Analysis of Supervised Classification Models with Implementations on Multipleclass Text Classification

Lihui Huang[1], Yiran Wang[2] and Jiewen Liu[3]

***Abstract*— In this project, we investigate the performance of several models, including Bernoulli Naive Bayes, Multinomial Naive Bayes, Support Vector Machine and so on, in the text classification problem about the comments from a famous online forum – Reddit. We use the unprocessed training dataset to train our models and test its performance on the test dataset which has its subreddits hidden. Some text preprocessing techniques (e.g. lemmatization, stop words and TF*IDF tokenization, etc.) are used before the evaluation. Then, methods like k-fold cross-validation, feature selection, ensembles and so on are utilized to test and improve the prediction accuracy of our models.**

From our experiments, we found that Multinomial Naive Bayes achieves relatively good accuracy on primitive data, but for other models need more modification from feature selection. By using feature selection to reduce the dimension, all our models' accuracy improved by a different but significant amount. In addition, our experiments shows that ensembles learning only augment the result of some models, and Multinomial Naive Bayes achieve the best overall performance.

I. INTRODUCTION

Background

Machine learning is a very useful and popular technique in resolving various problems. Classification and regression are among the most common problems. In this project, our task is to clean and preprocess our training data(reddit comments), choose the best model and find its parameters to best classify the comment from the testing dataset with its classes hidden.

Important Findings

Multinomial Naive Bayes[1] is a relatively stable algorithm after smoothing. Unvarnished model gives significant and good accuracy on the testing dataset, and also is fast to run. Its accuracy even doesn't change very much while other methods are introduced. Support Vector Machine's[2] accuracy in the cross-validation increases by at most 4% when bagging[3] is used, which is also intuitive that SVM model could benefit more from randomness and variance reduction by ensembles learning. Logistic Regression[4] comes thirds, which might indicate plain linear model might not suit well with complex text categorization problem.

Feature selection through chi-squared[5] scores help reduce the complexity and dimensions of the classification problem. A significant amount of run time is reduced and accuracy is increased for all models. Boosting doesn't behave well in our experiment since serious overfitting issue is introduced.

Related Work

In sentimental classification problem, the pioneering work including General Inquirer[6] combines the psychological experiments on people's mental state in response to different verbal behavior, and a series of computational procedures developed to quantify the use of some sentimental words and recurrent patterns to generate hypothesis and insights of the text. Later work becomes overall less complicated and start to partially rely on big data to generate the prediction from statistics. For example, the contribution by [7]Volcani and Fogel uses lexical impacts, which come from statistics, with synonyms, antonyms and related words taken into consideration, to predict the scale of emotion.

In more recent works, researchers begin to focus on more aspects of sentimental classification problem, like polarity of text, subjectivity and objectivity, etc. The research done by [8]Turney

[1]Lihui Huang (260821232): lihui.huang@mail.mcgill.ca
[2]Yiran Wang (260825557): yiran.wang3@mail.mcgill.ca
[3]Jiewen Liu (260825295): jiewen.liu@mail.mcgill.ca

developed quantitative ways to calculate the semantic orientation words as the mutual information between the given phrase and the word “excellent” or “poor” and make predictions according to the overall average difference from the mutual information, and his prediction model achieves significant accuracy in text from many domains like movie, travel and so on. The another famous experiment done by [9]Bo and Lillian explores the issue of classification of text into subjective and objective parts which facilitate greatly the understanding of opinion and information the text and is not restricted only to the analyzation of the sentimental level and attitude reflected by the text.

Many methods and models are introduced and investigated heavily to address the specific issue in sentimental classification problem. For instance, the Max Entropy and Support Vector Machine are proven to be useful and adopted in the removal or adjustment of neutral phrase, which is critical to the improvement of the classification accuracy. Strategies like semantic networks, word analysis and automatic identification for topic come with deep learning to better address the classification issue when it is narrowed down to some specific fields or called sub-problems and increase the prediction performance by increasing the specificity.

