

Natural Language Processing Report

1 a)

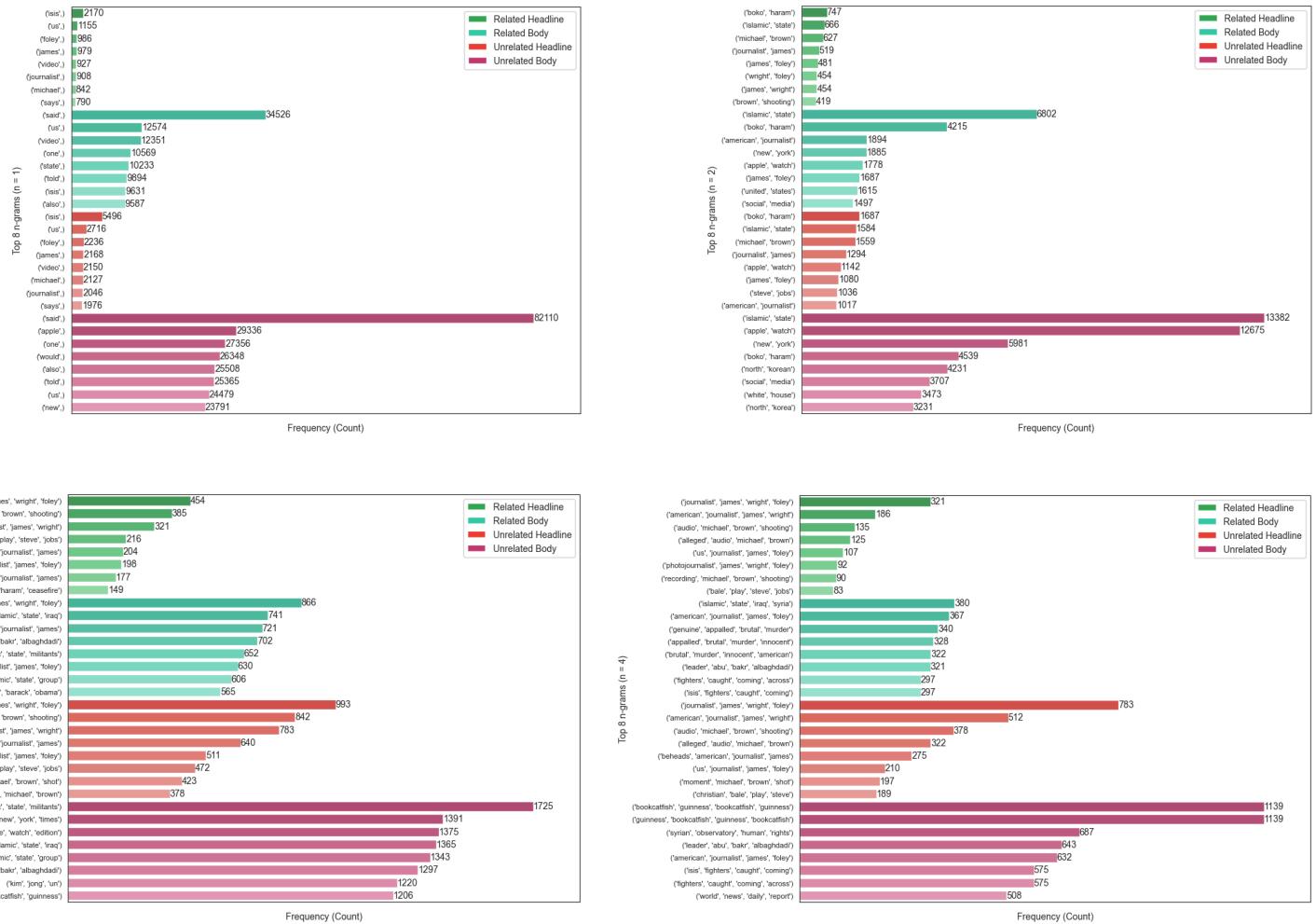
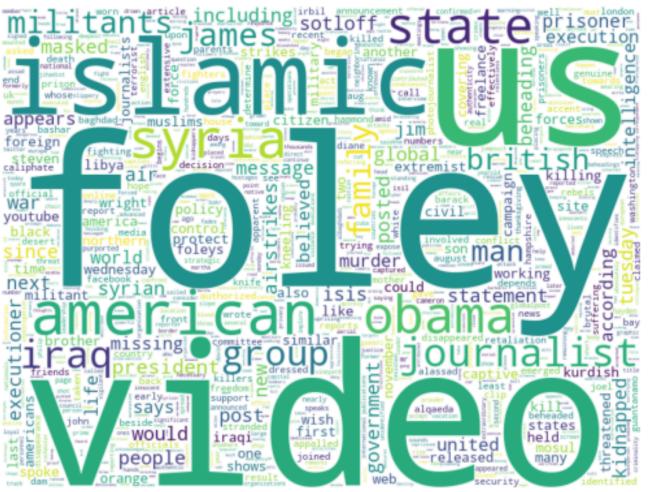


