FACT-CHECKING ON SCIENTIFIC CLAIMS

Statistical Software (16:954:577:01)
Team 16
Team Members:

Amrutha Karuturi
Taniya Satheesh Kulkarni

INTRODUCTION

According to recent statistics, nearly 5 billion people worldwide use social media, generating an immense volume of posts and interactions daily. While this connectivity has revolutionised communication, it has also become a breeding ground for misinformation. For instance, during the COVID-19 pandemic, numerous posts falsely claimed that simple home remedies could "cure" the virus, misleading countless individuals.

Compounding this issue is the significant gap between the everyday language of the general public and the complex terminology used in scientific communication, often resulting in critical information being "lost in translation." The lack of effective verification tools further exacerbates this problem. Although some social media platforms have implemented fact-checking algorithms, these systems are often limited in scope and lack the robustness needed to address the scale and nuance of misinformation.

As a result, misinformation can lead to serious consequences, such as individuals making harmful decisions, including consuming inappropriate or unsafe medications, based on false claims. Addressing these challenges requires more reliable tools and methods to ensure the accuracy of information shared across digital platforms.

Literature Review

Explainable Automated Fact-Checking for Public Health Claims (Kotonya, Neema et. al)

This paper presents a novel dataset designed for explainable fact-checking, which includes gold-standard explanations created by journalists, tailored specifically to public health claims. It then introduces a framework for generating both explanations and veracity predictions for public health claims, demonstrating that using in-domain data (data specific to public health) leads to better performance than general-purpose models.
Finally, the authors define three coherence properties for evaluating the quality of fact-checking explanations, which can be assessed both by humans and computationally, offering a robust method for evaluating explanation clarity and accuracy.

DATA

For this project, we utilize the PUBHEALTH fact-checking dataset developed by Neema Kotonya, which serves as a robust resource for evaluating the accuracy of public health claims. The dataset includes a collection of public health documents paired with associated claims, enabling fact-checking. It consists of the following 10 columns:

- **main_text:** Represents the full content of the public health document and serves as the premise.
- **claim:** Represents the hypothesis that we aim to verify.
- **label:** A categorical label that indicates the relationship between the claim and the document content. It has three possible values:
    - true: The claim accurately reflects the information in the document.
    - false: The claim contradicts the information in the document.
    - mixture: The claim contains elements that are both accurate and inaccurate with respect to the document.
    - unproven: The claim cannot be verified based on the available evidence in the document.
- **claim_id:** Represents a unique identification number for the claim
- **date_published:** The date on which the claim was published.
- **explanation:** Provides context and a detailed explanation of why the claim was assessed in a certain way, helping users understand the logic behind the decision.
- **fact_checkers:** Names of the people who fact-checked the claim
- **sources:** A list of URLs or references to articles, reports, or websites that support the fact-checking process and provide further details about the claim.
- **subjects:** The topics, themes, or key issues that the claim is related to.

The following is a snippet of the dataset:

	claim_id	claim	date_published	explanation	fact_checkers	main_text	sources	label	subjects
0	15661	"The money the Clinton Foundation took from fr...	April 26, 2015	"Gingrich said the Clinton Foundation ""took m...	Katie Sanders	"Hillary Clinton is in the political crosshair...	https://www.wsj.com/articles/clinton-foundatio...	false	Foreign Policy, PunditFact, Newt Gingrich,
1	9893	Annual Mammograms May Have More False-Positives	October 18, 2011	This article reports on the results of a study...		While the financial costs of screening mammogr...		mixture	Screening,WebMD,women's health

The dataset has been split into training, development, and testing sets.

DATA SET	NUMBER OF ROWS
TRAINING	9832
VALIDATION	1221
TESTING	1235

For this project, we focused on three key fields from the dataset: claim, explanation, and label.

