Final Project Report

CS 522, Akshaya Brian Tauro (A20502097)

May-05-2023

Section 1: Project goal and project set up

1) Project Goal: To evaluate the performance of classifiers based on their sentiment predictions towards certain category (like gender, race) of text.

2) Hypothesis: "A benchmark dataset can detect bias in the trained sentiment classifier towards specific categories of text" can be tested and proven true or false. Accept hypothesis if confirmed else reject.

3) Approach: Download, pre-process the twitter sentiment and EEC corpus dataset(Data pre-processing differ when using traditional classifiers and LLM based classifiers). Then, split the dataset in to train, validate and test set. Use the split to train, validate and test the classifiers, using the models, evaluate by measuring the bias on EEC (a benchmark dataset). Finally, compare and analyse the performance of the classifiers, to conclude on the classifier that produces accurate and fair results for sentiment analysis.

4) Setting up the following four classifiers SVM, Logistic regression, mBERT and RoBERTa to train on twitter sentiment dataset and to evaluate the bias and accuracy of these models using ECC Corpus dataset. The main difference involves in the way they are processing the data and learning methods. SVM and Logistic Regression are traditional classifiers whereas mBERT and RoBERTa are LLM based classifiers. The process of feature engineering in traditional classifiers involves manual effort to create features where domain knowledge is required whereas LLM classifiers use neural networks and pre-trained models to learn the features automatically from the input data.

5) Using the classifiers that were trained and validated on Twitter sentiment analysis dataset, evaluate its performance on EEC test set. We will evaluate each classifier based on the following metrics: accuracy, F1-score, recall and precision. To compare, we will list all the results in table format and display the results in graphical format, as necessary. The results will help us to compare the classifiers performance based on their sentiment predictions towards certain category.

6)

Dataset Name	Size of the dataset	# of positive examples	# of negative examples	# of Neutral
Twitter sentiment analysis	1578627	788442	790185	-
Equity Evaluation Corpus (EEC)	8640	2100 (joy)	6300 (sum of emotions with = anger, sadness, fear)	240 (Neutral)

Dataset - Twitter Sentiment Analysis	Training	Testing	Validation
	5k	2k	10k
Positive Class	2.5k	1k	5k
Negative Class	2.5k	1k	5k

7) We all have run the experiments using Google coLab notebook.

Section 2: Classifier training

Data	Size	Target	Training	Testing	Validation
Twitter Sentiment Analysis	1.6 million	0 for negative and 4 for positive	5k	2k	10k
Positive			2.5k	1k	5k
Negative			2.5k	1k	5k

Data Pre-processing

For SVM and LR classifiers, techniques my teammates used for pre-processing include tf-idf tokenization, stop word removal, lower casing, stemming, and removing HTML tags and punctuation, resulting in improved accuracy.

I used BERTTokenizerFast tokenizer with TFBertModel and RobertaTokenizerFast tokenizer with TFRobertaModel, the tokenizer is used to convert input text into a sequence of sub word tokens that can be fed into the corresponding model. BERTTokenizerFast uses the WordPiece algorithm whereas RobertaTokenizerFastuses the Byte-Pair Encoding(BPE) to tokenize text. WordPiece algorithm starts by breaking the input text into words and then applies a recursive sub word segmentation algorithm is used in BERTTokenizerFast that iteratively splits words into smaller sub words until the entire vocabulary is exhausted or the desired number of tokens is reached, and BPE algorithm iteratively replaces the most frequent pair of bytes/characters with a new token until the entire vocabulary is exhausted or the desired number of tokens is reached. The resulting tokens are represented as integer token IDs and can be fed directly into the corresponding models. For BERT and RoBERTa generally raw data is given as input as they are designed to handle raw text inputs without any additional pre-processing. Note, there was no issue when the input data was around 100 while training and validating the BERT and RoBERTa using twitter sentiment analysis dataset. However, I did face an issue when the input data was 15k. Hence, I ended up removing emojis, new line characters, links, mentions, non-utf8/ascii characters, hashtags, multiple spaces, and also converted the text to lowercase. Another point to be noted is, I did one-hot encoding to create an binary column for each category.