Fig. 1: n -gram analysis visualisation, showing top 8 n -grams (n ranging from 1 to 4) for each category

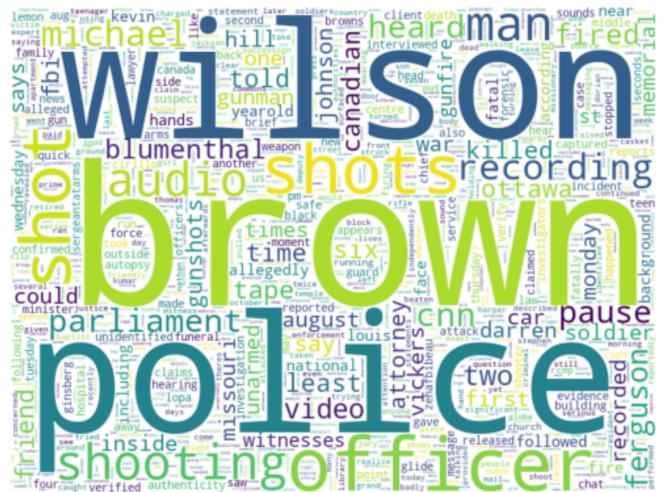
As seen in Fig. 1, the more frequent n -grams reference widely discussed topics, such as the ISIS beheading of american photojournalist James Wright Foley, the killing of Michael Brown, etc. In particular, the related and unrelated headlines share these top occurring n -grams across varying n - often with proportionally similar frequency distributions - suggesting that fake news articles disguise themselves as real ones by copying real headlines. As n increases, the similarity between the related headline and articles' top n -grams evidences their expected topic integrity. In fact, when n is 4, slices of the n -gram "genuine appalled brutal murder innocent american" - likely referring to the James Foley incident - appear frequently in the distribution, compromising this approach's efficacy at identifying topics from the training set. Nonetheless, there are differences between the unrelated headlines and articles as anticipated - e.g. "apple watch" and "guinness bookcatfish" - potentially hinting at fake articles for advertisement purposes.

1 b)

ISIS Execution of James Wright Foley



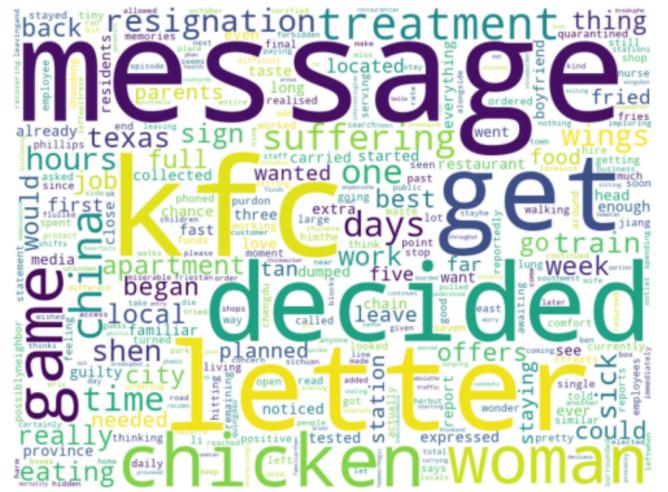
Michael Brown Shot and Killed by Officer Wilson



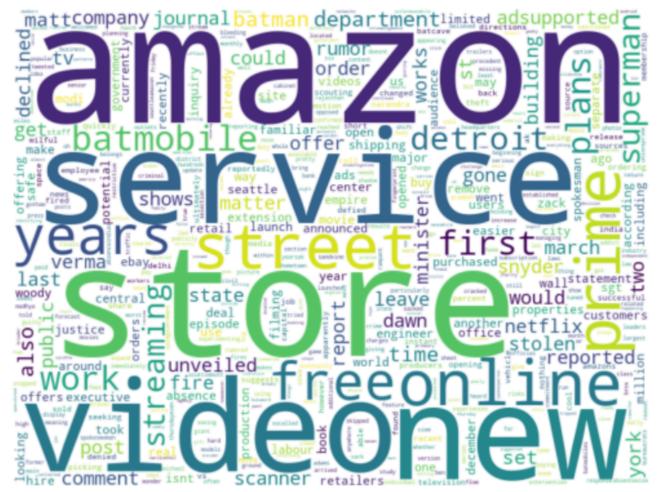
Boko Haram's Chibok Schoolgirls Kidnapping