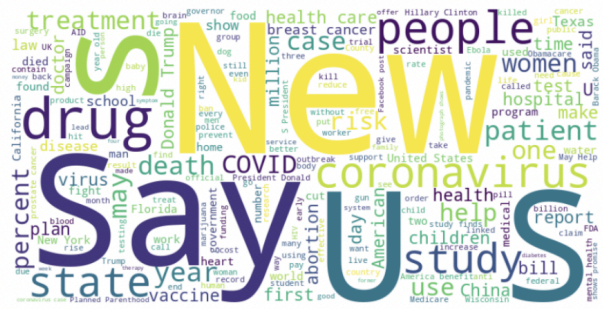
Data Cleaning and Preparation

The data was relatively clean; however, the following preprocessing steps were performed to streamline its usability:

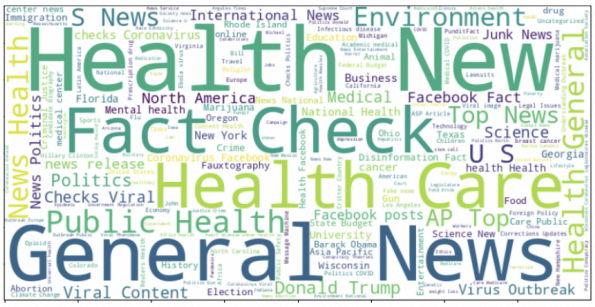
1) Removing rows with missing values.
2) Excluding the row where the label was “snopes”
3) Removing punctuations
4) Converting all text to lowercase to standardize test
5) Stopword removal
6) Label encoding

Exploratory Data Analysis

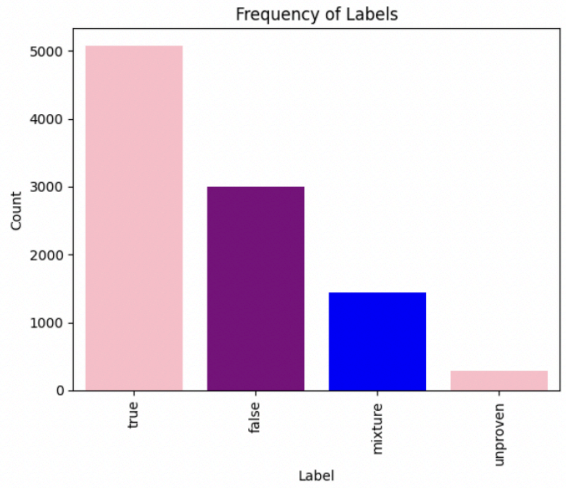
Common phrases in ‘claim’



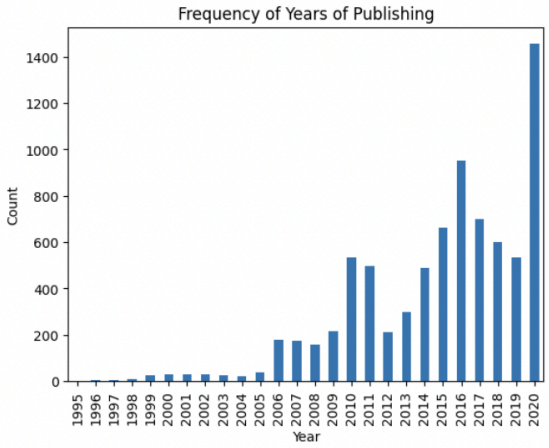
Common subjects



1)Frequency of label:



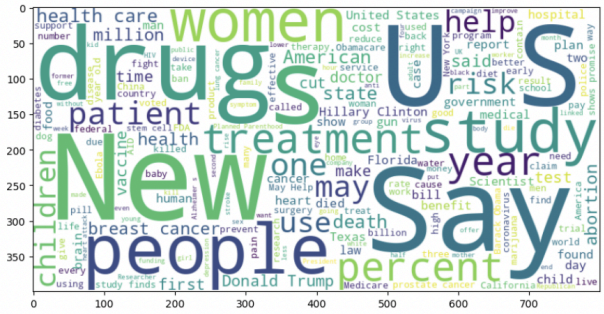
2)Year of publishing:



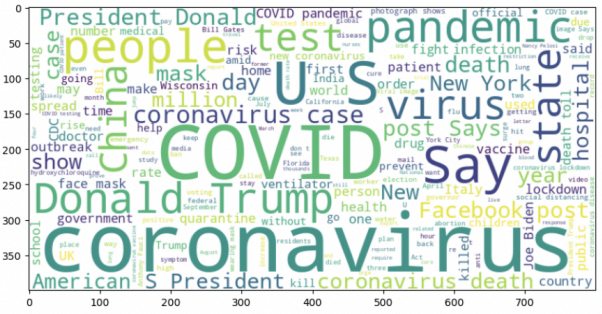
Given the substantial number of claims from 2020, we can anticipate a significant volume of claims related to the coronavirus.

