A Machine Learning Approach to Combat the Spread of Misinformation

Group 1

Amanda Chia Wan Ying (A0280523L)

Ang Kang Jie (A0226232R)

Danice Angelee C. Parel (A0280367Y)

Nattaya Silprakong (A0280554A)

Background

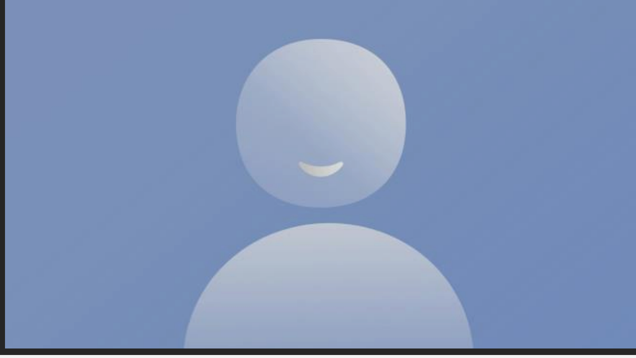
Fake news has become prevalent in today's digital age.

Due to the widespread use of social media for news consumption.

During 2016 US Presidential elections, fake election stories on Facebook garnered more engagements versus. news websites.
(8.7 million > 7.4 million)

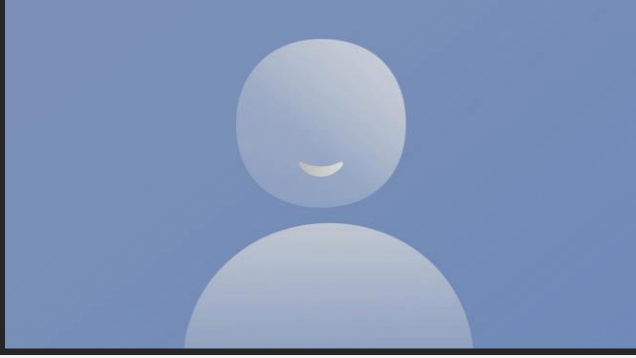
Proliferation of fake news poses significant political and social implications.

Background



Fact-checking to combat misinformation	**Manual fact-checking**	**Automatic fact-checking**
	+	-
	-	+

Objective



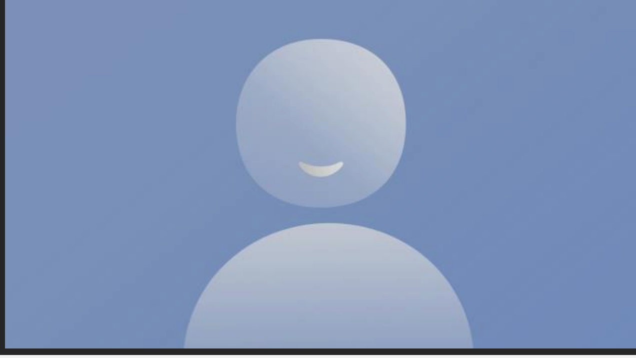
- Develop an effective **<u>fake news classifier</u>** to combat the spread of misinformation

1. Conduct a **<u>comprehensive survey</u>** of traditional and modern machine learning methods
2. **<u>Assess performance</u>** using key evaluation metrics: Precision, Recall, F1-Score, and Accuracy
3. Apply **<u>interpretable ML</u>** to provide transparency into classification decision, fostering greater trust in the system

Data

- **<u>Liar Dataset</u>** from Wang (2017). Used for fake news detection tasks
- 12,800 human labelled short statements from POLITIFACT.COM
- Statements from 2007 – 2016. Average statement length of 17.9 tokens
- Data is split 80-10-10 into train, validation, test
- 6 classes describing the severity of lies, ranging from "**pants-fire**" to "**true**".