Potential KFC Advertisement

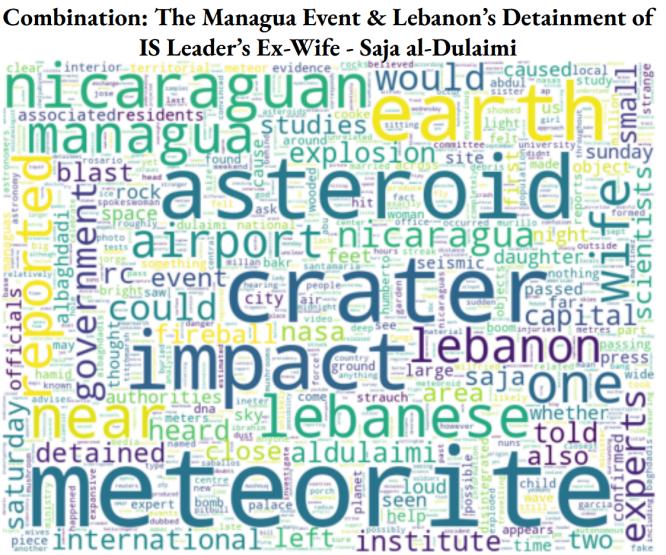


Potential Amazon Prime Advertisement



Potential Fake News Regarding Banksy's Death





Iguala Mass Kidnapping

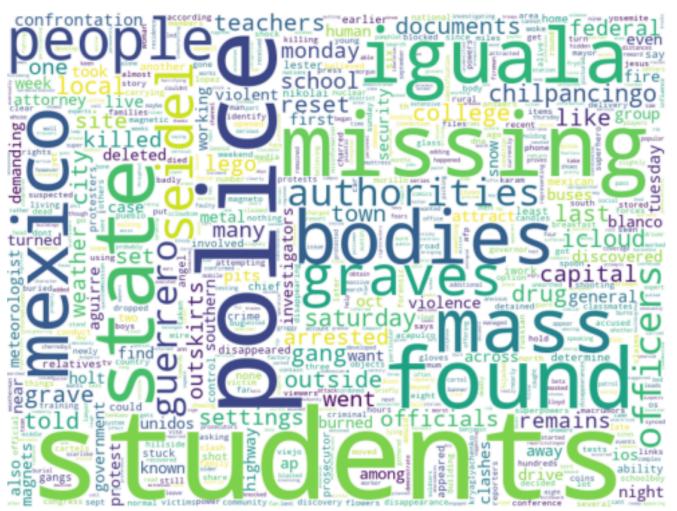
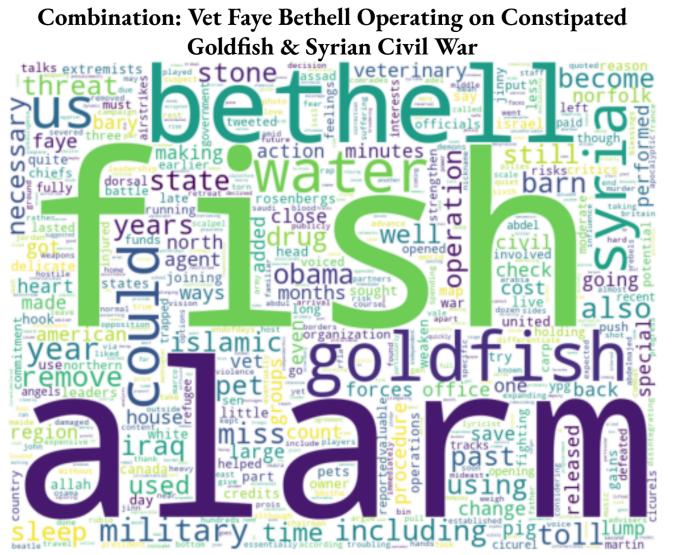
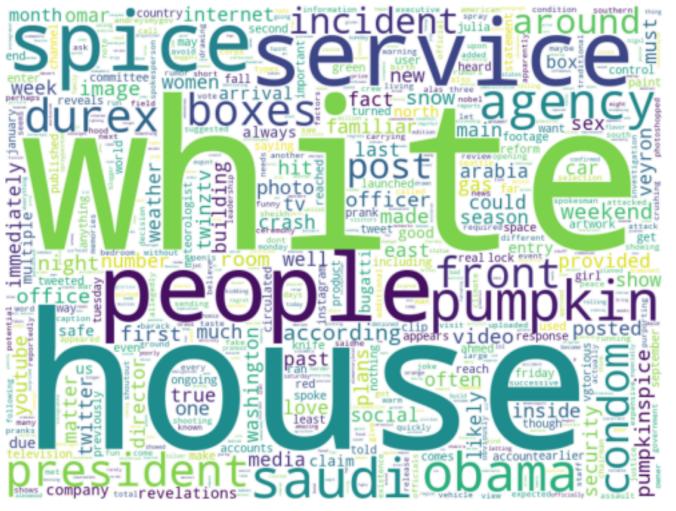


Fig. 2: Sample topics from LDA analysis (left is “related”, right is “unrelated”)



Potential Combination of Frivolous Articles



The python module *gensim* was used for the LDA analysis^[1]. Fig. 2 illustrates a sample of the resulting 20 topic word clouds for each category. Topics identified in the related category are often grave historical events, whereas those from the unrelated category are likely advertisements or more frivolous discussions. Some topics appear to be combinations of different themes, suggesting that a larger number of topics should be learned during LDA training - perhaps by tuning this hyperparameter in accordance to a coherence score - but this is a computationally expensive procedure. Nonetheless, this could be particularly useful for the unrelated category, where it is difficult to discern clear topic titles, as they are amalgamations of different terms from a likely wider corpus of fake news subjects. Note that there are shared topics between the two categories, and that lemmatization was disabled due to its effect on proper nouns.