II. DATASET AND SETUP

Our training dataset is composed of 70,000 user’s comments with its subreddit (class) specified. The data is unvarnished and contains many informal expression, syntax errors, non-English characters, non-informative pieces and so on. We first lemmatize all words in the comments to. recover the canonical form of the words and reduce the dimensions, remove all less or non-informative words and tokenize the rest. Until here we obtain 70798 words as our features. By observing the appearance of each words in the training set(Fig 1), we discovered that over the huge training dataset, 35398 words appeared only once(which is about half the feature size), while 11266 words appeared more than 10 times and 2237 words appeared more than 100 times. We then use the TF*IDF Vectorizer imported from

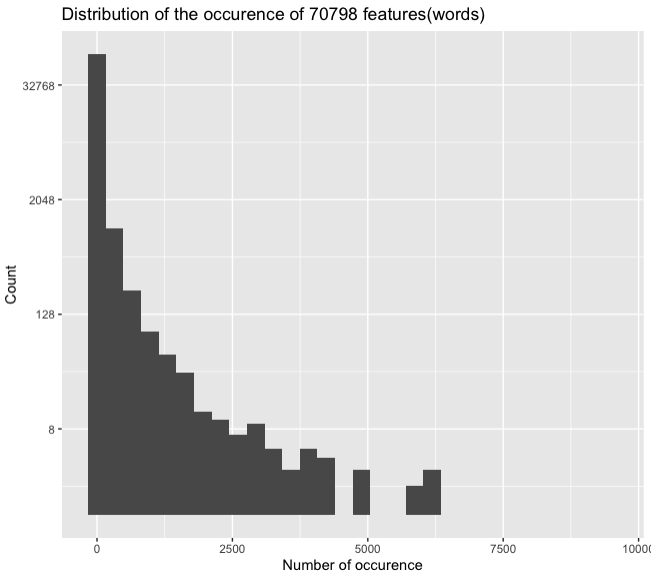

Fig. 1: Distribution summary of the appearance of tokens. The y axis is log2 scaled due to the huge difference in the data of the count. Some empty value of count (blank gap) indicates the quantity of feature(word) with that number of occurence falls is 0 or 1.

sklearn to vectorize our comments into a sparse matrix and use it as our training input. For the catagory of each comment, each catagory equally have 3500 training samples.

The testing dataset contains 30,000 un-preprocessed and uncategorized comments. The true subreddits(class) of the testing data are hidden. The same preprocessing technique is used in the testing dataset.

III. PROPOSED APPROACH

Feature Choice

The feature we chose to use is TF*IDF– term frequency–inverse document frequency – scores of the words that appear in our comments. The TF*IDF score, which is based on the formula:

$$w_{t,Corpus} = (\text{\#t in Corpus}) * log(\frac{\text{\#Docs in Corpus}}{\text{\#Docs with term t}})$$

is a good indication of the importance of a word in a text. The reason of choosing TF*IDF is that the data set contains abundance of separated document, and each of them have a label (20 kinds of label total). TF*IDF features usually performs well on this kind of data set. There are 70000+ features total, which are too many. Thus we used

SelectKbest method from sklearn to extract best features by ranking them according to Chi-square statistics. After experimenting, we found that out models performed best when we only keep 17000 best features.

Model Choice

To start, we briefly introduce the four models and the ensemble method we implemented.

Naive Bayes: Bernoulli and Multinomial Naive Bayes are both based on the idea that all features are independent of each other. P(X|Y) P(X1|Y) P(X2|Y) P(X3|Y) ····· P(Xn|Y)*P(Y). In our project, probability of each class is calculated based on its TF*IDF score value.

Logistic Regression: LR is a discriminative learning method that is trained with the data to the learn the conditional distribution P(Y|X) directly. To nd such a series of weights, so that linear approximation in $(W^T * X)$ maximize the most likelihood function, initial weights are adjusted through the gradient descend process.

Support Vector Machine: SVMs are a supervised learning models which aims to find a best hyperplane with (n-1) dimension to separate different samples(make a classification). In our experiment, Linear SVM is adopted to investigate the problem.

Ensemble methods: Ensembles are utilized to enhance the performance of our models. Ensembles methods achieve the performance by reducing the variance while not letting bias go too high, or combines the merits of 'complimentary' models to form a meta-classifier and produce the best result[10].

All tested models in this report are imported from sklearn library, while we also wrote a Bernoulli Naive Bayes from scratch. After trying several different hyper-parameter, we found that both LR and LSVM produce acceptable results. But, MultiNB and Ensemble Classifier performed much better than them.

