Meta Political Linguistic Bias Analysis

CS 505
Natural Language Programming

Team Members

	Name	Email
1	Zack Meeks	zmeeks@bu.edu
5	Gauravdeep Singh Bindra	gaurav57@bu.edu
6	Abdelazim Lokma	alokma@bu.edu

GitHub Repository: https://github.com/zmeeks/CS505

TABLE OF CONTENTS

Abstract

This research paper explores the use of Natural Language Processing (NLP) to classify the different levels of epistemic certainty in a text. Inspired by the study "Examining Political Rhetoric with Epistemic Stance Detection," we sought to replicate and extend its findings. Using BERT models fine-tuned on the FactBank 1.0 dataset, we achieved a promising accuracy of 95%. Our study concludes with insights into the dataset's limitations, the efficacy of our methodology, and the potential for future research in epistemic bias analysis using more event-centric texts.

“Epistemic bias” refers to a type of cognitive bias that affects our beliefs, understanding, and judgments about what is true or what constitutes knowledge. It can influence how individuals or groups perceive reality, process information, and make decisions. This bias is particularly relevant in the fields of epistemology (the study of knowledge), psychology, and social sciences. However, it’s difficult (or impossible) to utilize that sort of facticity in a linguistic way and so the term was chosen by the researchers in our cited research paper so as to distinguish it with the term “factual” or “facticity” – in the cited research paper epistemic bias is basically modality + polarity, and in our research paper it’s basically synonymous with modality and the definition of epistemic bias above essentially applies to modality unchanged.

Project Overview

Our original original goal was to see if we could use NLP to classify political biases of philosophical texts. Our idea was that we could use the texts of different translators of the same texts to see if it was possible to point to injections of different political biases into the translated texts. While researching the feasibility of this we stumbled upon the research paper “Examining Political Rhetoric with Epistemic Stance Detection” and we then decided to pursue replicating that. The authors of that paper have yet to post their jupyter notebooks or data, thus we tried to replicate and extend their findings from scratch.

The first stumbling block that we ran into was the fact that FactBank data that they referenced as having 77,000 token+ was far from obvious about how to use. The data consisted of 20 database tables. The research paper claimed to utilize the data-points for each sentence (an event, a source, and the epistemic-bias label). Our reading of the research paper assumed that “event” referred to the event-expression that a sentence included. What was not explicitly stated was the fact that the FactBank data set doesn’t include this event-expression data. Instead, what the FactBank dataset calls events, are more like words which are somehow associated to the event-expressions. We read the documentation twice and didn’t fully comprehend how the “events” consisting of nouns, adjectives, and verbs could be used to specify the event expressions that we really wanted to train any model on. However, it appeared that most events were preceded by a verb-phrase and we were able to develop a parsing strategy to make sense of what event expression was being referred to by just the verbs and so we cut the FactBank dataset down to only include the verbs. The research paper said that this dataset was curated by linguistic experts. Nevertheless, we found some interesting inconsistencies in the data: e.g. participles are sometimes labeled as verbs, but usually labeled as adjectives. If we had more time, we could easily add in many of the participles because the English language often drops implied verbs from sentences and that could give us more data points.

In addition to the mislabeled parts of speech, we could not figure out how adjectives and nouns describe events, and we were only able to work with verbs as they naturally applied to events themselves. For example, the sentence;

'So when Wong Kwan spent seventy million dollars for this house, he thought it was a great deal.'

Was sampled twice, once for the verb ‘thought’, and another for the noun ‘deal’, it is clear from this example that the word deal provides very little context on the epistemic bias of the event, when compared to the verb ‘thought’. As a result of this, we filtered all data instances that were related to non verb parts of speech, allowing us to focus only on the samples that could give our model the most relevant information.

FactBank categorizes the certainty of events using polarity (positive or negative occurrence) and modality (from possible to certain), adding an 'uncommitted' label for non-attribution or unknown source stance. Below is a table containing the labels and what they represent.