3) Word Clouds

Common phrases in claim pre-pandemic:



Common phases in claim post-pandemic:



Data collected after March 11, 2020, demonstrates a significant rise in the frequency of terms such as “COVID” and “coronavirus,” reflecting the global impact of the pandemic. This marked increase, compared to the pre-pandemic period, is unsurprising given the heightened focus on the virus in public discourse and media coverage.

EXPERIMENTS

1. BASELINE & DISTILBERT

We experimented with different combinations of tokenizers, namely:

- Bag-of-Words
- TF-IDF
- Word2Vec
- DistilBERT embeddings

We then used the tokenized data on the following models:

- Logistic Regression
- Simple Neural Network
- DistilBERT Sequence Classifier

While we implemented Logistic regression and a simple neural network model using BoWs, TF-IDF and Word2Vec, the DistilBERT Sequence Classifier made use of the DistilBERT embeddings.

Specifications

Tokenizers

- The **BoWs** representation (obtained using CountVectorizer()) was set to size of 5000 for each claim and document.

  The input size for the models with BoWs is 10000 (claim+document).
- The **TF-IDF** representation (obtained using TfidfVectorizer()) was set to size of 5000 for each claim and document.

  The input size for the models with TF-IDF is 10000 (claim+document).
- The **Word2Vec** embeddings (obtained using Word2Vec) were set to size of 300 for each claim and document.

  The input size for the models with Word2Vec is 600(claim+document).

Models

- The **logistic regression** from sklearn is used as a baseline model.
- The **simple neural network** has 3 layers with ReLU activations and softmax is applied on the output layer to give the probabilities.
- We used the **distilledbert-uncased** model for tokenization and sequence classification.
- The word embeddings were of dimension 768 which were the input to the fine tuned classifier for the **multi-class classification** task with 4 labels.

RESULT

Following are the evaluation metrics of all the models

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
BOWs Logistic	0.1517	0.4663	0.3199	0.1916
TF-IDF Logistic	0.6407	0.5702	0.5063	0.6489
Word2vec Logistic	0.5531	0.4749	0.5068	0.5803
BOWs NN	0.442	0.4534	0.4705	0.5051
TF-IDF NN	0.2879	0.4629	0.4055	0.3526
Word2vec NN	0.485	0.4395	0.4215	0.5028

Distilbert embeddings, Distilbert sequence classifier Epochs = 3	0.6586	0.5238	0.5314	0.6621
Distilbert (dropout=0.3) epochs =2	0.6853	0.4609	0.4591	0.6558
Distilbert (dropout=0.3), Epochs = 2, class weights	0.6009	0.5073	0.5577	0.6338