Data Pre-processing

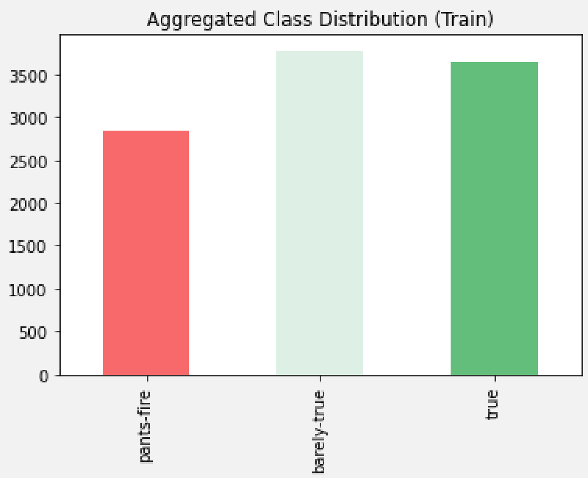


Class Aggregation

Classes aggregated into 3 groups to improve class balance and simplify interpretation

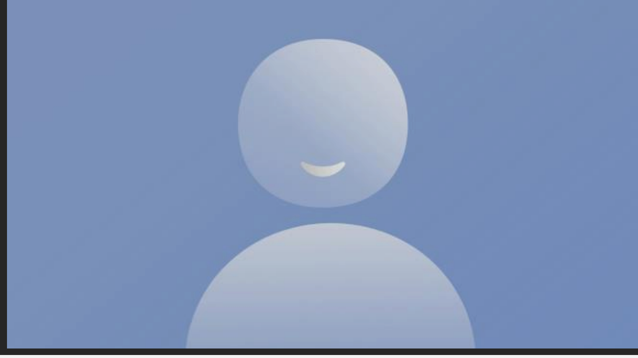
Post-aggregation, the class balance is improved – "fake news" accounts for 27.7%

Original Class	Grouped Class
pants-fire FALSE	pants-fire
barely-true half-true	barely-true
mostly-true true	true



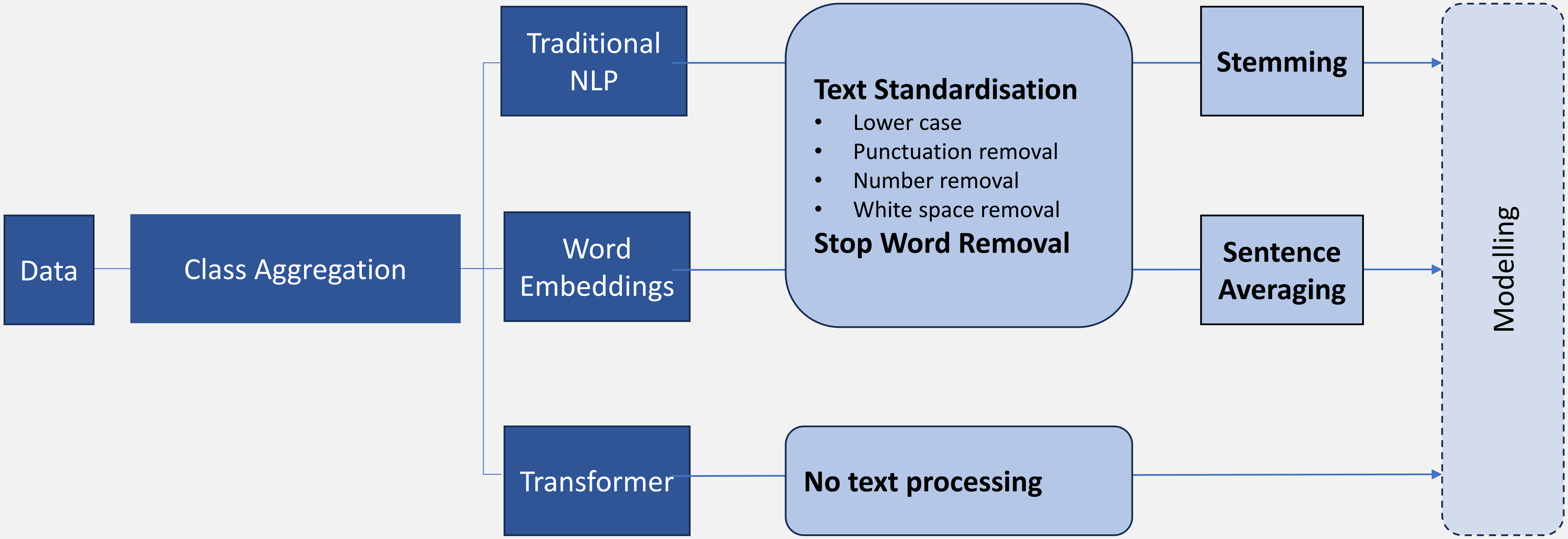
Statement	Label
Before World War II, very few people actually had health insurance.	true
The United States has a low voter turnout rate.	
Taxpayers subsidize 80 percent of each MARTA trip	Barely-true
Hillary Clinton said gun confiscation would be worth considering.	
Birth control pioneer Margaret Sanger was an active participant in the Ku Klux Klan.	Pants-fire
Many of the founding fathers were very actively involved in cockfighting.	

