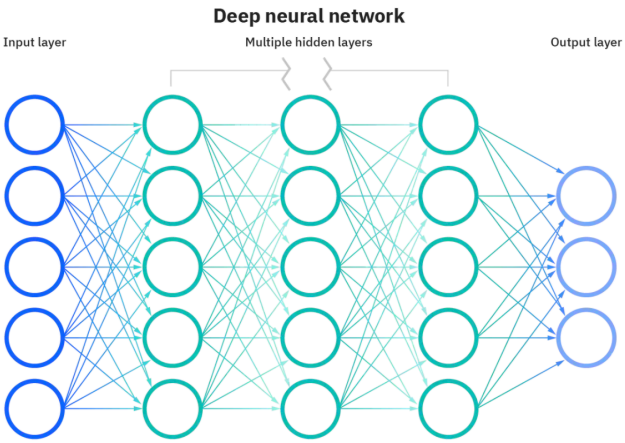

Figure 2. The architecture of deep neural network (DNN) [9]

neurons in a brain. Each connection can transmit signal like synapses to other neurons [8]. The figure below shows the architecture of deep neural networks, which consists of input layer with original input data, multiple hidden layers (with various hidden units) and the output layers which refers to the final classification.

In DNN, input data goes forward when the activation function (e.g. Sigmoid or ReLU function) in hidden units are activated. Backpropagation is used to train the NNs, where partial derivatives of defined loss function with respect to the weights (gradient) are computed based on the chain rule. And then makes an optimization of weights to minimize the loss function and reach the local optimum.

Convolutional neural networks (CNNs) are widely used in training large scale image data, as well as DeepVariant SNP data [10]. Different from traditional NNs, the architecture of CNN includes convolutional layer and pooling layer, where convolutional layer uses convolution kernels to extract the main features from original data and therefore increases the receptional field and pooling layer performs a downsampling operation (i.e. max or average pooling kernels) along the spatial dimensions.

Due to the complexity of CNN and corresponding huge amount of parameters to be trained, transfer learning is proposed to fine tune the new data on pre-defined model. The assumption is the pre-trained model is sufficiently close to application domain and fine-tuning starts from a point much closer to a local optimum. Tranfer learning allows rapid progress and improves the performance when modeling the new task.

2.3. Sequence classification

Sequence/Text classification and sentiment analysis are important tasks in NLP. In sequence classification, certain sequences are trained and assigned to labeled classes, which can be a library book, media articles and specific dataset (e.g. PubMed 200k RCT).

In order to perform sequence classification based on supervised ML algorithms, the most important thing is to extract the features and represent the text in structured vectors. And it includes the following procedures.

2.3.1. Sentence detection

Sentence detection is the method of detecting the start and end of one sentence. For example, detecting as a period (.) or (?) illustrates the ending point of one sentence and the beginning of another one as well.

2.3.2. Tokenization

Tokenization is a process of separating a piece of text into several unique units (tokens), which retains all the essential information about the data. Tokens can be either words, characters or subwords [12].

2.3.3. Stemming

Stemming is the process of removing suffixes and prefixes, and leaving the root of the word [11]. For instance, the root of the word “apple” and “apples” is “apple”.

2.3.4. Stop words

Stop words are those words that make useless contribution to the content of the document, such as and, the, or. Together with stemming, they aim to reduce the size of lexicon, and faster the computation time [11].

2.3.5. Represent text

a) One hot encoding: usually used in converting categorical variables into integer representation [13].

b) Raw counts: the number of counts that different words occur in a document. The raw counts will be transformed to indices which better reflects the relative importance of words [14].

c) Term frequency (TF):

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

In this formula, $f_{t,d}$ refers to the raw count of a term in the document, and the dominator is the sum of counts of all terms [15].

d) Term frequency – inverse document frequency (TF-IDF): unlike TF, TF-IDF focuses on term - inverse document frequency (IDF) except for the relative frequencies of occurrences of words. IDF refers to the logarithmically scaled inverse fraction of the documents which contain the word. The formula is shown [15].

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Where N is the total number of documents, and the denominator is the number of documents where term t appears.

Whereas TF-IDF is the product of TF and IDF [15]:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

A high TF and a low IDF (a term appears in more documents) refers to a high weight of the term. The weight can help to filter out common terms.

After generating the output of document text, various algorithms can be applied for text classification, for example, Naïve Bayes (NB), logistic regression, random forest, which will be illustrated in the following chapter.

