Fake News Detection using Natural Language Processing

Mohamed Ayman†
Computer Engineering
The American University in Cairo
mohamedayman15069@aucegypt.edu

Abdelrahman Shaaban
Computer Engineering
The American University in Cairo
abdelrahmanfawzy@aucegypt.edu

1. Introduction

Internet and social media have made the access to the news information much easier and comfortable. When people need to follow their event of interest, they just use their phones online. Unfortunately, some challenges have become apparent in disseminating information. More clearly, the prevalence of information on the Internet and social media have made intricate on users to distinguish between the fake and real news. As a result, mass media which has a great influence on the society may manipulate information in different ways. It means that people can easily produce mixed news of true and false information or false information completely. This has given motivation that there are several websites generating false information exclusively. Deliberately, they publish propaganda, disinformation, and hoaxes for sake of personal reasons, one of which is affecting public opinions on certain matters, political mostly. They amplify the effect of such information by publishing it on social media. A classic example of fake news is the mainstream news during U.S. 2016 president election (Tan et al., 2020).

Thus, there is a need for a solution to distinguish between fake and real news. Since artificial intelligence algorithms have started to work much better than before on classification problems, such as image recognition and sentiment analysis detection, and datasets of fake news have become ubiquitous, scientists believe that this problem can be addressed by means of machine learning and artificial intelligence (Gravanis et al., 2019). Thus, the researchers of the paper have been catalyzed to understand how machine learning can detect fake news. This paper discusses different types of machine learning algorithms for detecting fake news from text. In this sense, the research project will use multiple NLP models, such as BERT. Regarding datasets, LIAR, Fakeddit, Real News, and Snopes datasets are used. The goal of the research is to see which machine learning algorithm explained in the class will be the most accurate one for detecting the fake news.

2. Literature Review

Fake news detection is one of the most active research areas in recent years. In this section, we show some of the most well-known work in that area.

2.1 Fake News Detection Using Naive Bayes Classifier

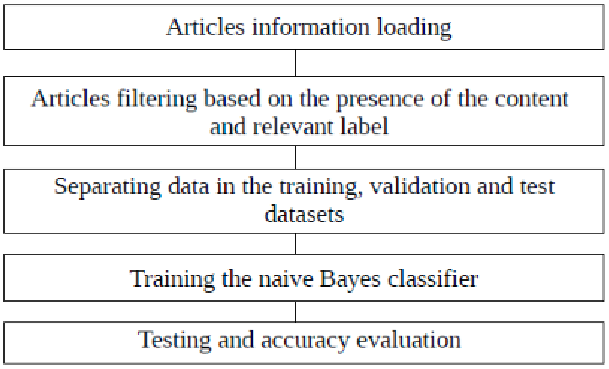
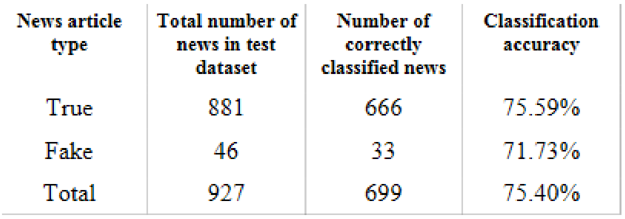

Articles information loading
Articles filtering based on the presence of the content and relevant label
Separating data in the training, validation and test datasets
Training the naive Bayes classifier
Testing and accuracy evaluation

Fig. 1. Generalized scheme of the naive Bayes algorithm training



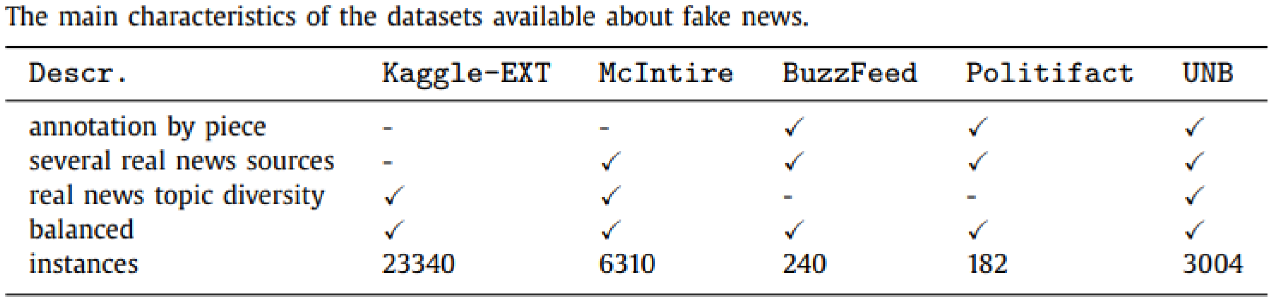
News article type	Total number of news in test dataset	Number of correctly classified news	Classification accuracy
True	881	666	75.59%
Fake	46	33	71.73%
Total	927	699	75.40%

