Gendered differences in DMs w/BERT

Ilsa Ahmed
UIUC
iahme5@illinois.edu

Abstract

This study investigates gendered differences in the use of discourse markers (DMs) in informal spoken British English using BERT, a popular transformer-based language model. Discourse markers such as "like," "uh," and "you know" play important roles in conversation and sociolinguistics. Utilizing the Spoken BNC 2014 corpus, which has been annotated for both linguistic and demographic data, we fine-tune BERT to classify DMs and predict speaker gender based on DM usage, and compare to baselines. The analysis evaluates BERT's performance in order to establish its sociolingusitic validity, and assessing the models ability to capture nuanced patterns in the corpus data as well as identifying areas of sociolinguistic bias in the model's behavior and other various limitations of the model architecture. Comparison of standard metrics and error analysis between the fine-tuned and out-of-the-box (OOTB) BERT models shows the impact of the corpus data and fine-tuning. In addition, we find what this says about language as a whole.

1 Introduction

Discourse markers (DMs) such as "like," "uh," and "you know" are fundamental elements of spoken language, serving as cues for discourse structuring, hesitation, and interpersonal social dynamics. DMs are highly sociolinguistically charged, with variations in their frequency and function often supposed to be linked to speaker demographics, particularly gender. More importantly, DMs are a relatively simple linguisitic phenomenon that can be categorized in a large set of data. Prior sociolinguistic studies have shown that women may possibly use certain DMs more frequently, while men might employ them differently as well as in certain contexts. However, the extent to which LLMs (Large Language Models), such as BERT, can replicate these sociolinguistic patterns can be explored further than it has.

This study addresses the question: can BERT identify gendered differences in DM usage, and how does its behavior align with established sociolinguistic trends, and what does this tell us about the model architecture? Using the Spoken BNC 2014 corpus, a richly annotated dataset of British English, we fine-tune BERT for the task of predicting the speaker's gender from utterances containing DMs from a pre-formulated list. To ensure robust evaluation, we compare the fine-tuned model against its out-of-the-box (OOTB) counterpart, which serves as a baseline for understanding the impact of domain-specific training, as well as a logistic regression model.

This research has three main objectives: (1) evaluate the performance of BERT in identifying DMs and predicting gender in utterances, using metrics such as accuracy, precision, recall, and F1-score; (2) analyze errors and biases in model predictions to assess their sociolinguistic validity; and (3) compare model behavior with findings from traditional sociolinguistic literature to identify patterns. By combining computational and sociolinguistic methodologies, this study contributes to the broader discussion of how machine learning tools can be utilized to study language use and variation.

In addition to performance evaluation, this paper critically examines the origins, data, and behavior of BERT as a computational model, focusing on its training data and assumptions. The results not only illuminate the potential and limitations of using BERT in sociolinguis-

tic research but also raise important questions about bias and representativeness in computational models. Through a rigorous analysis, this paper aims to bridge the gap between computational and traditional sociolinguistic paradigms.

