

DATS-6312 NLP for Data Science

Fake or Real News Classification

Prof. Amir Jafari

Individual Final Term Project

Pranay Bhakthula

1.Introduction:

The goal of our project is to classify real and fake news on the internet. As the scourge of “fake news” continues to plague our information environment, attention has turned toward devising automated solutions for detecting problematic online content. But, in order to build reliable algorithms for flagging “fake news,” we will need to go beyond broad definitions of the concept and identify distinguishing features that are specific enough for machine learning. We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

Do you trust all the news you consume from online media? Every news that we consume is not real. If you listen to fake news it means you are collecting the wrong information from the world which can affect society because a person’s views or thoughts can change after consuming fake news which the user perceives to be true. Since all the news we encounter in our day-to-day life is not authentic, how do we categorize if the news is fake or real?

A sort of sensationalist reporting, counterfeit news embodies bits of information that might be lies and is, for the most part, spread through web-based media and other online media. This is regularly done to further or force certain kinds of thoughts or for false promotion of products and is frequently accomplished with political plans. Such news things may contain bogus and additionally misrepresented cases and may wind up being virtualized by calculations, and clients may wind up in a channel bubble. Our project aims to train a model on fake and real news dataset, and predict.

2.Dataset Description:

There are two sorts of articles in the dataset: bogus and true news. The true articles were retrieved via crawling articles from Reuters.com; the dataset was compiled from real-world sources (News website). The phony news pieces were gathered from a variety of sources. The fake news items were gathered from untrustworthy websites that Politifact (a fact-checking organization in the United States) and Wikipedia had highlighted. The dataset contains a variety of articles on various themes, however the majority of the articles are about politics and world events. Two CSV files make up the dataset.

The first file, "True.csv," contains more than 12,600 reuter.com articles. The second file, "False.csv," has almost 12,600 items culled from various fake news sources. The following information is included in each article: the title, the text, the type, and the date the article was published. We focused on gathering articles from 2016 to 2017 to match the false news data acquired for kaggle.com. The information gathered was cleaned and processed, but the punctuation and errors found in the fake news were left in the text. The real news contains 21417 articles and we have two types in that one is world-news with article size of 10145 and the other is political news with 11272. The fake news contains 23481 articles and we have government-news with article size 1570, middle-east with size 778, US news with 783 size, left-news has 4459 size, politics has 6841, news has 9050 article size.

3.Data Preprocessing:

We performed the data cleaning steps to remove the unwanted words and simplify the text. Firstly, we removed the stopwords, followed by punctuations and lemmatizing the data. This helped us reduce the number of unique tokens in the

dataset, thus reducing the computational load. First we have cleared the urls and lowered the text and then removed contractions and punctuations and also made sure to remove special characters i.e., characters other than alphabets and now removed the stopwords at the end we applied stemming and lemmatization. The dataset was splitted to train, validation and test with a ratio of 0.8 :0.15:0.05.

4.Network and Models:

The Logistic Regression model and DeBERTa, RoBERTa and DistilBERT transformer models were chosen to be tested.

Logistic Regression: [Citation: **Authors: Alzen, J.L., Langdon, L.S. & Otero, V.K., <https://rdcu.be/cMzgW>**]

This sort of statistical analysis (also known as a logit model) is commonly used for predictive analytics and modeling, as well as machine learning applications. The dependent variable in this analytics approach is either finite or categorical: either A or B (binary regression) or a range of finite possibilities A, B, C, or D (multiple regression) (multinomial regression). By estimating probabilities using a logistic regression equation, it is employed in statistical software to comprehend the relationship between the dependent variable and one or more independent variables. This form of analysis can assist you in predicting the chances of an occurrence or a decision occurring.