For BERT, my teammate did another experiment using BertTokenizer, which is again a class from the Hugging Face's transformers library that uses WordPiece sub word segmentation technique, without performing data pre-processing, did pass raw input.

Evaluation and F1

F-1 is a commonly used metric to evaluate the performance of binary classification models on a given dataset. It is a measure of the harmonic mean of precision and recall of the model(src), which takes into account both false positives and false negatives. F-1 score ranges from 0 to 1, with higher values indicating better performance. It is a useful metric for assessing the overall accuracy and effectiveness of a binary classification system.

Metrics	**Before and After Hyperparameter Tuning**							
	SVM		**LR**		**BERT**		**ROBERTA**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Precision	0.72	0.7407	0.72	0.72	0.82	0.78	0.82	0.82
Recall	0.73	0.73	0.73	0.73	0.78	0.83	0.8	0.82
f1-score	0.71	0.73	0.728	0.73	0.8	0.81	0.81	0.82

Classifier training details

I used the 'bert-base-uncased' model which is one of the base models of BERT. It has 12 layers and 110 million parameters. The BertTokenizer is used to tokenize the input text into subwords that are fed into the BERT model. For RoBERTa, 'roberta-base' pretrained the base model on the English language using a masked language modeling (MLM) objective. The texts are tokenized using a byte version of Byte-Pair Encoding (BPE) and a vocabulary size of 50,000. The inputs of the model take pieces of 512 contiguous tokens that may span over documents. The beginning of a new document is marked with <s> and the end of one by </s>.

Hyperparameter tuning for BERT and RoBERTA		
• Parameters that I opted for hyperparameter tuning are **learning rate** and **batch size**. • **Reason:** According to the paper, reducing the learning rate and decreasing the batch size can enhance the network's training performance, particularly during the process of fine-tuning. Note: Though the referred paper is to tune CNN for the medical image dataset. I did try this analogy while tuning for BeRT and RoBERTa, and noticed the improvement in scores by 1%.	Before Tuning	Learning Rate: 2e-5 batch size = 40
	After Tuning (Best Value)	Learning Rate: 1e-5 batch size = 30

Below are the hyperparameter tuning for SVM and LR done by my team mates:

SVM	• For SVM my team mate did chose C, gamma and kernel for hyperparameter tuning. • Reason: From https://deepnote.com/@bhavesh-bhatt/svm-c-gamma-hyperparameter-ec7cdd4f-b499-4b4d-a320-f483e8099691 , C is used to impose a penalty on misclassified points. If the value of C is low, the penalty for misclassification is also low, resulting in a wider margin boundary. If C is high, then SVM attempts to minimize the number of misclassified points by reducing the margin width. Gamma controls the influence of a single training point, and is mainly utilized when the kernel is set to 'rbf,' not when it is set to 'Linear.' When using a linear kernel, the default value of 1 should suffice. Lower values of Gamma indicate a broader similarity radius, which results in more points being grouped together. Higher values of Gamma require points to be extremely close to each other in order to be considered part of the same group.	C: 0.1, 0.3, 0.5, 0.7, 0.9, 1, 5, 7, 10 kernel: 'linear'	C: 0.1, 0.3, 0.5, 0.7, 0.9, 1, 5, 7, 10 kernel: 'rbf' gamma: 0.1, 0.3, 0.5, 0.7, 0.9, 1, 5, 7, 10
	Best value	C: 0.7, kernel: 'linear'	
LR	• For logistic regression my team mate did choose C and penalty, as mentioned in the following source: https://www.projectpro.io/recipes/optimize-hyper-parameters-of-logistic-regression-model-using-grid-search-in-python, Smaller values of C lead to stronger regularization. To control the trade-off between overfitting and underfitting of the model, "C" and "penalty" were chosen as hyperparameters for tuning	penalty': ['l1', 'l2', 'elasticnet', 'none'], 'C': [0.001,0.01, 0.1, 1, 10, 100], solver': [liblinear]	
	Best Value	c: 1, penalty : l1	

Section 3. Bias Measuring and Evaluation

EEC is a dataset for measuring gender and race bias in NLP models. It contains 11 sentence templates with gender/race-associated words and words representing different emotions. The dataset uses African American and European American names and nouns representing females and males to evaluate gender and race bias. EEC aims to provide a standardized benchmark for evaluating the equity and fairness of NLP models across different domains.

