Navigating Truth: Leveraging NLP to Detect Fake News

王珩琨　饶艺　田丰　王璇　谢冰洁　张波睿

Nankai University
School of Statistics and Data Science

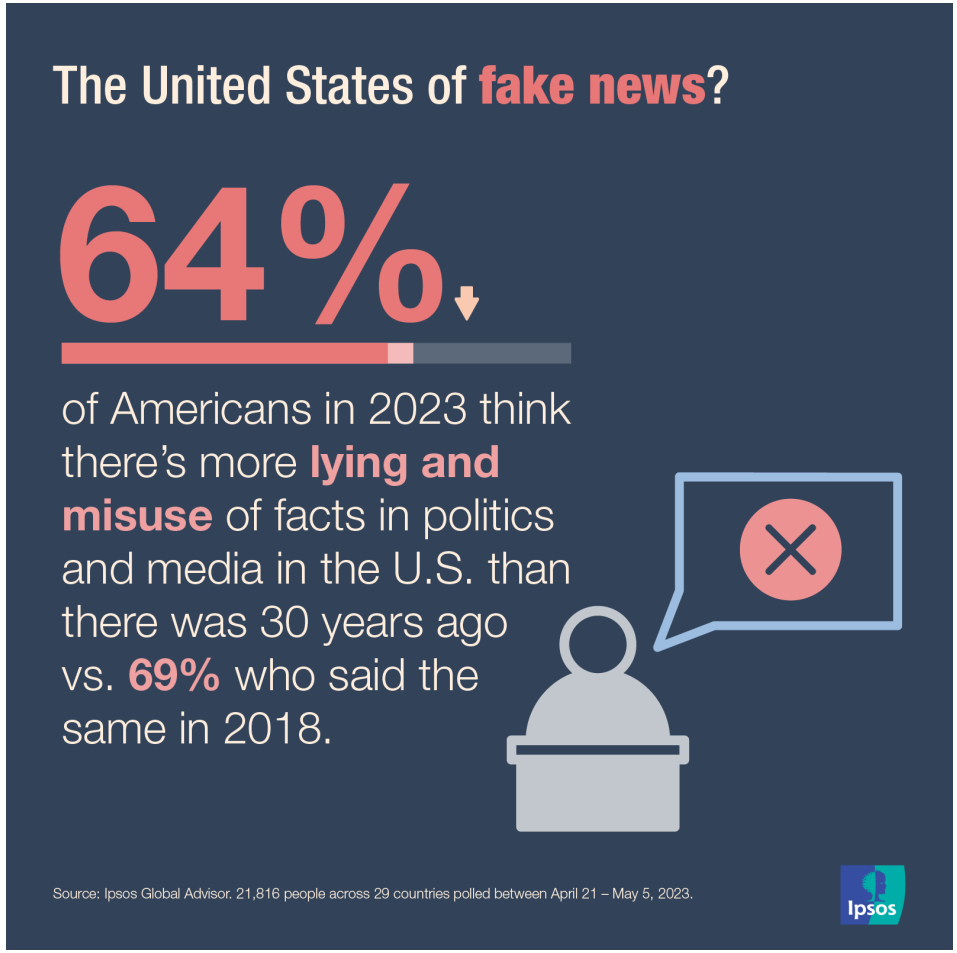
The second main building　2023.12

Contents

Background

More and More FAKE NEWS!

According to a survey by Ipsos Global Advisor, 64% of Americans in 2023 believe that there is more fake news now compared to 30 years ago.



Fake news in the age of AI

As the early January elections approach, pro-government media in Bangladesh are using artificial intelligence to fabricate news and videos in order to discredit the opposition Bangladesh Nationalist Party.

Our Works

We used the dataset from kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/, and considered six different methods of extracting features from the text:

- Statistical Features
  - CountVectorizer
  - TF-IDF
- Static Word Embeddings
  - Word2vec
  - GloVe
- Contextual Word Embeddings
  - all-mpnet-base-v2(MPNet)
  - text-embedding-ada-002(GPT3)

Data Processing

Overview

- Fake.csv 23481 rows $\times$ 4 columns labelled as 0
- True.csv 21417 rows $\times$ 4 columns labelled as 1

