Identifying Misinformation in Social Media and News Sources

Steven Sung, Peeti Sriwongsanguan

Abstract

Misinformation is a common problem in social media and news outlets that is tedious and difficult to identify. We propose a transformer model that automates this task by flagging claims based on article context. We use the WatClaimCheck dataset which contains aggregated data from professional fact checkers that contains claims, review articles, premise articles, and many more. The best performing model is the BERT model using review articles which obtained a F1 score of 0.94. However, we anticipate that it is overfitting. Therefore, we look at exploring alternatives such as DPR and time-series cross validation data.

Introduction

Misinformation on social media platforms has been running rampant in the past few years with several false claims and accusations. It carries many consequences including confusion, fear, and panic that can be harmful to individuals and society. For example, misinformation about the COVID-19 vaccine being dangerous caused people to not take the drug which resulted in many preventable deaths.

Identifying misinformation is often easier said than done. False claims are often disguised within legitimate information and propagate through different mediums like social media and news outlets. Additionally, people are more likely to believe in misinformation that aligns with their pre-existing beliefs and biases. As a consequence, catching fake news before it spreads is near impossible.

It is imperative for people to obtain news from reliable sources and make informed decisions. Therefore, an NLP model is proposed to detect and classify fake news which utilizes transformers and large language models to identify misinformation by analyzing the context of the articles and flagging claims.

Background

There are many types of misinformation such as click bait, political bias, and government propaganda that require extensive research and data collection. However, addressing all of them stretches beyond the scope and time of the course. Therefore, we will be focusing on fake news by analyzing the WatClaimCheck dataset that provides evidence refuting or supporting a claim.

WatClaimCheck

The main motivation of our research is based on a dataset from the paper “WatClaimCheck: A new Dataset for Claim Entailment and Inference” from the University of Waterloo, which is designed and curated for automated fact checking. The process of collecting data is very similar to a professional fact checker where they aggregate articles based on the relevance and context of the claim and write a review article to verify the veracity of the claim;

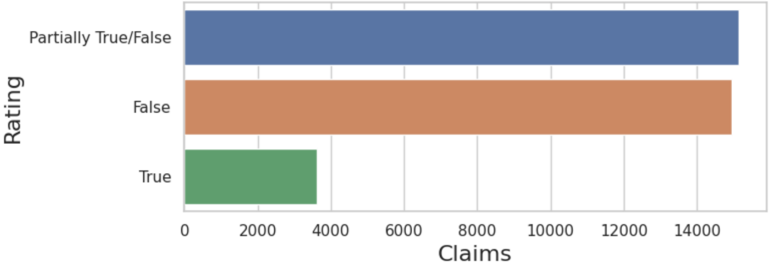
Transformers

Transformers will be the main focus of our research paper due to its widespread adoption and power. It is a type of deep neural network architecture that leverages attention to learn the relationship between words and tokens in a sequence (Vaswani et al. 2017). Transformers can be trained on the WatClaimCheck dataset by feeding it claims and articles to classify whether a claim is true or false.

Data

The data contains three json files for train, validation, and test, each containing metadata and labels. To transform the data into a pandas dataframe, the metadata and labels are exploded to retrieve the variables inside. The table of variables that we used to train the model is located in the appendix.

EDA



The distribution of the ratings, which determines the veracity of the claim, are highly skewed, showing more partially true/false and false compared to true statements. Therefore, using macro F1 score as the main metric would be suitable for their purposes. Categorical accuracy is also mentioned as a secondary metric for models that utilize an undersampled or oversampled variation of the dataset. Since our problem already contains labeled data, our objective is to train a model with the highest evaluation metric using F1 score since this is the main metric used in the WatClaimCheck paper.

Addressing Data Imbalance

The imbalanced dataset posed a challenge as it would make our model more biased, increasing the chances of our model classifying a claim as false or partially true/false. We attempted to address this problem by rebalancing the data using oversampling and undersampling techniques. However, we were unable to improve our baseline model performance. Therefore, we decided to use the untreated dataset, understanding that rebalancing would cause difficulties. We also recognized that future model development would require further investigation. To detect article veracity accurately, we will use the CNN model as a baseline to evaluate more complex models such as RoBERTa and BERT transformers.