To measure bias, we used the EEC dataset as input for four classifiers and made predictions on intensity scores, which were then averaged. We calculated the average intensity scores separately for African American females, African American males, European American females, and European American males. For gender bias, we computed the difference between the average scores for African American females/males and European American females/males. For race bias, we computed the difference between the average scores for European American/African American males and European American/African American females. This allowed us to quantify the extent of bias in the classifiers with respect to gender and race.

Below is the initial bias calculation from my team mates, using the intensity scores:

Template = <person subject> feels <emotion word>, emotion word='angry'				
	LG	**SVM**	**BERT**	**Roberta**
First names Male Avg	0.4282	0.3617	0.0017	0.0281
First names Female Avg	0.4345	0.3751	0.0016	0.0273
Noun Phrase Male Probabilities	'he' : 0.3562, 'this man' : 0.3877, 'this boy' : 0.4354, 'my brother' : 0.3648, 'my son' : 0.4412, 'my husband' : 0.4173, 'my boyfriend': 0.4326, 'my father' : 0.4333, 'my uncle' : 0.4041, 'my dad' : 0.4257	'he' : 0.2296, 'this man' : 0.3314, 'this boy' : 0.4038, 'my brother' : 0.2832, 'my son' : 0.3501, 'my husband' : 0.3463, 'my boyfriend': 0.3973, 'my father' : 0.4456, 'my uncle' : 0.3388, 'my dad' : 0.4221	'he' : 0.0017, 'this man' : 0.0016, 'this boy' : 0.0016, 'my brother' : 0.0013, 'my son' : 0.0013, 'my husband' : 0.0012, 'my boyfriend': 0.0013, 'my father' : 0.0013, 'my uncle' : 0.0013, 'my dad' : 0.0013	he' : 0.0390, 'this man' : 0.0252, 'this boy' : 0.0316, 'my brother' : 0.0136, 'my son' : 0.0117, 'my husband' : 0.0150, 'my boyfriend': 0.0124, 'my father' : 0.0137, 'my uncle' : 0.0163, 'my dad' : 0.0122
Noun Phrase Female Probabilities	'she' : 0.4901, 'this woman' : 0.3891, 'this girl' : 0.4476, 'my sister' : 0.3587, 'my daughter' : 0.4503, 'my wife' : 0.4338, 'my girlfriend' : 0.4372, 'my mother' : 0.4564, 'my aunt' : 0.4382, 'my mom' : 0.3595	'she' : 0.5051, 'this woman' : 0.3059, 'this girl' : 0.4155, 'my sister' : 0.2849, 'my daughter' : 0.4441, 'my wife' : 0.4090, 'my girlfriend' : 0.4080, 'my mother' : 0.4601, 'my aunt' : 0.4334, 'my mom' : 0.3431	'she' : 0.0016, 'this woman' : 0.0017, 'this girl' : 0.0017, 'my sister' : 0.0013, 'my daughter' : 0.0013, 'my wife' : 0.0013, 'my girlfriend' : 0.0012, 'my mother' : 0.0013, 'my aunt' : 0.0013, 'my mom' : 0.001	she' : 0.0435, 'this woman' : 0.0186, 'this girl' : 0.0250, 'my sister' : 0.0112, 'my daughter' : 0.0111, 'my wife' : 0.0142, 'my girlfriend' : 0.0121, 'my mother' : 0.0127, 'my aunt' : 0.0131, 'my mom' : 0.0119

It is important to note that the SVM and logistic regression classifiers have a higher natural variability in their scores compared to the LLM classifiers(BERT and RoBERTa). This means that the score deltas (differences in scores between pairs of sentences) may be influenced by this higher variability. Looking at the results, can see that for both male and female subjects, the LLM classifiers have much lower scores than the traditional classifiers. This indicates that the LLM classifiers may be less likely to exhibit bias in this task.