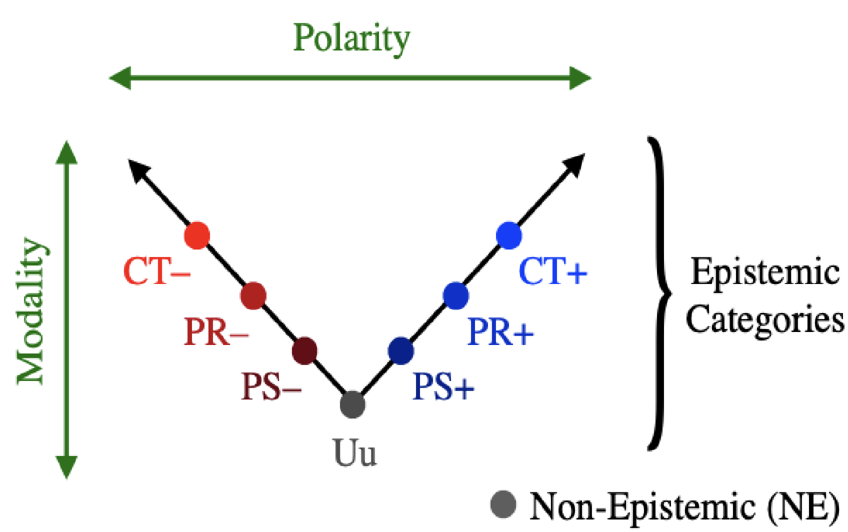


Fig1

CT+	Certainly positive, indicating the event definitely happened.
PR+	Probably positive, suggesting the event likely occurred.
PS+	Possibly positive, denoting a chance that the event happened.
CT-	Certainly negative, stating the event definitely did not happen.
PR-	Probably negative, suggesting the event likely did not occur.
PS-	Possibly negative, indicating there's a chance the event did not happen.
Uu	Fully underspecified, used when the source does not commit or know the factual status of the event.
NA	Non-applicable, when the factuality cannot be evaluated.

After an initial scan of our data, we noticed that many of the labels were sparse, which made the dataset harder to use for training. A decision was made to try to unify the labels into larger sets by eliminating their polarity, for example, joining PR+ and PR- into one label; PR. Data samples labeled 'NA' were removed from the dataset, as they provided no useful information for the model. Additionally, many events contained duplicate references that were underspecified, the removal of the 'Uu' label from the data allowed us to focus the analysis on clear, committed stances toward events and propositions, excluding instances where the source's stance is uncommitted or unknown. Our final labels used to train the model are shown below.

PS	PR	CT	Uu

Possible	Probable	Certain	Underspecified

To ensure the model is trained effectively, it was essential to get acquainted with the data it would learn from. The pie chart depicted below illustrates the various classes represented in our dataset.

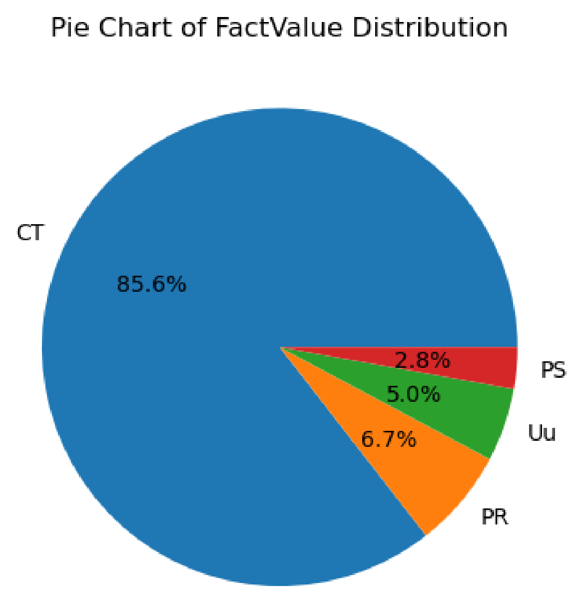


Fig. 2

The chart clearly shows a major bias for the CT class, when trained on this dataset, our model might develop a tendency to predict the majority class more frequently, which could lead to poor performance on less represented classes.

We used the FactBank dataset because the research paper stated that they had the best results with it. However, after getting pretty deep into this project and rereading the paper we realized that this dataset was narrower than our goals required – specifically, the original paper used FactBank because it was the only large epistemic bias dataset with respect to multiple sources/authors in English. However, most of the text we were looking at only has one source (the original author's as interpreted by the translator or the original author's characters as interpreted by the translator in the case of Plato's The Republic).

Another issue that we overlooked was that the philosophical texts we were attempting to analyze were not particularly conducive to "epistemic biases" in that they were mostly "eventless".

While events aren't necessary for epistemic biases (for example, Glaucon could have said that he was only of the opinion that people sought to be "just" for the sake of fear of being unjust, rather than assert it with total confidence. This may just be a consequence of writing style as many writers avoid saying things such as "it's my opinion that" when the fact that something is their opinion is obvious – this greatly increases the occurrence of "certainty" flavored modalities overwhelming the sentences in the texts that we chose to analyze. We chose the texts that we chose because we knew that they were referenced and cited frequently in political discourse and thus assumed they would be perfect for epistemic bias analysis. The translators that we chose were suspected of having different political biases. Regarding the Plato texts, Bloom was probably more conservative than Reeve is, and Ludovici was definitely more conservative than Kaufman was for the Nietzsche texts.

We fine tuned BERT models based on the FactBank dataset that we narrowed down. We also attempted to create our own dataset. It was here that we realized that these politically referenced texts were a bit too metaphysical to be good texts to analyze epistemic biases. We also started to train a RoBERTa model, but ran into some minor issues and gave up that search because our accuracy on BERT was more than high enough for our liking at 95%. Please see the model training section below for more details.