Data Pre-processing

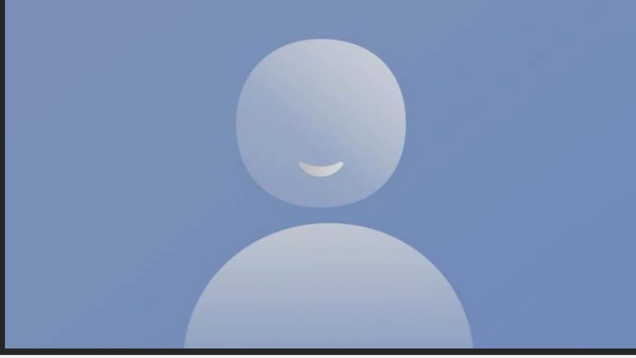


Depending on model choice…

Different levels of text processing



Machine Learning Methods



Traditional NLP

- Bag-of-Words (BoW) + LASSO
- Term-Frequency Inverse-Document-Frequency (TF-IDF) + LASSO

Word Embeddings

- Embeddings + LASSO
- Embeddings + CNN

Transformers

- DistilBERT
- Sentence Transformers

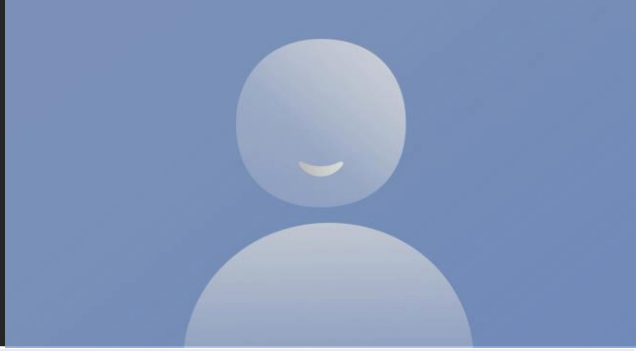
Traditional NLP – BoW and TF-IDF

BoW: Transforms textual data into fixed-length numerical vectors, disregards word order

TF-IDF: Computes word frequency against its frequency across the corpus

- Text preprocessing using Python’s Natural Language Toolkit (NLTK), Punkt tokenizer and Porter stemmer
- Maximum document frequency of 0.95 -> filters out extremely common words
- Tokenization included unigrams and bigrams -> captures some context
- Max_feature set to 1,000 -> reduces dimensionality and lowers overfitting risk

Word Embeddings



Traditional NLP | **Word Embeddings** | Transformers

Word Embeddings

- Word embeddings map words or phrases from raw text into a low-dimensional space
- Facilitate various NLP tasks by capturing contextual meanings and structural roles.

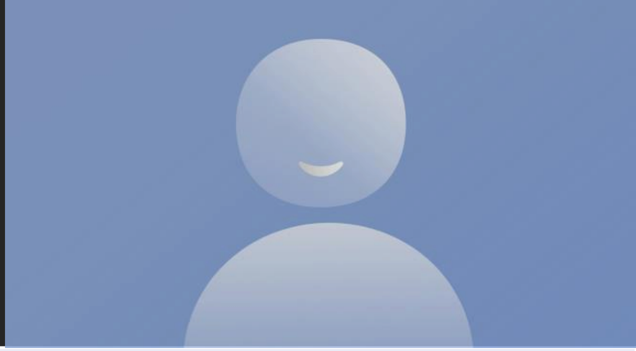
Global Vectors

- Balances semantic detail with computational efficiency, ideal for sentence-level analysis
- Utilizes glove.6B.100d, trained on 6 billion words, outputs 100-dimensional vectors