For male subjects, the highest probabilities were assigned to "my son" and "this boy" by all classifiers. For female subjects, the highest probabilities were assigned to "my mother" and "she" by all classifiers. Interestingly, the probabilities assigned to the specific noun phrases vary considerably between classifiers, highlighting the importance of understanding the limitations and variability of each method.

In general, these findings indicate that incorporating LLM classifiers could be a viable strategy for reducing bias in natural language processing applications. Nevertheless, my teammates conducted additional investigation and the details of our statistical analysis are outlined in Section 4.Bias Measure Analysis.

Section 4. Results Analysis, Error Analysis and Conclusion

F1 Scores Analysis:

As discussed, Precision and recall measure how well a classifier identifies positive instances in a dataset. Precision is the proportion of true positive predictions over all positive predictions, while recall is the proportion of true positive predictions over all actual positive instances.

The F1 score is a metric that combines both precision and recall, with a value ranging between 0 and 1, where higher values indicate better performance. A big difference between precision and recall leads to a lower F1 score because the metric considers both precision and recall, penalizing low values of either metric.

Metrics	SVM	LR	BERT	RoBERTa
Precision	0.7407	0.72	0.78	0.82
Recall	0.73	0.73	0.83	0.82
f1-score	0.73	0.73	0.81	0.82

The BERT and RoBERTa classifiers have higher precision scores compared to the SVM and LR classifiers, indicating that they make fewer false positive predictions. However, the precision score of RoBERTa is the highest among all classifiers, indicating that the ROBERTA classifier makes the fewest false positive predictions.

The BERT classifier has the highest recall score, indicating that it has the best ability to identify positive instances in the dataset.

The RoBERTa classifier has the highest overall F1 score, indicating that it has the best overall performance among all classifiers.

Resource Analysis:

Classifiers	Training time (Twitter Dataset)	Test time (EEC Dataset)	Google Collab			F1-Score
			CPU	GPU	Disk	
SVM	5 s	6 s	1.2/12.7 GB	-	0.45/107.7 GB	0.73
LR	2 s	0 s	0.7/12.7 GB	-	0.45/107.7 GB	0.73
BERT	960 s	120 s	6.1/12.7 GB	8.6/15.0 GB	24.5/166.8 GB	0.81
RoBERTa	1080 s	120 s				0.82

Note: BERT and RoBERTa were run in the same file. Hence, merged the cells.

From the table, can observe that SVM and LR classifiers require very less time and computational resources for training and testing, but their F1 scores are relatively lower than BERT and RoBERTa. On the other hand, BERT and RoBERTa classifiers required significantly more time and computational resources, but they achieved higher F1 scores. Also noted that, the resources required for the BERT and RoBERTa classifiers would include the pre-trained models as well so these models have a large number of parameters and require significant disk space and memory to load.

There is a clear trade-off between the time, computational resources, and the F1 scores of the classifiers. The SVM and LR classifiers are suitable for smaller datasets and simpler models where fast training and testing times are important, but they may not perform as well as BERT and RoBERTa on larger and more complex datasets.

Therefore, BERT and RoBERTa are suitable for larger datasets and complex models where accuracy is important and the computational resources and time required are not a constraint as we were able to do using google collab.

Bias Measure Analysis:

	SVM	**LR**	**BERT**	**RoBERTa**
Gender	Null hypo: True mean difference of female and male groups is zero Alter hypo: True mean difference of female and male groups is not zero p-value: 3.8901038469271175e-17 Since the p value < 0.05(significance level), there is a low probability of obtaining a result like this if null hypo was true. Hence, we reject the null hypothesis. For different emotions where p-value < significance level 0.05 Anger: p-value = 6.28524490750019e-10 sadness: p-value = 0.0003223063581841141 fear: p-value = 2.3929730246309286e-08 joy: p-value = 1.468536870867467e-06	For different emotions where p-value < significance level 0.05: Anger p-value: 0.0003024122652359377 Fear p-value: 0.0002800455998034942 3 joy p-value: 0.018701861376435296	For different emotions where p-value < significance level 0.05: Anger p-value: 0.04776152337545493 fear p-value: 0.02024785252947083 5	The best performing model is the Roberta model, as it accepts the null hypothesis for all 4 emotions. It says that the true mean difference for males and females is zero for all emotion types. Hence, we can conclude that this model is not easily affected by the inherent bias in the data.

