Detecting Bias in Lending Data with NLP Models

Stanford CS224N {Custom} Project - ***Option 1 Grading***

Senthil Selvaraj, Orion Darley
Department of Computer Science
Stanford University
sen07@stanford.edu, oriond@stanford.edu

Abstract

Consumer debt in the US reached over 13 Trillion Dollars [1] with home, auto, student and credit card loans being the major ones. Banks in the US are mandated by various regulations to have fair and consistent lending process. Despite all the earnest efforts, bias do exist in lending based on gender or race or zip code for comparable applications with similar income, credit and other relevant profiles. Our paper will provide a contextual word embedding analysis to identify bias using a BERT LSTM model that feeds into a fully connected network. We used zip codes as one of the input parameters to predict geographical bias. Our results so far with BERT prediction has 60 percentage accuracy and we will continue to fine tune the model to identify bias.

1 Key Information to include

- Mentor: Matt Lamm
- External Collaborators (if you have any):N/A
- Sharing project:N/A

2 Introduction

Primary goal of this paper is to establish a baseline through quantitative word embedding analysis on how to identify/detect the discriminatory practices that exist in consumer lending decisions in the US. Current detective and preventative controls on bias rely on banking governance oversight and audit / regulatory exams. This identifies areas where there is obvious violation that a loan is not approved for one application just because the race or gender is different. This paper will attempt to establish the "unconscious or inadvertent bias" through contextual word embedding analysis by studying data in sub-vector space.

With recent progress made in NLP word embedding space, we have seen evolution of contextual word analysis with models such as Word2Vec, GloVe, Fasttext, ELMo and BERT. Most of the work has been done to detect gender and racial bias on general areas of human interactions (from Twitter feeds or Wikipedia content). Our paper has leveraged those models to apply to a new use-case on consumer lending.

With the way we have uniquely designed our model, we first study the text description of consumer information in plain English at sub-vector level, then feed other input parameters to predict the output. Zip code is one of the main input parameters we used in our study as it represents a wide spectrum of ups and downs in our communities and reflects the differences in socio-economic status of people living in a particular area.

3 Related Work

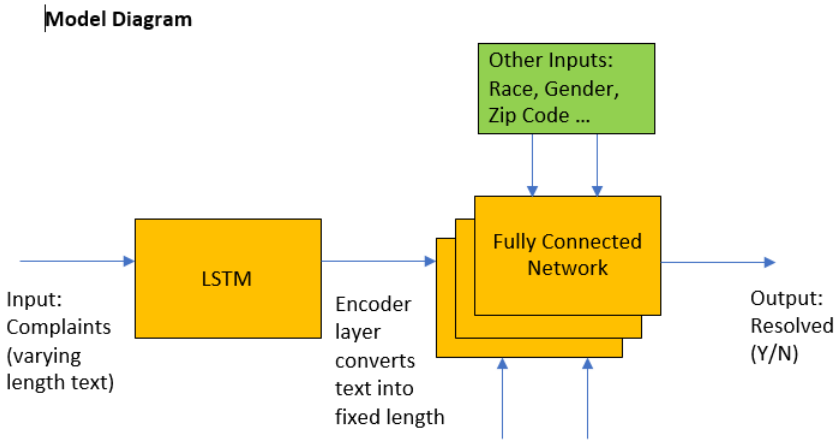
Word embedding is an interesting NLP area that represents word semantics in vector based representations. The 2016 paper by Bolukbasi et. al. [2] shed new light on the inherent bias that exist in data and how those biases are amplified with machine learning / word embedding models. The paper is titled on a biased analogy "Man-to-Computer Programmer is to Woman-to-Homemaker" to show how bias exist in word embedding. This can be attributed to the historical bias that exist (which are captured in the training data) or how models are built.