Table 1. The Classification results of the naive Bayes model

Granik and Mesyura (2017) worked on a simple approach for fake news detection using a naive Bayes classifier (Figure 1). It was trained and tested using a relatively small dataset of Facebook news posts, collected by BuzzFeed News, and achieved a classification accuracy of around 74% (Figure 2). The authors considered this as a decent result considering the relative simplicity of the model.

WOODSTOCK'18, June, 2018, El Paso, Texas USA
© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/1234567890

Table 2



The main characteristics of the datasets available about fake news.

Descr.	Kaggle-EXT	McIntire	BuzzFeed	Politifact	UNB
annotation by piece	-	-	✓	✓	✓
several real news sources	-	✓	✓	✓	✓
real news topic diversity	✓	✓	-	-	✓
balanced	✓	✓	✓	✓	✓
instances	23340	6310	240	182	3004

The main idea of the naive Bayes model they built is to treat each word of the news article independently. It is like using the naive Bayes theorem to calculate the probability of a news article being fake based on the article's words and the appearance of certain words that are known to be common in the fake news.

2.2 A benchmarking study for fake news detection

Georgios Gravanis and his team proposed a model for fake news detection using content-based features and machine learning algorithms. More specifically, they made use of linguistic features (e.g.words belonging in certain categories, part of Speech tags and others) in a combination with ML algorithms (the most popular ML classifiers).

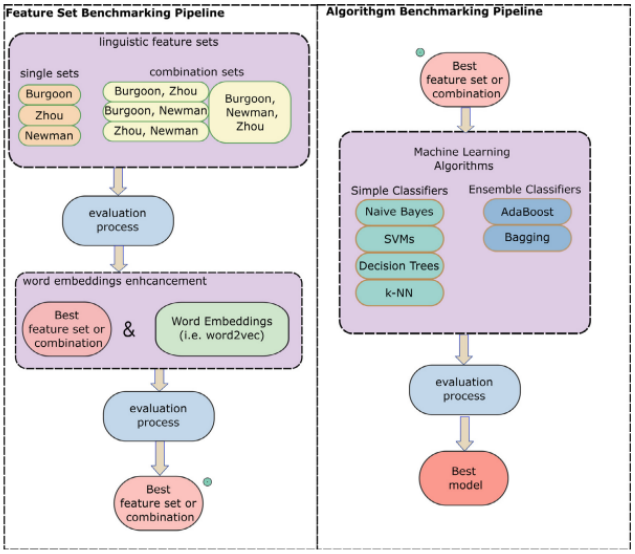
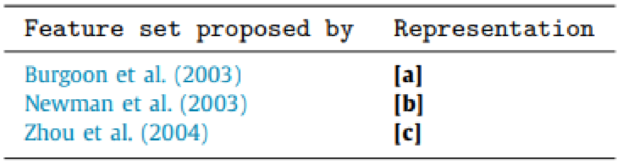


Fig. 2. Benchmarking pipelines followed in this study.

2.2.1 Feature Set Benchmarking pipeline

They created three tables for linguistic features. The features in the first table are comprised of four categories, namely grammatical complexity, vocabulary complexity, quantity, and Specificity and Expressiveness [a]. The second table divided the features to three categories, namely psychological processes, standard linguistic dimensions, and relativity [b]. The third table is comprised of nine categories: quantity, complexity, uncertainty, non-immediacy, expressivity, Specificity, and affect [c]. [a], [b], and [c] are represented as the number of tables as shown in the table 3 for simplicity.



Feature set proposed by	Representation
Burgoon et al. (2003)	**[a]**
Newman et al. (2003)	**[b]**
Zhou et al. (2004)	**[c]**

Table 3. Feature set representation for this study.