Sentence Embeddings

- Created by averaging all word vectors in a sentence to form a single representative vector
- Maintains essential semantic content for robust language understanding

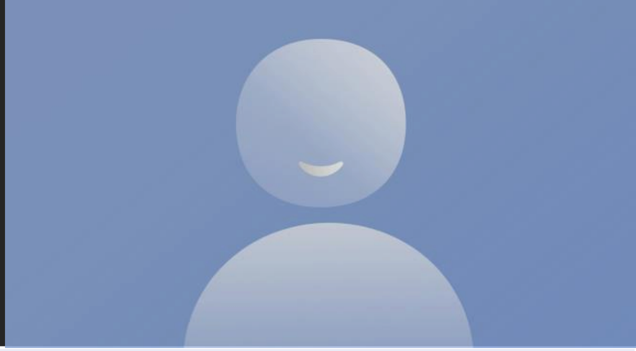
Word Embeddings – GloVE + Lasso



Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization.

- Logistic Regression with L1 regularization (LASSO) addresses the challenge of high dimensionality in NLP
- Incorporates a penalty term to the loss function to prioritize and retain only the most relevant features
- Provides a balance between model simplicity and predictive accuracy, making it ideal for text-based predictive analyses like sentiment analysis or topic classification.

Word Embeddings – GloVE + CNN

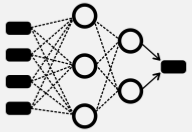


Traditional NLP | **Word Embeddings** | Transformers

GloVe embeddings are combined with Convolutional Neural Networks (CNNs) to efficiently capture local contextual features in text data.

Data Preprocessing

- Texts are analyzed for length distribution
- Tokenization and padding
- Vocabulary is limited to unique words
- Embedding layer initialized with GloVe matrix, fixed to preserve pre-trained semantic properties.



CNN Architecture and Training

- Alternating convolutional and max pooling layers extract text features efficiently
- Includes dense layers with dropout and early stopping for regularization and to prevent overfitting
- Trained using the Adam optimizer, 15 epochs, and categorical cross entropy for enhanced efficiency and generalizability

Transformers – DistilBERT

Transformers leverage **attention mechanisms** to capture linguistic dependencies and weigh the importance of words in sequences

- Architecture is parallelizable to reduce training time

DistilBERT Model

- Preserving 95% of BERT's performance with 40% fewer parameters
- Vocab size: 30,522 tokens. Max input size: 512 tokens
- Input text tokenized and padded. No text preprocessing required

Fine-tuning

- Last two layers are trained, while all other parameters are frozen
- Fine-tuned on a random sample of 1,284 training instances. 20 epochs at a low learning rate (5e-5).

Transformers – Sentence Transformers

Traditional NLP | **Word Embeddings** | Transformers

Transformer that computes semantically meaningful embeddings at the sentence level

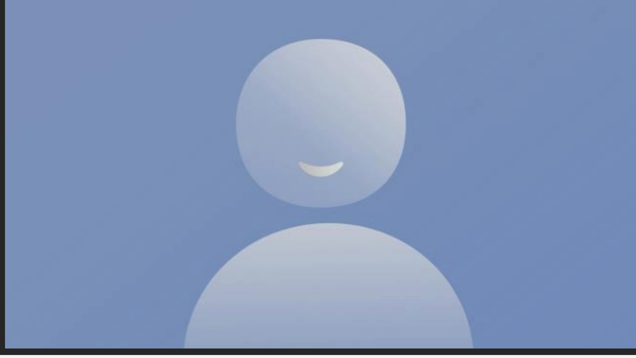
- Similar sentences are close together in the embedding vector space

SBERT

all-MiniLM-L6-v2 Model

- Pre-trained sentence transformer computes 384-dimensional embeddings for each sentence
- No tokenization, padding or text processing required
- Embeddings fed into LASSO model for classification