Truncating Article Sequence Length

Articles can contain words and sentences unrelated to the claim. For example, there are many social media text phrases like link and share as well as foreign languages. We found that removing 1000 characters from the beginning and ending improves the quality of the articles. Additionally, the articles have a large corpus, often ranging from 500 to 2000 tokens. By truncating the article length to around 500 characters, the main content of the review article is conserved.

Methods: Exploring CNN, RoBERTa, and BERT

Baseline Models: ANN, CNN, and RNN

As part of their initial exploration, we examined a variety of neural network architectures, focusing on ANN, CNN, and RNN. We trained the neural networks using different hyperparameters including different layer sizes and training with the variations of the sampled data. By experimenting our data initially with neural networks, we can gain a better understanding of what data preprocessing should be done.

Advanced Models: RoBERTa and BERT with CNN

Our architecture incorporates two powerful transformer models, RoBERTa and BERT, to enhance article veracity detection capabilities. By combining transformer attention mechanisms with CNN feature extraction capabilities, we aim to achieve more nuanced verification assessments. These models have demonstrated exceptional performance in natural language processing tasks and can detect robust article authenticity more accurately compared to its CNN counterpart. As part of the evaluation, key metrics will be compared to the established baseline, such as accuracy, precision, recall, and F1-score.

We compare different variations of transformers by changing the input categories and combining different categories such as claims, review articles, and premise articles. The information is passed through a tokenizer to map the passage into vectors to train our transformers. Once the transformers are trained and configured, the model outputs a softmax dense layer of three classes representing each of the labels in ratings.

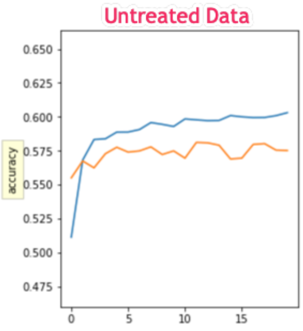
Dense Passage Retrieval

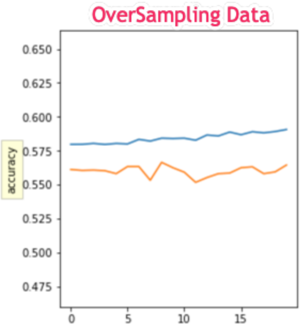
Review articles often have a high correlation with the claims themselves. However, this is not indicative of how facts are checked. Premise articles are collected before review articles are created to then state a claim. To address this problem, an intermediate Dense Passage Retrieval (DPR) model is trained on claims and review articles since review articles mainly contain evidence from premise articles. Usually, evidence is paraphrased to form a coherent argument in support of the claim veracity model (Khan et al. 2022). During inference, premise articles are passed into the model to calculate the cosine similarity with the claim.

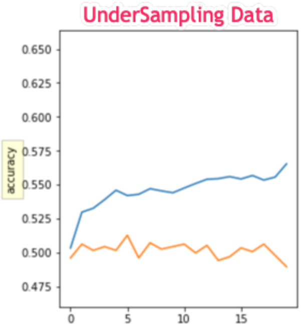
Results and Discussion

Baseline: ANN

In our initial baseline model, we used an ANN on the imbalanced, oversampled, and undersampled datasets. However, the model gave unsatisfactory results. After training for 10 epochs across all three datasets, the accuracy for all training and test sets remained at 50%. Due to the underwhelming performance in all untreated and treated imbalanced datasets, we decided to keep the untreated dataset and add layers to the baseline.


Untreated Data


OverSampling Data

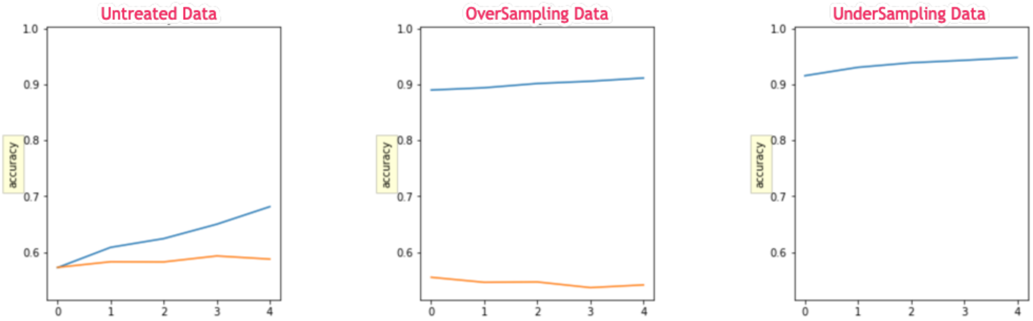

UnderSampling Data

We introduced additional layers before transitioning to transformers in order to enhance the baseline. Although the training performance significantly improved, reaching 96% accuracy, the validation accuracy remained low at 50%. In other words, more layers improved the model's ability to learn from training data, but did not translate into better generalization on unseen data. This is a clear sign of overfitting. We realized that our framework for article veracity detection needed further refinement and exploration as we prepared to incorporate more advanced transformer models. Graphs about the performance over 10 epochs can be found in the appendix.