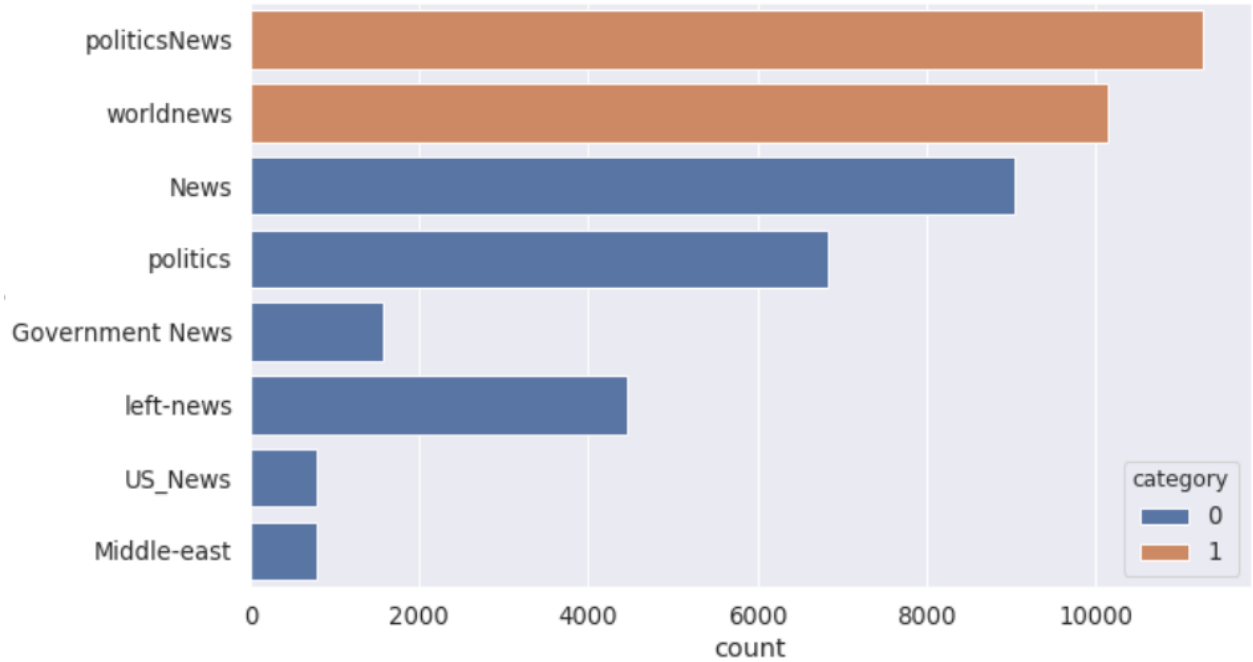
Index	Title	text	Subject	Date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017

Table: The first three rows of True.csv

Data Cleaning

1. Missing Value Check: There are 631 rows in the 'text' column where the string contains only spaces.

2. Column Fusion: The string contents of the 'title' and 'text' columns have been concatenated to provide more comprehensive information.



3. delete 'subject' and 'date' columns

Text Cleaning

1. Remove HTML tags
2. eliminate content within brackets such as timestamps [23:59 EST]
3. remove URLs
4. discard punctuation and stop words like 'the'
5. delete empty strings resulting from text cleaning