Race	Null hypo: True mean difference of African-American and European groups is zero Alter hypo: True mean difference of African-American and European groups is not zero p-value: 4.366344627083139e-06 Since the p value is less than 0.05(significance level), there is a low probability of obtaining a result like this if null hypo was true. Hence, we reject the null hypothesis. For different emotions where p-value < significance level 0.05 Anger p-value: 0.00017305026840960336 sadness p-value: 0.0035117785978371974 fear p-value: 0.0001487219874938079	For different emotions where p-value < significance level 0.05: Anger p-value: 2.748788346436684e-05 sadness p-value: 0.002534868726511252 Fear p-value: 2.49308621981525e-06	p-value: 0.04090316103239936 is less than significance level 0.05 for joy group	Accepts all 4 null hypothesis for the given 4 emotions.

Based on our results and t-test, SVM is biased than Logistic Regression significantly based on the statistical testing since LR does not reject null hypothesis for sadness.

The analysis conducted on the performance of four emotion recognition models using a dataset with inherent bias shows that the RoBERTa model performs the best. The t-test results indicate that the RoBERTa model does not exhibit gender and race bias for all four emotions. On the other hand, the SVM, LR, and BERT models show a significant difference between the performance of different demographic groups. These findings suggest that the RoBERTa model is less affected by the inherent bias in the data and could potentially provide more accurate emotion recognition results.

Below is the compared ground truth ‘labels’ with the predicted ‘labels’:

Metrics	**Error (%)**
LR	22.8
SVM	27.9
BERT	14.2
RoBERTa	4.7

Three Examples:

Training Data: Twitter Sentiment Analysis						**Test Data:** EEC Corpus Data **Template:** <person subject> found himself/herself in a/an <emotional situation word> situation.				
Name	**Race**	**Gender**	**Total**	**Positive**	**Negative**	**Emotion word**	**LR**	**SVM**	**BERT**	**RoBERTa**
Alphonse	AM	M	0	0	0	great	0.743513	0.821925	0.433953	0.579163
Jasmine	AM	F	1	0	1	great	0.727173	0.83654	0.594309	0.713498
Ellen	EA	F	5	3	2	great	0.742783	0.870032	0.515686	0.9670269
Frank	EA	M	8	3	5	great	0.700356	0.761466	0.651177	0.8531809
Alonzo	AM	M	0	0	0	funny	0.557495	0.650422	0.332386	0.748395
Shereen	AM	F	0	0	0	funny	0.580175	0.699728	0.489504	0.8351
Amanda	EA	F	8	7	0	funny	0.585782	0.686217	0.491428	0.8902802
Adam	EA	M	18	10	8	funny	0.551065	0.659736	0.604112	0.8451532
Ebony	AM	F	0	0	0	great	0.72943	0.826897	0.537418	0.9382157
Darnell	AM	M	0	0	0	great	0.743513	0.821925	0.577626	0.576272
Courtney	EA	F	3	3	0	great	0.700967	0.795171	0.573731	0.9307656
Andrew	EA	M	6	3	3	great	0.681657	0.724843	0.605125	0.9469501

Considering the natural score variability for traditional vs. LLM classifiers, as the SVM/LR classifiers tend to have higher variability in scores. Here, is the discussion on the differences in the classifier scores for each sentence in the above pair and providing explanations and analysis for them:

For Pair 1, the emotion word is "great," and the names in the pairs are Alphonse, Jasmine, Ellen, and Frank. The results show that all classifiers rated Jasmine positively, while the other names received mixed results. It is interesting to note that Ellen received higher scores from all four classifiers despite having three negative instances in the training data. This could be due to Ellen being a more common name in the training data or due to the content associated with the name Ellen being more positive overall. In contrast, Frank received lower scores, possibly due to being associated with negative content or being a less common name in the training data.