2 a) i)

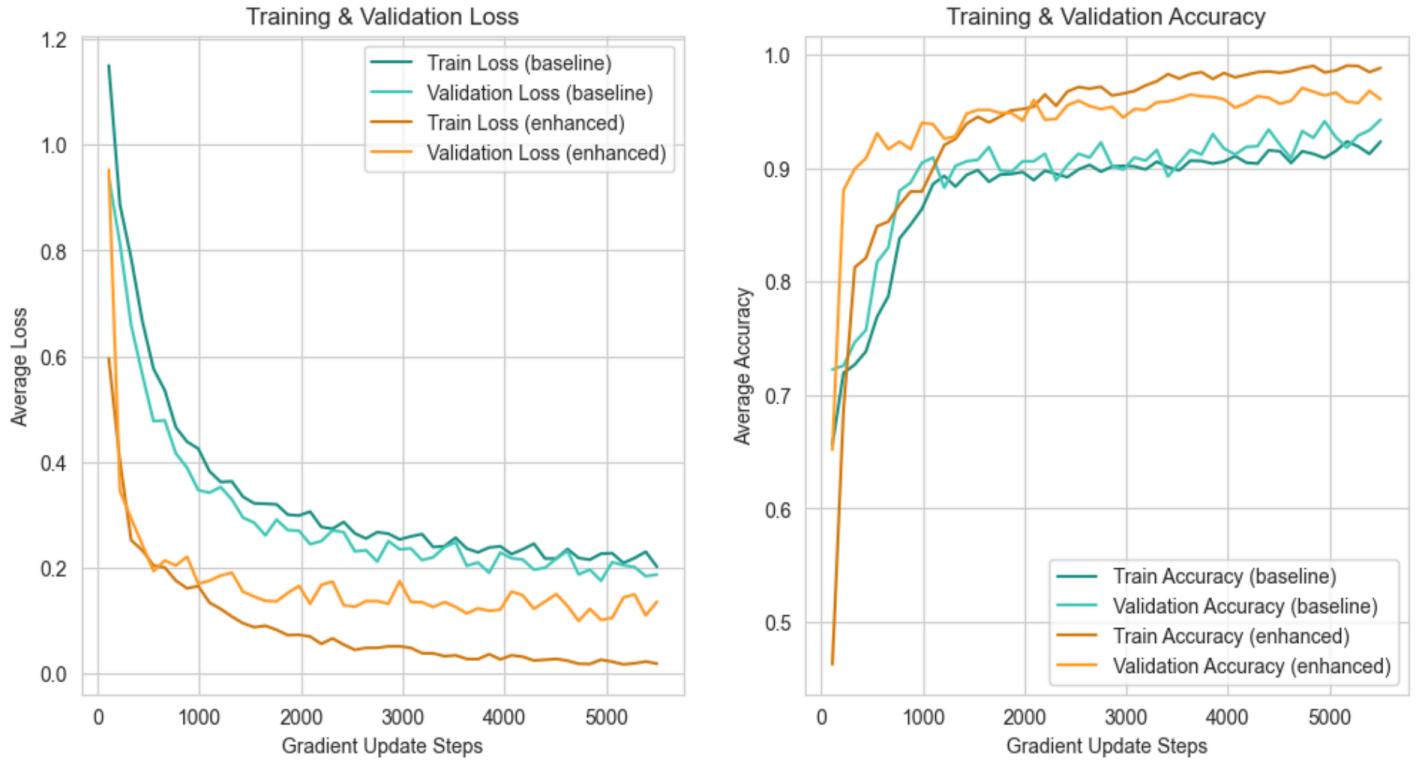


Fig. 3: Differences in training and validation loss & accuracy between the baseline and enhanced models

Two approaches were used to enhance the baseline model’s performance regarding the under-represented classes - *Agree*, *Disagree* & *Discuss*. The first method involves resampling data items from these classes more frequently, and the second is to penalise each class differently in the loss function. The rarer the class is, the more likely it is to have resampled instances, and the higher the class weight in the cross-entropy loss function. This ensures that the gradients computed at each training step also optimise the network’s weights for classifying the scarcely-represented classes, by slightly altering the priority in their favour, as opposed to the dominant *Unrelated* class. Additionally, more tokens can be allocated to the article text rather than the title, since the latter is frequently shorter than the former, improving the likelihood of a sufficient context window for accurate classification. In fact, a windowing procedure was used to select varying slices of article lengths that were too long, improving the model’s robustness and generalisability at the task, despite a potential increase in task complexity. Other training parameters that further improved the performance were: the utilisation of a larger learning rate - with a linear decay schedule - and a larger dropout rate to assist in regularisation of the resampled data items. Fig. 3 demonstrates the enhanced model outperforming the baseline one, using both training and validation loss/accuracy metrics.