There has been an increased number of research papers on this topic and gained popularity in data science since the 2016 paper. The most relevant one for our project is the recent paper by Saket* et. al [3] that leveraged the Word Embedding Association Test (WEAT) (Caliskan et. al [4]) evaluation to post-process traditional and contextualized word embedding to associate word embeddings between concepts captured in Implicit Association Test (IAT) [5]. This paper used "conceptor debiasing" principles to remove both racial and gender bias. The paper addressed how to capture the geometric direction of the word embedding and how to linearly separate them from gender neutral words.

Our approach is slightly different from the one used in this research paper, which used general pre-trained datasets to test different models from Word2Vec, GloVe, and BERT to improve accuracy of identifying bias and de-biasing them. Our project is focused on a business need with a real-life use case on consumer lending and to identify bias using word embedding quantitative analysis.

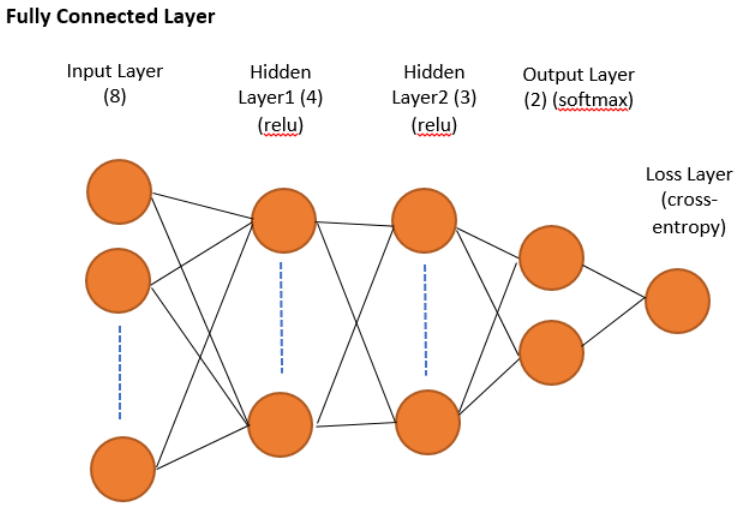
4 Approach

Our project objective is to explore consumer lending data and identify bias through contextual analysis. This will identify not only blatant violation in lending process but inadvertent bias (which we call unconscious) that exist in society. The project will be set up as a binary classification task. Though many papers have been published on gender and race bias, the project objective / execution on bank dataset is our original idea and we came up with our high level architecture model as shown in figure below.



We plan to use a BERT LSTM layer to sequence the input complaints narrative of varying length into a fixed length output through padding up to a maximum length. The Complaints dataset has four text fields that describe either consumer's challenges or company's response to those issues from consumers. The output from BERT is fed into a fully connected neural network as shown below. This network has an input layer, two hidden layers and two output units and a softmax output.

We plan to use BERT LSTM independently or with fully connected network; however we can also use any other simpler models to process the text narrative to accomplish our objective, which is to demonstrate that bias exists with good accuracy. Our approach is to explore the representations from LSTM and/or neural network can be used to classify geography through zip codes. This will suggest that output contains demographic information and can be used to prove any bias that may exist on certain geography.



5 Experiments

5.1 Data

We studied various datasets from government sites as banks are mandated to upload all loan details. We reviewed the HMDA [6] (Home Mortgage Disclosure Act) which has, for each record, details on loan, the property characteristics, the applicant demographics (including race and gender), and the lender information. This dataset does not have any text narrative field (in plain English words) that will describe the loan information.

We reviewed another dataset from CFPB [7] (Consumer Financial Protection Bureau) on consumer complaints and this has few text fields describing the nature of complaint. The data dictionary is shown below:

• Column A-Date received, B-Product, C-Sub-product,

• Column D-Issue, E-Sub-issue,

• Column F-Consumer complaint narrative, G-Company public response,

• Column H-Company, I-State, J-ZIP code, K-Tags,

• Column L-Consumer consent provided?, M-Submitted via, N-Date sent to company,

• Column O- Company response to consumer,

• Column P-Timely response?,

• Column Q-Consumer disputed?, R-Complaint ID

This dataset does not have race and gender but has other characteristics such as zip code, which can be used to detect geographical or class bias. We are using this dataset for our project and as future extension, explore other government / bank datasets to identify race and gender bias.

