Ensemble approach for Complaint Classification

Surayya Obaid, Anum Ilyas
Center for Computing Research Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi, Pakistan
*E-mail: anum.ilyas@juw.edu.pk

ABSTRACT

Businesses need to stay aware of their customer's perspectives or views about the offered products and services. User feedback is a common means of connecting with the users. Feedback and complaints are analyzed to extract different angles of information in order to fill up identified gaps and to work on downsides of business. Manually processing, classifying, and analyzing every single complaint is time-consuming and tedious. This problem can be addressed with the help of Machine Learning Algorithms. These algorithms offer a solution for a wide range of text classification issues. Complaint classification using machine learning algorithms has been used for years to make the feedback useful enough to reach key problems or issues regarding services. This paper covers an experiment conducted to observe the impact of using ensemble classifiers for complaint classification in contrast with single classifiers. The study was carried out using a dataset obtained from the Consumer Financial Protection Bureau, the U.S with 16217 instances of complaints. Performance of weak learners including Naïve Bayes, Logistic Regression and Support Vector Machine as well as ensemble classifiers including Random Forest and Adaboost is discussed in the paper. Adaboost with decision trees as base estimators, gave the best accuracy i.e. 85.02%.

Keywords: Complaint Classification; Ensemble Approach; Text Classification

INTRODUCTION

Text classification is highly applicative in many scenarios. Data generated in the form of reviews, responses, feedback, complaints, etc. from users end can be studied and inspected to extract patterns (Ullah et al. 2022). From product mining to sentiment analysis (Singh, Singh, and Singh 2017) to complaint classification, all forms of natural language processing come under text classification.

Complaint classification is one way of analyzing what major issues are in product or services any organization offers. Categorizing complaints in classes can be done by implementing machine learning algorithms. This type of problems lies in the category of text of classification.

Algorithms' performance varies with nature of problem where one may perform better than other one that gives accurate results in another domain. Selection of algorithm impacts on final results significantly. Complaints are registered using Natural Language regardless of its nature or platform. For classification, it is required to apply necessary protocols on the data prior to training to get better and precise results. Text classification usually involves certain preprocessing technique or combination of them before training of data such as tokenization, stop words removal, stemming, etc. (Kannnan and Gurusamy 2014).

In our experiment, we applied three machine learning algorithms after basic pre-processing. The dataset was then trained using ensemble learning algorithms (Dietterich 2002). These algorithms work by running a baseline algorithm multiple times to get results by taking an average of all or by voting. Ensemble learning algorithms are well suited for problems where better predictive performance is required such as low error in regression or high accuracy in classification (Suleymanov and Rustamov 2018). Ensemble Classifiers (Fayaz et al. 2020) are a better choice where single predictive models do not give satisfactory results. As complaint classification is division of text classification, presence of irrelevant, redundant and noisy features in the data is certain that eventually reduces overall performance of implemented algorithms (Field 2012). Using ensemble classifiers makes it possible to apply multistep and recursive learning process for categorization in regularized classes.

METHODOLOGY

The study's methodology is illustrated in Figure 1. The dataset, comprising over 16 thousand records, was selectively utilized to extract the desired subset of data. Prior to model training, NLP preprocessing steps were employed to enhance data quality. Initially, three weak learners—Naïve Bayes, Support Vector Machine, and Logistic Regression algorithms— were implemented. Subsequently, ensemble algorithms, namely Adaboost with varying numbers of estimators and Random Forest, were applied, as detailed in Table **1**.