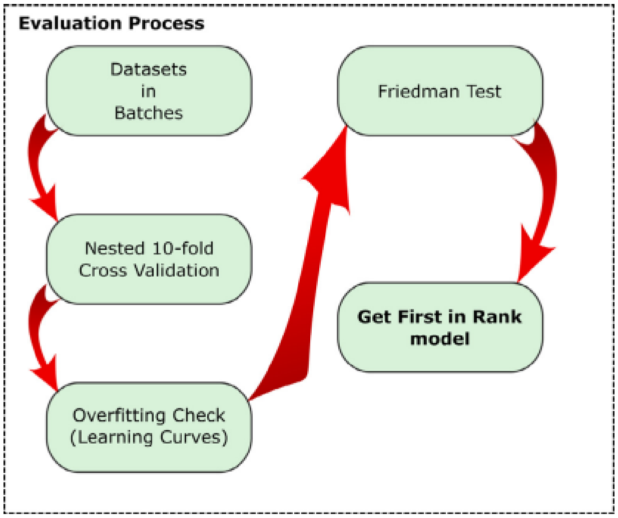


Fig. 3. The evaluation process we used for this study

From figure 2, the feature set benchmarking pipeline is comprised of different processes. In the first process, the Georgies Gravanies and his researchers succeeded in taking an advantage of the previous papers by experimenting the whole combination of three tables, together for sake of boosting the accuracy and the usage of linguistic features as presented in the table 2. The second process is "Evaluation Process". To ensure the equivalent evaluation of all feature sets and algorithms, they followed the steps as shown in figure 3. The figure 3 illustrates systematic steps to for equivalent evaluations for both features and algorithms. First, they split the dataset up to 1000 random batches. Second, in each dataset batch, they used 10-fold cross-validations. To avoid the overfitting, they referred to learning curves plots and got the optimum parameters values. Then, to compare the significance of multiple methods over several datasets, they used the Friedman Test. The third process is for word embedding. They utilized the Word2vec and GloVe for learning word embedding from raw text. When experimenting both, they did not notice any difference in terms of accuracy. Finally, they tried to use the same evaluation process to compare the best performance after enhancement between the combination of linguistic feature, word embeddings, and both together. They did not step at this point, but they created an unbiased fake news dataset with taking into consideration four important standards. First, Fake news articles should be reviewed by experts. Second, Real news should be published from credible organizations. Third, Fake news should originate from several sources. Fourth, it is a must to obtain several articles from a varied number of categories. From the table 2, it is apparent that they created UNB (unbiased dataset) for covering all mentioned description compared to the other datasets.

2.2.2 Machine Learning algorithms used in the study

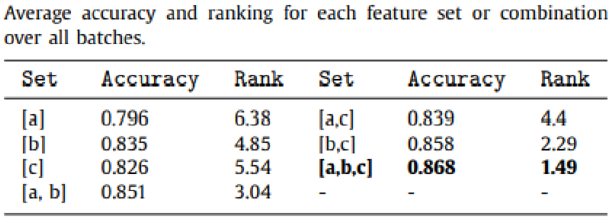
In this study, they compared four classifiers, namely SVMs, decision trees, Naïve Bayes, and K-NNs together with two ensemble methods (Bagging and AdaBoost) as shown in the figure 2 on the right.

2.2.3 The Results of the study

2.2.3.1 Experimenting all combinations of linguistic features.

Using the SVM classifier with a linear kernel using 70% of the data and tests the rest 30%. From the figure above, the reader can notice that the combination of a& b& c features achieves the best accuracy compared to the other combinations as presented in the table 4.

Table 4.



Average accuracy and ranking for each feature set or combination over all batches.

Set	Accuracy	Rank	Set	Accuracy	Rank
[a]	0.796	6.38	[a,c]	0.839	4.4
[b]	0.835	4.85	[b,c]	0.858	2.29
[c]	0.826	5.54	**[a,b,c]**	**0.868**	**1.49**
[a, b]	0.851	3.04	-	-	-

2.2.3.2 Experimenting [a][b][c] combination with word embeddings

The next step is the use of Word embedding, specifically Word2vec. They started with considering a pre-trained model (Google News corpus) to extract 300-word embeddings vector for each article. Then, they checked the performance with word embeddings alone and word embeddings with the three combinations of linguistic features and noticed that the [a][b][c] linguistic feature combination with embedding gives the best accuracy as presented in the figure 4.