Besides, deep learning method such as ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), can be used to perform pre-training task and to make a classification [16].

2.4. Sentiment analysis

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to the topic. The existing approaches includes knowledge-based techniques, statistical methods and hybrid approaches. Knowledge-based techniques used presence of affect words such as happy, sad, afraid, bored in order to classify the text into positive, neural or negative. Statistical methods inlude ML such as support vector machines (SVM), latent semantic analysis, and pre-trained deep learning method (BERT). Hybrid approaches uses both methods - knowledge based techniques and ML models to detect semantics [17].

Nowadays, more advanced methods attempt to recognize the multiple differentiated affective manifestations in text, which indicates emotions and opinions through analysis of language used for self-expression. For example, BERT model, which will be explained in the following chapter.

2.5. Mesurement: Precision, Recall and F1-score

The generation of evaluation metric of specific algorithm is an essential part of ML task.

Precision is defined as the proportion of positive identifications that is actually correct. And written in formula as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall(Sensitivity) answers the question what proportion of actual positives that is identified correctly. In formula:

$$\text{Recall} = \frac{TP}{TP+FN}$$

where TP is True Positives, FP is False Positives, FN is False Negatives

F1-score is a measure of the accuracy of a test. It is harmonic mean of Precision and Recall, defined as following:

$$\text{F1-score} = 2\frac{Precision*Recall}{Precision+Recall}$$

The highest possible value for F1-score is 1, meaning perfect precision and recall, and the lowest possible value is 0.

3. Methodology

3.1. Datasets

3.1.1. PubMed 200k RCT Dataset

PubMed 200k RCT is described in *Franck Dernoncourt, Ji Young Lee* [18]. The dataset is used for sequential sentence classification and consists of nearly 200,000 abstracts with the following classes: background, objective, method, result and conclusion. These abstracts are from randomized controlled trials and has 2.3 million sentences in total. PubMed 20k RCT is subset of 200k RCT, with reduced amount of entries.

3.1.2. Drug Review Dataset

The Drug Review Dataset is from the UCI Machine Learning Repository. The dataset provides patient reviews on specific drugs and a 10-star patient rating reflecting the overall patient satisfaction. The dataset is used in sentiment analysis of the review. The dataset is of shape (215063, 6) i.e. It has 215063 instances and 6 features [19].

Attributes of the dataset:

- drugName(categorical): name of drug
- condition(categorical): name of condition

- review(text): patient review
- rating(numerical): 10 star patient rating
- date(date): date of review entry
- usefulCount(numerical): number of users who found review useful

3.2. Machine Learning Models

3.2.1. Logistic Regression

Logistic regression is an important ML model for classification problems. It is a process of modeling the probability of a certain outcome given an input variable. Let x represent the vector of predictors for one case sample drawn from X (input values), then model the log-odds ratio for x belonging to each class as a linear function. Here we use Sigmoid function, It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. Since logistic regression is a linear method, and the log-odds ratio refers to:

$$\log(p(x)/1\text{-}p(x) = b0+b1*x$$

The coefficients of the algorithm is learnt from training data by maximum-likelihood estimation, which aims to seek the values for coefficients that minimize the error in the probabilities predicted by the model to those in the data [20].

3.2.2. Naïve Bayes

Naïve Bayes is also a group of probabilistic algorithm. It roots on conditional probability model using Bayes' theorem. It assumes that all properties independently contribute to the probability of one entry belonging to one class. $p(C_k|x_1,...,x_n)$ refers to the probability of given instance to be classified, $(x_1,...,x_n)$ represents n independent variables and k is the class entry. Based on the chain rule of conditional probability [21]:

$$p(C_k|x_1,...,x_n) = p(x_1|x_2,...,x_n,C_k)p(x_2|x_3,...,x_n,C_k)...p(x_{n-1}|x_n,C_k)p(x_n|C_k)p(C_k)$$

3.2.3. Random Forest

Random Forest is also called random decision forests. It is an ensemble learning method for classification and regression by generating multiple decision trees at training time [22]. Random Forest is more robust compared with individual decision tree because of the bagging idea (randomly select samples with replacement) and randomly select variables each time. This ensembling procedure greatly reduces the overall variance and can most of time get higher precision and accuracy compared to other ML models.