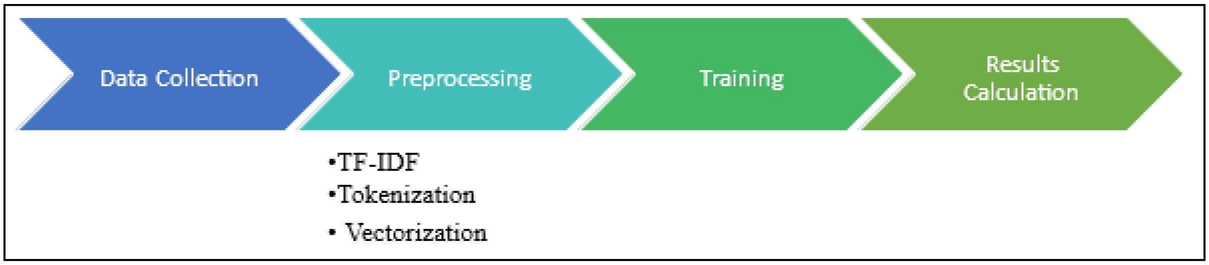


Figure 1. Methodology followed for the experiment.

Data Collection and Partitioning

Dataset we have used in this experiment is obtained from US Consumer Financial Protection Bureau. A subset of that dataset is used including more than 16217 records, in this experiment. We have used debt collection (4197), mortgages and loans (4740), credit cards (3272) and retail banking (4008). For training and testing, ratio of 80:20 is used. Common partitioning ratios for training and testing are 50:50, 67:33 and 80:20 for machine learning problems. In our experiment, we used 80:20 division for testing and training.

Preprocessing

For text classification, we have applied various preprocessing steps such as tokenization, stop words removal, special characters removal, stemming, and tokenization that are typically employed for natural language processing (Chai 2023) with the help of predefined functions of python libraries.