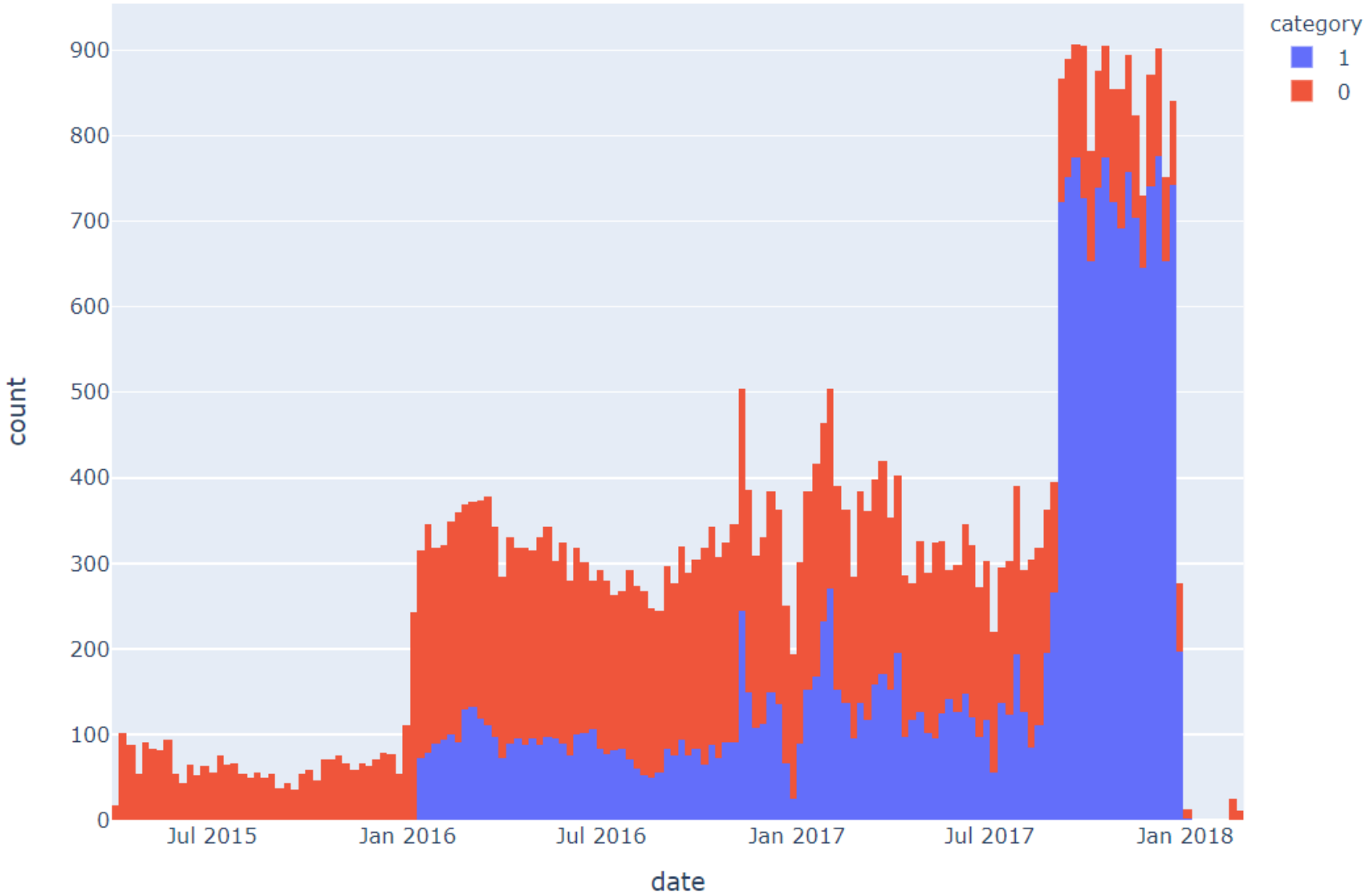
Display the data after cleaning

index	text	category
0	WASHINGTON (Reuters) head conservative Republi...	1
1	WASHINGTON (Reuters) Transgender people allowe...	1
2	WASHINGTON (Reuters) special counsel investiga...	1
3	WASHINGTON (Reuters) Trump campaign adviser Ge...	1
4	SEATTLE/WASHINGTON (Reuters) President Donald ...	1