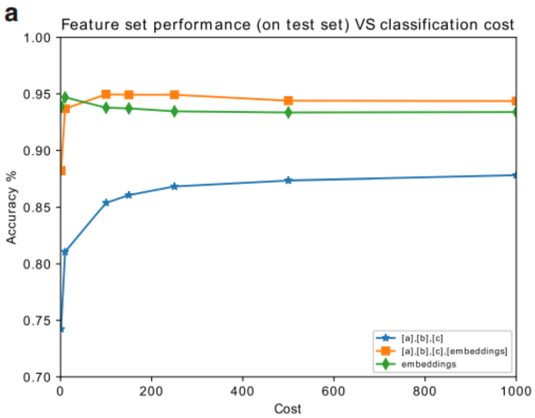


Fig 4. Performance of each feature set of combination vs C parameter.

2.2.3.3 Experimenting the Six Machine learning algorithms with [a][b][c] combination and Word Embeddings

The final step is that they experimented K-NN, Decision Tree, Naïve Bayes, SVM, AdaBoost, and Bagging for deciding which model is the best for detecting fake news with the dataset adjusted to [a][b][c] lingustic feature combination and word embeddings enhancement. SVM is the best in terms of accuracy as shown in figure 5.

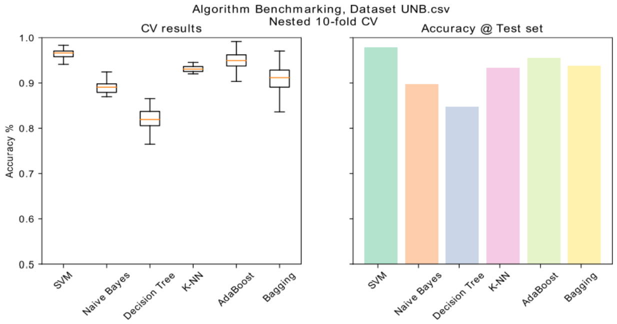


Fig 5. Algorithm evaluation on UNB Dataset.

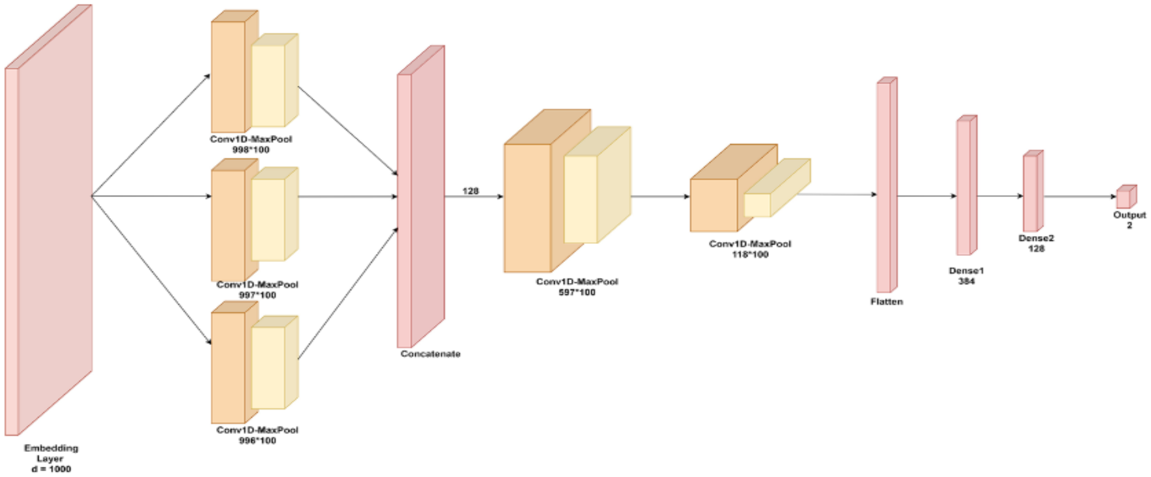
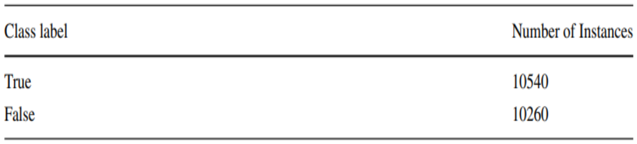


Fig. 6. FakeBERT Model

2.3 FakeBERT: Fake news detection in social media with a BERT-based deep learning approach

Kaliyar and Goswami (2021) developed a bidirectional deep learning approach for getting the relevant content of fake news, considering the semantic and long-distance dependencies in sentences. They could develop a combination of BERT (Bidirectional Encoder representations from Transformers) with multiple blocks of the deep Convolutional Neural Network (CNN) for building their classification model as shown in figure 6. They mainly used the vectors generated after word-embedding from BERT to be the input for the CNN layers due to its powerful learning ability in extracting the features and learning the representations from the dataset.