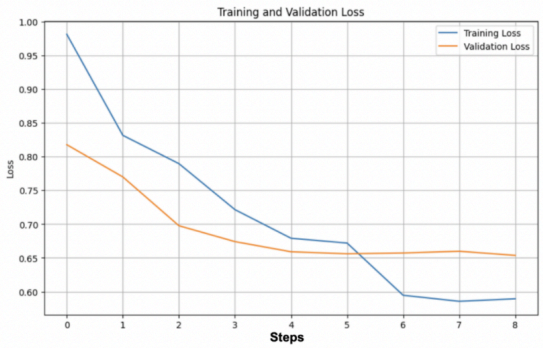
Result Analysis

- The performance of Logistic Regression was surprisingly better than Neural Networks when using TF-IDF and Word2Vec tokenizers.
    - TF-IDF often leads to a linear decision boundary, which suits Logistic Regression, as it excels in linear spaces.
    - Similarly, Word2Vec embeddings capture linear relationships between words, allowing Logistic Regression to effectively learn the decision boundary. On the other hand, Neural Networks, designed to capture non-linear patterns, may struggle with the sparse, linear nature of these embeddings. This explains why Logistic Regression outperforms Neural Networks in these cases.
- Next, we used a pre-trained distilBERT-base-uncased model:
    - First, we trained the model with 3 epochs and plotted training and validation loss, observed overfitting at the 3rd epoch.
    - To address overfitting we set dropout = 0.3 for regularization, still observed overfitting at the 3rd epoch and gave around the same results for training and validation loss. Although there isn't much improvement, the dropout model is less overfitting and more robust.
    - Now we trained the model with dropout to only 2 epochs, no overfitting was observed and accuracy increased from 0.65 to 0.68.
    - To address class imbalance issues, class weights were incorporated into the trainer. Higher weights were assigned to classes with fewer samples, while lower weights were given to classes with larger sample sizes while calculating the loss and updating parameters. This adjustment reduced bias in the model. Although it resulted in a slight decrease in accuracy score (~0.08), the model was less biased and fair across all classes.
    - The overall fine-tuning done here is training the model for our classification task, changing the no.of epochs and using dropout to address overfitting, using class weights to address imbalance datasets.

We can see the same in the following graphs:

Distillbert, 3 epochs

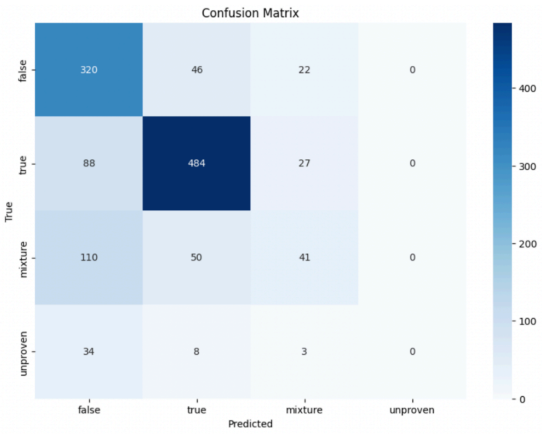
Distillbert(dropout = 0.3), 3 epochs



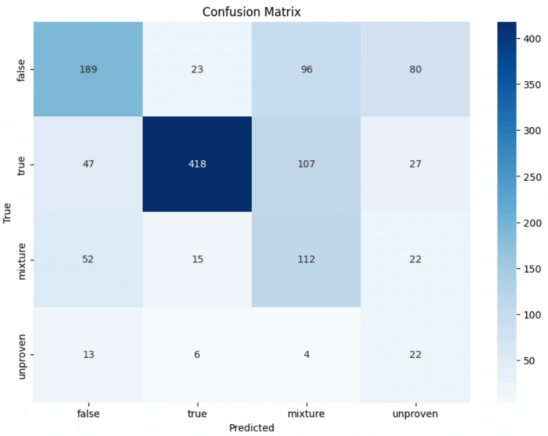
Distillbert(dropout=0.3), 2 epochs

Distillbert(dropout=0.3), 2 epochs, class weights

Distillbert(dropout=0.3), 2 epochs



Distillbert(dropout=0.3), 2 epochs, with class weights



2. SCIBERT WITH TOP-K SENTENCE RETRIEVAL

In addition to the baseline and DistilBERT models, we implemented SciBERT combined with a top-k sentence retrieval approach. The sentence retrieval process was conducted using Sentence-BERT (SBERT), which enabled the selection of the most relevant sentences for fact-checking.
SciBERT is a variant of the BERT model specifically designed to perform better on tasks involving scientific texts. Unlike the original BERT, which is trained on general-domain corpora, SciBERT is pre-trained on a large corpus of scientific articles from domains such as computer science and biomedical research.

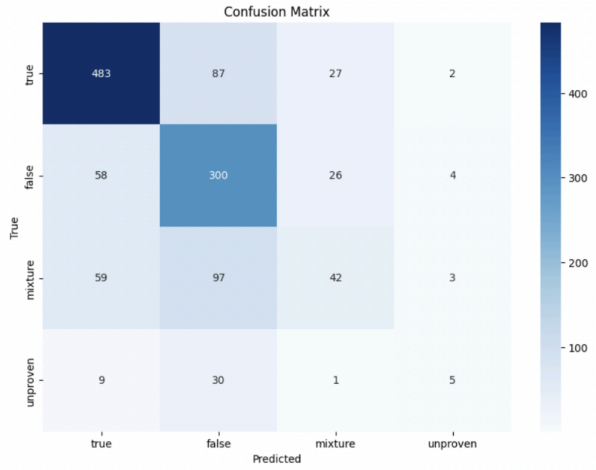
On the cleaned data we did the following:

- Class creation for dataset - Implemented a class for the dataset to handle tokenized inputs and corresponding labels.
- Sentence Retrieval - Used a pre-trained SentenceTransformer model (all-mpnet-base-v2) to encode sentences and retrieve the top-k most similar sentences to a given claim using cosine similarity.
  The ‘k’ selected was 3.
- Tokenization - Prepared the input data for a transformer model by tokenizing claims and their associated sentences.
- Training - We trained the data using SciBERT and experimented with various training parameters. However, our exploration was limited due to a lack of GPU resources. In addition to testing different parameters, we also investigated the impact of class imbalance on the model's performance.
    - To check the class imbalance issues, we apply the same method as seen in the DistilBERT case.
- Evaluation - The models were evaluated using a range of metrics including accuracy, precision, recall, F1 score, and confusion matrices.

RESULT

Epochs - 3, Batch_size - 32 (Model with the best result)

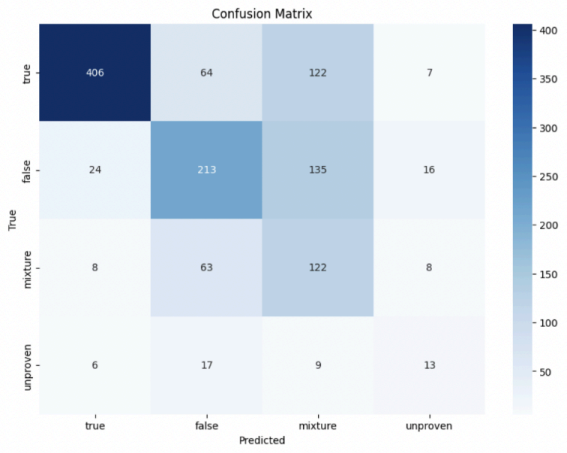
Confusion matrix without class balancing:



Classification report without class balancing:

	precision	recall	f1-score	support
false	0.58	0.77	0.67	388
mixture	0.44	0.21	0.28	201
true	0.79	0.81	0.80	599
unproven	0.36	0.11	0.17	45
accuracy			0.67	1233
macro avg	0.54	0.47	0.48	1233
weighted avg	0.65	0.67	0.65	1233

Confusion matrix with class balancing:



Classification report with class balancing:

Classification Report:

	precision	recall	f1-score	support
false	0.60	0.55	0.57	388
mixture	0.31	0.61	0.41	201
true	0.91	0.68	0.78	599
unproven	0.30	0.29	0.29	45
accuracy			0.61	1233
macro avg	0.53	0.53	0.51	1233
weighted avg	0.69	0.61	0.64	1233

Result Analysis

Class balancing improved the model's ability to identify minority classes ('mixture' and 'unproven'), which is important when dealing with imbalanced data.

Precision for these minority classes is still low, meaning the model is finding more instances of these classes but is also making more incorrect predictions.

While recall for minority classes improved with class balancing, there was a drop in accuracy (decrease by 0.06) and precision. This shows a trade-off between identifying more minority-class instances and keeping accuracy high for the majority class.

SUMMARY

- **Overall results:**
  - Logistic Regression (Baseline model) performed better than Neural Networks with TF-IDF and Word2Vec embeddings, as these methods suit linear models.
  - DistilBERT showed improved performance with dropout regularization and class balancing, although class balancing slightly reduced accuracy.
  - SciBERT demonstrated better recall for minority classes but at the cost of precision and overall accuracy, highlighting a trade-off between identifying more instances of minority classes and maintaining overall performance.
- **Improvements:**
  - Improving results could involve further hyperparameter tuning, experimenting with more epochs for DistilBERT, or using advanced class balancing techniques.

REFERENCES

1. Kotonya, N., & Toni, F. (2020). Explainable automated fact-checking for public health claims. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7740-7754. Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.623
2. Kotonya, N., & Toni, F. (2020). Explainable automated fact-checking for public health claims. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7740–7754. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.623