We filtered the database for last 5 years with 1.05 million records. We plan to use 800K for training set and 100K each for validation and test sets. However, we used a smaller number of records for milestone experiment. Here are the details of various parameters for our model:

• Inputs to LSTM: Columns D through G have text narrative of issues and responses.

• Inputs to Fully connected network (Appendix C): product and subproduct (columns B, C), company H, state I, zip J, consent provided L, submitted channel M.

• Output: column O to look for closed with explanation / monetary / non-monetary relief. Second output is column P for timely response.

As our project used a unique bank/government database that has never been used before in any research papers, we ran into basic data quality issues. We completed data pre-processing as listed below:

1. Loading data, removing outliers, labels with very little sample sizes, dropping unnecessary variables, revaluing categorical labels to numerical, concatenating text as predictors.

2. Normalizing the text column by transforming whitespaces to spaces and uncasing letters.

3. Artificially balance the dataset, for example by up-sampling or down-sampling each class. Both minority class up-sampling and majority class down-sampling approaches were used in variations of our approaches. Up-sampling is the process of randomly duplicating observations from the minority class in order to reinforce its signal whereas down-sampling involves randomly withdrawing observations from the majority class to prevent its signal from dominating the learning algorithm. For our experiment we concluded it would be most desirable to have a classifier produce high prediction accuracy over the majority class, while predicting reasonable accuracy for the minority classes, but both approaches were tried individually.

4.Tokenizing text or dividing the sentences into individual words. This also includes splitting all punctuation characters, removing special characters, removing misspelled words, and removing strings that contained pieces of words combined with random special characters from the text.

5. Adding CLS and SEP tokens to distinguish the beginning and the end of a sentence.

6. Dividing words into WordPieces based on similarity (i.e. transform “pulling” [“pull”,“ing”])

7. Mapping words in the text to indexes using the BERT’s own vocabulary which is saved in BERT’s vocab.txt file.

5.2 Evaluation method

Our research paper on gender and race bias used six sigma metrics to state null hypothesis. Previous research used datasets pre-trained on Wikipedia and Twitter feeds and evaluated models on cosine distance/similarity between target word pairs (ex. men or women) and attribute words (ex. occupation such as doctor or nurse).

Our model did not use this approach as we trained our own data from Government Complaints dataset. Our evaluation strategy is to use our model as a prediction machine. The input parameters passed into BERT and fully connected network and the output (ex. company’s response to the complaint) is predicted. This is evaluated by comparing with the actual results from the dataset on how the bank’s responded to and resolved customers’ complaints. Our goal is to improve the accuracy of prediction not limited to improving quality of text through data pipeline pre-processing, trying multiple modeling approaches, parameter and hyper-parameter optimization, etc.

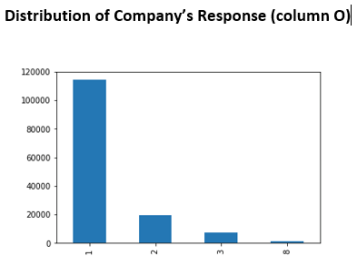
We also tested four different machine learning supervised classification models: Random Forest, Linear Support Vector Classifier (SVC), Multinomial Naive Bayes, and Logistic Regression. We used a near identical approach for text-preprocessing emphasized in the experiment section.

In order to return meaningful, we elevated the zip codes at state level. For this experiment, we selected seven states that either had a large number of observations or were geographically separated in the continental United States to include: TX, CA, NY, FL, GA, PA, IL.

5.3 Experimental details

Initially we used BERT with input parameters of "issue". In our training, BERT predicted the bank’s “response to the consumer” (response variable), which has multiple classification of issue closure: “close with explanation” (1), “closed with non-monetary relief” (2), “closed with monetary relief” (3), and “in-progress” (8). We use this output to predict the accuracy and compare them to actual results. This will be analyzed to see if the representation has any geographical bias in them.