How we started to create our own dataset: We hypothesized that given a sentence and an event, we could figure out the modality score in under 5 seconds such that we could more than double our data set in under 4 hours and possibly expand the combined dataset 3-fold by slightly perturbing each sentence twice. We ran into two problems with this that would make this task take SIGNIFICANTLY longer: the first one is that it's not at all obvious which verbs or participles should be used to flag events in a sentence, the second one being that most sentences in our philosophy texts don't even have what we understand to be events. E.g. "SOCRATES: It makes no difference, Polemarchus." doesn't have an obvious event; however one could easily see that there is nevertheless a certain amount of certainty flavored modality to the eventless sentence, which I'm not sure the FactBank data accounts for, but if it does, it's probably covered via further application of the datasets annotations beyond that which we included or understood.

We created some python scripts to randomly grab N sentences from a book and put them into a blank-line-delimited file. Another script then grabbed the contents of the first scripts output and listed all of the verbs and participles on new lines below the original sentence and then delimited everything again with blank lines. The idea was that these new files could quickly be populated with local epistemic bias label info. These tools could definitely be used to that end in the future, but it would behoove us to use texts with more events as text with events appears to have a greater likelihood of having a more diverse range of modalities (and polarities too) from which would make for more interesting comparisons. Perhaps instead of comparing different translations, we could instead compare political articles about the same events to see if the incidence rate of modality or polarity differ significantly or not.

While we were unable to use the philosophical texts for the purpose of epistemic bias analysis, we were nevertheless able to test the totality of our methodology in disconnected pieces. We were able to finetune the BERT model with our FactBank data set and verify that it worked better than a naive solution would predict (e.g. our dataset consisted of about 80% CT modalities, and we achieved accuracy of about 95%, thus we can say with confidence that the model is doing more than just outputting "CT" as the predicted label.)

As for our original intention of comparing philosophical texts we found four philosophy books: The Republic by Plato translated by Bloom and Reeve and The Will To Power by Niezsche translated by Ludovici and Kaufman. We originally wanted to use a Nietzshe book with less aphorisms as we thought that the writing style of aphorisms could lead to overfitting (as philosophical aphorisms tend to be written with more certainty-flavored modalities than average), but that was one of the only texts written both by Kaufman (whose footnotes read as classically liberal) and by Ludovici (a known fascist sympathizer) and so we switched to that text. We then separated all of the footnotes from the main texts and created eight new files. We then cleaned those eight files. We then ran several linguistic analyses on these texts including average sentence length, average word length, average number of syllables per word, sentiment analysis, and average depth of syntax-tree-parsings, as well as the respective standard-deviations. The least useful analysis was the out-of-the-box sentiment analysis. The most useful analysis was the average depth of syntax tree-parse depth (plus the standard deviation). The tree depth was roughly the same between Kaufman and Ludovici, but Reeve's translation was notably less complex (as measured in a deeper parse depth) than Bloom's translation, while Reeve's footnotes were even more notably more complex than Bloom's footnotes. We hypothesized that the conservative authors might write in less-readable fashion. We can't say with any confidence that that hypothesis has been confirmed, but it has yet to be denied. From our analysis, Bloom wrote sentences more than 50% longer than Reeve. We believe that the sentiment analysis wasn't useful due to the fact that most out-of-the-box sentiment analyzers are predominately trained on social media data and often require a lot of fine-tuning to come up with the rule-base relevant to our datasets. To aid in our completion in some of these tasks we created extra scripts (e.g. to aid in cleaning the data and to aid in running the syntax-tree-parsing-depth analysis).

Philosophical Dataset Descriptions

Dataset 1: Nietzsche
The Will to Power, Nietzsche, translation by: Kaufmann Text
The Will to Power, Nietzsche, translation by: Kaufmann Notes
The Will to Power, Nietzsche, translation by: Ludovici Text
The Will to Power, Nietzsche, translation by: Ludovici Notes

Dataset 2: Plato
The Republic, Plato, translation by: Bloom Text
The Republic, Plato, translation by: Bloom Notes
The Republic, Plato, translation by: Reeve Text
The Republic, Plato, translation by: Reeve Notes

Data Cleaning

Cleaning these texts was quite challenging due to the nature of how these were stored.
There were a lot of random symbols in the text that needed to be removed while some cleaning was mandated by training requirements.

Notable changes made

- (alphabet) : ‘ ‘
- _word_ : word
- [number] : number
- Removing bullet numerals
- Removing ‘Names:’ in dialogues
  This was done because we don’t really need the model to attribute those lines to a particular character. We are interested in understanding the epistemic bias of the author and the dialogue of all the characters is written by him.

MODEL CHOSEN : BERT

The BERT (Bidirectional Encoder Representations from Transformers) model, introduced by Google in 2018, revolutionized the field of natural language processing (NLP). It represents a significant departure from previous NLP models due to its deep bidirectional nature, allowing it to understand the context of a word based on all of its surroundings (both left and right of the word).

