AI_PHASE 2 PROJECT

BY

RAKSHITHA S(513421106036)

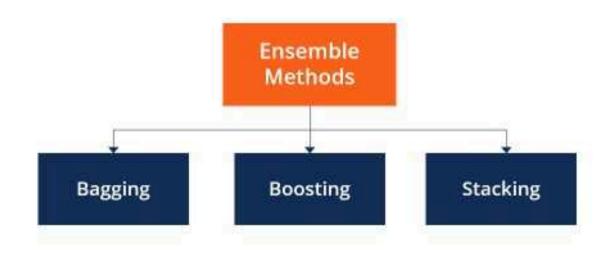
BE(ECE)

UNIVERSITY COLLEGE OF ENGINEERING OF KANCHIPURAM

ENSEMBLE METHODS

- Ensemble methods are machine learning techniques that combine multiple models or model instances to improve overall prediction accuracy and robustness.
- Instead of relying on a single model, ensemble methods leverage the outputs of multiple models to make more accurate predictions.
- Ensemble methods aim at improving predictability in models by combining several models to make one very reliable model. The most popular ensemble methods are boosting, bagging, and stacking.

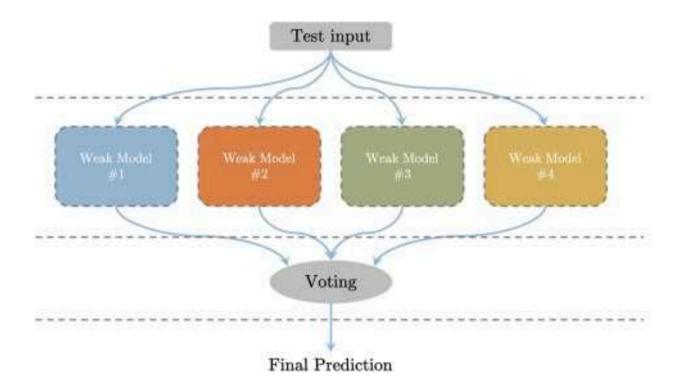
OVERVIEW ENSEMBLES METHODS



Ensemble methods in deep learning are used to improve the performance of neural networks and can take many forms including:

- Stacking: Training multiple deep learning models and utilizing the outputs of each model to train a
 "meta-model", a machine learning model that takes other models' outputs as inputs. The metamodel takes the base model predictions as inputs and learns how to best combine them to make the
 final prediction. This approach can enhance the model's predictive power and capture complex
 relationships in the data.
- Bagging: Training multiple instances of the same model on different subsets of data and combining the model outputs through averaging or voting. This approach can improve the model's generalizability.
- Model Averaging: Independently training multiple instances of the same deep learning model with different initializations (the initial values of the parameters or weights of a model before training), and averaging the model outputs to obtain a final prediction. This approach can reduce the impact of varying initializations among models and provide more stable predictions.
- Boosting, a very common ensemble method in classical machine learning is not prevalent in deep learning. Boosting entails combining weaker machine learning models, such as decision trees in classical machine learning, to create a single strong model. While there are some recent examples of boosting in deep learning, deep learning models are often capable of achieving high accuracy without the need for boosting.

FLOW CHART



DATASET:

Dataset Link: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

5.76+37 cmg	gative 1 gative 1 strong 0.6545 softline 0.6526 softline 0.6526 softline 0.6526 softline 0.6526 softline 1.5286 softline 1.5286 softline 1.5386	Had Flights Tan's Telf	8.7023 1	Virgin America	transfers Verentalytes (markling printing comings print (markling markling
5.76+27 resp 5.76+27 resp 5.76+17 pers 5.76+17 pers	gative 1 gative 1 gative 0.6437 gative 1 gative 0.6549 gat	Had Flights Tan's Telf	8.7023 1	Virgin America	warning for the first of the fi
5.76-37 (mg 5.76-37 (mg 5.76-17 pm 5.76-17 pm	gative 1 gative 1 softwe 0.6745 virts 0.654	Services	E NOAZ	Virgin America	provided to the provided to th
5.76+27 cong 5.76+27 cong 5.76+17 poor 5.76+17 cong 5.76+27 poor 5.76+27 poor 5.76+17 poor 5.76+	gative 1 patrixe 0.6792 patrixe 0.6539 patrixe 0.6539 patrixe 0.6539 patrixe 0.6539 patrixe 0.6539	Services	E NOAZ	Wingin America Wingin America Wingin America Wingin America Wingin America Wingin America Wingin America	maration comments of the policy of the polic
5.7%-27 reng b.75-17 pers 5.76-17 pers	igative 1. selfove 0.6545 vertal 0.6534 olifove 0.6534 olifove 1. selfove 1.	CAN'T TWE	e	Virgin America Virgin America Virgin America Virgin America Virgin America Virgin America	mandan spession prince
5.76-17 person 5.76-1			e	Virgin America Virgin America Virgin America Virgin America Virgin America Virgin America	prior prior (Prophesion Tupote Fate (St., Sept., years when
5.76-17 person 5.76-1			e	Virgin America Virgin America Virgin America Virgin America Virgin America Virgin America	prior prior (Prophesion Tupote Fate (St., Sept., years when
5.76-17 person 5.76-1			e	Virgin America Virgin America Virgin America Virgin America Virgin America Virgin America	prior prior (Prophesion Tupote Fate (St., Sept., years when
5.76 - 17 - max 5.76 - 17 - 100 5.76 - 17 - 100 5.76 - 17 - max 5.76 - 17 - max	0.634 0.639 0.639 0.633 0.633 0.633 0.633 0.633 0.634 0.63			Virgin America Virgin America Virgin America Virgin America Virgin America	prime officephonen Tupote Fato NR Jout you tube
5.76+17 poor 5.76+17 poor	O.655 0 O.655 0 D.655 0 D.655 0 D.655 0 D.655 0 D.655 0 D.655 0 D.665 0		e	Virgin America Virgin America Virgin America Virgin America	Theretores TrusteTates 1010_foot_yeoutselee
5.76-17 pos 5.76-17 cos 5.76-17 pos 5.76-17 pos 5.76-17 pos 5.76-17 cos 5.76-17 cos 5.76-17 pos 5.76-17 pos 5.76-17 pos 5.76-17 pos 5.76-17 pos 5.76-17 pos 5.76-17 pos	1 1 1 1 1 1 1 1 1 1		ø	Virgin America Virgin America Virgin America	TupiteTate nit_but_youtube
5.78-17 tons 5.76-17 pers 5.76-17 pers	outral U.9799 ottove 1 ottove 0.6451		0	Virgin America Virgin America	nin_but_youtube
5.76-17 pms 5.76-17 pms	rettive i rettive 0.6453			Virgin America	
5.7E+17 pm 6.7E+17 pm	eltive 0.6451			2 2 10 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	PhypierCamilian
B.76+17 pms 5.76+17 pms 5.76+17 oug b.76+17 pms b.76+17 pms 5.76+17 pms 5.76+17 pms 5.76+17 pms	0.645%			Deliconation of the con-	
5.7E+17 pm 5.7E+17 oug 5.7E+17 pm 5.7E+17 oug 5.7E+17 oug 5.7E+17 pm 5.7E+17 pm 5.7E+17 neg				Wirgin America	HYDAH Carmicae
5.7E+17 rang 5.7E+17 pros 5.7E+17 res 5.7E+17 pros 5.2E+17 pros 5.2E+17 resg	and the second s			Wright America	PRODUCTION CONTRACTOR
5.76+17 por 5.76+17 ceg 5.76+17 per 5.26+17 por 5.26+17 por 5.76+17 reg	neltitive 2			Virgin America	NUMBER OF STREET
5.76+17 ong 5.76+17 pm 5.26+27 pm 5.75+27 ong	getter 0.6842	Late Flight	0.3004	Virgin America	Arthur towarder meksak
5.76+17 pm 5.26+27 pm 5.75+17 mag	mittive 8			Wilden America	Intillers and delivery to the control of the contro
5.26+37 per 5.75+27 reg	gettee 3.	Bed Flant		Virgini America	Too office more sta-
5.75+27 ring	nitive: 1			Vergin America	follow for several triving
	distribution by			Virgin America	PML plence
	945798 0.670%	Can't full	10.093.5	Wirgin America.	- WATEGREE
S. PEXLY DOS				Virgin America	SET LES
S.PEKEF per	sittive 1			Wingto America	Elvinotiech.
S.PE+3.P ORG	utral L			Wingin America	cjiyach 210m.
5.7E+17 neg	gative 1	Customer	5.8557	Virgin America	a present distance
S. PRICET, NOR	gatter E	Eusteiner		Winglin America	barrena S.R.
5.7E+17 nos	Author E	Can't Tell	District	Wirgin America	improdutts/symm
S. PE-SE cinic	istral 0.6854			Wingin America	Adaminigar
5.7E+17 neg		Bart France	- 1	Virgin America	Believ Name Species 9-2-3.
5-PE+37 neu	THE RESERVE AND ADDRESS OF THE PARTY OF THE	STOWN VENUE	. 0	Vergin America	Terrorrent/spotwice.
5.7E - LT com		Flight Book	- 4	Wilder America	perdienalichier
B.TE+EF OWN		The second second		Virgin America	PERFERREDUE.
5.75+3.7 con		Eustomer	- 4	Virgin America	Encohoste L
5.7E+17. neg	The state of the s	Eustomer	3	Virgin America	amandulumerariy
5.70 x 3.7 puts		DESCRIPTION DE	3.7	Wirgin America	Assests Tabtures Tolore
B. Phys.F ries				Virgin America	retrigge continues.