3.2.4. BERT

For the above three algorithms, they use word based approaches for tokenization and representation. But for BERT (Bidirectional Encoder Representations from Transformers), it makes use of Transformer, includes an encoder that reads the input data and a decoderfor producing the prediction for the task [23]. Two corpora BookCorpus and English Wikipedia were used for pre-training of the English BERT model. Masked language modeling (MLM) allows bidirectional pre-training. It will replace some tokens with a mask token [MASK]. After that, the model is trained to replace the [MASK] tokens with the correct words. The pre-trained BERT model is capable of performing NLP tasks [24].

3.2.5. ELECTRA

Unlike MLM pre-training, ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) uses an approach called “replaced token detection" . This approach uses another transformer neural network- generator that attempts to trick the model by replacing random tokens with fake tokens. The generator is trained with maximum likelihood to predict maked words and feeds into another discriminator. The discriminator shares the same input word embeddings. After pre-training, the discriminator is fine-tuned on following NLP tasks [16].

3.2.6. Comparison of algorithms

The first three algorithms are used in supervised sequence classification based on vector representations and sentence lables. For logistic regression, it is easy to implement and interpret, but there is assumption that linear class separation possible with not too large class overlap, the number of regressors should be smaller than the number of data points, and there should be no colinearities among independent variables.

For Naïve Bayes, when assumption of independence holds, then it performs better than logistic regression. And Naïve Bayes also works well for categorical input variables compared to numerical ones [25].

For Random forest, variable importance is easy to generate by permutation method. And it has many extensions, for example, proximity of data points, missing values imputation and so on. But if there are more features with small fraction of relevant variables in the dataset, it means there is high possibility to reach overfitting.

For BERT and ELECTRA, they use deep learning based transformer neural networks to train the text data. ELECTRA is used in sequence classification on PubMed 200k RCT dataset, whereas BERT is used to perform sentiment analysis on Drug Review Dataset in our lab course.

4. Results

4.1. Data Understanding

The goal of data understanding is to understand the attributes, missing values, inaccuracies, imbalance and relationships among features. For textual features, we also need to focus on the length of sentence, number of tokens and representation methods (e.g. TF-IDF).

4.1.1. PubMed 200k RCT Dataset

4.1.1.1. Class distribution

Firstly, we check the class distributions of the dataset, and found that the dataset is imbalanced, which is shown in the Figure 3 below:

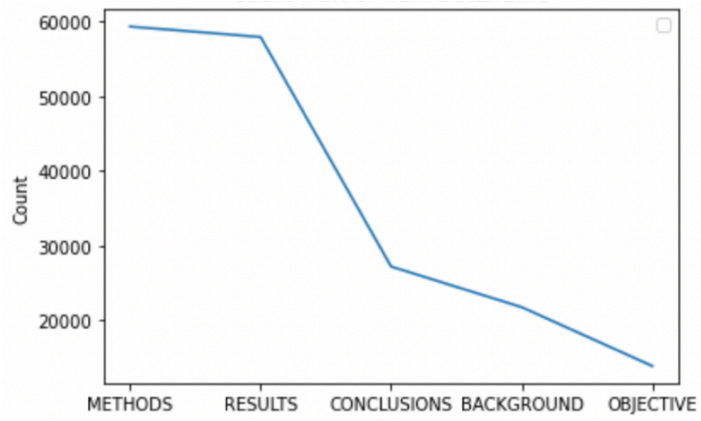

Figure 3. The class distribution of PubMed 200k RCT dataset.

"Methods" has the large amount of entries (59353), followed by "Results", nearly 57900 entries. There is sharp decline for two classes: "Conclusions" and "Background", with the number of 27168 and 21727 respectively. And for "Objective", it has the lowest number of counts, namely 13839.

4.1.1.2. Generate wordcloud

The wordcloud is a visual representation of word frequency. The more commly terms which appears in each class is shown in Figure 4.