Model Evaluation – Overall



Multi-class Evaluation Metrics on Test Set

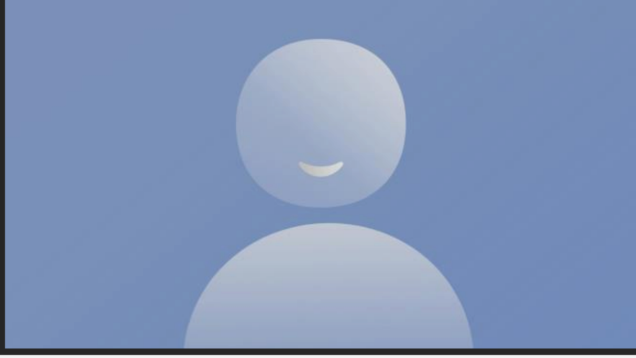
Model	Macro Precision (%)	Macro Recall (%)	Macro F1Score (%)	Accuracy (%)
BoW	40.4	40.2	40.1	41.3
TF-IDF	40.6	40.4	39.9	42.2
GloVe + LASSO	43.5	42.2	41.9	44.0
GloVe + CNN	40.8	39.1	38.2	40.4
DistilBERT	42.5	41.9	42.1	42.7
all-MiniLM-L6-v2	**43.6**	**43.1**	**43.0**	**44.4**

1. all-MiniLM-L6-v2 leads in all metrics, reflecting advanced contextual understanding.

2. GloVe+LASSO considerable improvement, showing the value of semantic embeddings.

3. GloVe+CNN surprisingly underperformed traditional methods, likely reflecting choice of shallow architecture.

Model Evaluation – Fake news only



Evaluation Metrics for pants-fire on Test Set

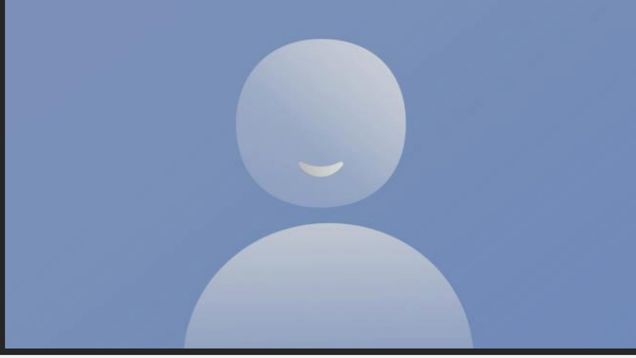
Model	Precision (%)	Recall (%)	F1Score (%)
BoW	35.0	28.7	31.5
TF-IDF	34.3	21.6	26.5
GloVe + LASSO	**41.6**	24.0	30.4
GloVe + CNN	36.5	29.2	32.5
DistilBERT	39.1	**33.9**	**36.3**
all-MiniLM-L6-v2	39.6	28.9	33.4

1. DistilBERT excels in 'pants-fire' detection with top recall and F1 score – identifies majority of fake news without sacrificing on false positive detection

2. GloVe+LASSO achieved highest precision – require less manual labour for fake news verification.

3. Emphasis on recall as key business metric for effective fake news mitigation strategies. Undetected fake news (false negative) more detrimental than manual verification (false positive)

Model Explainability



- Ensure **<u>trust and accountability</u>** – help users understand rationale behind fake news classification
- Explanations reveal valuable insights into the **<u>fit and fairness</u>** of classifiers
  - Language models can inherit human prejudices and cultural associations

- Two explanation methods
  1. **Feature Importance** with LASSO Intrinsic Explainability (BoW, TF-IDF)
  2. **News Content** explanation with attention mechanism (DistilBERT)

Model Explainability – BoW and TF-IDF

Feature	Coefficient
tax break	1.13
muslim	1.11
protest	0.93
everybodi	0.90
statist	0.88
income tax	0.88
realli	0.84
rep	0.84
tax increas	0.83
sell	0.82