2 Data

In order to accomplish the task of evaluating BERTS abilities in regards to gendered language the Spoken BNC 2014 corpus is used due to its rich metadata. The Spoken BNC (British National Corpus) 2014 is a collection of transcripts of colloquial British English conversation, with both a tagged and untagged dataset of those conversations, only the tagged dataset is used. The corpus was collected with the consent of participants, they all recorded their own conversations knowingly and sent it to the researchers for them to process and tag, the metadata for each speaker is self-reported. The Spoken BNC 2014 dataset is used to evaluate how well BERT can perform in an informal context, or rather a more natural context. Not only is it informal rather than formal english. It is British rather than U.S english. While the two Englishes share more similarities than not, BERT has millions of parameters that can be swayed by minor details. The corpus contains metadata on gender, age, age group, nationality, place of birth, first language, current living, education, occupation, class, and each individuals contributed utterances to the corpus. While the data is mostly diversified, coming from the general public, it skews female, an initial check was done with the rest of the metadata, checking instances of each demographic, sex was the simplest classification method. The corpus contains eleven and a half million words of informal conversation among the British public, meaning an optimal corpus to study instances of discourse markers, a largely spoken phenomenon. Each word is tagged with a POS marker within the .xml files in the tagged folder of the corpus. DMs come from a wide range of speech elements, but most commonly in the data they were tagged with RR which meant general adverb, and UH meaning interjection, so this is the criteria we used to identify discourse markers for the computational models. There are some possible inconsistencies in the tagging, and there are lots of different tags. For example the discourse marker phrase "so like" is often tagged as RG, degree adverb RR, general adverb ("right_RR **so_RG like_RR** what_DDQ the_AT manufacturers_NN2 want_VV0 you_PPY to_TO") but in another, very similar context, it is tagged with RG, degree adverb, JJ, general adjective ("right_RR **so_RG like_JJ** thumbs_NN2 up_II21 to_II22 the_AT guy_NN1"). Due to such variances we will be focusing in on the individual word data rather than phrases, we compiled a list of the most used Discourse Markers in the dataset, specifically choosing ones with different gender connotations. The .xml files which hold the tagged data have a "who" tag which contains a number for each speaker, this number is cross referenced within the Corpus metadata file in order to extract the gender of the speaker of any given sentence, and this will be how we map DM usage to gender. The process of cleaning the data was not intensive, however one thing is that there are more female contributions than male, and this could skew BERT to predicting female the majority of the time. It can be argued that downsampling female utterances can make BERTs predictions less accurate, but for the sake of fairly assessing each sexes DM usage, it is important to downsample as to not confound results and make sure each gender has an equal amount so the content of the utterance outweighs quantity. In the Spoken BNC corpus, there are a couple of self-report options for the speakers' sex. First, and most common, is male and female. The option to write in a non-binary identity was included in the self report, but it was not used, instead the non-binary option "X" was only used when several individuals were speaking at once, and it was decided to not include that in the dataset used to fine-tune as that would clutter the data. In addition, there is UNKFEMALE and UNKMALE, which while had no additional metadata support, did have gender included, which is what is needed. We also randomize and in-

terleave the order in which BERT processed each sentence, as females and males clumped together often, as is the case with natural conversation, and in case BERT ran out of memory while fine-tuning, then it wouldn't have ended by processing a large quantity of either sexes produced sentences, and instead maintained a balanced intake throughout. The tokenization process for the data, before fine-tuning, was also necessary in order for the format to be correct for BERTs needs. The tokenization is done in a Jupyter Notebook, we extract each utterance as a token which each must include instances of discourse markers, specific discourse markers i.e ["yeah", "oh", "erm", "mm", "yes", "no", "well", "so", "just", "like", "um", "right"]. In the Jupyter notebook we parse each .xml file of the corpus for such utterances, and cross reference the "WHO" tag in the metadata file which identifies the speaker and therefore the true sex, [Figure 2] a sample of the dataset is shown, please note the cutoff is only for display purposes and the full sentences are indeed contained within the cleaned dataset. M= 0.0 and F = 1.0. Before getting started we further analyze the spoken BNC data, to ensure some sort of patterns in DM usage, and surely enough there are differences in DM usage between the sexes. In the case of BERT, even the smallest biases must be picked up on, to ensure a fair model. The list of DMs which we choose from consist of both boosters, and hedgers. Popular culture, which is backed by research, suggests males use boosters, and females hedgers more often. This seems to be consistent with the corpus data, with certain booster DMs being used more by males such as "you know" and "right". The data is visualized as a bar graph in [Figure 3]. We expanded the list of common DMs from our previous paper, in order to better encapsulate gender relations, here is the full final list of DMs we search for in the corpus:

["absolutely", "yeah", "quite", "clearly", "utterly", "proper", "dead", "first", "second", "third", "oh", "indeed", "I mean", "mind you", "erm", "you know", "okay", "anyway", "uh", "mm", "yes", "no", "well", "so", "just", "like", "um", "right", "totally", "maybe", "sort of", "kind of", "perhaps", "I guess", "it seems", "could be", "possibly", "might", "fairly", "almost", "at least", "pretty much", "in a way", "more or less", "to some extent", "roughly", "probably", "relatively", "as far as I can tell", "more likely than not", "sometimes", "definitely", "really", "for sure", "undoubtedly", "certainly", "always", "forever", "completely", "extremely", "without a doubt", "in fact", "of course", "never", "entirely", "surely", "quite literally", "positively", "unequivocally", "to be honest", "by all means", "as a matter of fact"]