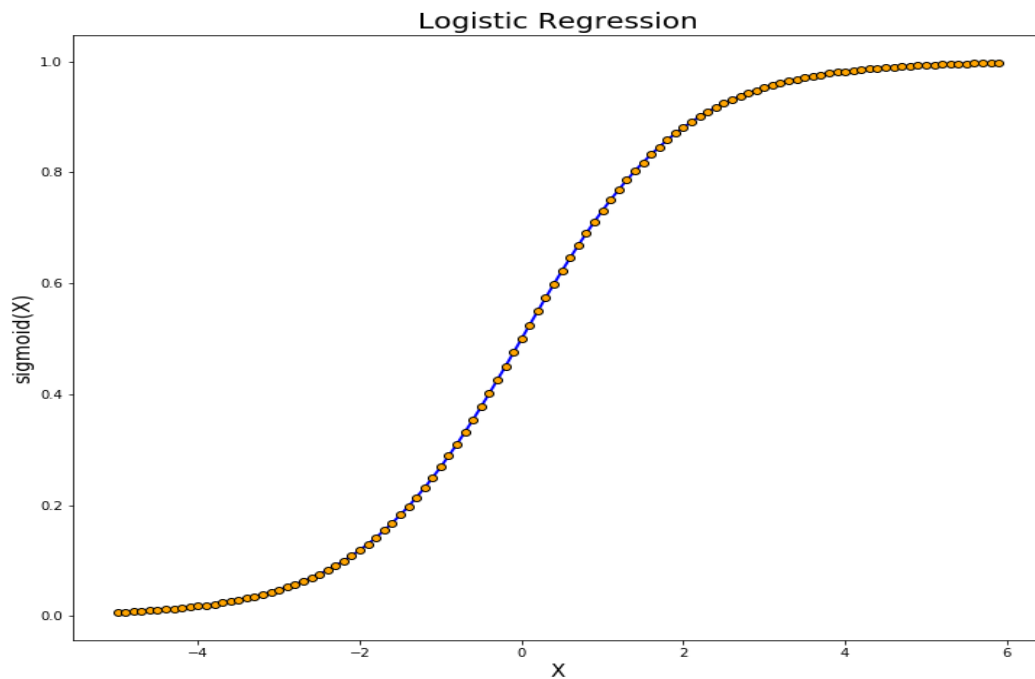


Figure1: Logistic regression

DistilBERT: [Citation: Authors: Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, [arXiv:1910.01108](https://arxiv.org/abs/1910.01108)]

DistilBERT is a technique for pre-training a smaller general-purpose language representation model that can then be fine-tuned to perform effectively on a range of tasks, similar to its larger counterparts. While most previous research focused on utilizing knowledge distillation to develop task-specific models, we can lower the size of a BERT model by 40% while preserving 97 percent of its language understanding skills and being 60% faster by employing knowledge distillation during the pre-training phase. Additionally, it provides a triple loss that combines language modeling, distillation, and cosine-distance losses to take use of the inductive biases learned by larger models throughout the training process.

