Truth Tracker:

Tools and Techniques for Identifying Fake News

Our Team

Yi Liu

Ph.D. student in Computing & Data Sciences

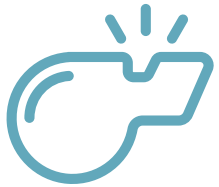
Zhuoyan Ma

Ph.D. student in Marketing

Ruozhu Wang

Ph.D. student in Operations Management

Introduction



Problem

The proliferation of fake news on social media is accelerating, contributing to cognitive biases and irrational decision-making.

Motivation

Expert-based manual fact-checking is time-consuming and costly.

Methodology

1. Textual analysis
2. Advanced NLP & Deep learning
3. Large language models(LLMs): Prompting engineering

This project aims to develop reliable methods for detecting fake news, fostering a more informed and truth-based information landscape.

Data: LIAR PLUS

1. **Target label:**

   True-related: Half-true, Mostly-true, True.

   False-related: Pants-fire, False, Barely-true.

2. **Statement (text):**

   The entire PolitiFact article.

3. **Subject (categorical, non-ordinal):**

   Subject of the statement (e.g., health care, social security, economy, etc).

4. **Job title (categorical, non-ordinal):**

   Job title of the speaker (e.g., President, state representative, etc).

5. **State info (categorical, non-ordinal):**

   US state where the speaker is based (e.g., Massachusetts, New York, Florida, etc).

6. **Party affiliation (categorical, non-ordinal):**

   The party affiliation of the speaker (e.g., republican, democrat, etc).