Class label	Number of Instances
True	10540
False	10260

Table 5. Fake news dataset with the class labels

For the training and testing, they used a relatively small (around 21,000 samples) collection of the fake and real news that was propagated during the time of the U.S. Presidential election 2016 with the number of instances as shown in table 5. Their classification results demonstrated that their model could outperform many existing models with an accuracy of 98.9% as illustrated in figure 7.

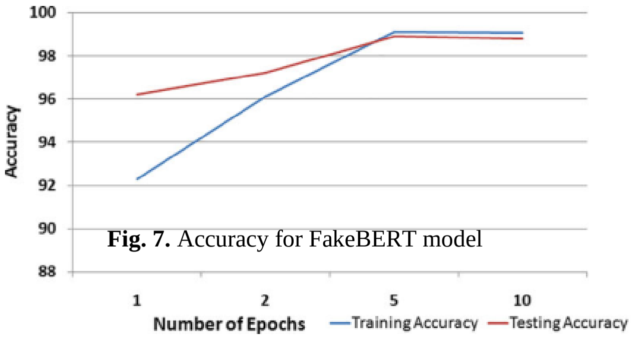
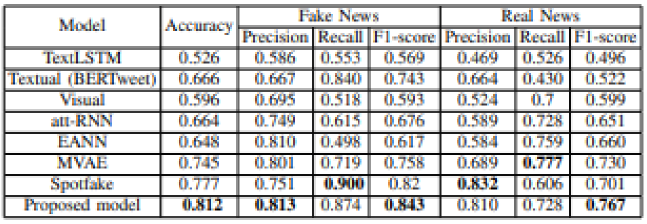


Fig. 7. Accuracy for FakeBERT model

2.4 Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection

Tuan and Minh (2021) presented a multimodal model for detecting fake news fusing the features of textual and visual data. As the different modalities of the news articles can show different aspects of it and complement each other in detecting the authenticity of the news, they believed that can help in a better classification. They used a pre-trained BERT model for the text features and a VGG-19 pre-trained model for the images' features. They proposed a scale-dot product attention mechanism to get the relationship between the two components' features, in addition to a self-attention mechanism on images to detect if all parts of each image are related. They used the MediaEval 2016 dataset for the training and evaluation with around 17,000 unique tweets, each with its associated image. They could obtain a 80.8 of accuracy and 80% of F1-score, which outperformed the state-of-the-art method on a public Twitter dataset by 3.1% accuracy as shown in table 6. The model they proposed was built using four major parts as demonstrated in figure 8. The first part is to extract the text features using BERT embedding combined with a 1-D Convolution Neural Network (CNN). They used a BERT model pre-trained on Tweet data, with the ability of its different hidden layers to capture different kinds of information of the text and its context. They used 1-D CNN layers to extract more information from different sets of the embedded word vectors. The second part is to extract the



Model	Accuracy	Fake News			Real News		
		Precision	Recall	F1-score	Precision	Recall	F1-score
TextLSTM	0.526	0.586	0.553	0.569	0.469	0.526	0.496
Textual (BERTweet)	0.666	0.667	0.840	0.743	0.664	0.430	0.522
Visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
MVAE	0.745	0.801	0.719	0.758	0.689	**0.777**	0.730
Spotfake	0.777	0.751	**0.900**	0.82	**0.832**	0.606	0.701
Proposed model	**0.812**	**0.813**	0.874	**0.843**	0.810	0.728	**0.767**

Table 6. Performance of the paper proposed model vs other models

image features using the pre-trained VGG-19 model. The third part is to capture the common features of the first two parts using the scaled dot-product attention mechanism besides using the self-attention mechanism on the images to measure how the content of each image relates to each other. The final part is all about fusing the collected features and passing them into a fully connected layer for the final step of classification.

