

Fake_News_Detection

Classification



Michalina Hulak



Overview

The spread of fake news has become a major issue in today's world. False information spreads rapidly through social media and other online platforms, leading to widespread misinformation and confusion. Fake news can have serious implications, such as spreading misinformation, influencing public opinion, and even affecting election results. Therefore, it is important to have a reliable and accurate system for identifying fake news.






The aim

The aim of this project is to develop a machine learning model that can detect fake news.

To achieve this, I used a datasets of news articles, where each article is labeled as either real or fake. I explored various natural language processing (NLP) techniques to extract meaningful features from the text data, such as word embeddings, bag of words, SentenceTransformer.



Datasets

There are 3 datasets. All come from kaggle.com:

WELFake_Dataset



	title	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1	NaN	Did they post their votes for Hillary already?	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	SATAN 2: Russia unveils an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1

Fake & Real



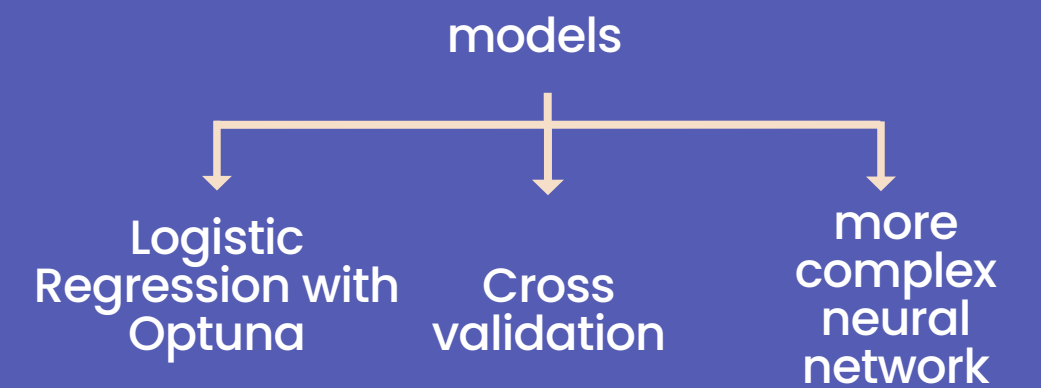
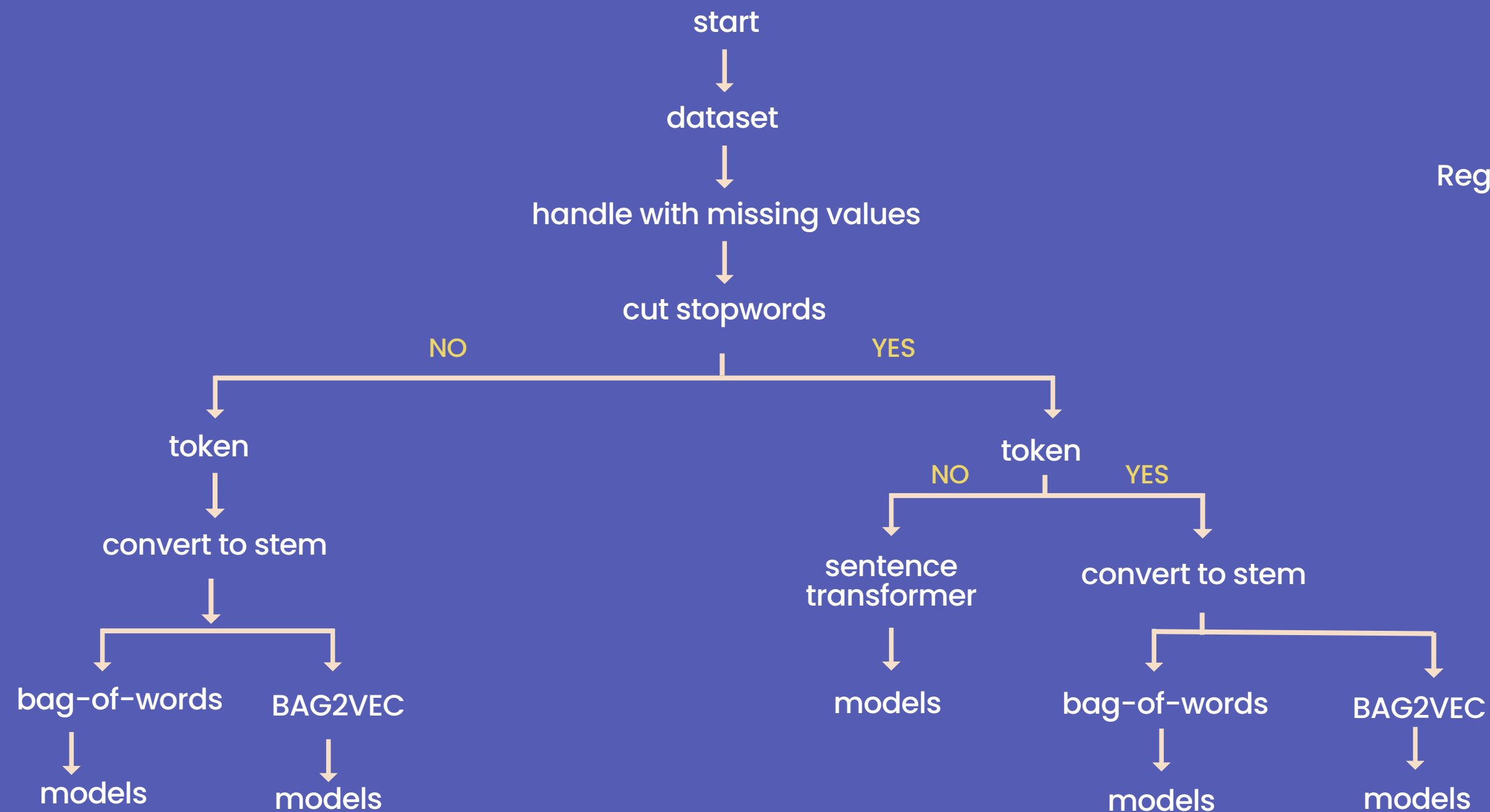
	title	text	subject	date	label
0	With Trump win, Republican chairman Priebus em...	NEW YORK (Reuters) - Donald Trump's White Hous...	politicsNews	November 9, 2016	1
1	Mattis signs orders to send additional troops ...	WASHINGTON (Reuters) - U.S. Defense Secretary ...	politicsNews	August 31, 2017	1
2	MEMBERS: EP #5 – DRIVE BY WIRE: 'Taxi to the U...	MEMBERS can join host Patrick Henningsen and ...	Middle-east	April 5, 2017	0