44889 rows $\times$ 2 columns

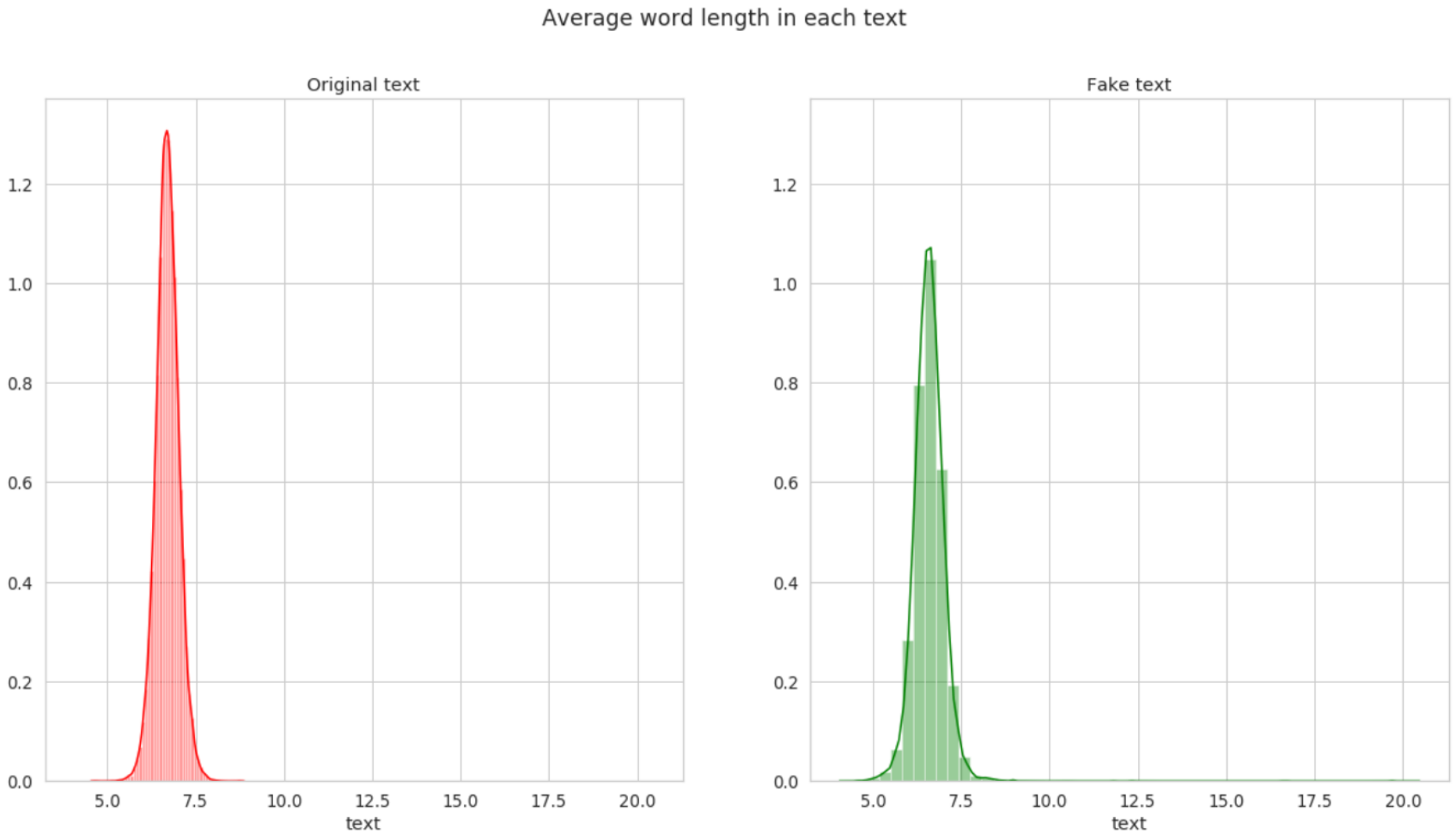
EDA

The Temporal Distribution of News Volume

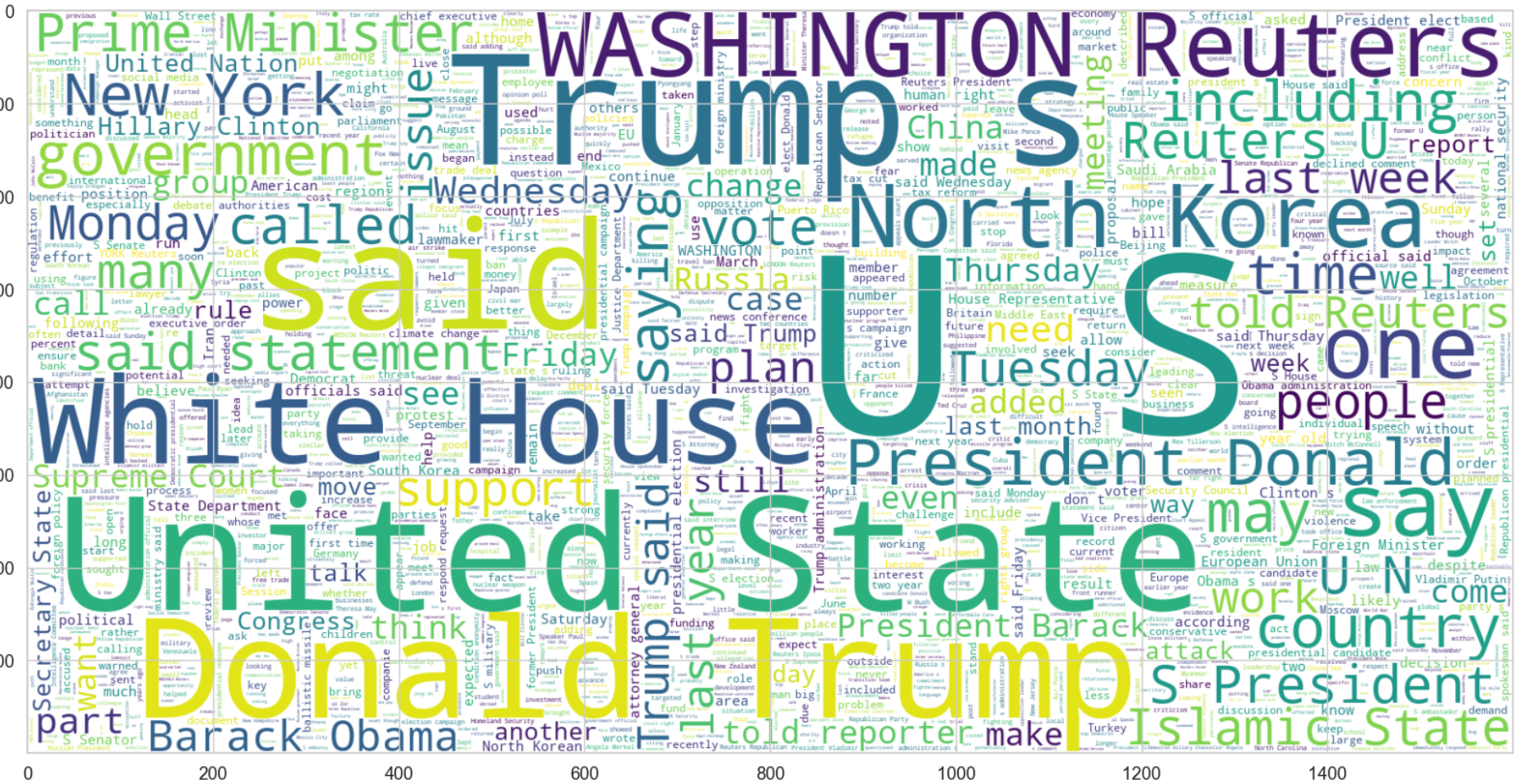


words in texts

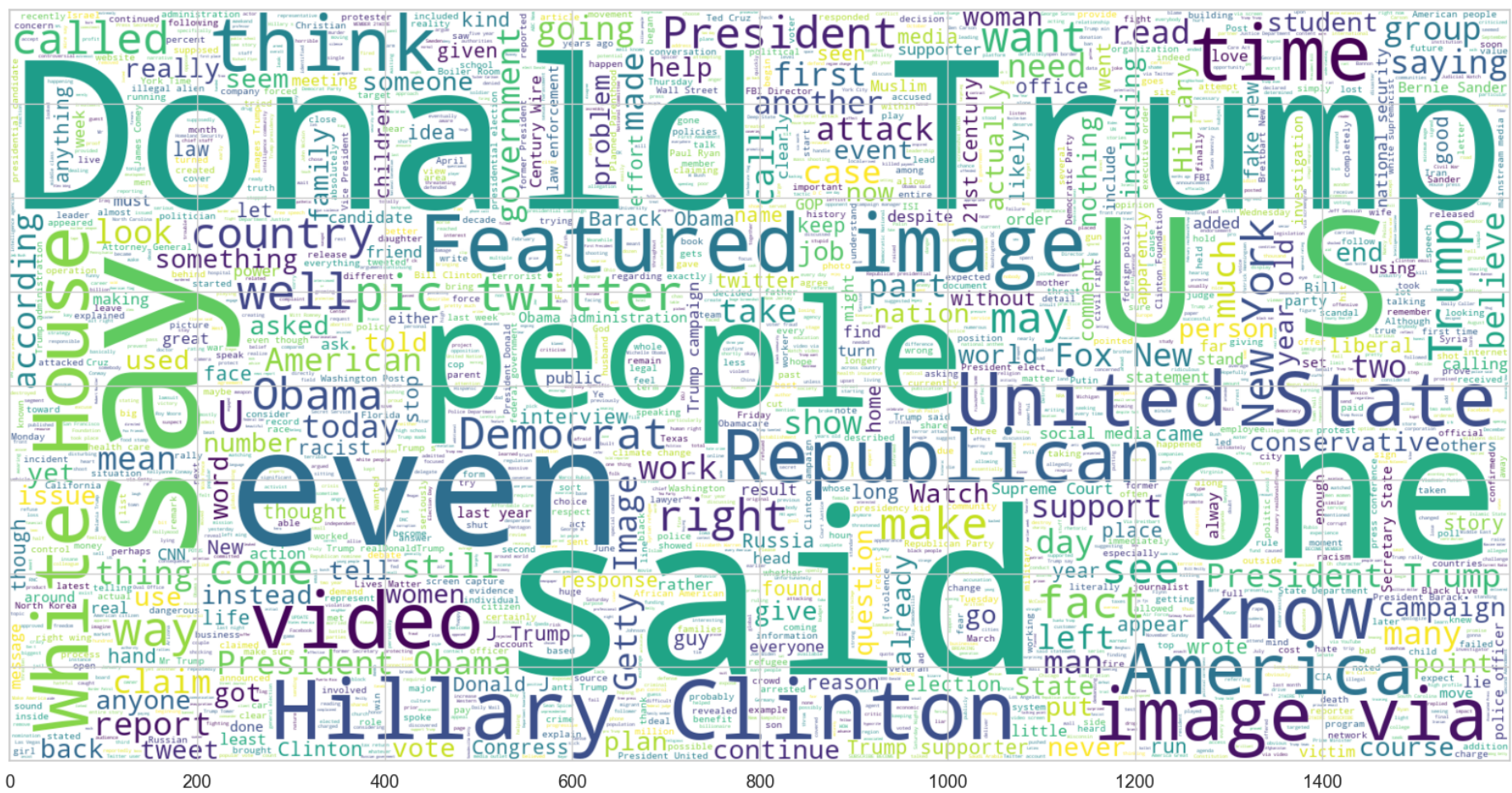
average word length



WordCloud for real text



WordCloud for fake text



Text Analysis Methods: A Comparative Study

Statistical Features

- CountVectorizer
- TF-IDF $\quad tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$

Definition (Term frequency)

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ is the frequency of term t in document d.

Definition (Inverse document frequency)

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

where:

- N is the total number of documents in the corpus $N = |D|$.
- $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears (i.e., $tf(t, d) \neq 0$). If the term is not in the document, then $idf(t, D) = 0$.

CountVectorizer vs. TF-IDF

Method	SVC			lightgbm			MLP(128/64/32/1)		
Embedding	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy
CountVectorizer(5000 dims)	**0.9966**	**0.9968**	**0.9967**	**0.9972**	**0.9980**	**0.9976**	**0.9942**	**0.9944**	**0.9943**
TF-IDF(5000 dims)	0.9961	0.9946	0.9953	0.9977	0.9972	0.9974	0.9884	0.9874	0.9879