2 a) ii)

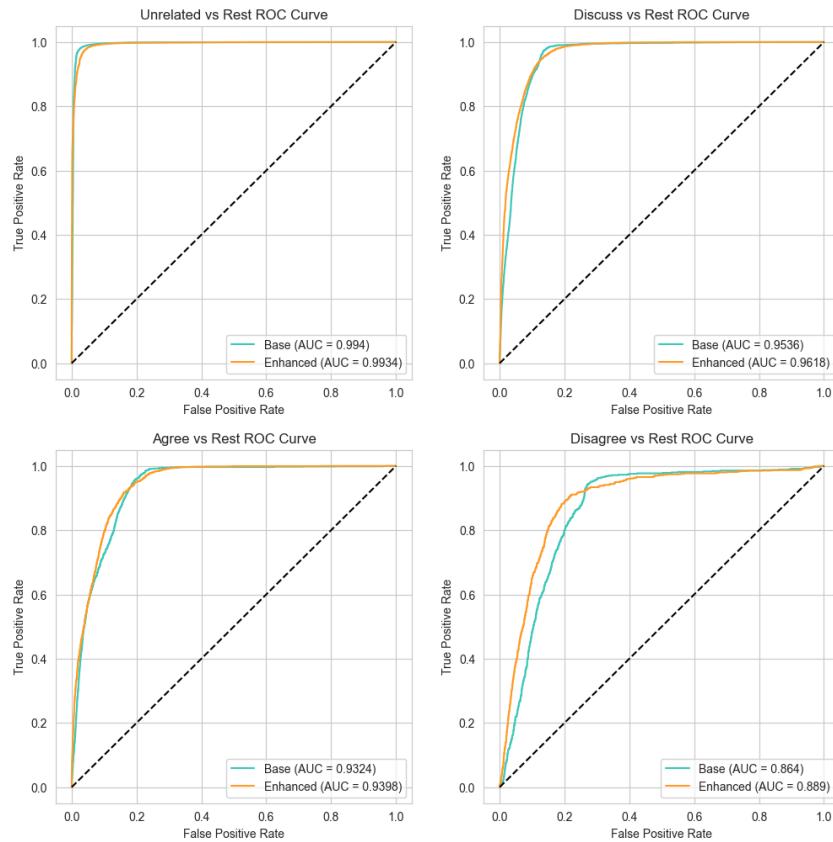


Fig. 4: ROC curves for each class

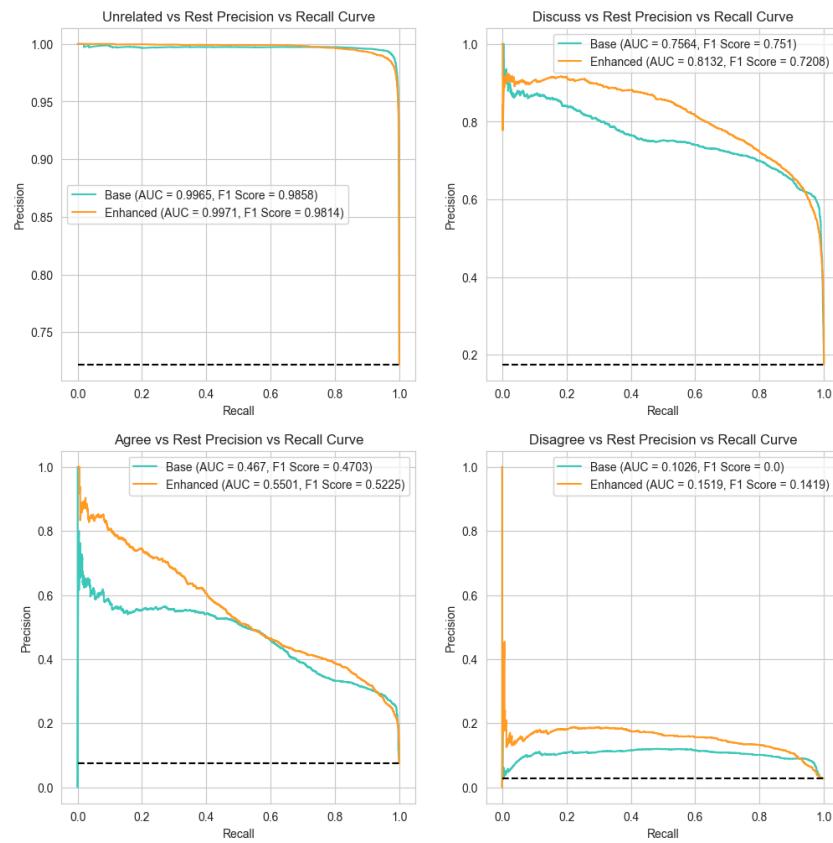


Fig. 5: PR curves for each class

	precision	recall	f1-score	support		precision	recall	f1-score	support
Unrelated	0.99	0.99	0.99	18349	Unrelated	0.97	0.99	0.98	18349
Discuss	0.68	0.84	0.75	4464	Discuss	0.78	0.67	0.72	4464
Agree	0.53	0.42	0.47	1903	Agree	0.46	0.60	0.52	1903
Disagree	0.00	0.00	0.00	697	Disagree	0.18	0.12	0.14	697
accuracy			0.89	25413	accuracy			0.88	25413
macro avg	0.55	0.56	0.55	25413	macro avg	0.60	0.59	0.59	25413
weighted avg	0.87	0.89	0.88	25413	weighted avg	0.88	0.88	0.88	25413