Fake news net



	title	news_url	source_domain	tweet_num	real
0	Kandi Burruss Explodes Over Rape Accusation on...	http://toofab.com/2017/05/08/real-housewives-a...	toofab.com	42	1
1	People's Choice Awards 2018: The best red carp...	https://www.today.com/style/see-people-s-choic...	www.today.com	0	1
2	Sophia Bush Sends Sweet Birthday Message to 'O...	https://www.etonline.com/news/220806_sophia_bu...	www.etonline.com	63	1
3	Colombian singer Maluma sparks rumours of inap...	https://www.dailymail.co.uk/news/article-33655...	www.dailymail.co.uk	20	1
4	Gossip Girl 10 Years Later: How Upper East Sid...	https://www.zerchoo.com/entertainment/gossip-g...	www.zerchoo.com	38	1

The algorithm



5 roots * 3 models = 15models
5 * 3 models * 3 datasets = 45 models



Results

1. For WELFake_Dataset dataset

Models – 1st root
* cut stopwords: no
* tokenize: yes
* convert to stem: yes
* bag-of-words

Root	model	train accuracy	val accuracy	test accuracy
1st	Logistic Regression with Optuna	1	0.9114	0.9097
	Cross validation	0.9333	0.9673	0.967
	More complex NN	0.995	0.9733	0.9792
2nd	Logistic Regression with Optuna	0.5	0.4933	0.4992
	Cross validation	1	0.9119	0.9117
	More complex NN	0.9479	0.9448	0.9384
3rd	Logistic Regression with Optuna	1	0.767	0.7677
	Cross validation	0.85	0.8538	0.8547
	More complex NN	0.8896	0.8686	0.8751
4th	Logistic Regression with Optuna	0.5	0.9196	0.9254
	Cross validation	1	0.9586	0.9603
	More complex NN	0.9972	0.9671	0.9685
5th	Logistic Regression with Optuna	1	0.8171	0.8223
	Cross validation	0.8	0.8539	0.8783
	More complex NN	0.9183	0.9121	0.9174

The model from 1st root, which uses more complex NN, achieves the highest test accuracy of 0.9792, followed closely by the 5th root, model, which also uses more complex NN, with a test accuracy of 0.9174. Both of these models have perfect or near-perfect train and val accuracy, indicating that they are not overfitting to the training data. Therefore, based on the test set performance, we would choose the first root model, which uses more complex NN and achieves the highest test accuracy of 0.9792.