Elaborating Our Baseline: CNN

As part of our initial exploration, we studied and explored several deep learning architectures, focusing on Convolutional Neural Networks and Recurrent Neural Networks. Interestingly, the CNN outperformed the ANN for our dataset. In spite of our expectation that adding more CNN layers would improve performance, neither a one-layer nor a two-layer model performed significantly better. In order to benchmark the performance of more complex models like RoBERTa and BERT transformers, we selected a single CNN layer as our baseline text classification model. The graph in the appendix shows the graphs per epoch.

The validation metric, however, appears to stop improving as the training metric improves. This indicates that our baseline model performs well on training data, but does not generalize well on new or unknown data. To reduce overfitting, we train advanced models that utilize attention in the next section in order to understand the context behind the claim and review articles.

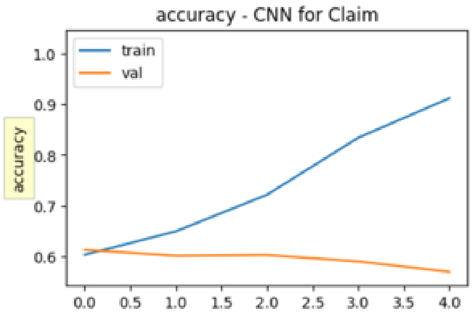


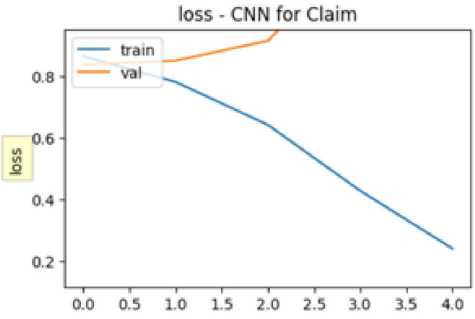
Transformer

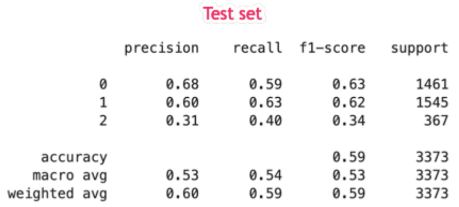
Model 1: BERT Claim Model

In the first model, the claim-targeted model was trained over five epochs and resulted in an accuracy of 93.11%. However, the performance on the test set was challenged with an overall accuracy of 59%. In the classification report, it was evident that there were challenges in correctly identifying the three classes, especially with the minority class of factual claims, which had relatively lower precision, recall, and F1-scores.

The validation loss increases in our claim model because claims themselves are statements with no context surrounding them. Therefore, we can anticipate the model to perform poorly without supporting evidence from the premise or review article. Additionally, the accuracy is similar to the baseline CNN because the claims are fairly short where each claim averages 16 words while the longest claim is around 100 words. The attention mechanism allows transformers to perform better than CNNs with longer sequences.





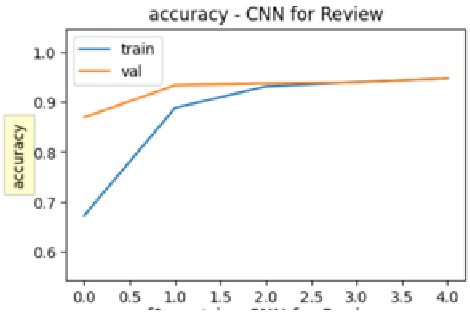

Test set

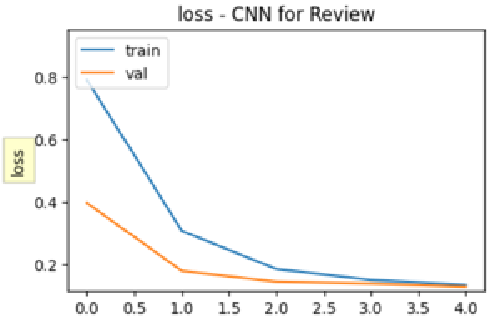
```
              precision    recall  f1-score   support

           0       0.68      0.59      0.63      1461
           1       0.60      0.63      0.62      1545
           2       0.31      0.40      0.34       367

    accuracy                           0.59      3373
   macro avg       0.53      0.54      0.53      3373
weighted avg       0.60      0.59      0.59      3373
```