Key Features of BERT

- Bidirectional Context: Unlike traditional models which read text sequentially (either left-to-right or right-to-left), BERT reads the entire sequence of words at once. This means it gains a deeper understanding of context and word relationships.
- Transformer Architecture: BERT is based on the Transformer, an attention mechanism that learns contextual relations between words in a text. This architecture differs significantly from the RNNs and LSTMs commonly used in previous NLP models.
- Pre-training and Fine-tuning: BERT involves two stages of training:
    - Pre-training: The model is trained on a large corpus of text (like Wikipedia and BooksCorpus) on two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP).
    - Fine-tuning: The pre-trained BERT model is then fine-tuned with additional output layers for specific tasks like classification, question answering, etc., without extensive task-specific architecture modifications.
- Large Scale and Deep Learning: BERT models are typically very large. The original BERT (BERT Base) has 110 million parameters and BERT Large has 340 million.

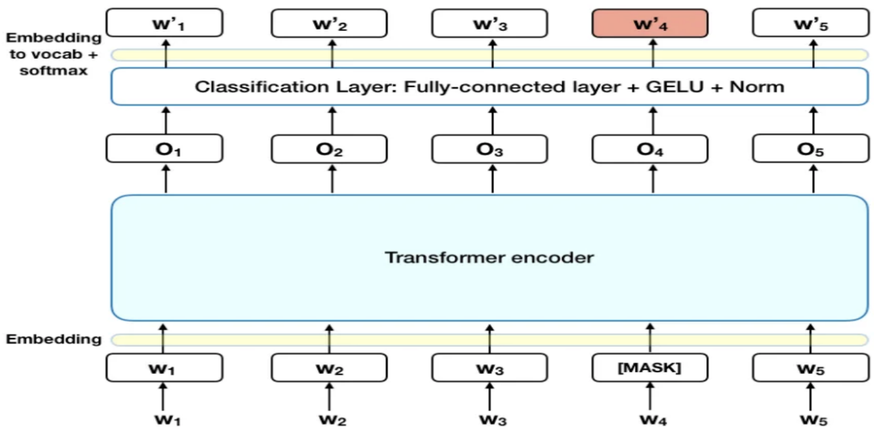
Advantages

- Contextualized Word Embeddings: BERT provides deep context for word embeddings, understanding the meaning of words based on all of their surroundings.
- State-of-the-Art Performance: It has set new records in various NLP benchmarks.

Challenges

- Resource-Intensive: BERT requires significant computational power for both training and inference, making it challenging to deploy on resource-constrained devices.
- Overfitting in Small Datasets: Its large size can lead to overfitting when fine-tuned on smaller datasets.

BERT Architecture



Reason for choosing BERT

BERT has been pre-trained on a vast corpus of text from the internet, which equips it with a rich understanding of language before it is even tailored to our specific task. This pre-training step means that BERT has already learned an immense amount about language structure and variety, which we can then fine-tune to our classification task with relatively little additional data.

By utilizing BERT, we leverage its deep understanding of language context and structure, which allows us to achieve more accurate results in our classification tasks. This leads to models that perform with a high degree of understanding and subtlety, capturing the intricacies and variations in the text that other models may miss.

Running The Model

Version 1: We only trained it on the ‘sent’ column, with the factValue column as the Label