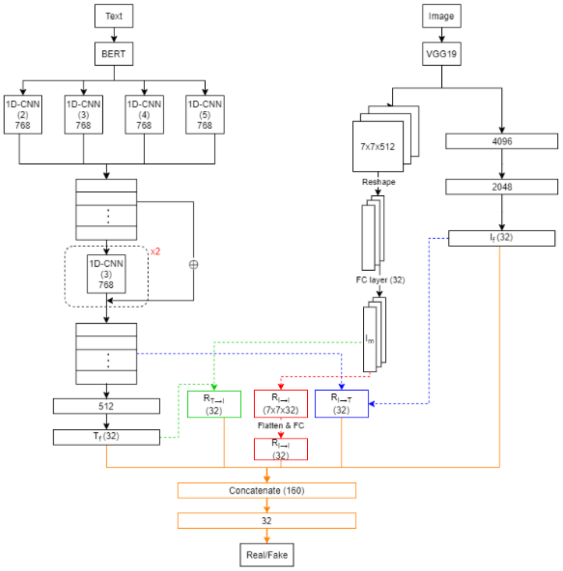
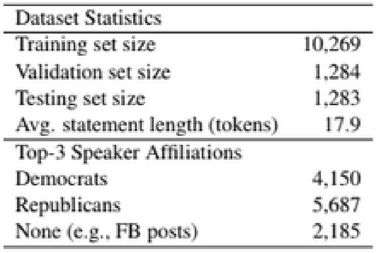


Fig. 8. Model Architecture.

3. Datasets Survey

In this section, we survey well-known datasets related to the chosen area of research and decide which one we will use.

3.1 LIAR Dataset

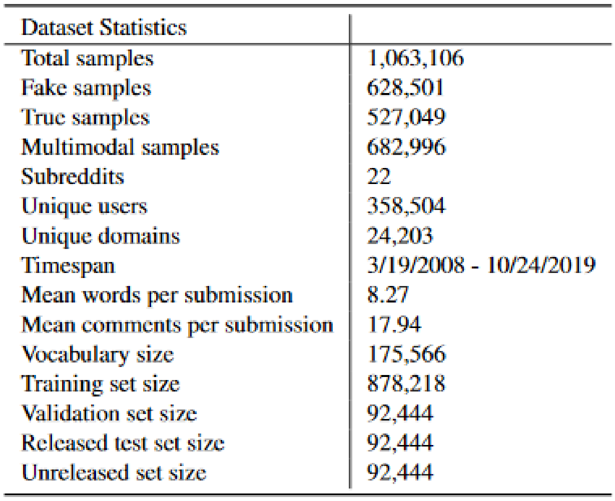


Dataset Statistics	
Training set size	10,269
Validation set size	1,284
Testing set size	1,283
Avg. statement length (tokens)	17.9
Top-3 Speaker Affiliations	
Democrats	4,150
Republicans	5,687
None (e.g., FB posts)	2,185

Table 7. The LAIR Dataset Statistics

LIAR dataset is one of the most well-known and publicly available datasets that have been used widely in the research area of fake news detection since 2017. It contains 12.8K of manually labeled short statements in various contexts from the website “Politifact.com” which provides detailed analysis reports and links to source documents for each case. LIAR authors considered six fine-grained labels, as shown in table 7, for the truthfulness ratings: pants-fire, false, barely true, half-true, mostly true, and true. All the classes are equally distributed, ranging from 2,063 to 2,638 except for the first class (pants-fire) with only 1,050 samples, which makes this dataset relatively well-balanced. LIAR contains seven features including the news’ statement, subjects, speaker, speaker’s job title, the state’s information, the party affiliation, and the context. LIAR’s limitations include its lack of the time features, the associated visual components, and the social context.