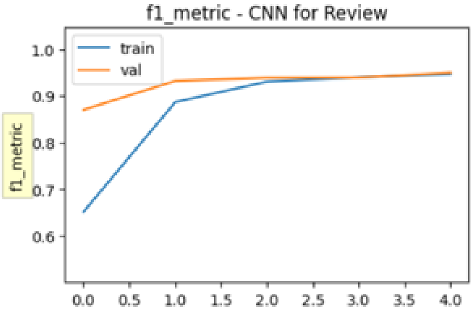
Model 2: BERT Review Article Model

The second model focuses on the veracity of review articles. The training set achieved an impressive accuracy of 95.42%. Based on the test set, the model performed well with an accuracy of 94%. As can be seen from the classification report, the model is able to detect the veracity of review articles with high precision, recall, and F1-score for all classes.

Here, we can see a significant improvement by utilizing the review article. Compared to the first model which only had around 10 words for the training data, we expanded it to 500 characters. This allows the model to train on more sequences relevant to the claim. The review article also contains key words that support the claim, so the context and combination of words can also contribute to the veracity of the article.







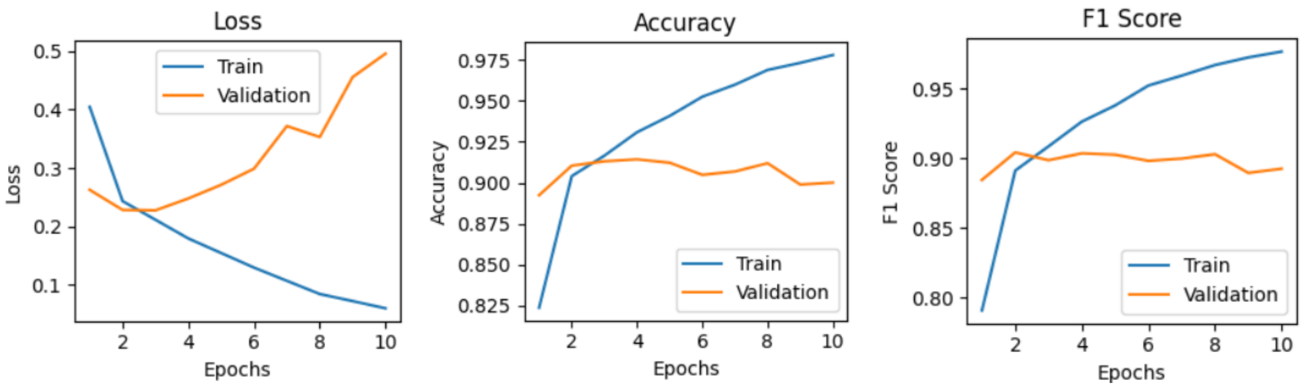
Model 3: BERT Claim and Review Article Model

The third model, which targets both claims and reviews articles, reached an accuracy rate of 96.04% on the training set. In all classes, the test set performance demonstrated a good balance between precision, recall, and F1-score. Again, the classification report shows how well the model can distinguish between classes. We see similar results to the review article because the claim is within the review article, thus creating high correlation between the two variables.

Model 4: RoBERTa Claim and Review Article

On top of the BERT models, we explored different transformers such as RoBERTa which was mentioned in the WatClaimCheck paper. The model was trained similarly to BERT using claim and review articles as the main inputs but contains extremely varied results. The RoBERTa model overfits with the validation loss increasing after 2 epochs. Additionally, the F1 score and accuracy seem to stay consistent throughout each epoch, signifying that the model is not learning. In the research paper, they obtained a 0.741 in macro-F1 whereas ours was around 0.9 with review articles.