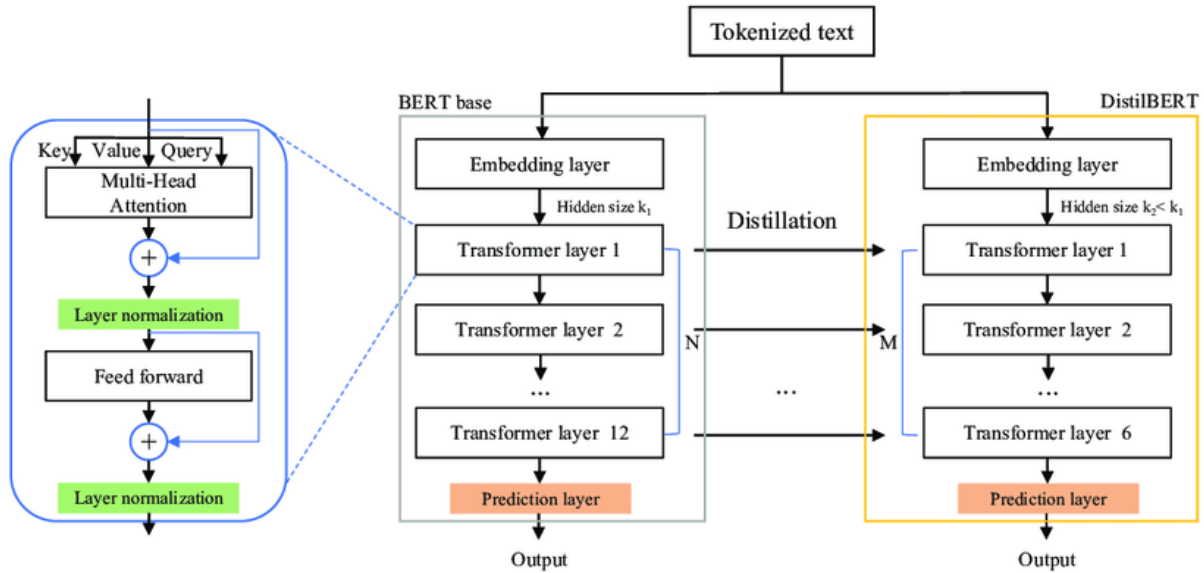


Figure 4: DistilBERT transformer

5. Model Metrics:

We looked at the following metrics in deciding how effective our models were:

F1-score:

F1-score weights precision and recall [blog.floydhub.com].

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Accuracy:

Accuracy is simply what percentage of observations were classified correctly [blog.floydhub.com].

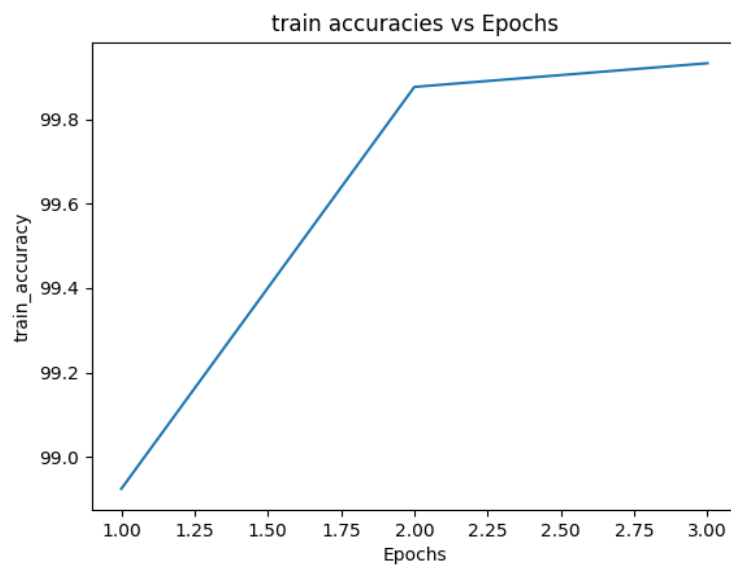
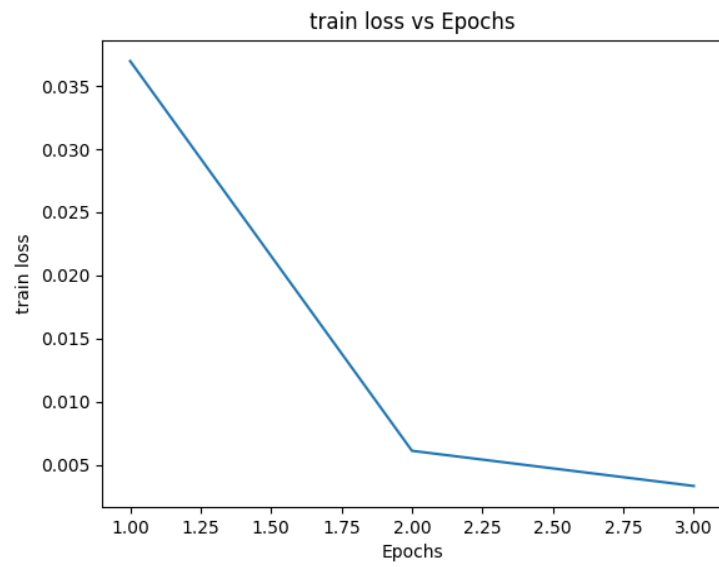
Non-Classical Model Results