3 Methods

3.1 Data

For each of the three models, the data preparation process is similar besides the two BERT models needing a csv format. First, we parse the corpus for each utterance with an instance of a DM from our pre-selected markers which we found through prior research and analysis of the corpus. We import pandas to create a dataset in which to append these instances, with a speaker column (i.e: S0032) followed by the text (ex: "oh how did you sleep?"). We read the metadata .csv file using pandas, to extract the sex of each speaker, the data is delimited by tab. We merge this metadata to our initial dataset. For training, males are assigned 0.0 and females 1.0. We then tokenize the data using torch (in the tokenizedata function). Now this is where we downsample the female instances to match male instances, first separating the classes, and making sure female instances are the same as male, 320133 male, and 320133 female. We shuffle, and then interleave (male after female) the data. The process diverges here.

3.2 models

As a baseline, we use a logistic regression model that accomplishes the binary classification task with no outside training, a purely statistical model. The model code is implemented in python in logreg.ipynb. The relationship between discourse markers and gender, we try to confine it to a linear relationship using this model. It uses a sigmoid function, rather than

trying to imitate some natural truth about the world, it uses something that is "good enough". The standard metrics, as well as a confusion matrix is outputted. We can see, it did not do very well, but considering the nuance of the dataset, it did better than expected. We see in fig 4 the recall, precision and f score are all similar for both gender classes. This model did slightly better than random guessing, it achieves a similar precision between males and females (0.57, 0.58). Recall for males is slightly higher than females, the overall accuracy is 0.58. It did slightly better than half the time. But, the confusion matrix betrays heavy overlap, as the logistic regression model can not truly see gender in a way that a model with embeddings could, suggesting a highly non-linear relationship between DM usage and gender. The errors outputted in logregmisclassified.csv were largely random, and it doesn't seem either gender is predicted more than the other, errors are in fig 11.

Next, we use both a standard out of the box, (ootb) BERT model in order to see how BERT was trained initially, in a highly generic way, it will give insight into the intentions of BERTs creators and its sociolinguistic context. This model performs slightly worse than the logistic regression model, this is due to several reasons inherent to the design of BERT. When given this task, every time, BERT will default its predictions to one of the labels, either male (0.0) or female (i.0). OOTB bert was not trained for this binary classification task it is simply going on its generic tendencies, it defaults to a specific class like male to classify the majory of the tokens, and when it picks up a rare signal it will occaisonally classify that as the other class, this is not a bad method for BERT as it does give two distinct categories, but it is not what we are looking for, we need BERT to pick up on specific patterns in the data which we gleaned exist through pre-analyzing the data, how certain DMs trend masculine vs feminine, this is not what OOTB Bert is capable of doing, despite being a pre-trained model, its sociolinguistic abilities are lacking due to its rather generic data of wikipedia and books. In addition to this prediction task, we also have a mask prediction task for BERT in maskpredict.ipynb. This task is unrelated to the corpus, we take a sentence, filled with certain discourse markers, and see what bert predicts as the gender. This task is what bert was more suitably trained for, instead of the binary classification task which would better suit a fine-tuned-model:

Sentence:

[DM]Like, I am just a [MASK]

Predictions: ['girl', 'kid', 'guy', 'virgin', 'job', 'boy']

[DM]Surely, I am just a [MASK]

Predictions: ['man', 'girl', 'virgin', 'minor']

This lends itself to the idea that certain DMs are correlated to gender, now there are other elements of this sentence that may skew gender, but, it seems the changing of the DM led to this change in predictions, but there are many similarities between them as well. This is why the task of identifying gender can be so challenging for these models, it is a fine balance. Admittedly, BERT has patterns from its training data that does as evidenced above, include gendered DMs, which is why fine-tuning is the next step, as BERT is showing some semblance of gender identification, within a task it was initially trained for (mask predicting). Code for this task, and seeing berts prediction for every DM in our list, is on github, in maskpredict.ipynb.