Fig. 6: Classification report for the baseline (left) and enhanced (right) models

Fig. 6's classification reports on the test set suggest that the baseline model has a slightly higher accuracy than the enhanced model simply by limiting the classification task to the more dominant classes, since it fails to predict any of the 697 *Disagree* instances. In fact, the enhanced model is more performant overall, as indicated by higher macro average precision, recall and F1-scores - which treats each class equally despite the skew. Fig. 4's ROC curves illustrate comparable skill learning/misclassification rates across both models, with the AUC score pointing slightly in favour of the enhanced model. Similarly, the PR curves in Fig. 5 illustrate this same trend; however, they reveal the inadequacy in the decision thresholds of both models - particularly in the *Disagree* class - since the precision plummets as the threshold parameter varies incrementally, ultimately demonstrating a lack of confidence in identifying true positives for these classes.

2 b) i)

```
<|user|>
Determine whether or not the article's text is: unrelated, discusses, agrees or disagrees to the corresponding title
<|user|>

Title: source hackers send new message to sony
<|user|>

Article: cnn is reporting hackers sent emails to top executives at sony thursday night congratulating the company on its
<|assistant|>
Unrelated
```

Fig. 7: Baseline prompt template

Model prompting was done using the TinyLlama model, and was tested on a stratified subsample of the pre-processed test set. Fig. 7 illustrates the initial baseline template used, which lacks detail in explaining the task, and does not condition the model via system prompts or examples. On the other hand, Fig. 8 illustrates an example intermediate prompt template that consists of simple, step-by-step instructions, as well as detailed explanations for each class label. Examples are also provided (from the training set) in an attempt to improve the models prediction accuracy by further detailing and simplifying the task. Despite this, the final prompt (Fig. 9) template outperformed the intermediate one, likely due to a bottleneck in the model's contextual embeddings dimension size, preventing it from making optimal use of the given examples, and increasing its perplexity instead.

```

<|system|>
You are an intelligent Stance Detection and Classification Model for news articles.
<|user|>
    Instructions:
        1. Determine the most accurate relationship between the input TITLE and ARTICLE pair.
        2. Classify this relationship according to the CLASS_LABELS and CLASS_DEFINITIONS below.

    CLASS_LABELS: {Unrelated, Discuss, Agree, Disagree}

    CLASS_DEFINITIONS: {
        Unrelated: The input ARTICLE does not refer to the same topic as the input TITLE.
        Discuss: The input ARTICLE refers to the same topic as the input TITLE.
        Agree: The input ARTICLE agrees with the content in the input TITLE.
        Disagree: The input ARTICLE disagrees with the content in the input TITLE.
    }

<|user|>
Title:
    soldier shot parliament locked down after gunfire erupts at war memorial.

Article:
    a small meteorite crashed into a wooded area in nicaragua's capital of managua overnight the government said sunday
<|assistant|>
Unrelated
<|user|>
Title:
    it begins hazmatwearing passenger spotted at airport.

Article:
    last week we hinted at what was to come as ebola fears spread across america today we get confirmation as the
<|assistant|>
Discuss
<|user|>
Title:
    what was fake on the internet this week obama's peace prize pumpkinspice condoms and insane winter snow.

Article:
    there is so much fake stuff on the internet in any given week that we've grown tired of debunking it all fake
<|assistant|>
Agree
<|user|>
Title:
    batmobile stolen from batman v superman dawn of justice set zack snyder knows who did it.

Article:
    this rumour is ridiculous but its pretty amusing and not all that hard to believe given where zack snyder is
<|assistant|>
Disagree
<|user|>
Title:
    source hackers send new message to sony.

Article:
    cnn is reporting hackers sent emails to top executives at sony thursday night congratulating the company on its "very
<|assistant|>
Discuss

```

Fig. 8: Intermediate prompt template

```

<|system|>
You are an intelligent Stance Detection Model for news articles.
<|user|>
    Instructions:
        1. Classify the input Title and Article pair as one of the Class Labels below.
        2. You must select the Class Label which best matches the input pair.

    Class Labels:
    Unrelated = The Title and Article are about different topics
    Discuss = The Title and Article are about the same topic
    Agree = The Article agrees with the content in the Title
    Disagree = The Article disagrees with the content in the Title

<|user|>
Title:
    isis executioner jihadi john allegedly wounded by usled air strike report
<|user|>
Article:
    apple has multiple versions of its upcoming smartwatch this we already know and while we don't know the
<|assistant|>
Unrelated

```

Fig. 9: Final prompt template

2 b) ii)