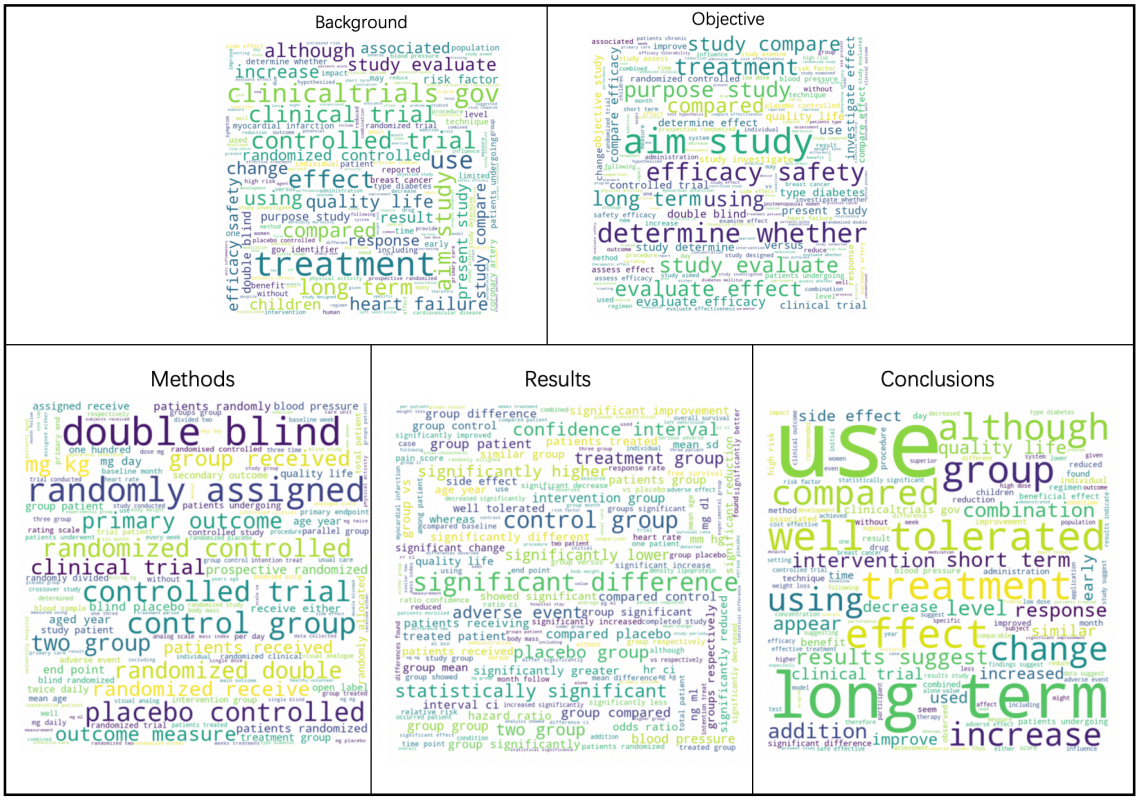

Figure 4. Wordclouds of different classes in the PubMed 200k RCT dataset.

For “Background”, the words with highest frequency is “treatment”. And in the category of “Objective”, it is “aim” and “study”. In “Methods”, “double blind” occurs with the top one frequency, which means this method is selected by most of studies in clinical trials. And for “Results”, we cannot clearly discriminate the prominent words in this class, whereas “significant difference” can be regarded as major terms compared with others. In “Conclusions”, “use” and “long term” are the most visible ones, which may reflect certain intervention method and time period of certain diagnosis.

4.1.2. Drug Review Dataset

There are 899 missing values in class “Condition“, which accounts for 0.56% of all entries. The missing values are deleted since the occupation is low.

4.1.2.1. Analysis of each category

In Table 1, it shows the top five popular drugs. Levonorgestrel accounts for largest amount of the whole dataset (3631), followed by Etonogestrel (3321). And all of these drugs are hormonic drugs.

Table 1. The top five popular drugs in Drug Review Dataset.

Drug names	Counts
Levonorgestrel	3631
Etonogestrel	3321
Ethinyl estradiol / norethindrone	2750
Nexplanon	2156
Ethinyl estradiol / norgestimate	2033

When it comes to "Condition", there are totally 885 different conditions in the dataset. The top five commonest conditions are shown in Table 2.

Table 2. a) Commonest conditions and b) largest number of conditions per drug

a)

Condition	Counts
Birth Control	28788
Depression	9069
Pain	6145
Anxiety	5904
Acne	5588

b)

Drug name	Number of conditions
Prednisone	38
Gabapentin	27
Doxycycline	24
Neurontin	21
Metronidazole	21

Most of drugs are related to condition – birth control in the dataset, which has largest amount of entries (28788), followed by depression (9069). The number of conditions of Prednisone is larger than other drugs, with the number of 38. Prednisone belongs to a class of drugs known as corticosteroids. To be specific, arthritis, blood disorders, breathing problems, severe allergies, skin diseases, cancer, eye problems, immune system disorders and so on.

As for the "Rating" and "Usefulcount", we can see there is relationship between the two classes. In Figure 5, the drug with high rating has largest usefulcount, which might due to the fact that patient usually check for rating information when they choose certain drugs.