Finally we fine-tune the BERT model in order to have BERTs prediction abilities better suited to the task, and to get into its inner workings. We use fine-tuned bert for the binary classification task. We train BERT using the interleaved data, a male sample followed by a female sample, in order to maintain balance. The fine tuning is done in Google colab with a borrowed GPU. The reasoning behind fine-tuning BERT to the task of gender (m/f) classification based on DMs is that BERT is known to be capable of binary classification tasks, i.e sentiment analysis of movie reviews (positive or negative) however it is generally accepted words have inherent sentiments, it is arguable that words have inherent gender, this is what we are trying to explore. After fine-tuning we look at the results [fig 7], and compare it with the ootb models' results [fig 8]. We know

that ootb BERT is not capable of this binary classification task, despite its inherent biases towards gender, as if it is not trained on this particular task it will default to one gender due to the classifier head. It is not because it thinks every sentence is a male utterance, in fact if the dataset begins with a female utterance, the results are flipped. Now, when we compare that to fine-tuned BERT there is now more of a balance, albeit a couple issues arose during training. As we can see in fig 8 BERT overfitted the data beginning in epoch one and two, with increased validation loss (which was high in the first place) this is due to several possible reasons. The signal may not be strong enough, between gender and discourse markers, at least within the SpokenBNC 204 dataset. BERT has about one hundred and ten million parameters, and discourse markers are a relatively simple thing, and taking a glance at the results, we can see the gender-dm relationship is sort of sparse. When BERT faces something simple like DMs, it overfits, and memorizes patterns rather than making more meaningful generalizations. We can look at the error output to confirm, when BERT sees "like" in a text, it does not truly consider the context of the entire sentence, and labels it as female. BERT is better at predicting female utterances, as we see in Fig 8. It has a recall of 0.73, while for males, it is a measly 0.46. I do not believe that ootb model can meaningfully assess the gender-dm relationship, it was just included to further explain BERTs architecture, but we can instead assess these results against the unbiased logistic regression model. If we look at the confusion matrices, we see that BERT has a slightly improved performance in contrast to the logreg model. it has fewer misclassifications. However, it is a marginal improvement. A sample set of errors from BERT fine tuned are contained in fig 9.

4 Discussion

Based on the results of our logistic regression model, our out-of-the-box BERT-base-uncased model and its mask prediction task, and our BERT model which was fine-tuned for the binary classification task on the Spoken BNC Corpus, in addition to our manual data analysis, and error analysis of each model, we find the differences in Discourse Marker usage among genders is significant but in a pretty small way. The bert fine-tuned model did the best in its prediction task. Based on prior research into male/female speech pattern dynamics we find that it is expected for males to use booster DMs, and females to use hedge DMs in addition to using more DMs overall, this idea is explored in the influential paper Language and Womans place by Robin Lakoff. However, this is research that is also supported by popular culture, and may be out dated or inaccurate overall in other contexts. language and woman's place summarizes key attitudes towards women and their language use, from its time, and it is still relevant now in many ways. Robin Lakoff characterizes women's speech as polite, and indirect. This is where Iw believed the gender split would occur, and what our fine-tuned BERT would pick up on. The hedging nature of female speech, usage of markers such as "like, sort of" etc. There are also boosters i.e "totally, really" included in female speech as Lakoff writes, but these are read as emotionality rather than confidence and directness. It appears, this perception of female speech is highly influenced by social factors. This paper, many would say is out dated. When we get to the reality of the issue, the link is weak, between speech and gender, from what our findings of this particular british spoken english dataset has found. Women's speech contributes to their lack of authority, that is very much true, but perhaps it is not womens speech in of itself, but the fact that the speech comes from a woman. Lakoffs methods were not computational in nature, she used anecdotes, her own speech, those of her friends, but there is still truth in her claims. She concludes that it is important to identify those elements of gendered speech which reflect true inequity, which I do agree with. When we analyzed the data before feeding into the models, we saw that in some cases gendered DM patterns arose, but the urestion now is to see if that really matters in any material sense. BERT says

not really. It is known that BERT contains gender biases, which is explored thoroughly in Investigating Gender Bias in BERT(Bhardwaj et al.) Creating a model tailored to a specific task typically diminishes biases, which is our output from our fine-tuned bert model was not as biased as it could have been, rather it was a bit more confused. In the case of Bhardwaj they created a BERTde model which was a model where gender directions were mostly removed form the embeddings to reduce BERTs bias. But it can be argued, these gender directions are important to the inherent structure of the model. I think it is pretty important to think about if we want language models to model human language, or to model something else enitrely. Taking away biases from a model has numerous implications, like noiis, which must be bandaged with various algorithsm, such as in the case of bhardwaj where they came up with A one. Human language has biases. Other papers are referenced below exploring bias in transformer models. Overall, BERT found patterns and it picked up, but it had higher accuracy with females rather than males. To discuss limitations, there may have been a signal which was too weak for BERT to pick up on, the data in Fig3 visualizes this, there are differences in DM usage, but it is not the most significant. This weak signal, and lack of discriminatory features in the corpus utterances, is what led to BERTs lack of achievment. In addition, as BERT excells at using context, perhaps a sentence was not enough for BERT to discriminate and choose a gender, but in informal language sentences tend to be short and uick. Overall, at least in british english, DMs are not strongly gendered, they are, just not very strongly.