Static Word Embeddings

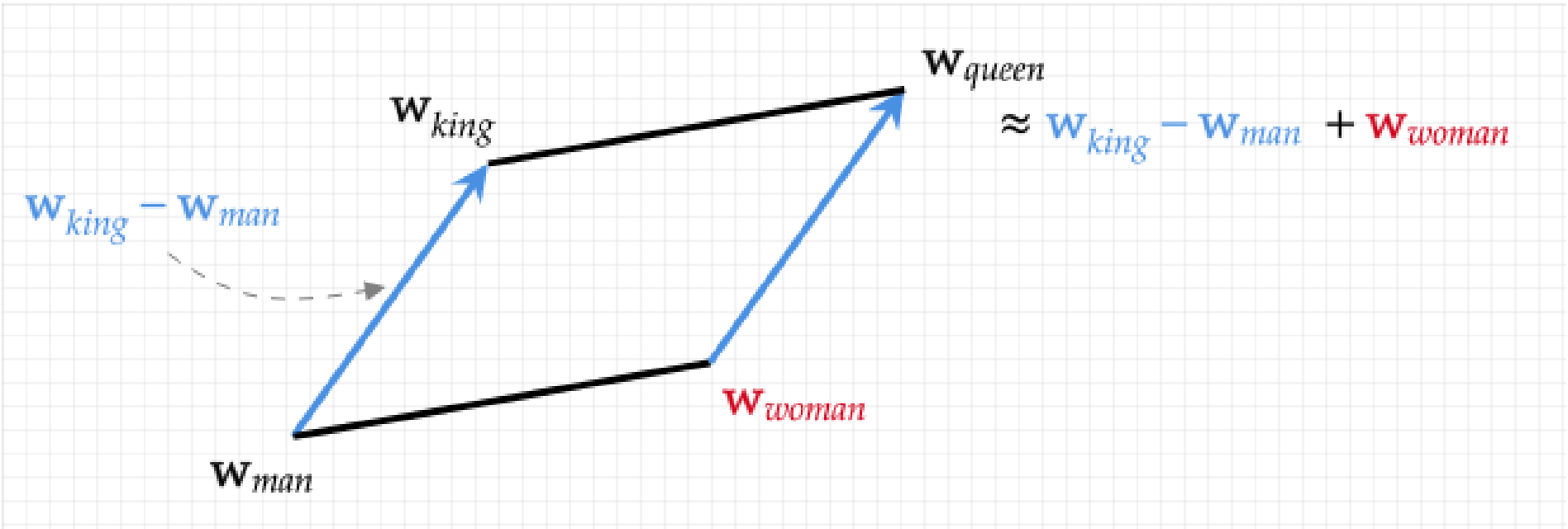
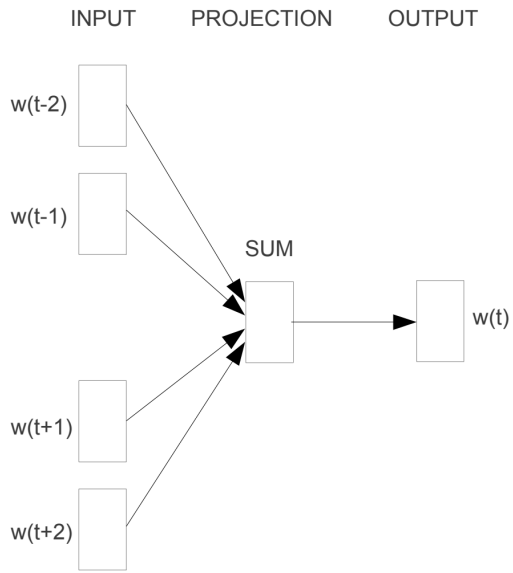


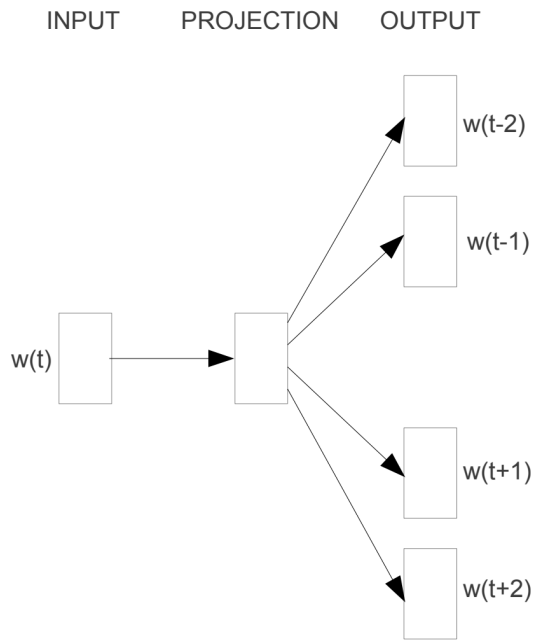
Figure: $queen \approx king - man + woman$

Word2vec (Mikolov et al., 2013)



CBOW

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \sum_{-n \le j \le n, \neq 0} \log p\left(w_{j+1} \mid w_t\right)$$



Skip-gram

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t | w_{t-n}, \ldots, w_{t-1}, w_{t+1}, \ldots,$$

$$w_{t+n})$$

GloVe (Pennington et al., 2014)

GloVe Model Objective Function

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

$$f(x) = \begin{cases} (\frac{x}{x_{\max}})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

Co-occurrence Matrix Example

	Word1	Word2	Word3
Word1	X_{11}	X_{12}	X_{13}
Word2	X_{21}	X_{22}	X_{23}
Word3	X_{31}	X_{32}	X_{33}

Note: X_{ij} denotes the number of times Word$i$ co-occurs with Word$j$ within a certain context window.