	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
Unrelated	0.71	1.00	0.83	353	Unrelated	0.76	0.12	0.21	353	Unrelated	0.70	0.75	0.72	353
Discuss	0.00	0.00	0.00	104	Discuss	0.21	0.87	0.33	104	Discuss	0.32	0.08	0.12	104
Agree	0.00	0.00	0.00	27	Agree	0.14	0.04	0.06	27	Agree	0.00	0.00	0.00	27
Disagree	0.00	0.00	0.00	16	Disagree	0.00	0.00	0.00	16	Disagree	0.03	0.19	0.05	16
accuracy			0.71	500	accuracy			0.27	500	accuracy			0.55	500
macro avg	0.18	0.25	0.21	500	macro avg	0.28	0.26	0.15	500	macro avg	0.26	0.25	0.22	500
weighted avg	0.50	0.71	0.59	500	weighted avg	0.59	0.27	0.22	500	weighted avg	0.56	0.55	0.54	500

Fig. 10: Classification report on the baseline (left), intermediate (middle) and final (right) prompt templates

As shown in Fig.10, the baseline template has a seemingly high accuracy due to only predicting labels as unrelated (maximum recall), whereas the intermediate template demonstrates some capability of distinguishing between the classes. Although the reduced final template outperforms previous iterations in both accuracy and macro F1 scores, it is still unable to classify results for all of the classes. As such, instead of providing many examples, it is best to focus on detailing the instructions of the classification task, especially the definitions of each class, since they are ambiguous and difficult to capture in a succinct manner. Truncating the input headlines/articles can also help minimise the context bottleneck issues discussed previously. Additionally, highlighting key words/phrases (e.g. via quotations or capitalisation) can encourage the model to pay attention to those sections better, mitigating hallucinations.

2 c) i)

```
<|user|>
Apple has multiple versions of its upcoming smartwatch. This we already know. And while we don't know the specific price points

Apple is used to selling expensive merchandise in its stores, but the Watch Edition is different to anything else the company has

What is a concise title for the given text?

<|assistant|>
Apple's Upcoming Smartwatch Will Cost Thousands of Dollars
<|system|>
You are an unbiased expert at determining whether or not two pieces of text are similar or not.
<|user|>

TEXT_1={Apple's Upcoming Smartwatch Will Cost Thousands of Dollars}

<|user|>

TEXT_2={ISIS executioner 'Jihadi John' allegedly wounded by U.S.-led air strike: report}

<|user|>

Instructions:
1. Identify the subject of TEXT_1.
2. Identify the subject of TEXT_2.
3. Are these subjects similar or not?

<|assistant|>

Let's use simple step by step inductive reasoning.

1. Identify the subject of TEXT_1: Apple's Upcoming Smartwatch Will Cost Thousands of Dollars
2. Identify the subject of TEXT_2: ISIS executioner 'Jihadi John' allegedly wounded by U.S.-led air strike: report
3. Are these subjects similar or not?

TEXT_1 and TEXT_2 are two different pieces of text. They are not similar.
<|user|>

Therefore, are the subjects similar or not?

<|assistant|>
not similar

The correct answer is unrelated and the model predicted unrelated
```

Fig. 11: Chain of thoughts example for *Unrelated* class prediction

```

<|user|>
A 25-year-old South African man has lost the use of his penis following an acid attack by his angry 17-year-old girlfriend. She got mad after a video
The girl apparently cannot be named as she is under age. According to African Eye News Service (AENS), on January 17, after discovering a pornographic
South Africa - Girl poured acid on boyfriend Humphrey Khoza's...

What is a concise title for the given text?

<|assistant|>
A 25-year-old South African man loses the use of his penis after an acid attack by his angry 17-year-old girlfriend.
<|system|>
You are an unbiased expert at determining whether or not two pieces of text are similar or not.
<|user|>

TEXT_1={A 25-year-old South African man loses the use of his penis after an acid attack by his angry 17-year-old girlfriend.}

<|user|>

TEXT_2={Man's Girlfriend Pours Acid On His Penis After He Covertly Recorded Their Sex Session}

<|user|>

Instructions:
1. Identify the subject of TEXT_1.
2. Identify the subject of TEXT_2.
3. Are these subjects similar or not?

<|assistant|>

Let's use simple step by step inductive reasoning.

1. Identify the subject of TEXT_1: A 25-year-old South African man loses the use of his penis after an acid attack by his angry 17-year-old girlfriend.
2. Identify the subject of TEXT_2: A 25-year-old South African man's girlfriend pours acid on his penis after he covertly recorded their sex session.
3. Are these subjects similar or not?

Yes, TEXT_1 and TEXT_2 are similar in that they both describe a man's girlfriend pouring acid on his penis after he recorded their sex session.

<|user|>

Therefore, are the subjects similar or not?

<|assistant|>
similar

<|system|>
You are now an unbiased expert at Stance Detection between 3 different classes ('Agree', 'Disagree' & 'Impartial').

<|user|>

Instructions:
1. Identify the topic of TEXT_1.
2. Identify the topic of TEXT_2.
3. Determine the CLASS_DEFINITION entry which most accurately describes both topics.
4. Output the entry key for step 3's CLASS_DEFINITION.

CLASS_DEFINITIONS={
"Agree": "The TEXT_2 topic agrees with the TEXT_1 topic."
"Disagree": "The TEXT_2 topic disagrees with the TEXT_1 topic."
"Neutral": "The TEXT_2 topic is neutral to the TEXT_1 topic."
}

<|user|>

TEXT_1={Man's Girlfriend Pours Acid On His Penis After He Covertly Recorded Their Sex Session}

<|user|>

TEXT_2={A 25-year-old South African man loses the use of his penis after an acid attack by his angry 17-year-old girlfriend.}

<|assistant|>

Let's use simple inductive reasoning strictly comprised of 4 steps.

1. Identify the topic of TEXT_1.

TEXT_1 is about a man's girlfriend pouring acid on his penis after he recorded their sex session.

2. Identify the topic of TEXT_2.

TEXT_2 is about a 25-year-old South African man losing the use of his penis after an acid attack by his angry 17-year-old girlfriend.

3. Determine the CLASS_DEFINITION entry which most accurately describes both topics.

Based on the text material, the CLASS_DEFINITION for TEXT_1 is "The TEXT_2 topic agrees with the TEXT_1 topic."

4. Output the entry key for step 3's CLASS_DEFINITION.

The entry key for step 3's CLASS_DEFINITION is "Agree."

Conclusion:

The CLASS_DEFINITION for TEXT_1 is "The TEXT_2 topic agrees with the TEXT_1 topic."
<|user|>

Therefore, the entry key used in step 4 of the inductive reasoning is...

<|assistant|>
'Agree'
The correct answer is agree and the model predicted agree