5 Tables and Figures

5.1 Hyperlinks

https://github.com/mahmed1312

6 Citation

Lakoff, R. (1973). Language and Woman's Place. Language in Society, 2(1), 45–80. http://www.jstor.org/stable/4166707

```
gender
0.0    299745
1.0    299745
Name: count, dtype: int64
```

Figure 1: Balanced Dataset

	speaker	text	sex	gender
176663	S0084	yeah	F	1.0
671957	S0198	oh it must have been like oh my god got ta tak...	F	1.0
181437	S0510	to but the sound guy was oh I do n't know he w...	F	1.0
700427	S0570	I du n no	M	0.0
365541	S0588	and and those things do n't matter like	F	1.0
466083	S0618	so you made a robot ?	F	1.0
784575	S0521	Holocaust yeah	M	0.0
181636	S0654	yeah and also silk is not man-made it 's made ...	F	1.0
101514	S0013	she said she 'd like a ginger tea	F	1.0
681710	S0591	have a handicap of fifty or something like that	M	0.0
449127	S0014	oh okay	F	1.0
50179	S0390	hey stop firing questions at me I do n't know ...	F	1.0
718228	S0024	yeah	F	1.0
676727	S0619	I do n't know	F	1.0
72682	S0255	so I need to take --ANONnameF 's cake apparent...	F	1.0
418810	S0024	you made me sound horrible	F	1.0
482556	S0325	I my mum and dad do n't like getting I have on...	F	1.0
310467	S0439	yeah so it 's been a positive day all round re...	F	1.0
280567	S0086	yeah it 's really well written well there 's a...	M	0.0
650833	S0558	--ANONnameM said to me he there was talk that ...	F	1.0
418816	S0144	yeah	M	0.0
564521	S0525	well yeah a couple of weeks ago	F	1.0
167747	S0083	mm	F	1.0
528982	S0402	see now we 've given we 've given it away	M	0.0
592552	S0604	yeah but it it wait it	M	0.0
735424	S0655	I had coffin on Facebook and if you noticed th...	M	0.0
435776	S0623	I do n't know what time er just after four I t...	F	1.0
327040	S0012	oh	M	0.0
735146	S0008	you know you take the --UNCLEARWORD	M	0.0
505984	S0405	wow erm right er er let 's let 's discuss what...	M	0.0

Figure 2: Sample from dataset

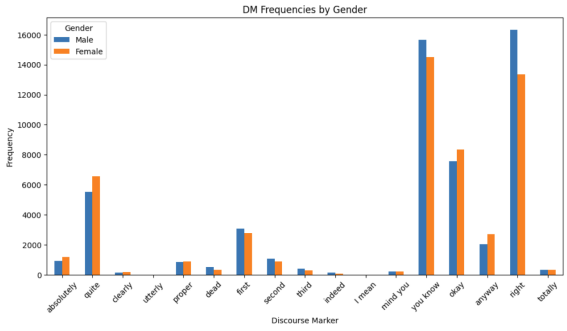

Figure 3: bar graph of DM instances by gender

```
Logistic Regression Model Metrics:
              precision    recall  f1-score   support

        Male       0.57      0.62      0.59     59949
      Female       0.58      0.53      0.56     59949

    accuracy                           0.58    119898
   macro avg       0.58      0.58      0.58    119898
weighted avg       0.58      0.58      0.58    119898

Confusion Matrix:
[[37110 22839]
 [27972 31977]]
```

Figure 4: Logistic Regression Model Stats

```
Sample 43:
Text: --UNCLEARWORD the cheap no the cheapest one was two hundred and fifty quid
True Label: 0.0, Predicted Label: 1
Correct: False
```