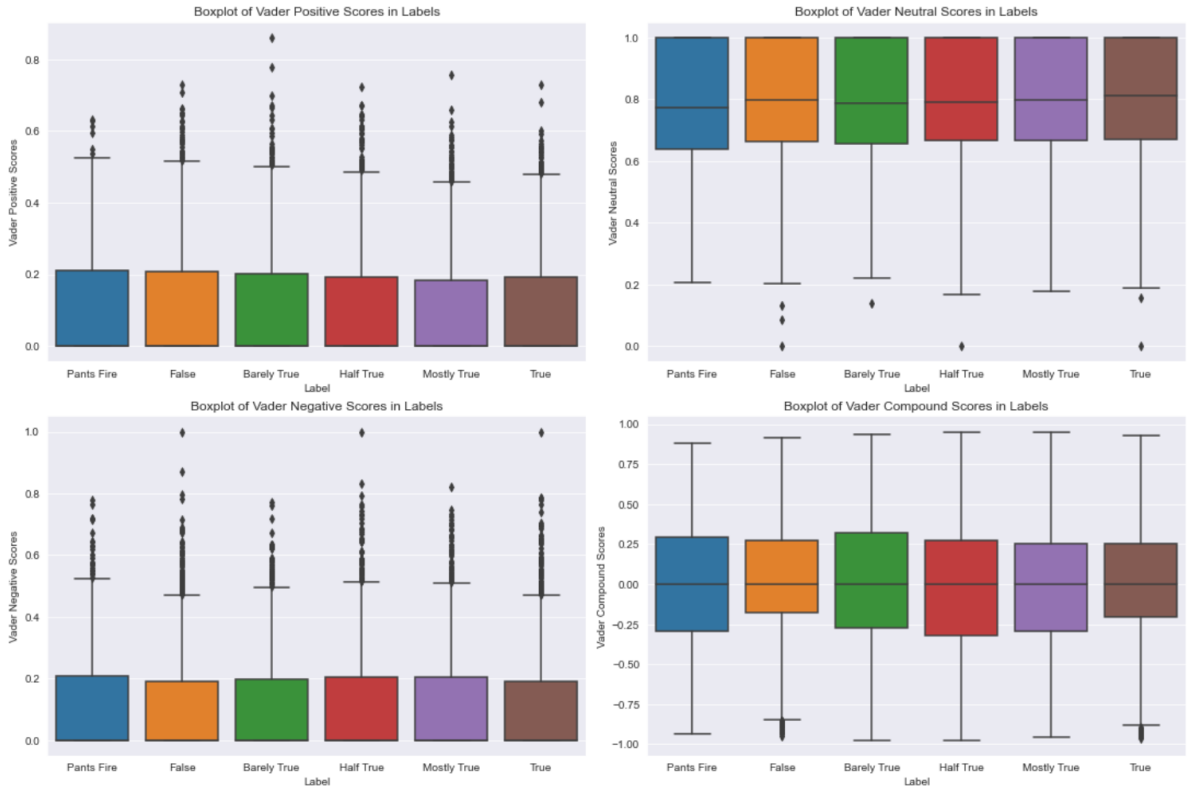
7. **Context (text):**

   The context, such as the venue/location of the statement.

8. **Justification (text):**

   The reasoning statement.

Emotion analysis of Statements vs. Label

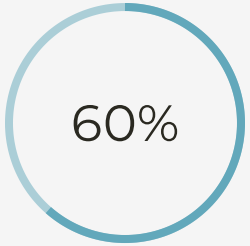


Valence Aware Dictionary and sEntiment Reasoner (VADER):

- The sentiment score (negative/positive/neutral) of a statement is calculated by summing up sentiment scores of each VADER-dictionary-listed word in each statement.
- There is not no obvious trend in emotions of each across different labels.

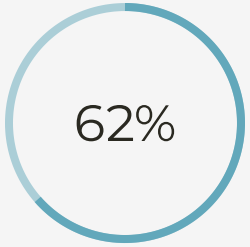
Binary Classification Results

Deep Learning Based Techniques



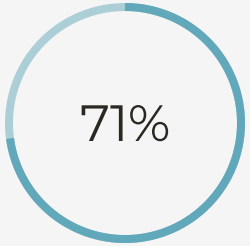
Logistic Regression

Statement



Support Vector Machine

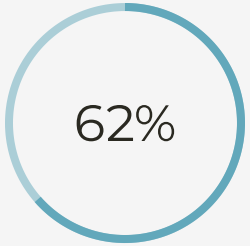
Statement



Random Forest

Statement + meta data

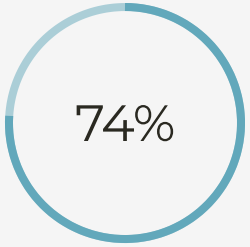
Machine Learning Based Techniques



BERT

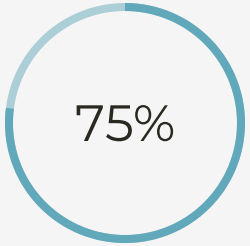
lr = 2e-5 (Chiorrini er al. 2021)

Statement



XGBoost

Statement + meta data

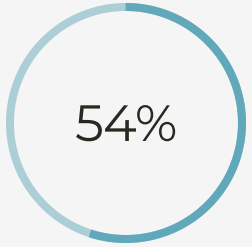


RNN

(with Word2vec embedding)