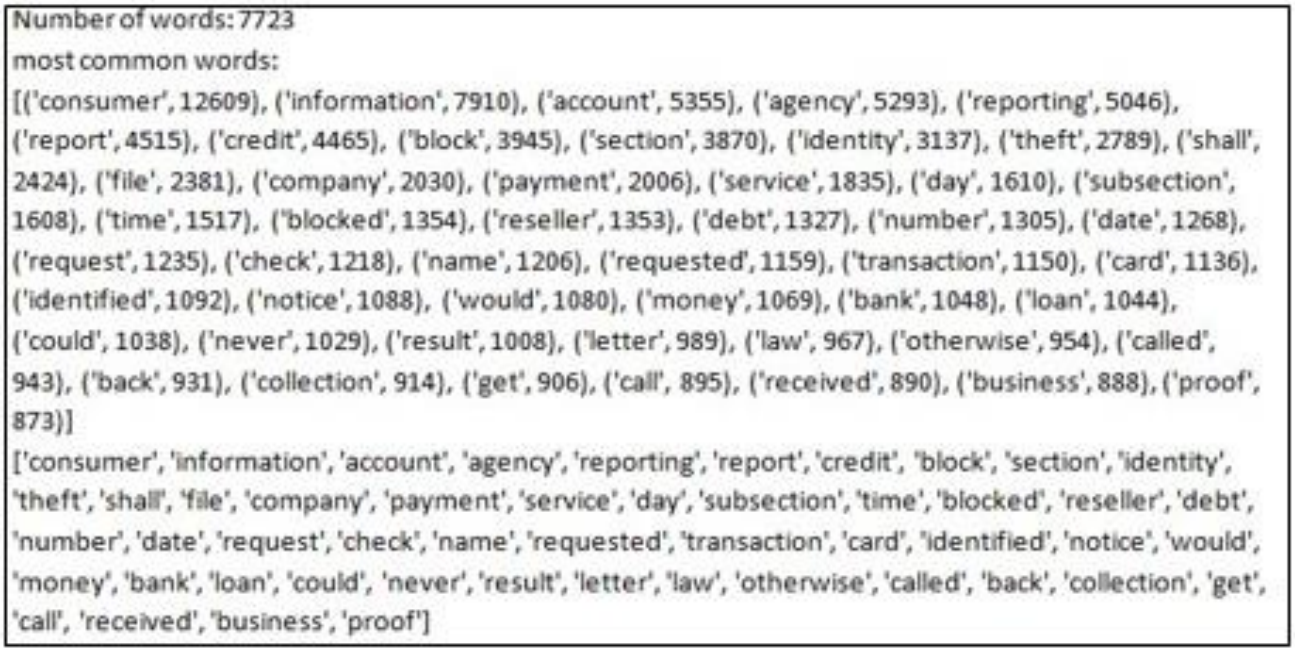
Features Selection

This process is done by multiple means such as count vector and Term Frequency- inverse Document Frequency. Count Vector is a matrix notation of the dataset in which every row represents a document from the corpus, every column represents a word from the corpus, and every cell represents the frequency count of a particular token in a particular document. TF-IDF score represents the relative importance of a word in the document and is calculated as the number of times that token ‘w’ occurs in an article ‘a’, summed across all the articles in a particular class.

RESULTS AND DISCUSSION

After preparatory steps such as TF-IDF vectorization, lemmatization, tokenization and stop words removal, list of most common words was generated as shown in Fig. 1. Weak Learners including Naïve Bayes, Logistic Regression and Support Vector Machine (SVM) were applied before ensemble classifiers. Support Vector Machine (SVM) outperformed all weak classifiers with accuracy score of 85.0% as shown in Table. 1. Naïve

Bayes did not perform well although it is considered good classifier. Naïve Bayes assumes every feature equally important and independent which makes it faulty in some scenarios.



```
Number of words: 7723
most common words:
[('consumer', 12609), ('information', 7910), ('account', 5355), ('agency', 5293), ('reporting', 5046),
('report', 4515), ('credit', 4465), ('block', 3945), ('section', 3870), ('identity', 3137), ('theft', 2789), ('shall',
2424), ('file', 2381), ('company', 2030), ('payment', 2006), ('service', 1835), ('day', 1610), ('subsection',
1608), ('time', 1517), ('blocked', 1354), ('reseller', 1353), ('debt', 1327), ('number', 1305), ('date', 1268),
('request', 1235), ('check', 1218), ('name', 1206), ('requested', 1159), ('transaction', 1150), ('card', 1136),
('identified', 1092), ('notice', 1088), ('would', 1080), ('money', 1069), ('bank', 1048), ('loan', 1044),
('could', 1038), ('never', 1029), ('result', 1008), ('letter', 989), ('law', 967), ('otherwise', 954), ('called',
943), ('back', 931), ('collection', 914), ('get', 906), ('call', 895), ('received', 890), ('business', 888), ('proof',
873)]
['consumer', 'information', 'account', 'agency', 'reporting', 'report', 'credit', 'block', 'section', 'identity',
'theft', 'shall', 'file', 'company', 'payment', 'service', 'day', 'subsection', 'time', 'blocked', 'reseller', 'debt',
'number', 'date', 'request', 'check', 'name', 'requested', 'transaction', 'card', 'identified', 'notice', 'would',
'money', 'bank', 'loan', 'could', 'never', 'result', 'letter', 'law', 'otherwise', 'called', 'back', 'collection', 'get',
'call', 'received', 'business', 'proof']
```

Figure 2. List of common words found in complaints

Figure **2**. depicts list of common words that were found while training. Ensemble learning algorithms were applied afterwards that included Random Forest and Ada-boost. At first, number of estimators were limited to 40 with Decision tree as base estimator which gave 82% prediction accuracy. The algorithm performance was improved by increasing the count of estimators but base estimator remained same as before. This slight change refined accuracy score by 3 percent yielding 85.02%, highest as compared to all.

Table 1. Accuracy Score of Weak Learners (WL) and Ensemble Classifiers (EC).

	Classifiers	Accuracy Score%
WL1	Naïve Bayes	57.33
WL2	Logistic Regression	83.2
EC1	Random Forest	82.13
WL3	Support Vector Machine	85.0
EC2	Adaboost (100 Estimators, Decision Tree)	85.02

Table **2** summarizes category-wise classification report which shows precision, re-call and F1 score of all four categories for each algorithm. Naïve Bayes gave best results for precision i.e. 1 for three categories (Mortgages and loans, Credit cards and Retail banking) while recall was observed lowest for Mortgages and loans (0.03), lowest among all.