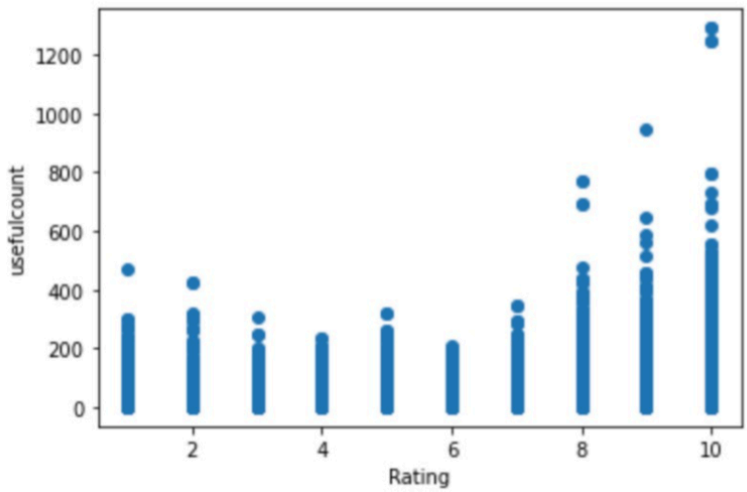

Figure 5. Rating v.s. UsefulCount

4.1.2.2. Texual feature extraction

Here we take one example for detecting the length of sentence, tokenization, stemming and remove the stop words.

"It has no side effect, I take it in combination of Bystolic 5mg and Fish oil" (Review of Valsartan)

- The length of sentence: 79
- Tokenization: [‘it’, ‘has’, ‘no’, ‘side’, ‘effect’, ‘i’, ‘take’, ‘it’, ‘in’, ‘combination’, ‘of’, ‘bystolic’, ‘5’, ‘mg’, ‘and’, ‘fish’, ‘oil’]
  Number of tokens: 17
- Stemming: [‘it’, ‘ha’, ‘no’, ‘side’, ‘effect’, ‘i’, ‘take’, ‘it’, ‘in’, ‘combination’, ‘of’, ‘bystolic’, ‘5’, ‘mg’, ‘and’, ‘fish’, ‘oil’]
- Stop word removal: 'ha effect combination bystolic 5 mg fish oil'

4.1.2.3. POS tagging

[('it', 'PRP'), ('has', 'VBZ'), ('no', 'DT'), ('side', 'NN'), ('effect', 'NN'), ('i', 'NN'), ('take', 'VBP'), ('it', 'PRP'), ('in', 'IN'), ('combination', 'NN'), ('of', 'IN'), ('bystolic', 'JJ'), ('5', 'CD'), ('mg', 'NN'), ('and', 'CC'), ('fish', 'JJ'), ('oil', 'NN')

4.1.2.4. TF-IDF

Two Reviews:

- Review 1(ID: 206461) “It has no side effect, I take it in combination of Bystolic 5mg and Fish Oil”

- Review 2(ID: 80520) “Reduced my pain by 80% and lets me live a normal life again”

Vocabulary: ['80', 'again', 'and', 'by', 'bystolic', 'combination', 'effect', 'fish', 'has', 'in', 'it', 'lets', 'life', 'live', 'me', 'mg', 'my', 'no', 'normal', 'of', 'oil', 'pain', 'reduced', 'side', 'take']

Vectorize text:

- review_1 = [0,0,0.175, 0, 0.246, 0, 0.246, 0.246, 0.246, 0.246, 0.492, 0, 0, 0, 0, 0.246, 0, 0.246, 0, 0.246, 0.246, 0, 0, 0.246, 0.246]
- review_2 = [0.295, 0.295, 0.210, 0.295, 0, 0, 0, 0, 0, 0, 0, 0.295, 0.295, 0.295, 0.295, 0, 0.295, 0, 0.295, 0,0, 0.295, 0.295, 0, 0]

4.2. Sequence classification

In order to perform sequence classification, different algorithms were used for training PubMed 20k RCT dataset. Figure 6 shows the performance metrics of all of them on test data.