Figure 5: Out of the box BERT classifying noise as female

```
Sample 2000:
Text: mm hm
True Label: 1.0, Predicted Label: 0
Correct: False


Metrics for Small Sample:
              precision    recall  f1-score   support

         0.0       0.50      1.00      0.67      1000
         1.0       0.50      0.00      0.01      1000

    accuracy                           0.50      2000
   macro avg       0.50      0.50      0.34      2000
weighted avg       0.50      0.50      0.34      2000

Confusion Matrix:
[[997   3]
 [997   3]]

Detailed results saved to 'bert_ootb_small_sample_results.csv'
```

Figure 6: Metrics for small sample OOTB BERT

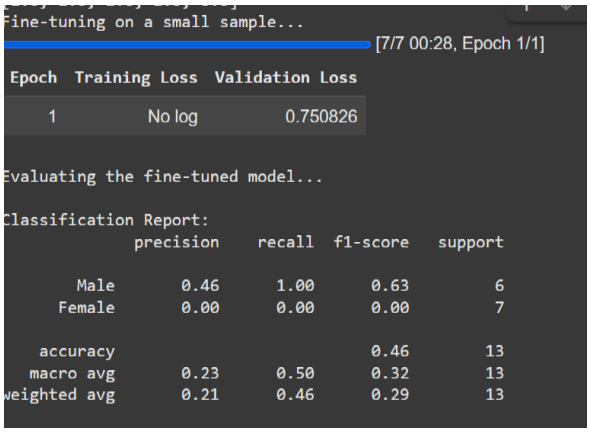

Figure 7: Small sample fine-tuned

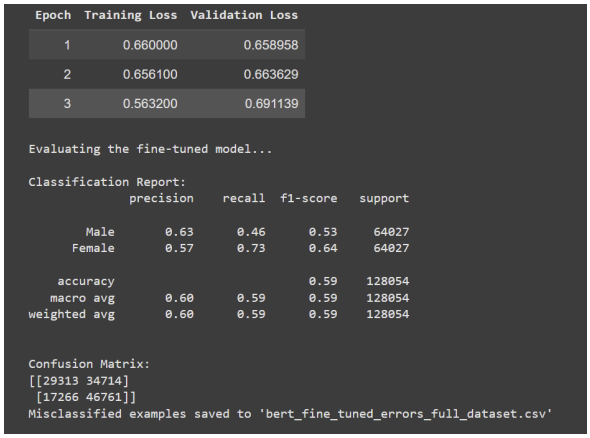

Figure 8: Fine Tuned Model Stats

```
Evaluating model on a small sample...
Starting predictions on a small sample...
Sample 1:
Text: you 're probably right
True Label: 0.0, Predicted Label: 0
Correct: True

Sample 2:
Text: yeah yeah she was fine
True Label: 1.0, Predicted Label: 0
Correct: False

Sample 3:
Text: yeah
True Label: 0.0, Predicted Label: 0
Correct: True

Sample 4:
Text: I was immediately my hackles were up immediately I thought it was one thing to come and watch a wedding from ou
tside but to participate --UNCLEARWORD but you all seem cold about it and I just think
True Label: 1.0, Predicted Label: 0
Correct: False

Sample 5:
Text: and then they could n't er and so that 's --UNCLEARWORD job with this thing that takes the bone part out
True Label: 0.0, Predicted Label: 0
Correct: True

Sample 6:
Text: we 'll see in a minute I just need to concentrate while I 'm adding
True Label: 1.0, Predicted Label: 0
Correct: False
```