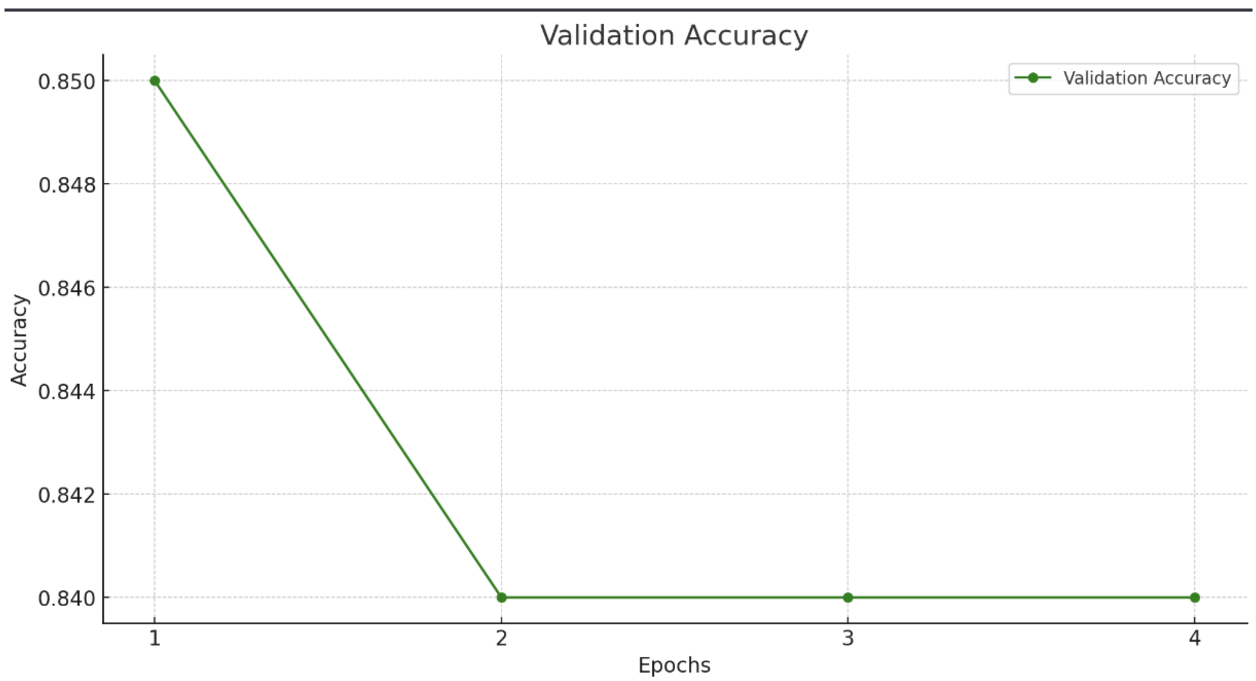
	sent	eText	factValue
388	Blockbuster shares closed yesterday at $18.75,...	closed	CT
1391	Mr. Junius said Nashua's "intention is to rema...	said	CT
1732	Replied State Department deputy spokesman Rich...	Replied	CT
2369	The quarantine hopes to staunch the flow of Ir...	going	CT
742	For the nine months, the company reported a ne...	reported	CT
1352	More people are hurt.	hurt	CT
225	America West wouldn't identify the entity, but...	said	CT
2014	The Justice Department has emphasized that the...	emphasized	CT
2451	The western military alliance invited the thre...	join	Uu
326	As the yen soared in recent years, Sansui's de...	became	CT

```
======== Epoch 1 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:07.
  Batch    80  of    159.    Elapsed: 0:00:14.
  Batch   120  of    159.    Elapsed: 0:00:22.

  Average training loss: 0.61
  Training epcoh took: 0:00:29

Running Validation...
  Accuracy: 0.85
  Validation Loss: 0.52
  Validation took: 0:00:01

======== Epoch 2 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:07.
  Batch    80  of    159.    Elapsed: 0:00:15.
  Batch   120  of    159.    Elapsed: 0:00:22.

  Average training loss: 0.42
  Training epcoh took: 0:00:30

Running Validation...
  Accuracy: 0.84
  Validation Loss: 0.45
  Validation took: 0:00:01

======== Epoch 3 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:08.
  Batch    80  of    159.    Elapsed: 0:00:15.
  Batch   120  of    159.    Elapsed: 0:00:23.

  Average training loss: 0.31
  Training epcoh took: 0:00:31

Running Validation...
  Accuracy: 0.84
  Validation Loss: 0.46
  Validation took: 0:00:01

======== Epoch 4 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:08.
  Batch    80  of    159.    Elapsed: 0:00:16.
  Batch   120  of    159.    Elapsed: 0:00:24.

  Average training loss: 0.25
  Training epcoh took: 0:00:31
```

```
Running Validation...
  Accuracy: 0.82
  Validation Loss: 0.47
  Validation took: 0:00:01

Training complete!
Total training took 0:02:05 (h:mm:ss)
```

Above are the screenshots of the metric outputs during the runs of the model. We can notice that the model starts overfitting and the validation accuracy starts decreasing from Epoch 4.

Validation Loss stabilizes so further training is not required



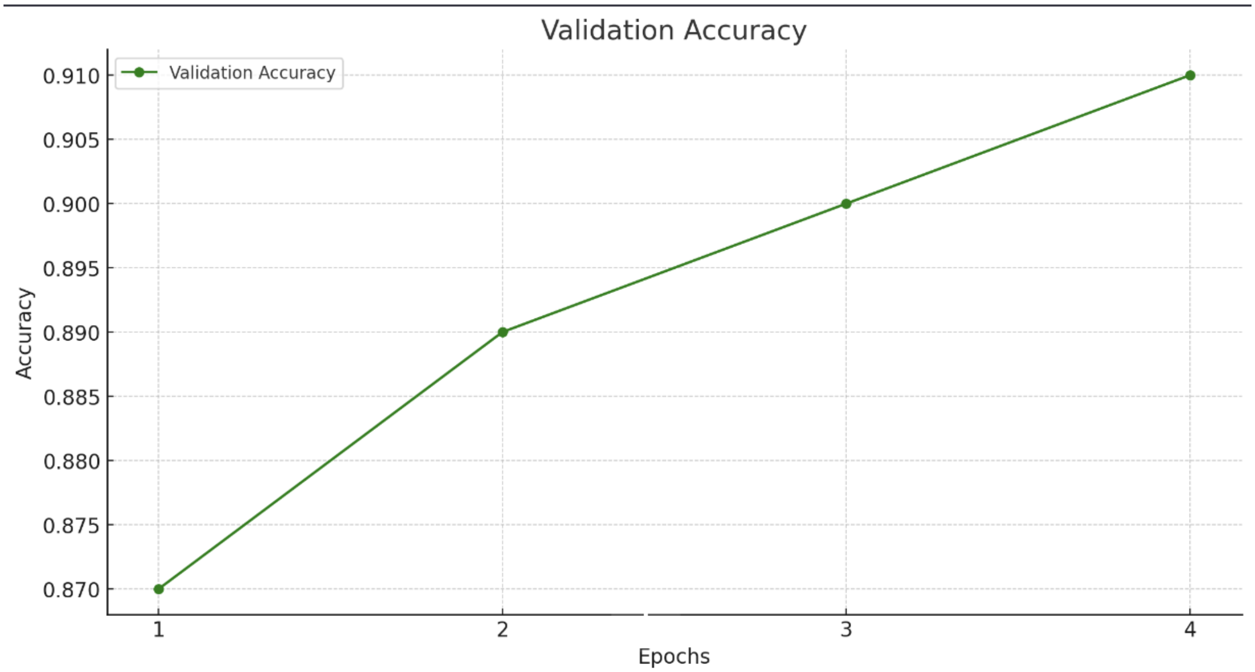
Going by the validation loss, the model at Epoch 2 should generalize the best

