Identifying Deceptive Content: A Study on Clickbait and Fake News Detection

Shashank Rangarajan, Chia-Yu Tung, Rishabh Ghosh, Michael Guastalla, Edwin Wang
Department of Computer Science, University of Southern California
{sr87317, ctung, ghoshris, guastall, kuanchun}@usc.edu

1 Progress report

Reiterating our project's goal, we aim to improve the existing state-of-the-art NLP methods for more accurate detection and classification of fake news and clickbait. And our objectives can be summarized as follows:

1. Implement existing state-of-the-art models for clickbait and fake news detection.
2. Investigate if models for fake news detection can augment clickbait detection and vice-versa.
3. Compare the performance of the state-of-the-art models on various publicly available datasets and study their results.
4. Explore the relationship between fake news and clickbait to enable more fine-grained prediction in the models.

And so far, we have been able to make progress towards objectives - 1, and 2. Here is the concrete details of the progress we have made till now.

1.1 Implementing state-of-the-art models for clickbait detection

We have implemented LSTM and GRU models for detecting clickbait. To train and test these models, we used the Clickbait Dataset(on kaggle, Unknown) and the Clickbait-detector(Mathur, 2017) datasets. We compared our results with those of another model that used the same datasets, to see if our approach yielded similar or better results.

We experimented with different combinations of these two datasets in four cases. These cases involved using one dataset to train and the other to predict, combining the two datasets to train and predict for only one dataset, and so on.

We found that our LSTM model produced similar results when compared to another model. However, we observed an interesting trend in our experiments. When we trained our model on the Clickbait Dataset(on kaggle, Unknown) and tested it on the Clickbait-detector(Mathur, 2017), our performance was lower. On the other hand, when we trained our model on the Clickbait-detector(Mathur, 2017) and tested it on the Clickbait Dataset(on kaggle, Unknown), we achieved better performance.

The Clickbait-detector(Mathur, 2017) dataset was collected from Reddit, which makes it closer to real-life scenarios. In contrast, the Clickbait Dataset(on kaggle, Unknown) consists of news articles. This difference between the two datasets gives us an important point for future investigation. Despite having a smaller amount of data, the Clickbait-detector(Mathur, 2017) dataset yielded better results during training. We can further explore the reasons behind this discrepancy.

Moreover, we acknowledge that two datasets may not be sufficient for our research. However, including other datasets proved to be a challenge, given their varying contents. Some datasets contained only titles, while others included text bodies or even more information. This made it difficult to combine these datasets in a meaningful way for our research.

1.2 Implementing state-of-the-art models for Fake news detection

For fake news detection we implemented the same model used in the paper titled "r/Fakeddit"(Nakamura et al., 2019). Using the same dataset of 1,063,106 samples that was defined in this paper, we implemented the same text based classification that was described in the paper for classifying fake news in 2, 3, and 6 way classification. The paper uses the titles of reddit posts for classification; the dataset also contains metadata and comment data, but these were not used in classification. The score of posts was used during preprocessing to filter out noisy samples. We initially used BERT(Devlin et al., 2019) to generate sentence embeddings for the titles in our dataset. We then used a simple feed forward network with one hidden layer of size 224 to produce class labels.

Using this method we were able to achieve an accuracy of 0.86 on 2 way classification, the same score reported in the paper. We have also tried replacing the BERT model with RoBERTa(Liu et al., 2019), which has shown promising results so far, but requires more fine tuning. The paper also explores combining images and text for multimodal classification, but for now we have decided to focus solely on text.

This apart, we also tried finetuning the model on the LIAR dataset (Wang, 2017), but we saw that the performance of model goes down to 65% accuracy which is relatively poor compared to 75% reported by (Yang et al., 2019) or 89% reported by (Aslam et al., 2021). As a next step, we are implementing models that are specifically shown to work well for LIAR dataset, and we hope that these models can be a meaningful addition to the ensemble we are working towards.

With these results, there is also room to explore different model architectures and hyperparameters to further improve our acuracy on the data. We also plan to validate these results on similar fake news datasets to ensure that our model generalizes well to data taken from different sources.

1.3 Investigation of unified models that augment fake news and clickbait detection

We also parallelly tried to see if we can unify the models that were created for clickbait and fake news detection by streamlining the inputs and pipelining all the models to use the output layers to create an ensemble model, but we ran into issues due to the diverse preprocessing steps adopted for different datasets. We will continue to work on this avenue and perform experiments to see if unifying fake news and clickbait can augment each other's detection.

2 Risks and Challenges

So far, we anticipate the following risks and challenges as we perform more experiments:

1. Diverse sources of data come with inherent biases, and different types of texts - news, Reddit, social media, etc. which means that handling all these different data sources meaningfully also becomes more challenging

2. For our 4th objective, we still haven't found a suitable dataset that has fine-grained true labels for clickbait and fakenews

3. Some of our datasets, such as fakeddit, rely on techniques of distant supervision to generate training and validation data so it is possible that the labels contain a lot of noise.

4. Pipelining all data from different sources through a set of models that were trained differently is proven to be challenging as stated in section 1.3

We also anticipate other challenges as we dive deep on our 4th objective on how we can create models that can make fine-grained predictions.

3 Mitigation Plan

Following are some mitigating steps we intend to take to counter the risks and challenges mentioned above:

1. To handle the diverse nature of our data, and to streamline these data through our model pipelines efficiently, we intend to create a common pre-processing pipeline for all datasets. Here, we plan to also integrate pre-trained encodings from BERT models, and other cleaning techniques we've learnt to create through our homeworks.

2. We also plan to combine data from multiple channels and tasks to create a comprehensive dataset to study the fine-grained nature of prediction between fakenews and clickbait. We will adopt the Fake news corpus (Corpus, 2017) dataset to build our models

3. Thus far, we have primarily used accuracy as a metric when comparing our results with the papers we are implementing. Going forward we could also include other metrics such as f1 score, recall, precision to fine tune our models and potentially find flaws in our approach or data.

4 Individual Contributions So Far

Task	Member (last name)
Implement state-of-the-art clickbait detection model	Ghosh Tung
Implement state-of-the-art fake news detection model	Guastalla Rangarajan
Unify the clickbait and Fakenews models	Wang

References

Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil. 2021. Fake detect: A deep learning ensemble model for fake news detection. *complexity*, 2021:1–8.

Fake News Corpus. 2017. Fake news corpus. https://github.com/several27/FakeNewsCorpus.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Saurabh Mathur. 2017. clickbait-detector. https://github.com/saurabhmathur96/clickbait-detector/tree/master/data.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection.

Clickbait Dataset on kaggle. Unknown. Clickbait dataset on kaggle. https://www.kaggle.com/datasets/amananandrai/clickbait-dataset?datasetId=609158.

William Yang Wang. 2017. Liar. https://www.cs.ucsb.edu/~william/data/liar_dataset.zip.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5644–5651.