The distribution shown here. Notice the first category of issue closure is skewed and we will consider adjustments in data. The initial results with the single variable were poor, so more variables were explored and further model iterations were made. We extended the range of inputs to the model by adding another input parameter "sub-issue" and concatenated this to existing input "issue". Concatenating these two input text variables provided a broader description of the customer's concerns with their financial providers. Another input included was "product id", containing information on the type of product the dispute was directly referring to. We also followed the general text-preprocess and modeling guidelines outlined (see experiments section) and concatenated the "product-id" with "sub-product". "Zip code" was also used but produced similarly poor results.



Considering the model produced poor results for this response variable, we changed the response from "response to the consumer" to "zip code". This also produced low f1-scores in the 0-5 percent range. The "zip code" distribution was highly imbalanced, so observations under 40th percentile were remove to help establish a more balanced dataset. Ultimately this response variable was substituted for "state" which modeling approaches produced the best results on. The more significant results are shown next section. This experiment and results put us closer to our project objective to identify geographical bias using zip codes.

ML Model Experiment: In addition to the aforementioned deep learning model, we tested the four supervised ML classifiers as explained in previous section. This provided us a comparative analysis to validate our objective and to pick the best-fit model. Zip codes results are at a very narrow level and we cannot see meaningful differences in lending decisions to identify potential bias, so we aggregated them at state level for 7 select states. The results are shown in next section.

5.4 Results

For our experiment, we tried few iterations by varying the parameters - sample size, input column (s) and so on. Hyperparameters for this model are as follows: sample size of 25,000, sentence sequences to be at most 128 tokens long, batch size of 32, learning rate of 0.0001, number of training epochs = 3, warmup proportion equal to 0.1. Results are to improve prediction of the company response (column O) and measured in terms of evaluation accuracy, false positives/negatives and loss, true positives/negatives and number of steps. Accuracy reached just over 80 percent The details are shown below:

Exp1: First trial with 25K training set, one column input.

'eval$_accuracy'$: $0.80179447,'false_negatives'$: $7091.0,'false_positives'$: $0.0,'loss'$: $0.6497155,'true_negatives'$: $28685.0,'true_positives'$: $0.0,'global_step'$: 13415

Exp2: Added 2nd variable product+subproduct Pred on classes 1,2,3 Random sample of 25k

LEARNING$_RATE = 0.001NUM_TRAIN_EPOCHS = 5$

'eval$_accuracy'$: $0.80877095,'false_negatives'$: $1016.0,'false_positives'$: $0.0,'loss'$: $0.59981775,'true_negatives'$: $4297.0,'true_positives'$: $0.0,'global_step'$: 2490

Training Set Shape : (15937, 3) Validation Set Shape : (5313, 3) Test Set Shape : (3750, 2) Index(['issue2', 'product2', 'y'], dtype='object')

Training took time 1:06:47.828785 K80 GPU

Exp3: Pred on classes 1,2,3,8 Sampled first 125000 rows after filter

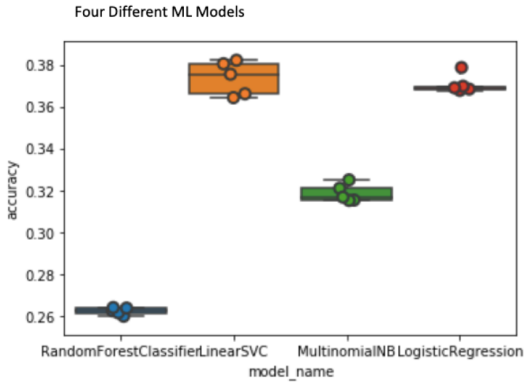
LEARNING$_RATE = 0.0001NUM_TRAIN_EPOCHS = 3.0 * 4$

'eval$_accuracy'$: $0.78346664,'false_negatives'$: $4060.0,'false_positives'$: $0.0,'loss'$: $0.6981902,'true_negatives'$: $14690.0,'true_positives'$: $0.0,'global_step'$: 28125