Version 2: When we use the focus word along with the sentence.

```
======== Epoch 1 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:08.
  Batch    80  of    159.    Elapsed: 0:00:15.
  Batch   120  of    159.    Elapsed: 0:00:22.

  Average training loss: 0.59
  Training epcoh took: 0:00:30

Running Validation...
  Accuracy: 0.84
  Validation Loss: 0.46
  Validation took: 0:00:01

======== Epoch 2 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:07.
  Batch    80  of    159.    Elapsed: 0:00:15.
  Batch   120  of    159.    Elapsed: 0:00:22.

  Average training loss: 0.30
  Training epcoh took: 0:00:30

Running Validation...
  Accuracy: 0.87
  Validation Loss: 0.41
  Validation took: 0:00:01

======== Epoch 3 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:08.
  Batch    80  of    159.    Elapsed: 0:00:15.
  Batch   120  of    159.    Elapsed: 0:00:23.

  Average training loss: 0.18
  Training epcoh took: 0:00:31

Running Validation...
  Accuracy: 0.89
  Validation Loss: 0.43
  Validation took: 0:00:01
```

```
======== Epoch 4 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:08.
  Batch    80  of    159.    Elapsed: 0:00:16.
  Batch   120  of    159.    Elapsed: 0:00:24.

  Average training loss: 0.12
  Training epcoh took: 0:00:32

Running Validation...
  Accuracy: 0.89
  Validation Loss: 0.42
  Validation took: 0:00:01

Training complete!
Total training took 0:02:06 (h:mm:ss)
```

We can see from the graph above that the lowest validation loss was recorded at the second epoch.



The highest validation accuracy is 91% at Epoch 4. The Epoch 2 model would probably still generalize the best. Especially in multi-class models, we shouldn't just go by the accuracy, but we should consider validation loss to be the main metric in deciding how long to train or which model to use.

ADDRESSING CLASS IMBALANCE

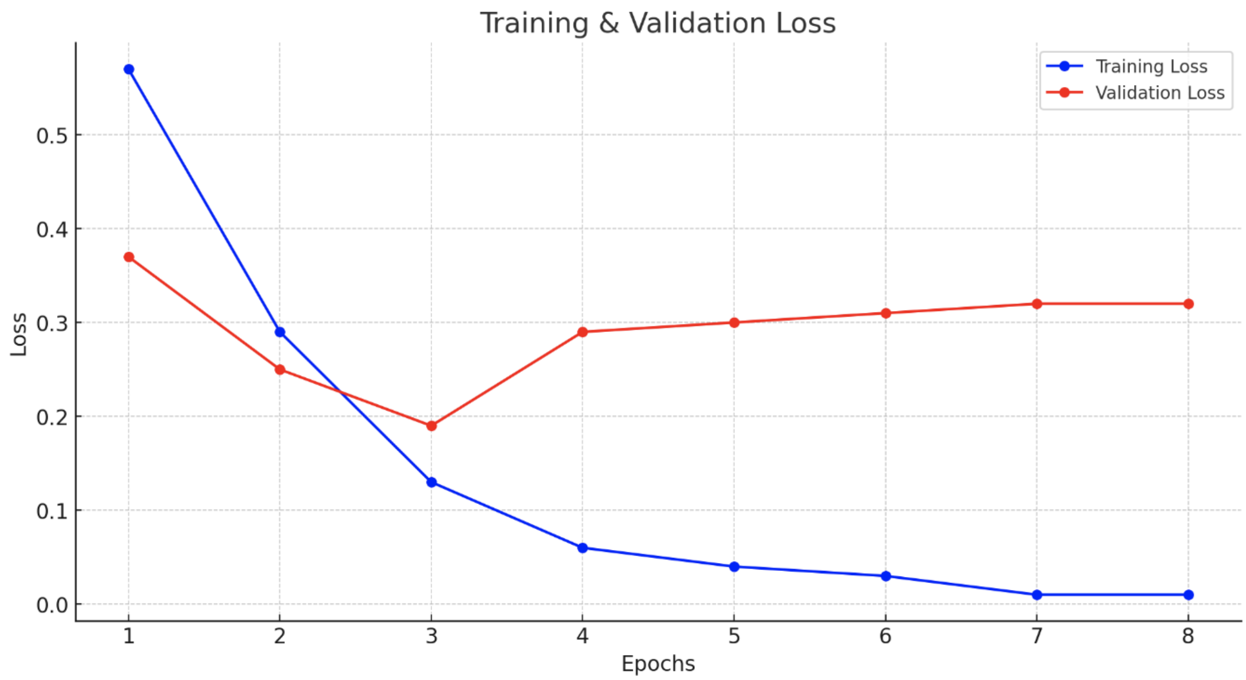
```
======== Epoch 1 / 4 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:08.
  Batch    80  of    159.    Elapsed: 0:00:16.
  Batch   120  of    159.    Elapsed: 0:00:24.

  Average training loss: 0.08
  Training epcoh took: 0:00:31

Running Validation...
  Precision: 0.90
  Recall: 0.91
  F1 Score: 0.91
  Validation Loss: 0.38
  Validation took: 0:00:01
```