Results

2. For Fake & Real dataset

Models – 1st root
* cut stopwords: no
* tokenize: yes
* convert to stem: yes
* bag-of-words

Root	model	train accuracy	val accuracy	test accuracy
1st	Logistic Regression with Optuna	1	0.9336	0.9367
	Cross validation	1	0.9957	0.9957
	More complex NN	0.9996	0.9952	0.996
2nd	Logistic Regression with Optuna	0.75	0.5254	0.5294
	Cross validation	0.9778	0.9791	0.9661
	More complex NN	0.9909	0.9896	0.9889
3rd	Logistic Regression with Optuna	0.9167	0.8962	0.886
	Cross validation	1	0.92	0.9194
	More complex NN	0.9409	0.9243	0.9323
4th	Logistic Regression with Optuna	1	0.9757	0.9784
	Cross validation	1	0.9956	0.9972
	More complex NN	0.9993	0.9949	0.9964
5th	Logistic Regression with Optuna	1	0.9533	0.9547
	Cross validation	0.8667	0.9502	0.957
	More complex NN	0.9795	0.9742	0.9788

Among the five root and models, the first and fourth root appear to be the best, having the highest accuracy scores across all three methods. Specifically, the model achieved perfect training accuracy and very high validation and test accuracy scores, indicating that it has effectively generalized to unseen data. The more complex neural network outperformed the other methods in terms of accuracy scores, with the logistic regression with Optuna and cross-validation showing lower performance in some cases. Overall, the results suggest that a more complex neural network could be a good choice for this dataset

Results

3. For Fake news net dataset

Models – 4th root
* cut stopwords: yes
* tokenize: yes
* convert to stem: yes
* bag-of-words

Root	model	train accuracy	val accuracy	test accuracy
1st	Logistic Regression with Optuna	0.6111	0.7531	0.74
	Cross validation	0.7485	0.7919	0.7966
	More complex NN	0.9446	0.8783	0.8819
2nd	Logistic Regression with Optuna	0.7222	0.7205	0.7147
	Cross validation	0.825	0.7184	0.7232
	More complex NN	0.8057	0.7989	0.8002
3rd	Logistic Regression with Optuna	0.9444	0.7276	0.7172
	Cross validation	0.6564	0.7385	0.7423
	More complex NN	0.8172	0.7757	0.7873
4th	Logistic Regression with Optuna	0.7222	0.7414	0.75
	Cross validation	0.9	0.7833	0.8115
	More complex NN	0.8485	0.793	0.8919
5th	Logistic Regression with Optuna	0.5	0.5019	0.5021
	Cross validation	0.675	0.7245	0.7154
	More complex NN	0.8101	0.8039	0.8085

After analyzing the results of the different models, we can see that the most accurate model in terms of test accuracy is the 4th one, which uses Logistic Regression with Optuna and has a test accuracy of 0.75. The more complex neural network in this case did not perform as well as the logistic regression model. It is also worth noting that the 1st model with the most complex neural network achieved the highest training accuracy, but did not perform as well on the validation and test sets. Therefore, based on the given results, we would choose the 4th model with Logistic Regression as the best one.

Conclusions

1. For 1. WELFake_Dataset and 2. Fake & Real dataset, the same more complex NN model was chosen, which does not require removing stopwords in data preparation and uses a bag-of-words approach.
2. For the Fake news net dataset, a logistic regression model with hyperparameters selected by Optuna was chosen. In this case, removing stopwords was beneficial, and bag-of-words was also used in preprocessing.
3. There is no one correct way to transform data and build a model that would be universal for every dataset. Each dataset should be treated separately as a unique problem.