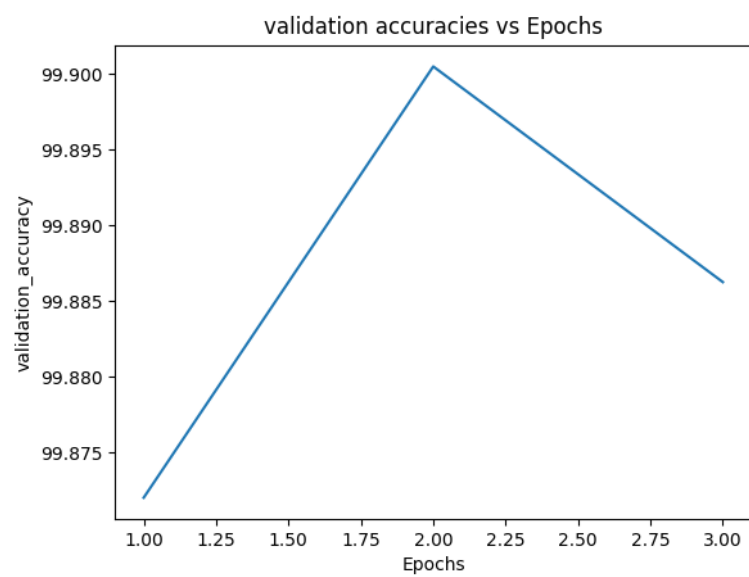
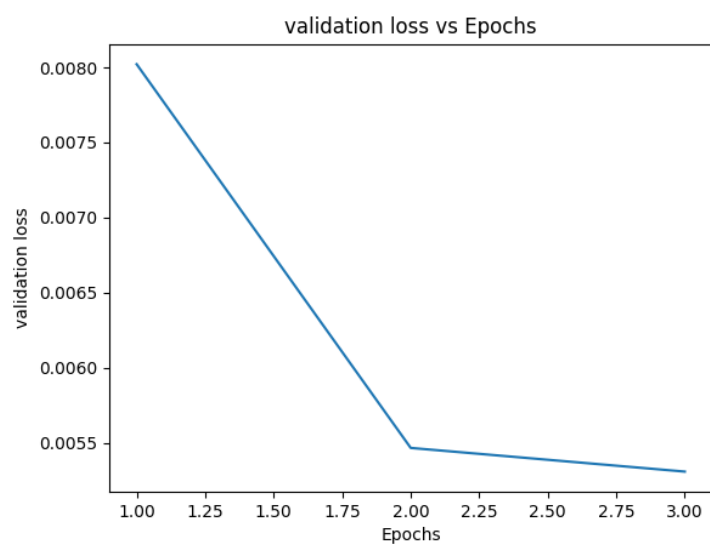
Model	Optimizer	Epoch	Batch Size	Max_Len	Learning Rate	Train Accuracy	Validation Accuracy
DistilBERT	AdamW	4	3	30	0.001	52.3	52.29
DistilBERT	AdamW	2	32	256	0.01	52.28	52.36
DistilBERT	AdamW	3	32	256	1.00E-05	99.93	99.89
Model	Precision		recall		f1-score		Total F1 Score
	0's	1's	0's	1's	0's	1's	
DistilBERT	0.52	0	1	0	0.69	0	
DistilBERT							
DistilBERT	0.98	1	1	0.98	0.99	0.99	0.98951

Logistic Regression final model f1-score=0.9868544600938968

DistilBERT final model f1-score =0.989517819706499

DistilBERT final model





8. Conclusions:

By comparing the results of our experiments we found out that the non-classical models (DeBERTa, RoBERTa, DistilBERT) outperformed the classical model (Logistic Regression). We found that the RoBERTa model is the best model of the transformers with f1-score = 0.9916.

In the future work, we should try to improve the architecture to give the results faster. We can also use higher computing power so that we can experiment with more hyperparameters such as batch size, epochs, optimizers, learning rate and find the model with higher accuracy.

9. Percentage of code:

Percentage of my code = ~62%

Percentage of code taken from internet = ~38%

10. References:

<https://www.analyticsvidhya.com/blog/2021/07/detecting-fake-news-with-natural-language-processing/>

<https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/>