Resolving the Spread of Fake News by Utilizing Natural Language Processing Tools to Fact-Check Information

Sam Edwards and **Ross Tebbetts** and **Naseela Pervez** and **Hang Yang** and **Daniel Bai**
University of Southern California

1 Dataset Exploration

This report provides a detailed description of the tasks performed so far, which will be a base for carrying out the experiments that we have planned for this project. The team has worked in accordance with the feedback provided on the proposal and carried out data exploration and baseline models. We have successfully narrowed down a dataset and completed the data preparation.

Before we moved on with applying any models, narrowing down to a single dataset was a priority. The team made sure the to perform a deep and wide data exploration before selecting the dataset.

1.1 Fake News Challenge Dataset

The first dataset that the team explored is the (Fake News Challenge Dataset). While exploring the dataset, we detailed that the dataset has the columns of - Headline, BodyID, Stance. The dataset was solely focused on the news headlines. The Stance column has four unique classes - unnrelated, discuss, agree, disagree. The flow of the model would be to extract evidences from the corpus of webpages of the news sites. The distribution of the classes is given in the below Table 1

Stance	Count
unrelated	36545
discuss	8909
agree	3678
disagree	840

Table 1: Distribution of classes in the Fake News Challenge Dataset

With this data, we have created a training dataset which will be used to train a bidirectional LSTM model using Pytorch, using GloVe embeddings. This will be the initial prediction model used in the development of this project.

1.2 SciFact

The second dataset that we explored was the Scifact dataset. This dataset aims at verifying scientific claims. We performed a data exploration of the training dataset and the training data has the following columns - id, claim, evidence and cited-doc-ids. Scifact also provides a corpus. The task aimed at the dataset is to extract the evidences that support or refute the Scientific facts from the given corpus. Since we wanted our task to encompass both evidence extraction as well as fact verification we realised that we will not proceed with utilizing this dataset.

1.3 Fact Extraction and Verification(FEVER)

The final dataset that we explored as the part of the project was the FEVER(2018) dataset. The dataset has had some changes ever since the original edition, but we explored the very first FEVER edition. The dataset is considered a state of art dataset for Fact Extraction and Verification tasks. The dataset contains the following columns - id, verifiable, label, claim, evidence. The verifiable aspect is a result of feature extraction. It specifies given the evidences for the claim and whether or not the claim is verifiable. The verifiable column has two unique classes - VERIFIABLE and NOT VERIFIABLE.

VERIFIABLE	Count
VERIFIABLE	109810
NOT VERIFIABLE	35639

Table 2: Distribution of classes in the verifiable(FEVER)

The distribution of the classes can be seen in the Table 2. The label column specifies the results of the Fact Verification. It has 3 classes - SUPPORTS, REFUTES, NOT ENOUGH INFO. The distribution of the classes can be seen in the Table 3. The claims column specifies the claims and the

evidence column has a list of evidences that are annotated. Some of the initial inferences from the data exploration were as follows:

Maximum length of the claim: 614

Minimum length of the claim: 10

Maximum length of the evidence: 251

Minimum length of the evidence: 1

Since the dataset is versatile, it includes a shared task of fact verification and claim and has a large number of training samples(145449), the team has decided proceed with using this dataset.

2 Dataset preparation

The evidences and claims in the FEVER(2018) are both text data. Since the focus of the project is to use NLP techniques, we are going to built models that use both the columns of the dataset. The evidences were annotated using a large corpus of wikipages. FEVER(2018) provided the corpus that has been used to retrieve the evidences. Before we started working on the dataset, we pre-processed and prepared the dataset for our use. The evidences for some claims can be less while the others can be more. In the dataset preparation, we have considered a maximum of 5 evidences. Every evidence is annotated as a list having annotation-id, evidence-id, url and sentence id. The wikipedia pages were used to create a json file with url as the id and the list of sentences in the given text. The final dataset that we have has = id, verifiable, claim, evidences(list of text of evidences at most 5)

	id	verifiable	claim	evidences
0	3	VERIFIABLE	Chris Hemsworth appeared in A Perfect Getaway.	['0\tChris Hemsworth -LRB- born 11 August 1983...
1	7	VERIFIABLE	Roald Dahl is a writer.	['0\tRoald Dahl -LRB- -LSB- langpron˙rou.ald _...
2	8	VERIFIABLE	Roald Dahl is a governor.	['0\tRoald Dahl -LRB- -LSB- langpron˙rou.ald _...
3	9	VERIFIABLE	Ireland has relatively low-lying mountains.	['0\tIreland -LRB- -LSB- ˙aɪərlənd -RSB- Éire ...
4	10	VERIFIABLE	Ireland does not have relatively low-lying mou...	['0\tIreland -LRB- -LSB- ˙aɪərlənd -RSB- Éire ...

Figure 1: An image of the final dataset

3 Risks and challenges

There are a number of risks and challenges that we can come across the project. However, presently we are focusing on two main points - **imbalanced dataset** and **length of claims**.

label	Count
SUPPORTS	80035
REFUTES	29775
NOT ENOUGH INFO	35639

Table 3: Distribution of classes in label(FEVER)

We can see from the previous sections that we have approximately 8:2:3 ratio of the supports: refutes: not enough info. This may cause the model to have a higher accuracy but a lower F1 and AOC scores.

The second challenge that we have in mind while building the baseline and state of art models is the length of claims. Some claims are more than 500 characters. This means that information can be lost in these claims.

We are also debating over if this should be a multioutput (label and verifiable) problem or a single output(label) problem.

4 How we plan to address the challenges?

To address the challenge of the dataset inbalance, we are planning to apply upsampling techniques for increasing the number of refutes and not enough info samples. We will first build our baseline LSTM without upsampling and then apply the upsampling technique to the transformer model that we plan on using.

To address the challenge of length of claims, we are still exploring the techniques. We will probably see the performance of the transformer model before we explore more on this, and decide which direction to take.

To address the third challenge, some of the team members are working on exploring more on the literature while others are working on the baseline models. We have three ways to go about it - predict verifiable and use it as an input to another model to predict the label, predict both verifiable and label, predict only label using verifiable as an input.

5 Individual Contributions

Naseela Pervez: I worked on the data exploration and data preparation of the FEVER(2018) dataset.

Hang Yang: I worked on the data exploration and data preparation of the Scifact dataset.

Sam Edwards: I worked on the initial development of a bidirectional LSTM Model with Glove embeddings

Daniel Bai: I worked on the Risk and Challenges and "How we plan to address the challenges?" sections of the Status Report

Ross Tebbetts: I worked and functioned as the team lead, managing a git repository for the team and assisting in the development of LSTM Model.

References

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing &amp Management*, 60(2):103219.

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A. Hale, and Rada Mihalcea. 2022. Matching tweets with applicable fact-checks across languages.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news.

PyTorch. 2023. Nlp from scratch: Classifying names with a character-level rnn¶.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.