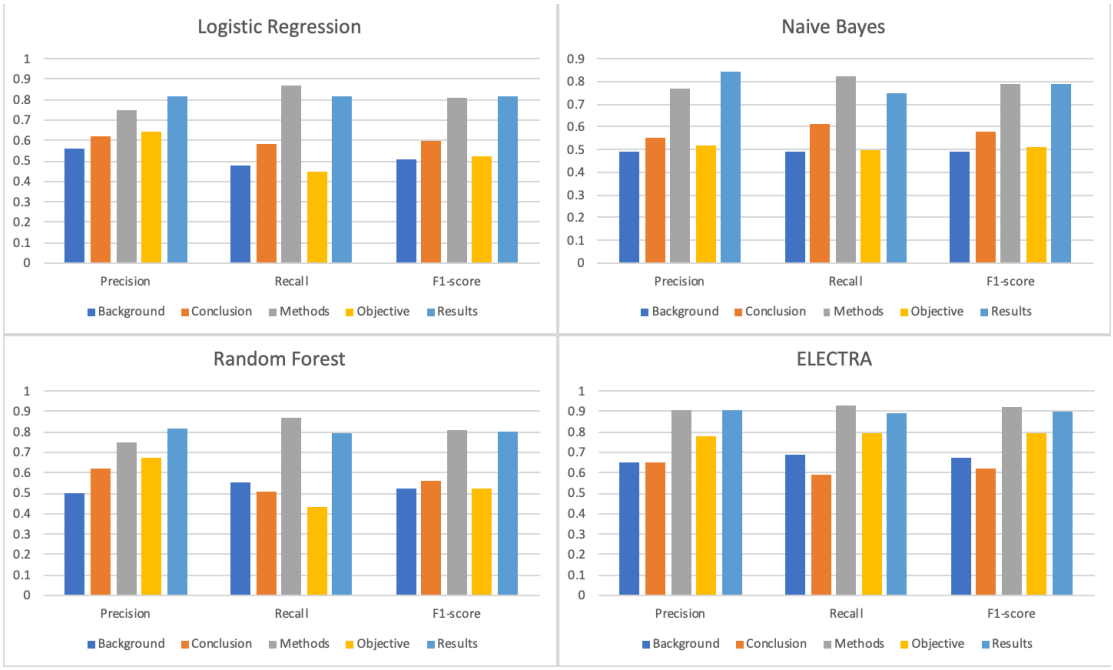

Figure 6. Comparison among different algorithms based on the performace metrics of given classes

For the comparison of different algorithms, ELECTRA performs best since precison, recall and F1-score among classes are overall higher than other algorithms. This means that deep learning model is better than traditional ML algorithms when it comes to solving the sequence classification problem in PubMed dataset.

For different categories, “Methods” and “Results” have larger precision, recall and F1-score compared to other classes. This illustrates that models with current hyperparamters are capable of classifying both groups well, but there is possibility for better performance of other groups when tuning the hyperparamters in the models.

In Figure 7, the accuracy scores of different algorithms are given. It is consistent with the result in Figure 6 that ELECTRA has the highest accuracy score of 0.84. ELECTRA can predict unseen data with better performance.

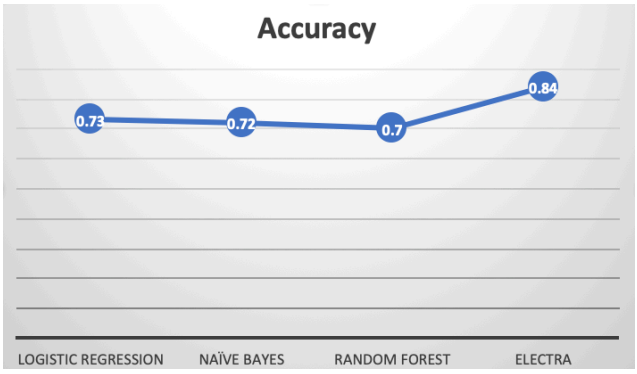

Figure 7. Comparison among different algorithms based on accuracy score

In order to check the further prediction classification performace of ELECTRA on new text – [*This result looks good!*], an new classifier was generated and checked the final score based on TextClassificationPipeline. The result showed that the text is from Label 3 (“Result”) with the score of 0.714. This illustrated the ELECTRA algorithm has the robust and better performance on unseen data.

4.3. Sentiment analysis

In order to perform sentiment analysis on Drug Review Dataset, BERT model is used to identify positive, negative and neutral reviews on the effects of the drugs. First of all, Rating scores are converted to three categories: Positive (>=7.0), Netrual (between 4.0 and 7.0) and Negative (<=4). The classes are imbalanced with larger fraction of positive ones (over 100,000 cases) whereas negative and neutral classes have nearly 40,000 and 15,000 entries separately, which is shown in Figure 8.