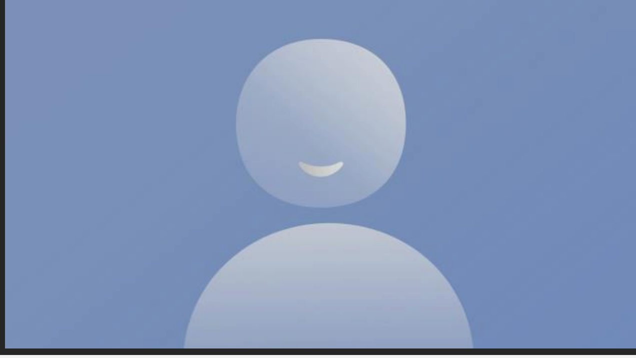
Top 10 Feature Importance using BoW

Feature	Coefficient
muslim	2.18
rep	1.86
obama	1.65
tax increas	1.61
protest	1.57
illeg	1.52
scott walker	1.46
doctor	1.44
care law	1.37
talk	1.35

Top 10 Feature Importance using TF-IDF

- Indicate word tokens that are most strongly associated with statements that were labeled as Pants-fire
- Provides transparency into how the model decides which statements are Pants-fire, aiding interpretability

Model Explainability – DistilBERT



- **Integrated Gradients** (IG) to attribute predictions to input text
- **Benefits**
  1. Local explainability
  2. Highlights specific input text driving classification decisions
- IG Process
  1. Interpolate word embeddings between zero-pad and input text
  2. Compute gradients for each interpolation
  3. Integrate gradients over interpolation path to obtain attribution score

Attribution of words in fake news

Predicted Label	Attribution Label	Attribution Score	Word Importance
pants-fire (0.18)	true	-2.69	[CLS] birth control pioneer margaret sang ##er was ##an active participant in the ku k ##lux klan . [SEP]

Predicted Label	Attribution Label	Attribution Score	Word Importance
pants-fire (0.23)	true	-2.76	[CLS] on an income cap for recipients of the popular hope scholarship [SEP]

Predicted Label	Attribution Label	Attribution Score	Word Importance
pants-fire (0.35)	true	0.00	[CLS] many of the founding fathers were very actively involved in cock ##fighting . [SEP]

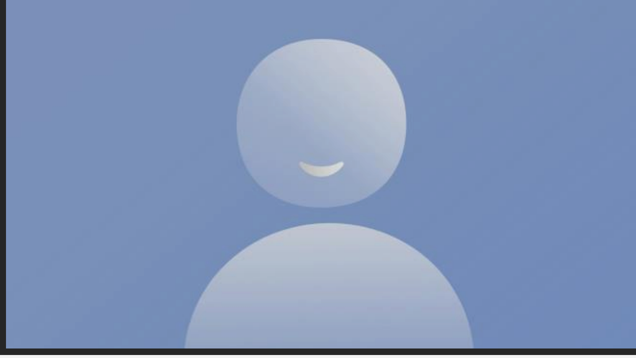
Attribution of words in real news

Predicted Label	Attribution Label	Attribution Score	Word Importance
true (0.48)	true	1.99	[CLS] says the unemployment rate for college graduates is 4 . 4 percent and over 10 percent for non ##coll ##ege - educated . [SEP]

Predicted Label	Attribution Label	Attribution Score	Word Importance
true (0.66)	true	2.44	[CLS] the united states has a low voter turnout rate . [SEP]

Predicted Label	Attribution Label	Attribution Score	Word Importance
true (0.44)	true	2.17	[CLS] before world war ii , very few people actually had health insurance . [SEP]

Conclusion



Findings

1. Transformer models like all-MiniLM-L6-v2 surpass traditional ML in context understanding for fake news detection.
2. DistilBERT’s high recall indicates strong potential in critical misinformation scenarios.
3. Justified the importance of recall over precision in mitigating the spread of fake news.
4. Explored model explainability methods to improve trust and accountability in predictions

Future Work?

1. Incorporate multimodal data to enhance detection of sophisticated misinformation.
2. Assess models’ robustness against adversarial attacks and evolving fake news strategies.
3. Continue leveraging ML advancements to bolster information integrity and public discourse.


THANK YOU!