Training Set Shape : (75000, 2) Validation Set Shape : (18750, 2) Test Set Shape : (31250, 1)

Training took time 3:50:55.134098 P5000 GPU

Results of ML Models:
The results are displayed below with their precision, recall, f1-scores We are able to get a precision and accuracy around 30 to 40 percent for these states to predict the zip code.



	precision	recall	f1-score	support
TX	0.34	0.37	0.35	1179
CA	0.40	0.52	0.46	1708
NY	0.32	0.25	0.28	709
FL	0.35	0.43	0.38	1284
GA	0.36	0.27	0.31	784
PA	0.35	0.11	0.17	438
IL	0.39	0.18	0.25	498
accuracy			0.36	6600
macro avg	0.36	0.31	0.31	6600
weighted avg	0.36	0.36	0.35	6600

6 Analysis

Our project started with an objective to identify bias in consumer lending information in the US. Based on the Complaints dataset we have chosen, we set out to focus on geographical bias with zip code as a parameter. We tried various modeling approaches – independent BERT models, BERT integrated into a fully connected network, and machine learning classifiers.

Our first sub-task in these experiments is to focus on predicting company's response as output and comparing this with actual results. The deep learning approach achieved poor results with overall accuracy of 80 percent, however each deep learning model's recall and precision remained zero (zero true positives) after multiple versions and iterations. As explained in section 5.3 Experimental Details, we changed the output from predicting consumer responses to zip codes and then aggregated to states to get the best results.

Even though we believe we are directionally correct, we need to do more work on data quality. Consumer lending with CFPB/HMDA datasets are not pre-trained for NLP analysis, hence more work is needed in this area. Next steps would also include exploring deep learning models alternative to BERT and integrate the fully connected layer in order to achieve better results.

The machine learning models showed improved results over the deep learning approaches. The input is paired with the response variable(s) zip code and/or state with the intent on discovering possible existing bias exist in the customer complaints predictor variables. The Support Vector Classifier (SVC) was the highest performing model producing an f1-score at 35 percent, with precision scores between 32 and 40 percent and recall between 11 and 52 percent. We think this may because of the way SVC classifier works by creating a "best fit" hyperplane that divides the data points.

Random Forest classifier has an accuracy score at lower end compared to the four supervised ML algorithms as we think this algorithm works with multiple decision tree, aggregates them to give a stable prediction. Logistic Regression model also predicted accuracy at a higher range. Multinomial Naïve Bayes model has plotted accuracy that is between the above models given its assumption of conditional independence between every pair of features between zip codes and product IDs.

As mentioned for deep learning, ML classifiers models also need more work on data quality with a better understanding of decisions made for various zip codes. We also need to train the models with a larger dataset.

7 Conclusion

As pointed out in previous sections, we set out to do a unique word embedding use case with consumer lending and detecting bias by training our model with government dataset. We learnt how much data quality is important for training NLP word embedding models by conducting a heavy cleanup and preprocessing of the data. We tweaked the parameters of BERT and fully connected layers to achieve around 80 percent accuracy but the true positives are not convincing. The ML classifier models give us a prediction accuracy in 30-40 percent range.

As future work, we recommend to improve the quality of the dataset used for this experiment. Model hyperparameter optimization along with architectural deisgn should be further examined to achieve an improved accuracy for this dataset to predict if geographical bias exists with zip code or state as one input. Additional research can be extended to identify race/gender bias by exploring other government datasets.

References

[1] Federal Reserve Bank. https://www.federalreserve.gov/releases/g19/current/.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[3] Saket Karve, Lyle Ungar, and João Sedoc. Conceptor debiasing of word representations evaluated on weat. *arXiv preprint arXiv:1906.05993*, 2019.

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[5] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[6] Home Mortgage Disclosure Act Data Collection. https://www.consumerfinance.gov/data-research/hmda/.

[7] Consumer Financial Protection Bureau Complaint Database. https://www.consumerfinance.gov/data-research/consumer-complaints/.