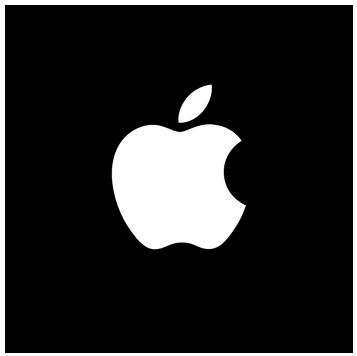
Word2vec vs. GloVe

Method	SVC			lightgbm			LSTM(128/64)+dense(32/1)		
Embedding	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy
Word2vec(100 dims/per word)	**0.9816**	**0.9756**	**0.9785**	**0.9767**	**0.9806**	**0.9787**	0.9950	0.9980	0.9965
glove(100 dims/per word)	0.9494	0.9504	0.9499	0.9498	0.9608	0.9555	**0.9987**	**0.9990**	**0.9988**

Contextual Word Embeddings

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.



Pretrained Transformers

Figure: HuggingFace

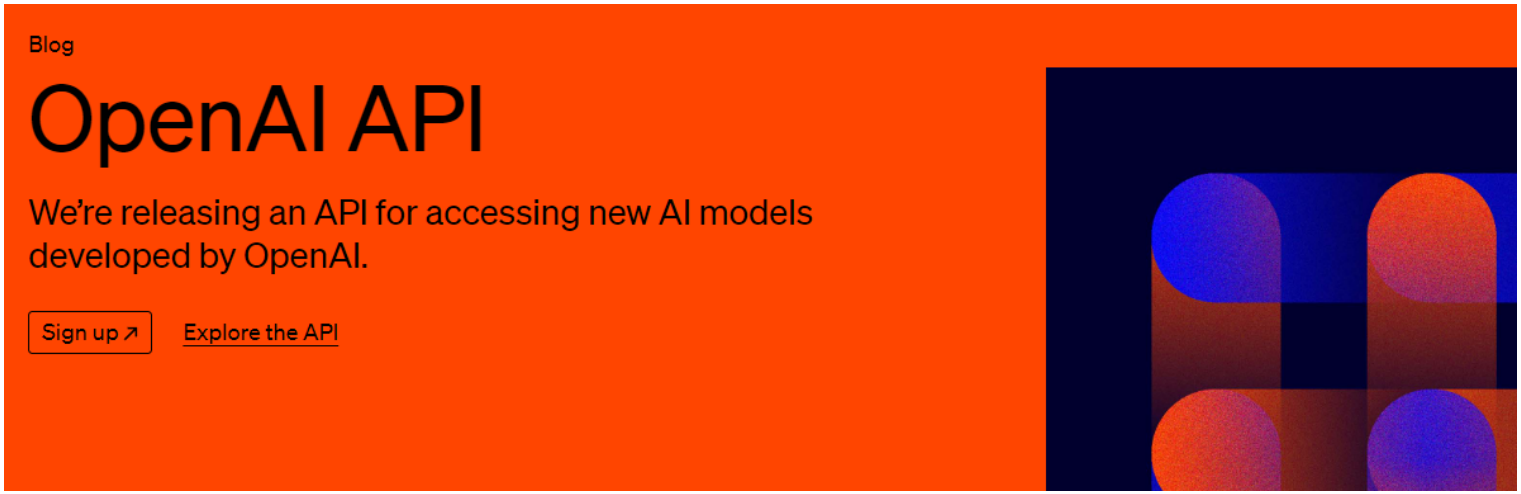


Figure: OpenAI API

all-mpnet-base-v2 vs. text-embedding-ada-002

Method	SVC			lightgbm			MLP(128/64/32/1)		
Embedding	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy
all-mpnet-base-v2(768 dims)	0.9844	0.9843	0.9849	0.9765	0.9800	0.9783	0.9644	0.9897	0.9776
text-embedding-ada-002(1536 dims)	**0.9993**	**0.9991**	**0.9992**	**0.9937**	**0.9938**	**0.9938**	**0.9988**	**0.9998**	**0.9993**

Summary

Summary

- On the 'Fake and Real News' dataset, the majority of feature extraction methods achieved an accuracy exceeding 99%.
- The dataset is of high quality but not sufficiently large in scale, resulting in minimal differences among various feature extraction methods. Further experiments with more data are required to validate the effectiveness of these methods in distinguishing between true and false news.

Thanks!