For MultiNB, if we choose a relatively small alpha(smoothing level), the training accuracy will be dramatically high, but the overfitting problem will occur. Therefore, we chose alpha=0.22 after experimenting.

For the ensembles methods, they performed well too when applied to our models. But we still

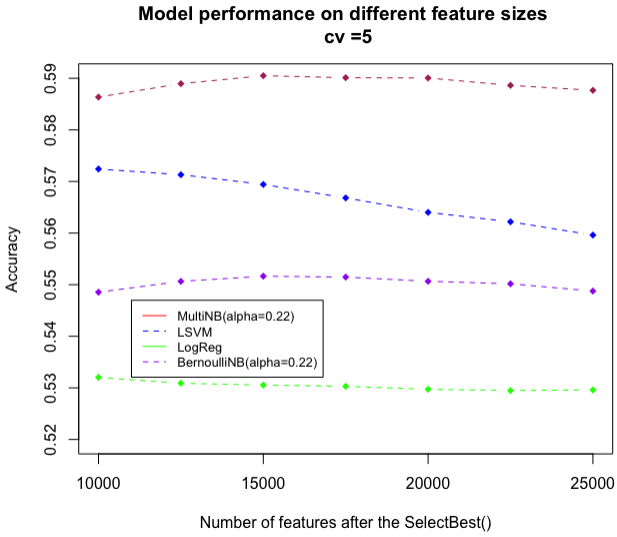

Fig. 2: Comparison of each model over different kbest selected feature size.

encountered a terrible overfitting problem in this model. The test accuracy is significantly lower than training accuracy, where the cross validation accuracy is around 59% while the test accuracy is only 57%.

IV. RESULTS

We start by comparing the accuracy of different base models. Here(Fig 2) we include four supervised classification models of Bernoulli naive Bayes, Multinomial Naive Bayes, Linear SVM and Logistic Regression. Overall, Multinomial Naive Bayes has the best performance since it reaches a highest cv accuracy among the four models, following is Linear SVM, Bernoulli Naive Bayes and Logistic Regression. However, when we submitted our prediction using Multinomial Naive Bayes, our accuracy score dropped by 1.5% compared to our cross validation accuracy. This can be due to overfitting on the training set. We believe that our model performance would enhance if we can obtain larger training dataset.

We also evaluate the models by observing their training runtime. Table I suggests that Logistic Regression has the fastest training time of 21.188 seconds when training with 17000 features, while Bernoulli Naive Bayes has the slowest training speed with 31.04 seconds. Yet we notice that Logistic Regression has the relatively

Models	LSVC	LogReg	MultiNB	BernoulliNB
Runtime(seconds)	31.007	21.188	25.592	31.04
Accuracy	0.568	0.530	0.590	0.552

TABLE I: Runtime and accuracy comparison of supervised classification models with feature selection of size 17000.

low accuracy. Though the model is fast, we do not recommend to use Logistic Regression for multiclass classification problems.

In our experiment we also developed some interesting founding regarding bagging method using Support vector machine. Support vector machine is a popular algorithm in finding the best hyperplane when solving the classification problem. After the dimension reduction, we try to utilize this model in dealing our issue of text classification. We expect to find a reliable linear separation to our existing training data, so we adopt LSVM version of it.

In our investigation, we find that the voting through the bagging method markedly increases the accuracy of our LSVM model. The accuracy based on cross validation rises from around 0.570 given by the unvarnished LSVM to around 0.590 at most, while bagging doesn't enhance and weaken the performance of the other models like Multinomial Naïve Bayes and Logistic Regression(decrease by 0.0-1.6%). The result suggest that variance of LSVM model is effectively cut down by our bagging and LSVM benefit from the randomness and variance reduction, and also the possible reason behind the result might be that Multinomial Naïve Bayes and Logistic Regression are relatively 'stable' algorithms which gives low variance, and the introduced randomness outweighs the likely advantages brought by the decreased variance.

Enough attention should be raised in the parameter adjustment in the bagging method especially when it comes to number of samples. According to Fig3, Low sample number such as 6000 results in underfitting while high sample number of 20000 results in overfitting(only randomness increased but not much variance reduced). Bagging method is less versatile to the number of estimators changes(Fig4).