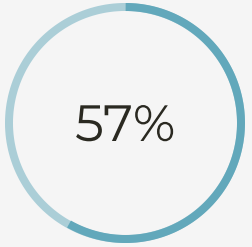
Statement + meta data

LLM exploration



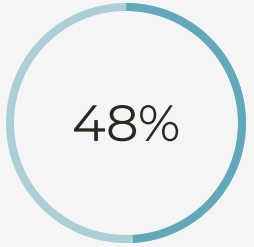
Zero-shot prompting

Statement + <justification >



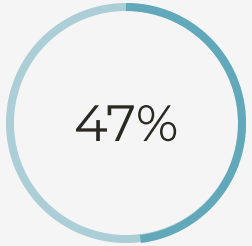
Zero-shot Chain-of-Thought

Statement + <justification >



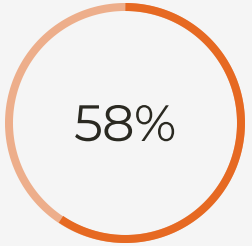
Zero-CoT from textual description

Statement + <justification >



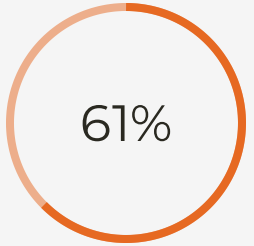
Zero-CoT from commonsense

Statement + <justification >



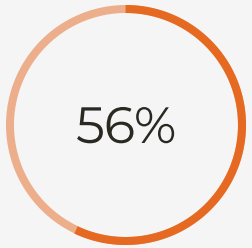
Zero-shot prompting

Statement + <justification > + meta data



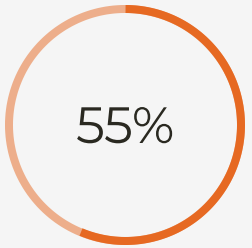
Zero-shot Chain-of-Thought

Statement + <justification > + meta data



Zero-CoT from textual description

Statement + <justification > + meta data



Zero-CoT from commonsense

Statement + <justification > + meta data

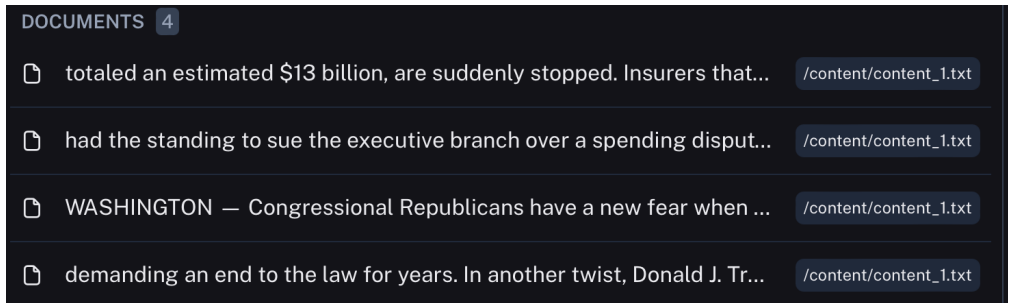
LLM gain more predictive power from meta-data / social context, including the the fake news frequency of certain topics, sources, and publisher

RAG Exploration (>95% accuracy)

Due to the complexity of the intention or mechanisms behind fake news, like unsupported evidence, emotion exaggeration, false connection, context manipulation, echo chambers, it's hard for human to do fact checking using a general reasoning process as well.

We further explore LLM's performance detecting fake news due to unsupported & contradictory evidence.