We ran multiple iterations of this model and discovered that each one did have a decreasing validation loss. One reason why we think that RoBERTa may differ greatly from BERT is because RoBERTa utilizes dynamic masking generated during training whereas BERT uses static masking. The review articles fed into the model were naively truncated, so they may have masked unwanted phrases such as web links and bullet points that are irrelevant to the claim. Another reason why it may have overfitted is because RoBERTa is trained on a much larger corpus than BERT. Therefore, it might not have been able to generalize nuances of the article.



Model 5: RoBERTa Claim and Premise Article

The Dense Passage Retrieval (DPR) is a model that obtains relevant information from the premise article to support the claim using cosine similarity. However, this variation of the model performs worse than its review article counterpart. The DPR may have trouble obtaining relevant information from the premise article that supports the claim. This is similar to the metrics the authors of the WatClaimCheck dataset experienced where their model had a F1 Score of around 0.58 whereas ours was around 0.5. The difference between the evaluation metric can be explained by how the DPR is initially trained. DPR contains positive and negative articles in order to optimize the negative log likelihood. However, we only provided one instance of positive and negative per claim which can worsen the model performance. Despite its poor performance, this model is most ideal in a practical world where the DPR retrieves relevant information from the web and creates similarities between keywords.

Conclusion

We trained a transformer model to identify misinformation within social media and news outlets and found that using review articles provided the best results. Because the transformer models are few-shot learners, it can perform well without training on multiple epochs as shown between model 2 and 4. We explored different models using variations of input combinations with claim, review articles, and premise articles and saw that just incorporating review articles performed the best.

The model we decided to use is the BERT model with review articles. This model achieved the highest F1 score at 0.95 and accuracy of 0.95 on the test data. Not only did it achieve the highest accuracy, but it also had a max sentence length of 500 words, striking a balance between capturing relevant information and computation efficiency. This is different from the model that the WatClaimCheck authors stated since they utilized RoBERTa and different algorithms to obtain their results, specifically using RoBERTa, DPR, and prequential data.

Despite the BERT model with review articles performing the best, it only works when a review article has been created after gathering the information from the premise articles. This would cause issues in the practical sense because our model would identify words closest to the claim and evaluate the context, despite being true or false. In our next steps, we are looking at improving the DPR model to remove the reliance on review articles which may be a reason for our overfitted models. We also aim to add prequential data stated by the original authors of the WatClaimCheck dataset to mimic how information passes into models over time.

References

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (Cornell University). https://arxiv.org/pdf/1810.04805v2

Khan, K. A., Wang, R., & Poupart, P. (2022). WatClaimCheck: A new Dataset for Claim Entailment and Inference. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2022.acl-long.92

Steven Sung, S., & Sriwongsanguan, P. Identifying Misinformation in Social Media and New Sources [Computer software]. https://github.com/sysung/w266-final-project

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. arXiv (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Appendix