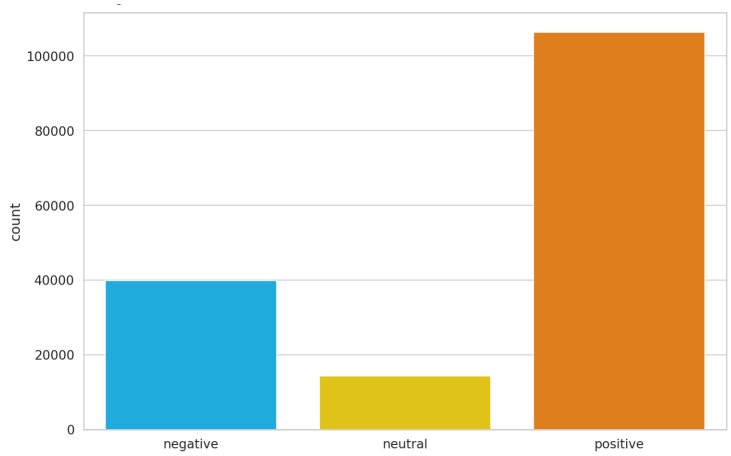

Figure 8. The class distribution after converting Rating score to three categories

Due to the computational limitation, only one epoch was trained in BERT model. Table 3 shows the precision, recall and F1-score results fort he sentiment analysis using BERT. The accuracy of the model is 0.86, and the macro average precision, recall and F1-score equal 0.71, 0.67 and 0.67 individually. It can be seen that positive class has highest metrics value and reach over 0.90, whereas for neutral class, the recall is only 0.19 and F1-score is 0.26. This means that the model can classify positive ones with high precison, followed by negative reviews and neutral ones were poorly classified.

Table 3. The performance metrics of sentiment analysis by BERT

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.79	0.87	0.83	13428
Neutral	0.43	0.19	0.26	4793
Positive	0.91	0.95	0.93	35250
Accuracy			0.86	53471
Macro avg	0.71	0.67	0.67	53471
Weighted avg	0.84	0.86	0.84	53471

Figure 9 below shows the confusion matrix for the test dataset. For positive class, 33353 reviews were predicted correctly (94.2%), 1160 were classified as negative ones (3.0%) and 2020 were predicted as neutral reviews (5.8%). For negative reviews, 11735 (79.1%)were correctly predicted, 1861 were incorrectly classified as neural and 1239 were classified as positive. For neutral class, only 912 reviews (43.3%) were correctly classified.

Figure 9. The confusion matrix for sentiment analysis on test dataset

5. Conclusion

5.1. Data Understanding

The PubMed 200k RCT Dataset is imbalanced among classes and each class has its own representative words.

The most common drugs in DrugReview Dataset are hormonal drugs. And Prednisone has larger number of conditions to use. For Rating and UsefulCount, they have certain relationships.

5.2. Sequence classification

According to the results for performance metrics of the traditional ML algorithms, logistic regression has the highest predicting accuracy (0.73) compared to Naive Bayes and Random Forest. But the differences among them are small, since the scores for the others are 0.72 and 0.70 separately. When it comes to the deep learning model-ELECTRA, the increasing of performance is larger, with the accuracy score of 0.84. This means that for the current setting of hyperparametres, ELECTRA works best and performs robust.

5.3. Sentiment analysis

Sentiment analysis performs well on positive reviews, followed by negative ones. It is because of the imbalance of the three classes. The current model can be used to

classify positive or negative reviews with precision score of 0.91 and 0.79. And it would be a problem when distinguishing neural reviews, since the precision is only 0.43, which would be not reliable.

5.4. Strengths and Limitations

For sequence classification, logistic regression is easy to implement. But when it comes to generate best performance, we should take ELECTRA model. ELECTRA has the limitation of high computaional cost compared to traditional ML algorithms.

For sentiment analysis, the prediction works well on positive reviews, which means the model is powerful if there is enough data to be trained. As for the limitations, the lack of computational power is a big issue for training large dataset using BERT. And we chose the default hyperparameters as mentioned in previous publication, it might not suit for our dataset. Lastly, the dataset we used in highly imbalanced, which will strongly effect the prediction result.

5.5. Outlook

For sequence classification, the performace of traditional ML algorithms can be improved by tuning hyperparameters, since some of the hyperparameters are default setting in current models.

For sentiment analysis, in order to improve the performance of BERT for the dataset, following aspects should be taken into consideration.

1. Run more epochs to minimize the loss and increase accuracy.
2. Balance the current dataset, e.g. increase the fraction of neural and negative groups.
3. Fine-tuning the hyperparameters for the model.