In order to guarantee that our high model scores were not a result of over predicting the biased classes. We relied on other metrics such as precision, recall and f1 and they gave similar results. These metrics provide a more comprehensive picture of the model's performance, especially in datasets with class imbalances.

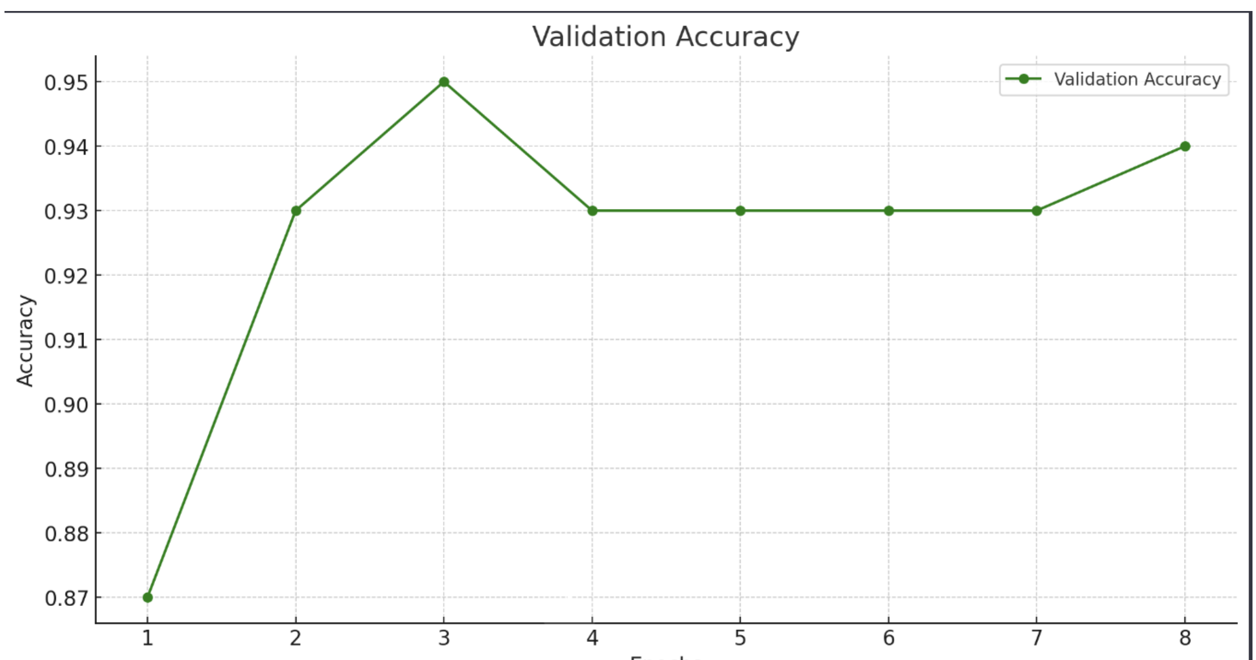
By taking average = ‘weighted’ while calculating these metrics, we are not letting class imbalance affect the accuracy results.

OPTIMISING PERFORMANCE

Then we increased max_length to 128 and it ended up increasing the validation accuracy. To allow the model to be able to learn the higher length, increased the number of epochs to 8.



The minimal loss happens at Epoch 3, and its less than the lowest loss in earlier hyper-parameter variations.