Table 2. Category-wise classification report.

	Precision	Recall	F1-score
	Naïve Bayes		
Debt collection	0.91	0.15	0.26
Mortgages and loans	1.00	0.03	0.05
Credit cards	1.00	0.04	0.08
Retail banking	1.00	0.04	0.08
	Logistic Regression		
Debt collection	0.84	0.73	0.78
Mortgages and loans	0.88	0.76	0.82
Credit cards	0.75	0.60	0.67
Retail banking	0.76	0.75	0.75
	Support Vector Machine		
Debt collection	0.82	0.80	0.81
Mortgages and loans	0.86	0.84	0.85
Credit cards	0.71	0.68	0.69
Retail banking	0.73	0.80	0.77
	Random Forest		
Debt collection	0.84	0.70	0.76
Mortgages and loans	0.93	0.67	0.78
Credit cards	0.80	0.58	0.68
Retail banking	0.71	0.63	0.67
	Adaboost		
Debt collection	0.87	0.85	0.80
Mortgages and loans	0.83	0.85	0.87
Credit cards	0.79	0.83	0.79
Retail banking	0.85	0.85	0.83

CONCLUSION

Addressing complaints or reviews on products and services can significantly improve quality of services. Automatic classification can help in identifying key areas that need attention to improve customer satisfaction. Our study showcases the findings of experiments we conducted over US Consumer Financial Protection Bureau dataset. The experiment comprised a comparative analysis of the performance of weak learners; Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM), as well as ensemble classifiers such as Random Forest and Adaboost. Among all, Adaboost and Support Vector Machine gave the best accuracy score of around 85%. Random Forest, Logistic Regression and Naïve Bayes accuracy scores were 82.13%, 83.2% and 57.33% respectively.

REFERENCES

1. Ullah, Mohammad Aman, Korimunnesa Munmun, Fatematuz Zohra Tamanna, and Md Shahnur Azad Chowdhury. 2022. "Sentiment Analysis Using Ensemble Technique on Textual and Emoticon Data." 2022 International Conference on Innovations in Science, Engineering and Technology, ICISET 2022: 255–59
2. Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh. 2017. "Optimization of Sentiment Analysis Using Machine Learning Classifiers." Human-centric Computing and Information Sciences 7.

3. Kannnan, S, and Vairaprakash Gurusamy. 2014. “Preprocessing Techniques for Text Mining.” International Journal of Computer Science & Communication Networks (February 2015): 7–16.
4. Dietterich, Thomas G. 2002. “Ensemble Learning.” In The Handbook of Brain Theory and Neural Networks, Second Edition., http://mitpress.mit.edu.
5. Suleymanov, U., and S. Rustamov. 2018. “Automated News Categorization Using Machine Learning Methods.” IOP Conference Series: Materials Science and Engineering 459(1).
6. Fayaz, Muhammad et al. 2020. “Ensemble Machine Learning Model for Classification of Spam Product Reviews.” Complexity 2020.
7. Field, Andy. 2012. “Logistic Regression Logistic Regression Logistic Regression.” Discovering Statistics Using SPSS: 731–35.
8. Chai CP. Comparison of text preprocessing methods. Natural Language Engineering. 2023;29(3):509-553. doi:10.1017/S1351324922000213
9. Salar-García, M. J., Walter, X. A., Gurauskis, J., de Ramón Fernández, A., &Ieropoulos, I. (2021). Effectofiron oxide content and microstructuralporosityonthe performance ofceramicmembranes as microbial fuel cellseparators. Electrochimica Acta, 367. https://doi.org/10.1016/j.electacta.2020.137385
10. Zhang, Z., Zhang, Q., Ren, C., Luo, F., Ma, Q., Hu, Y. S., Zhou, Z., Li, H., Huang, X., & Chen, L. (2016). A ceramic/polymer composite solidelectrolyteforsodiumbatteries. JournalofMaterialsChemistry A, 4(41). https://doi.org/10.1039/c6ta07590h