Note: all the prediction accuracies are generated using K-fold cross validation with 5 folders to ensure that the evaluation result are

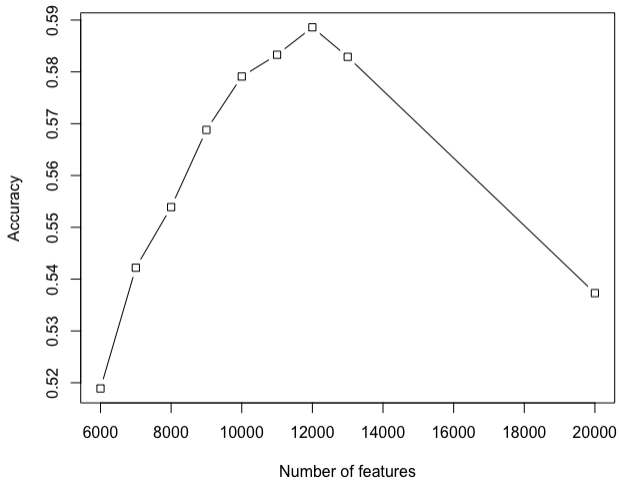

Fig. 3: Effect of number of features on bagging used in LSVM. (n_estimators = 24).

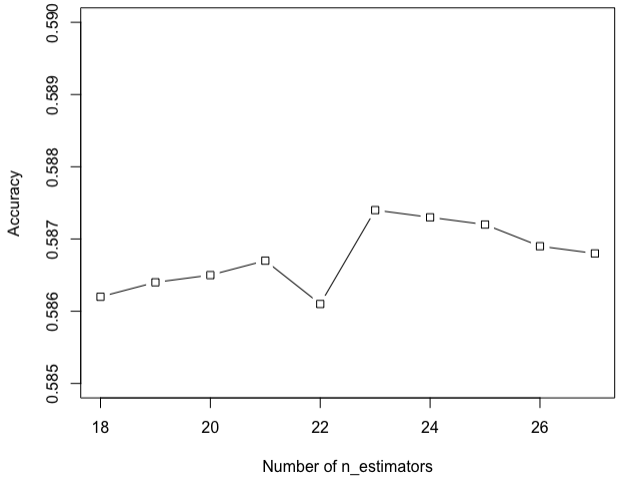

Fig. 4: Effect of number of estimators on bagging used in LSVM (number of features = 10000).

relatively accurate.

V. DISCUSSION AND CONCLUSION

In this project, we find that the Multinomial Naive Bayes gives the best prediction and is computationally fast to run compared to other models. However, there are still many possibilities to give the better prediction by better choice of parameters in the models, adopting more complicated algorithm to preprocess our data and even using deep learning.

Ensemble methods are useful if they are applied properly. Stacking avoids overfitting by reducing variance. Bagging could be useful under certain models and conditions like LSVM by the randomness introduced as well as the reduction in the variance. In the meantime, negative impact will be brought in if we do not take the property and traits of our model into consideration.

In addition, much attention are required for the underfitting and overfitting issues enough even if the change of some parameter doesn't influence the result much.

VI. STATEMENT OF CONTRIBUTIONS

1) Lihui Huang. Implementation: responsible for feature extraction and model experiments. Write-up: responsible for the Dataset and setup, Results and Statement of contributions.
2) Yiran Wang. Implementation: mainly responsible for model experiments. Write-up: responsible for the Proposed approach.
3) Jiewen Liu. Implementation: responsible for programming the Bernoulli Naive Bayes model and model experiments. Write-up: Contributed to the whole report write up, mainly responsible for the Abstract, Introduction, Related Work and Discussion and Conclusion.

REFERENCES

[1] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[2] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[3] Cen Li. Classifying imbalanced data using a bagging ensemble variation (bev). In *Proceedings of the 45th annual southeast regional conference*, pages 203–208. ACM, 2007.

[4] Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[5] Henry Oliver Lancaster and Eugene Seneta. Chi-square distribution. *Encyclopedia of biostatistics*, 2, 2005.

[6] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[7] Yanon Volcani and David Fogel. System and method for determining and controlling the impact of text, November 13 2003. US Patent App. 10/376,680.

[8] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[9] Bo Pang and Lillian Lee. 4.1. 2 subjectivity detection and opinion identification. *Opinion mining and sentiment analysis*, 2008.

[10] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.