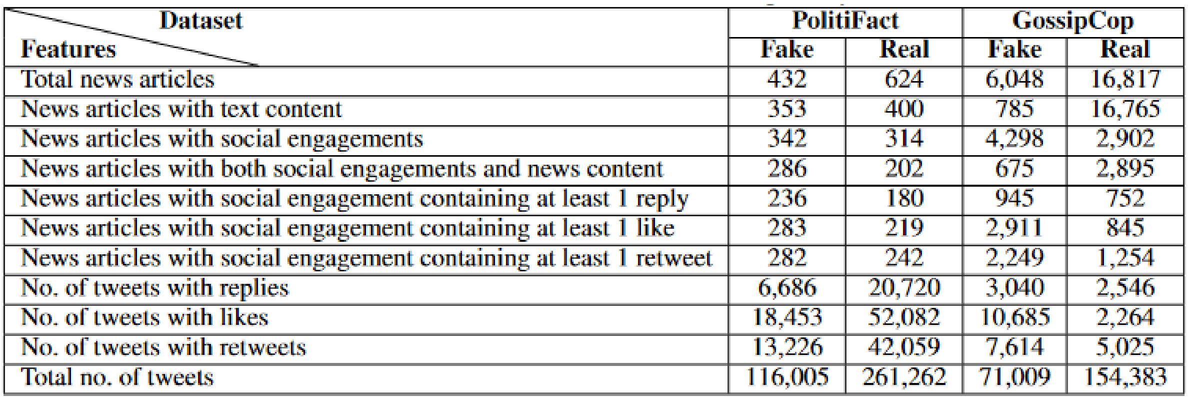
3.2 Fakeddit Dataset



Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Table 8. The Fakeddit Dataset Statistics

Fakeddit is a multimodal, online available dataset, collected from the website Readdit, a social news and discussion website where users can post submissions on various subreddits. It consists of over 1 million samples from multiple categories of fake news. The samples are labeled according to 2-way, 3-way, and 6-way classification categories. The dataset has a multimodal subset of 682,996 samples, which include each news’ headline along with the news piece’s attached image. For the 2-way classification, the dataset is almost balanced with 527,049 true samples and 628,501 fake samples as shown in table 8. The authors recommend using the 6-way labels: true, satire, misleading content, manipulated content, false connection, and imposter content, along with the clean-title feature and its associated image from the public dataset for better training and evaluation. This dataset is very large and balanced to be used without any challenges for building a multimodal model, although it lacks the time features and the news’ context.

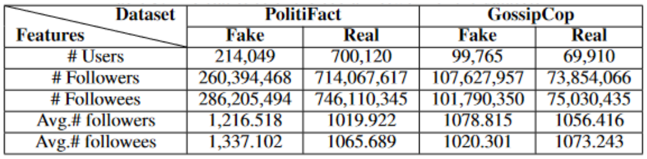


Dataset / Features	PolitiFact		GossipCop	
	Fake	Real	Fake	Real
Total news articles	432	624	6,048	16,817
News articles with text content	353	400	785	16,765
News articles with social engagements	342	314	4,298	2,902
News articles with both social engagements and news content	286	202	675	2,895
News articles with social engagement containing at least 1 reply	236	180	945	752
News articles with social engagement containing at least 1 like	283	219	2,911	845
News articles with social engagement containing at least 1 retweet	282	242	2,249	1,254
No. of tweets with replies	6,686	20,720	3,040	2,546
No. of tweets with likes	18,453	52,082	10,685	2,264
No. of tweets with retweets	13,226	42,059	7,614	5,025
Total no. of tweets	116,005	261,262	71,009	154,383

Table 9. Statistics of the FakeNewsNet repository

3.3 FakeNewsNet Dataset

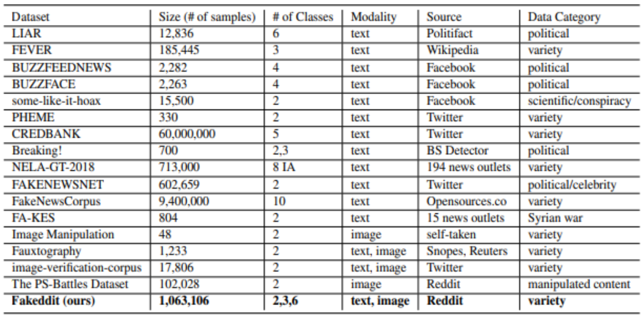
FakeNewsNet dataset contains two comprehensive datasets collected from two fact-checking websites: GossipCop and PolitiFact including news contents with labels, along with social context and dynamic information. For the news content, this dataset includes various information such as headline, body text, images, author information and attached links. For the social context, it includes the social engagement to the fake and real news pieces, the second order user behaviors to theses posts of the news, such as retweets and comments, and all of that along with the metadata of the users' profiles and their network information. It contains around 17,500 real news articles and 6500 fake ones. This dataset works well for developing a model that considers multiple aspects such as the news context and the users' network contribution for the fake news detection. However, tables 9 and 10 show that it is not balanced well and very complex to be used.



Dataset / Features	PolitiFact		GossipCop	
	Fake	Real	Fake	Real
# Users	214,049	700,120	99,765	69,910
# Followers	260,394,468	714,067,617	107,627,957	73,854,066
# Followees	286,205,494	746,110,345	101,790,350	75,030,435
Avg.# followers	1,216.518	1019.922	1078.815	1056.416
Avg.# followees	1,337.102	1065.689	1020.301	1073.243

Table 10. Statistics of the social networks of datasets.