References

1. Machine learning: what it is and why it matters. URL: https://www.sas.com/en_us/insights/analytics/machine-learning.html (visited on 21.09.2021)

   ###Aimen Yang *et al*. Review on the application of machine learning algorithms in the sequence Data mining of DNA. *Front.Bioeng.Biotechnol*. http://doi.org/10.3389/fbioe.2020.01032
2. Applications of Machine learning. URL: https://www.javatpoint.com/applications-of-machine-learning (visited on 21.09.2021)
3. Karen Cardozo Vera. Artifical intelligence & Machine learning in life sciences. URL: https://www.modis.com/en-be/insights/blog/artificial-intelligence-and-machine-learning-in-life-sciences/ (visited on 21.09.2021)
4. Sentiment Analysis — Definition of Sentiment Analysis by Oxford Dictionary on Lexico.com also meaning of Sentiment Analysis. URL: https://www.lexico.com/definition/sentiment_analysis (visited on 21.09.2021)
5. Sentiment Analysis: The Go-To Guide. URL: https://medium.com/@rachel_39895?p=74bd3dfb536c (visited on 21.09.2021)
6. Thu Dinh. Detecting side effects and evaluating effectiveness of drugs from customer's online reviews using text analytics and data mining models. URL: https://www.mwsug.org/proceedings/2019/IN/MWSUG-2019-IN-064.pdf (visited on 24.09.2021)
7. Richard S. Segall, Qingyu Zhang and Mei Cao. Web-based text mining of hotel customer comments using SAS® text miner and megaputer polyanalyst®. *Swdsi* 2009.
8. Neural network. *Wikipedia*. URL: https://en.wikipedia.org/wiki/Neural_network (visited on 25.09.2021)
9. Neural Networks – What are neural networks? URL: https://www.ibm.com/cloud/learn/neural-networks (visited on 25.09.2021)
10. Mark DePristo and Ryan Poplin. DeepVariant: Highly Accurate Genomes With Deep Neural Networks. *Google AI Blog*. 2017.

11. Javed Shaikh. Machine Learning, NLP: Text Classification using sckit-learn, python and NLTK. *Towards data science*. 2017. NRL: https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a (visited on 25.09.2021)
12. Aravind Pai. What is Tokenization in NLP? *Analytics Vidhya*. 2020. URL: https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/ (visited on 24.09.2021)
13. NLP-progress: Dependency parsing. URL: http://nlpprogress.com/english/dependency_parsing.html (visited on 24.09.2021)
14. Text Mining (Big data, unstructured data) – Transforming word frequencies. *Statistica*. URL: https://statisticasoftware.wordpress.com/2012/09/06/text-mining-big-data-unstructured-data/ (visited on 24.09.2021)
15. Tf-idf. *Wikipedia*. URL: https://en.wikipedia.org/wiki/Tf%E2%80%93idf (visited on 25.09.2021)
16. Google AI Blog: More efficient NLP model pre-training with ELECTRA. URL: https://ai.googleblog.com/2020/03/more-efficient-nlp-model-pre-training.html (visited on 25.09.2021)
17. Sentiment analysis. Wikipedia. URL: https://en.wikipedia.org/wiki/Sentiment_analysis#Methods_and_features (visited on 25.09.2021)
18. Franck Dernoncourt, Ji Young Lee. PubMed 200k RCT: a dataset for sequenctial sentence classification in medical abstracts. *International Joint Conference on Natural Language Processing (IJCNLP)*.2017.
19. Drug Review Dataset(Drugs.com) Data Set. *UCI Machine Learning Repository*. 2018. URL: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+(Drugs.com) (visited on 25.09.2021)
20. Logistic Regression for Machine Learning. Jason Brownlee. *Machine Learning Mastery.* 2016. URL: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ (visited on 25.09.2021)
21. Naive Bayes classifier. Wikipedia. URL: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (visited on 25.09.2021)
22. Random forest. Wikipedia. URL: https://en.wikipedia.org/wiki/Random_forest (visited on 25.09.2021)

23. Rani Horev. BERT Explained: State of the art language model for NLP. *Towards data science*. URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (visited on 25.09.2021)
24. James Briggs. ELECTRA is BERT – Superchanged. Towards data science. URL: https://towardsdatascience.com/electra-is-bert-supercharged-b450246c4edb (visited on 25.09.2021)
25. Sunil Ray. 6 easy steps to learn Naive Bayes algorithm with codes in Python and R. *Analytics Vidhya*. URL: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (visited on 25.09.2021)