For Pair 2, the emotion word is "funny," and the names are Alonzo, Shereen, Amanda, and Adam. Here, Amanda received the highest positive scores from all four classifiers, despite having seven negative instances in the training data. This could be due to Amanda being a more common name in the training data or being associated with more positive content. Alonzo and Shereen received lower scores, possibly due to being associated with negative content or being fewer common names in the training data.

For Pair 3, the emotion word is "great," and the names are Ebony, Darnell, Courtney, and Andrew. In this case, Ebony and Darnell received similar scores from all classifiers, possibly due to their association with positive content or being more common names in the training data. Courtney received higher scores, possibly due to being associated with more positive content or being a more common name in the training data. Andrew received lower scores, possibly due to being associated with negative content or being a less common name in the training data.

In terms of the bias, it is difficult to determine whether it comes from the pre-training of the large language models, the fine-tuning step, or both. It is important to note that the scores for BERT and RoBERTa tend to be closer to 0 or 1 compared to SVM/LR, which could be due to how the models work.

Other thoughts and observations:

In Pair 2, the scores for the African American individuals are consistently lower than those for the European American individuals, regardless of gender.
In Pair 3, the African American individual received lower scores than the European American individual, again regardless of gender. These results suggest that the classifiers may be biased towards European American individuals.
It's important to note that these results are based on a small sample size, and more analysis would be needed to determine whether the observed bias is significant and consistent across different datasets. It's also possible that other factors, such as the content of the tweets or the specific words used, may have influenced the results.

Regarding the impact of race on the bias in the sentiment analysis classifiers, it's difficult to determine from these results whether the bias comes from the pre-training of the large language models, the fine-tuning step, or both.

Based on the results of the t-tests, can say that there is a significant difference in the mean emotion scores between male and female groups, and between African-American and European groups for all four emotions. This suggests that the data contains inherent biases, and the emotions are affected by gender and race. Therefore, the null hypothesis, which suggests that the true mean difference between these two groups is zero, can be rejected. However, the RoBERTa model accepts the null hypothesis for all four emotions, indicating that it is less affected by the inherent bias in the data compared to the other models (SVM, LR, and BERT). Therefore, the RoBERTa model can be considered the best performing model in terms of being less affected by gender and race biases. However, it is important to note that the t-test results do not necessarily imply causation and that other factors could influence the emotions being expressed. Nevertheless, the results provide valuable insights into the potential biases in the data and the effectiveness of different models in dealing with them.

Conclusion:

This project provides insight into the performance of sentiment classifiers and highlights the challenges of measuring bias in NLP models.

Our results through extensive experimentation and analysis, provide two key insights First Measuring bias using pairs of sentences is advantageous as it is closer to how the model is used in practice, Second Insight is that it is difficult to detect bias in LLM-based models due to variability in small scores.

Section 5. Tasks per team member

Student Name	Task	Comments
Mehak Preet Singh	Logistic Regression, mBERT	Bert (Exp 1), LR, Hyper parameter tuning on BERT expt 1, intensity scores on EEC test data, Bias Measure tables based on gender and race, Bias analysis
Akshaya Brian Tauro	RoBERTa	Setup BERT(Exp2) and RoBERTa, and data pre-processing. Train the setup models, and hyperparameter tuning. Intensity scores on EEC test data for both (RoBERTa, BERT). Bias analysis.
Varsha Swaminathan	Data Pre-Processing and Bias Measure	Data set pre-processing (twitter), SVM and LR hyperparameter tuning, T-testing and concluding hypothesis, Bias analysis

Dimple Kanakam Sai	Logistic Regression, mBERT	Bert (Expt 1) and LR, Hyper parameter tuning on LR, time calculations for the models, Bias analysis
Bhavesh Rajesh Talreja	Data Pre-Processing and Bias Measure	Data set pre-processing (twitter), SVM and LR hyperparameter tuning, F1 score, precision, recall score. Intensity scores on EEC test data bias (LR, SVM), T-testing and concluding hypothesis, Bias analysis.
Emeline Fratacci	SVM	SVM model, f1 scores, hyperparameter tuning on SVMs, Bias analysis