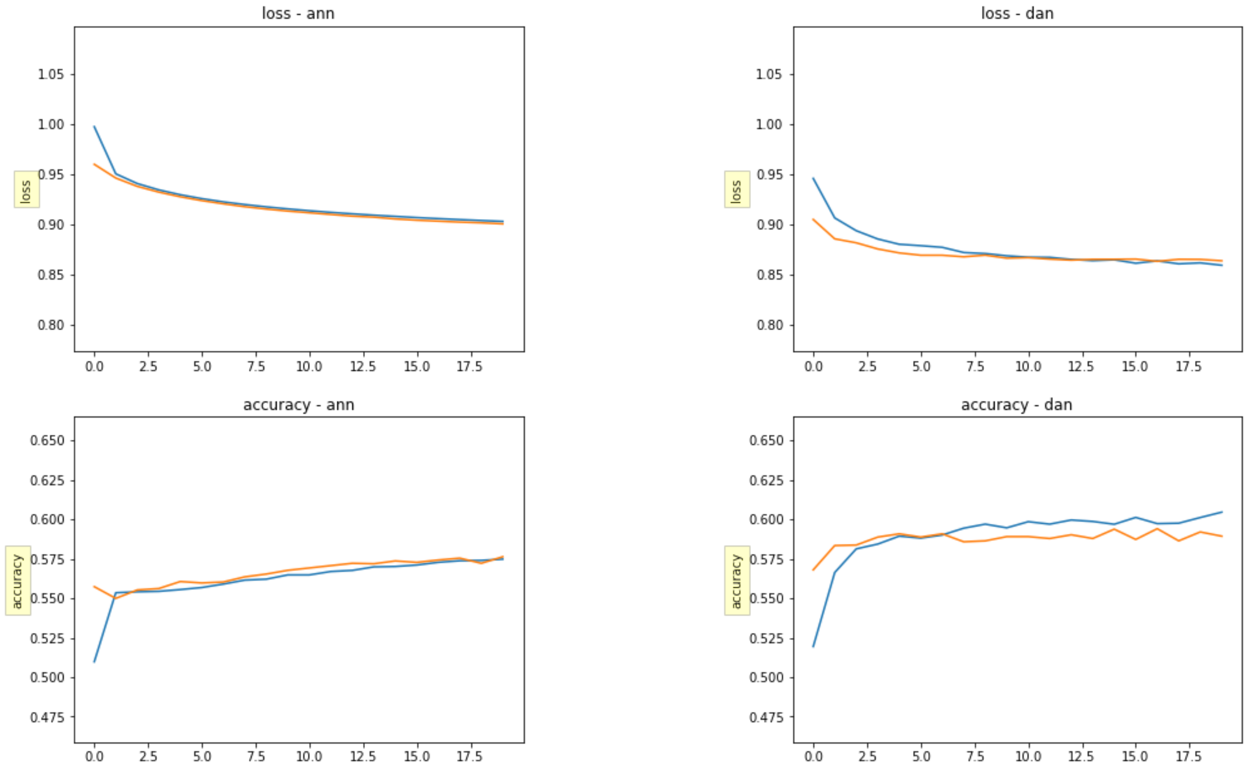
Data

Variable	Description	Type
claim	The headline statement that will be judged by its veracity	string
premise_articles	Documentation supporting or contradicting the claim. Each key is the URL and each value is the article's content	dictionary
rating	Label of veracity of claim (0 is false, 1 is partially true/false, 2 is true)	integer
treview_article	JSON file with review article content	string

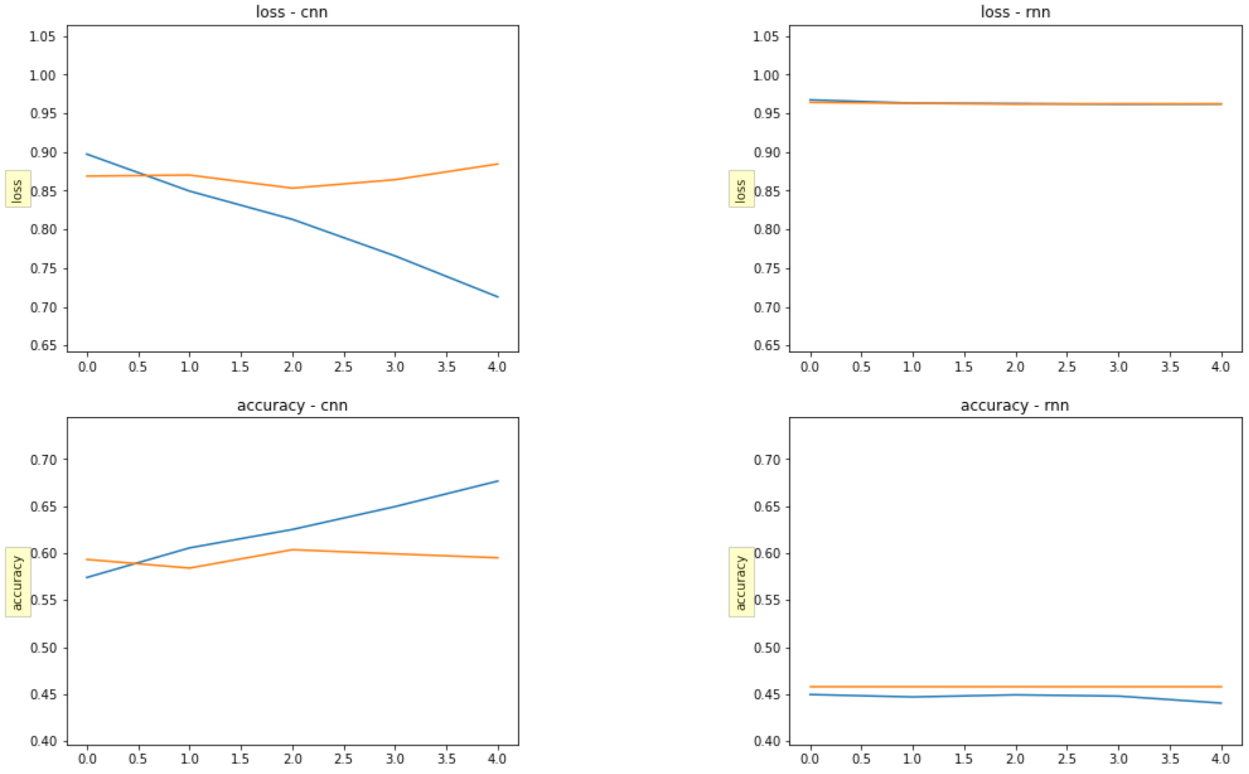
Methods: Exploring CNN, RoBERTa, and BERT