Figure 9: ootb Bert only Predicting Male


```
{'text': 'mm', 'true_label': '0.0', 'predicted_label': '1'}
{'text': "oh god okay I 'll do the threes then I 'll do the kitchen then I 'll do the", 'true_label': '0.0', 'predict
ed_label': '1'}
{'text': 'how the fuck do you know that ?', 'true_label': '1.0', 'predicted_label': '0'}
{'text': 'yeah yeah', 'true_label': '0.0', 'predicted_label': '1'}
{'text': 'I think I wa- I really wan na read that', 'true_label': '0.0', 'predicted_label': '1'}
{'text': "and plus if he did do it he there 's no way he would have been able to cover it up", 'true_label': '0.0', '
predicted_label': '1'}
{'text': 'well were you thinking about that then ?', 'true_label': '0.0', 'predicted_label': '1'}
{'text': 'mm', 'true_label': '0.0', 'predicted_label': '1'}
{'text': '--ANONnameM would you like another scone ?', 'true_label': '0.0', 'predicted_label': '1'}
{'text': '--UNCLEARWORD hall yes yeah yeah just past the --UNCLEARWORD', 'true_label': '1.0', 'predicted_label': '0'}
{'text': "no they 're not", 'true_label': '1.0', 'predicted_label': '0'}
{'text': 'yeah', 'true_label': '0.0', 'predicted_label': '1'}
{'text': "I do n't know", 'true_label': '0.0', 'predicted_label': '1'}
{'text': "I do n't know shall I get it ?", 'true_label': '0.0', 'predicted_label': '1'}
{'text': 'yeah', 'true_label': '0.0', 'predicted_label': '1'}
{'text': 'no well you know like the the trend', 'true_label': '1.0', 'predicted_label': '0'}
{'text': "I do n't know yet I 'll --UNCLEARWORD and have a think", 'true_label': '0.0', 'predicted_label': '1'}
{'text': "yeah but that 's", 'true_label': '1.0', 'predicted_label': '0'}
{'text': 'yeah how much is it ? ten is it or twenty ?', 'true_label': '0.0', 'predicted_label': '1'}
{'text': "well I heard she 's got she 's got two she 's got two weeks to per- before she goes to Tottenham so I hope
she I hope she manages to stay --UNCLEARWORD", 'true_label': '0.0', 'predicted_label': '1'}
{'text': "no we do n't", 'true_label': '1.0', 'predicted_label': '0'}
{'text': 'uhu', 'true_label': '0.0', 'predicted_label': '1'}
{'text': 'really ?', 'true_label': '0.0', 'predicted_label': '1'}
{'text': "someone 's gon na have to help me here I 'm getting pretty full", 'true_label': '1.0', 'predicted_label':
'0'}
{'text': 'oh yeah yeah', 'true_label': '0.0', 'predicted_label': '1'}
{'text': "he 's I talked to him no I 'm not saying that I I was talking to him last week I- I went up last week all h
e wanted to do was watch Davis Cup TV the Davis Cup on TV and in the end I got --ANONnameF did I tell you ?", 'true_l
abel': '1.0', 'predicted_label': '0'}
```


Figure 10: Bert fine-tuned errors


```
text,true_label,predicted_label
he shouted he shouted at --ANONnameM and --ANONnameM was like taking pictures like the photographer,0.0,1.0
yes,0.0,1.0
just fucked it up,0.0,1.0
there 's already lots of military cuts in --UNCLEARWORD so I do n't know how that will pan out when countries get even smaller,1.0,0.0
and it 's it 's just boom boom boom boom boom boom boom boom boom --UNCLEARWORD there 's there 's there 's he puts no feeling into at all bu
yes that 's right and cheaper probably too to France,1.0,0.0
another one,1.0,0.0
to be honest no if there 's anything --UNCLEARWORD,1.0,0.0
oh yes,0.0,1.0
well okay it 's like I du n no one o'clock in the afternoon,0.0,1.0
mm,0.0,1.0
oh god they might be from,0.0,1.0
Cos the sea 's just been coming in making the islands smaller and smaller,1.0,0.0
organising someone else instead of himself,1.0,0.0
yeah it 's nice,0.0,1.0
yeah yeah,1.0,0.0
I like the little bike,0.0,1.0
yeah,1.0,0.0
```


Figure 11: Logistic Regression Misclassifications

Bhardwaj, Rishabh, et al. “Investigating Gender Bias in BERT.” Cognitive Computation, vol. 13, no. 4, July 2021, pp. 1008–18. Springer Link, https://doi.org/10.1007/s12559-021-09881-2.

Li, Bingbing, et al. Detecting Gender Bias in Transformer-Based Models: A Case Study on BERT. arXiv:2110.15733, arXiv, 15 Oct. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2110.15733.

Sileo, Damien, et al. DiscSense: Automated Semantic Analysis of Discourse Markers. arXiv:2006.01603, arXiv, 2 June 2020. arXiv.org, https://doi.org/10.48550/arXiv.2006.01603.

Nkhata, G. (2022). Movie Reviews Sentiment Analysis Using BERT. Graduate Theses and Dissertations Retrieved from https://scholarworks.uark.edu/etd/4768