4. Solution Overview.



Dataset	Size (# of samples)	# of Classes	Modality	Source	Data Category
LIAR	12,836	6	text	Politifact	political
FEVER	185,445	3	text	Wikipedia	variety
BUZZFEEDNEWS	2,282	4	text	Facebook	political
BUZZFACE	2,263	4	text	Facebook	political
some-like-it-hoax	15,500	2	text	Facebook	scientific/conspiracy
PHEME	330	2	text	Twitter	variety
CREDBANK	60,000,000	5	text	Twitter	variety
Breaking!	700	2,3	text	BS Detector	political
NELA-GT-2018	713,000	8 IA	text	194 news outlets	variety
FAKENEWSNET	602,659	2	text	Twitter	political/celebrity
FakeNewsCorpus	9,400,000	10	text	Opensources.co	variety
FA-KES	804	2	text	15 news outlets	Syrian war
Image Manipulation	48	2	image	self-taken	variety
Fauxtography	1,233	2	text, image	Snopes, Reuters	variety
image-verification-corpus	17,806	2	text, image	Twitter	variety
The PS-Battles Dataset	102,028	2	image	Reddit	manipulated content
Fakeddit (ours)	**1,063,106**	**2,3,6**	**text, image**	**Reddit**	**variety**

Table 11. Statistics of the social networks of datasets

Fake news has become a problem of great impact on societies and communities due to continuous, intense fakesters content distribution. In this sense, we have thought about how we can solve this problem by constructing a very accurate model to detect fake news. From thorough research, we have decided to use the Fakeddit dataset for discernible reasons. Fakeddit dataset is a very large, well-balanced, and representative dataset that is available online and was labeled by experts manually. Also, the dataset provides three label-based categories for each sample: 2-way, 3-way, and 6-way classification with an extensive variety compared to many datasets as shown in table 11. That allows us to train for fake news detection at a high level, in addition to a more fine-grained classification. The two-way classification detects whether a piece of news is true or false. The three-way classification determines whether it is completely true, fake but contains text that is true, or false. The six-way is a more fine-grained classification as shown in figure 9. We intend to start by experimenting with multiple NLP models using only the news' text for the 6-way classification. This will be followed by integrating the associated images of the multimodal dataset to build a multimodal fake news classifier, experimenting with different fusion and attention techniques if the time allows.

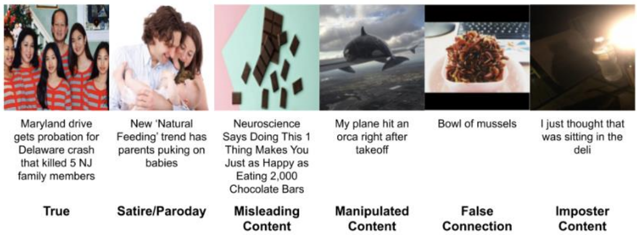

Maryland drive gets probation for Delaware crash that killed 5 NJ family members
New 'Natural Feeding' trend has parents puking on babies
Neuroscience Says Doing This 1 Thing Makes You Just as Happy as Eating 2,000 Chocolate Bars
My plane hit an orca right after takeoff
Bowl of mussels
I just thought that was sitting in the deli
True
Satire/Paroday
Misleading Content
Manipulated Content
False Connection
Imposter Content

Fig. 9. Dataset examples with 6-way classification labels

REFERENCES

[1] Gravanis, Georgios & Vakali, Athena & Diamantaras, Kostas & Karadais, Panagiotis. (2019). Behind the Cues: A benchmarking study for Fake News Detection. Expert Systems with Applications. 128. 10.1016/j.eswa.2019.03.036.

[2] Kaliyar, Rohit & Goswami, Anurag & Narang, Pratik. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications. 80. 10.1007/s11042-020-10183-2.

[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[4] Nakamura, Kai & Levy, Sharon & Wang, William. (2019). r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection.

[5] Shu, Kai & Mahudeswaran, Deepak & Wang, Suhang & Lee, Dongwon & Liu, Huan. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data. 8. 171-188. 10.1089/big.2020.0062.

[6] Tuan, Nguyen & Minh, Pham Quang Nhat. (2021). Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection.

[7] Wang, William. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. 422-426. 10.18653/v1/P17-2067.