Model	Input Variation	Maximum Sequence Length	Epochs
BERT	Claim (Headline)	250	5
BERT	Review Article	500	5
BERT	Claim + Review Article	700	5
RoBERTa	Claim + Review Article	512	10
DPR + RoBERTa	Claim + Claimant + Premise Article	512	10

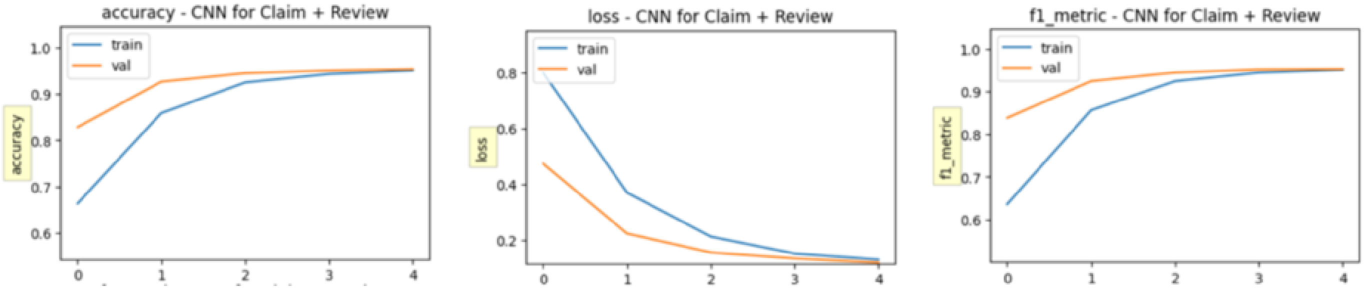
Baseline Model: ANN and DAN



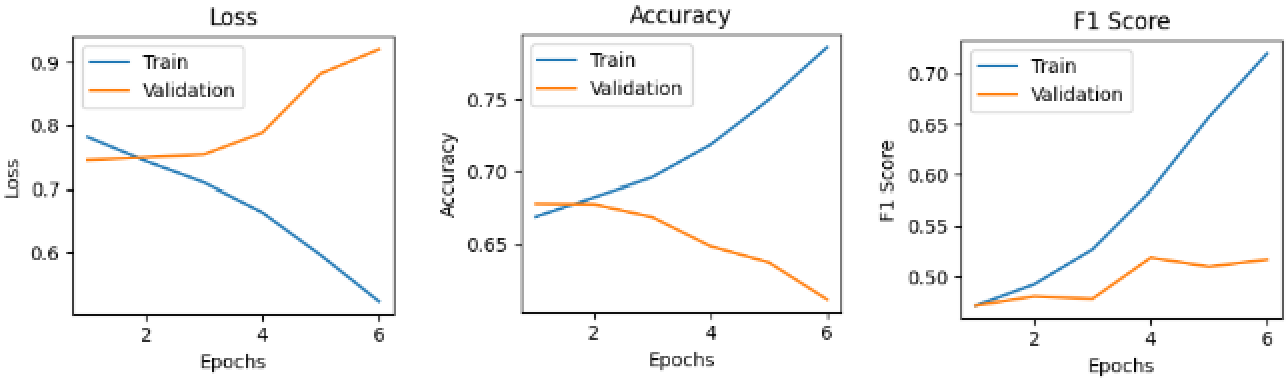
Baseline Model: CNN and RNN



Model 3:BERT Claim and Review Article Model



Model 5: RoBERTa Claim and Premise Article


Loss
Train
Validation
Loss
Epochs
Accuracy
Train
Validation
Accuracy
Epochs
F1 Score
Train
Validation
F1 Score
Epochs