Peak validation accuracy is 95% which is pretty good.

```
======== Epoch 3 / 8 ========
Training...
  Batch    40  of    159.    Elapsed: 0:00:13.
  Batch    80  of    159.    Elapsed: 0:00:27.
  Batch   120  of    159.    Elapsed: 0:00:40.

  Average training loss: 0.01
  Training epcoh took: 0:00:53

Running Validation...
  Precision: 0.93
  Recall: 0.94
  F1 Score: 0.93
  Validation Loss: 0.32
  Validation took: 0:00:02
```

The other metrics also went up for this optimal combination of hyper-parameters

Hyperparameter Configuration

Below is an image displaying the hyper-parameter values corresponding to the optimal results

Hyperparameter	Value
max_length	128
batch_size	16
num_labels	4
learning_rate (lr)	2e-5
epsilon (eps)	1e-8
epochs	8
num_warmup_steps	0
num_training_steps	num_batches * epochs
gradient_clipping	1.0

Conclusion and Future Work

This research uses NLP to classify epistemic certainty levels in text, achieving 95% accuracy with BERT models on the FactBank 1.0 dataset. "Epistemic bias" here refers to modality, highlighting cognitive biases in perceptions of truth and knowledge. While we were not ultimately successful in classifying the translated philosophical texts for epistemic bias, we made great progress in training our BERT model on the FactBank 1.0 dataset. Additionally, we advanced a methodology for generating local datasets quickly with the scripts that we created. If we wanted to advance this project further we could utilize those scripts for generating text-specific datasets. It might also behoove us to check out some of the single author epistemic bias datasets as well as choosing texts that have more events than *The Republic* or *The Will to Power* did – perhaps comparing news stories of the same event from different news sources with different political biases would be good.

Individual Contributions

Zachary Meeks

1. Designed the original projects aims
2. Researched philosophy books that would make good candidates for this experiment
3. Converted book pdf files into text files and separated footnotes from main texts
4. Acquired dataset from University of Pennsylvania
5. Spent hours with Abdelazim trying to figure out how to get event-expressions out of the 20 database tables
6. Created scripts to generate our own datasets from local texts
7. Helped others with stumbling blocks, etc

Abdelazim Lokma

1. Analyzed the FactBank 1.0 dataset, importing the data, cleaning and processing it.
2. Calculated and extracted metrics on philosophical translations for report.
3. Conducted an exploratory data analysis on the FactBank 1.0 dataset and the Translated philosophical works to gain a deeper understanding of them.

Gauravdeep Singh Bindra

1. Wrote script to clean Plato and Nietzsche text and notes files.
2. Researched NLTK metrics that could be useful for our project.
3. Wrote code for fine-tuning BERT model
4. Calculated validation metrics for model performance

Appendix and References

1. "Examining Political Rhetoric with Epistemic Stance Detection." arXiv. Accessed [Your Access Date]. https://arxiv.org/pdf/2212.14486.pdf.

2. Example explaining :

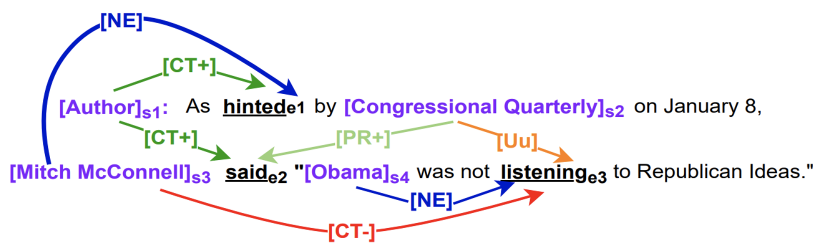


Figure 1: Illustrative example, simplified and adapted from a sentence in the Mass Market Manifestos corpus. There are four sources (s1–s4) and three events (e1–e3) with $4 \times 3 = 12$ labels between them; all epistemic stances are shown, but most non-epistemic (NE) labels are hidden for clarity. §1 and §3 describe the labels.

Consider an example shown in Figure 1, where the author of the text (s1) quotes a speculation from the Congressional Quarterly (s2) about what Mitch McConnell (s3) said concerning Obama (s4). In this example, while the author of the text believes that the Congressional Quarterly hinted something about McConnell (thus, exhibiting a *certainly positive* (*CT*+) stance towards the event (e1), she remains *uncommitted* (*Uu*) about the quoted event (e3) that McConnell describes (edge omitted for visual clarity). Of course, this event is asserted as *certainly negative* (*CT*-) by McConnell, the speaker of the quote. The Congressional Quarterly suggests that Mitch McConnell made a statement (a *probably positive* (*PR*+) stance towards e2) while remaining *uncommitted* towards what he said. Finally, *Obama*'s own beliefs about whether he paid attention to Republican ideas are not expressed in this sentence; thus, s4 (Obama) has a *non-epistemic* label toward the listening event (e3).

3. . Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT Explained: State of the Art Language Model for NLP." Towards Data Science. Accessed [Your Access Date]. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

4. . Vaswani, Ashish. "Transformers Lecture." YouTube. https://www.youtube.com/watch?v=xN0HTuIMibE&t=4850s.

5. . Saurí, Roser, and Martha Palmer. "A Factuality Profiler for Eventualities in Text." Language Resources and Evaluation 43, no. 3 (2009): 227-268. https://link.springer.com/article/10.1007/s10579-009-9089-9.

6. . "FactBank: A Corpus Annotated with Event Factuality." Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2009T23.