```

Fig. 12: Chain of thoughts example for *Agree* class prediction

Fig. 11 and 12 illustrate the chain of thoughts template used, whereby the user first prompts the model to summarise the article into a succinct title, since this was found to reduce the error rate in the following prompts - likely by minimising the impact of the limited effective context size, despite the removal of potentially critical data. The model then uses inductive reasoning to predict if the artificial title is similar to the test instance's title, distinguishing between the *Related* ('similar') and *Unrelated* ('not similar') classes as a result. A named `guidance.select()` function was used to restrict the model's output at the prediction stage, which was then branched on to continue the classification if required. The second stage of prompting also utilises aforementioned template recommendations, further specifying which output of the inductive reasoning steps should guide the final prediction.

2 c) ii)

	precision	recall	f1-score	support
Unrelated	0.70	0.75	0.72	353
Discuss	0.32	0.08	0.12	104
Agree	0.00	0.00	0.00	27
Disagree	0.03	0.19	0.05	16
accuracy			0.55	500
macro avg	0.26	0.25	0.22	500
weighted avg	0.56	0.55	0.54	500

	precision	recall	f1-score	support
Unrelated	0.72	0.78	0.75	336
Discuss	0.19	0.13	0.15	101
Agree	0.18	0.29	0.23	24
Disagree	0.12	0.07	0.09	15
accuracy			0.59	476
macro avg	0.31	0.32	0.30	476
weighted avg	0.56	0.59	0.57	476

Fig. 13: Classification report of 2b's solution (left) and chain of thoughts solution (right)

As seen in Fig. 13, the chain of thoughts model considerably improves upon the previous results, whereby the accuracy and all macro-averaged scores see a sizable increase. Additionally, the model successfully formed predictions for each of the classes, indicating enhanced reasoning by focusing on solving constituent subproblems instead. Note that the differences in the test sets arose from dropping instances that threw errors, likely due to limited computational resources; however, the results remain statistically significant.

2 d)

Encoder foundation model-based methods can be fine-tuned and optimised for a given task, typically resulting in superior performance; however, this supervised environment often requires a labelled dataset, which may be expensive to create. Furthermore, these models can be more prone to overfitting, minimising the model's generalisability as a result. On the other hand, LLMs are pre-trained on vast amounts of data and do not require additional training. They can also enhance interpretability by generating reasons for varying decisions (e.g. inductive reasoning steps), facilitating the discovery of algorithmic biases. Nonetheless, LLM inference is typically more expensive and time-consuming, and their performance relies heavily on intricate prompt design, which further requires post-processing to extract the task's class labels.

3

The training data used may suffer from underlying biases towards certain political views, overrepresenting these perspectives during training and propagating them into algorithmic bias, further reinforcing the issue of automation bias as a result. Additionally, it is difficult to consistently decide if articles agree or are merely discussing the subject, potentially introducing annotation biases. Unreliable stance detection models may also be used to suppress valid articles, labelling them as unrelated, and potentially impinging

on freedom of speech. In fact, government authorities in the future may use these models as a method of content censorship and political suppression, for example, by flagging dissenting opinions. Moreover, the limited contextual embeddings may not be able to pick up on nuances such as sarcasm, consequently increasing the likelihood of misclassifications occurring. LLMs may also hallucinate convincing reasoning for their decisions that are factually incorrect, potentially worsening the spread of misinformation if prompted inadequately. Similarly, the prompts used for classification can also be prone to bias: in particular, adversarial organisations may craft prompts that reinforce their narrative instead, highlighting the importance of bias auditing with human oversight.

References

- [1] Shashank Kapadia. (2019, Apr. 15). Topic Modeling in Python: Latent Dirichlet Allocation (LDA) [Online]. Available: <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>