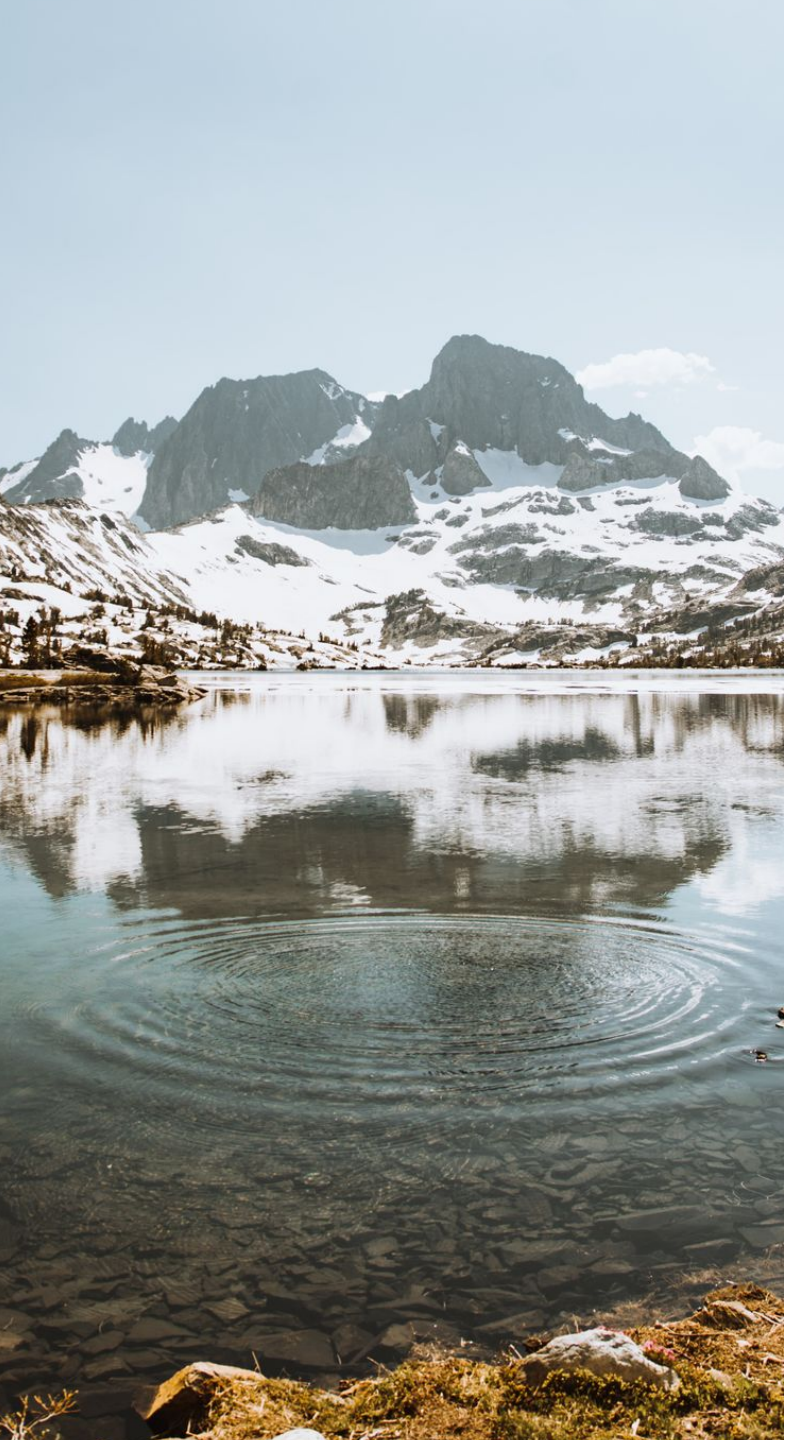
```
{
  "query": "Question: Evaluate the varicity of the statement
delimited by angle\n    brackets. If it was unsupported by or
contradict to the evidence, it's more likely to be false, return 0
along with a reason;\n    otherwise, it's more likely to be true,
return 1 along with a reason. \n    <Statement: The Trump
administration plans to eliminate all health insurance subsidies
immediately after inauguration, leaving millions without any health
coverage.>\n    Your answer should be in JSON format, whose keys
are veracity and reason"
}
```



LLM's response: { "veracity": 0, "reason": "The statement is not supported by the evidence. The documents indicate discussions on potential options and resolutions post-inauguration, but there is no explicit plan mentioned to eliminate all health insurance subsidies immediately after inauguration." }

Human's evidence: While the content discusses the potential halting of subsidies, it mentions that this would be part of a negotiation and legal review process, not an immediate action.

Once we specified one type of fake news and have the citation materials behind, LLM augmented with RAG will be a powerful tool to check the truthfulness and reliability of a news statement.



Statement itself alone is not enough to detect fake news.

Fake news is aimed to mimic the truth and the intention behind it is complicated.

Adding meta-data can help a lot.

Especially characteristics of users who wrote the statement.

RAG seems to be the most promising detection method.

Linking with external knowledge base improves the classification accuracy to more than 90%.

Thank you!

Any questions?