Financial Complaint Text Classification

Prof Ashwini M Joshi (Guide), Sarthak S, Sameer Kulkarni , Sanket Kattimani, Sumukha MK
Department of Computer Science and Engineering, PES University.

Problem Statement

The problem that we are trying to solve is to automatically classify financial complaints that should be routed to their respective departments using Machine Learning and Deep Learning Algorithms.

Financial complaint classifier application will mainly help bank employees who are under consumer complaint department to easily classify the complaints they receive to respective departments of the bank.

Background

The authors does not describe how they handled class imbalance problem.

Word-level CNN can be useful and requires less training time compared to Character level CNN.

Bidirectional-LSTM model can be applied as it handles context better from sequential data like text.

Dataset and Features / Project Requirements / Product Features

Dataset: Consumer Financial Protection Bureau government website

There are 9 categories in total

Techniques used : Data augmentation with Synonym replacement

Project Requirements: Complaint Text as input, Server should be running continuously.

Features : Display the category as output , Friendly User Interface.

Design Approach / Methods

User need to enter financial complaints in english Pre-processing techniques : lower case convertion,stopword removal and lemmatization

Data is Converted to TF-IDF format and applied to ML and DL models(uni-gram and bi-gram).Same data is used for ensemble models such as Random Forest and XG Boost.

Final model used is Hard Voting Classifier model of Logistic Regression,Naive Bayes and Linear SVM with 1,00,000 sample complaints.

Results and Discussion

Model	Unigram Accuracy	Bigram Accuracy
Multinomial Naive Bayes	79.99%	82.94%
Decision Tree	71.33%	65.35%
Linear SVM	83.51%	85.35%
Logistic Regression	83.03%	81.71%
K-Nearest Neighbours	73.76%	69.10%

Model	Accuracy with 100 estimators	Accuracy with 500 estimators
Random Forest	83.83%	84.78%
XG Boost	82.37%	84.68%

Model	Accuracy
Word level CNN	84.25
LSTM	83.72
Bidirectional LSTM	84.13

Hard Voting Classifier of Linear SVM, Random Forest and XG boost gave highest accuracy of 85.91% but took more time to process complaint. Hard Voting Classifier with 1,00,000 sample complaints with Linear SVM, Naive Bayes and Logistic Regression gave accuracy of 82.46% and also processed complaints much faster.

Summary of Project Outcome

Using data augmentation increased model performance.

Bi-gram models performed better than uni-gram models but increased model complexity.

Ensemble models performed better than traditional ML and DL models in terms of both accuracy and time.

The given complaint text is categorized to valid department of the bank.

Conclusions and Future Work

The choice of the model is based on user need based on speed and accuracy.

The future work could be improvising speed while keeping accuracy as high possible and building fully working software for the finance companies.

Future Work can also include accepting other languages and not only English, improving the accuracy of prediction even more.

References

M. Ali Fauzi,"Automatic Complaint Classification System Using Classifier Ensembles",Telfor Journal, Vol. 10, No. 2, 2018 [4]

The future Pavels Goncarovs,"Active Learning SVM Classification Algorithm for Complaints Management Process IEEE[5].

Prof. Ashwini M Joshi · Sarthak S · Sameer Kulkarni · Sumukha MK · Sanket S Kattimani