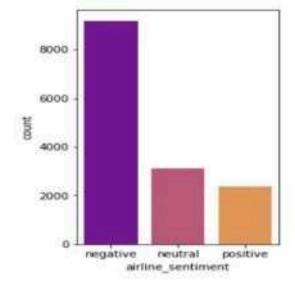
- tweets = pd.read_csv('Tweets.csv')
- · Let's look at features included in dataset:
- tweets.head()
- tweets.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
tweet id
                                14640 non-null int64
airline sentiment
                                14640 non-null object
airline sentiment confidence
                                14640 non-null float64
negativereason
                                9178 non-null object
negativereason confidence
                                10522 non-null float64
airline
                                14640 non-null object
airline sentiment gold
                                40 non-null object
                                14640 non-null object
name
negativereason gold
                                32 non-null object
retweet count
                                14640 non-null int64
text
                                14640 non-null object
                                1019 non-null object
tweet coord
tweet_created
                                14640 non-null object
tweet_location
                                9907 non-null object
user timezone
                                9820 non-null object
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

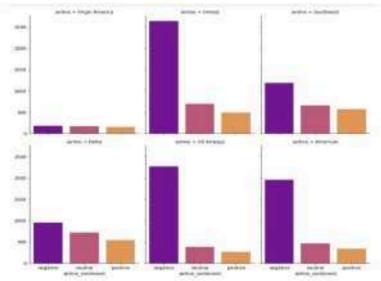
plt.figure(figsize=(3,5))
 sns.countplot(tweets['airline_sentiment'],

• order=tweets.airline_sentiment.value_counts().index,palette='plasma

plt.show()



 g = sns.FacetGrid(tweets, col="airline", col_wrap=3, height=5, aspect =0.7) g = g.map(sns.countplot, "airline_sentiment",order =tweets.airline_sentiment.value_counts().index, palette='plasma') plt.show()

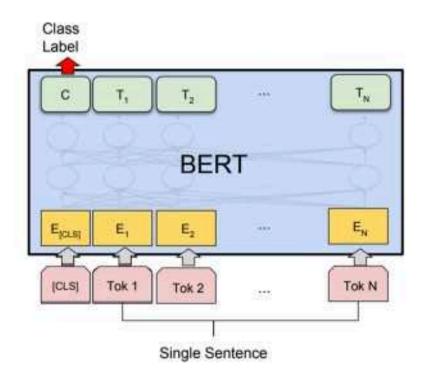


- To do sentiment analysis, we need to import a few libraries. Since this is a classification problem, I use LGBMClassifier.
- from lightgbm import LGBMClassifier
- We need to convert these tweets (texts) to a matrix of token counts.
- from sklearn.feature_extraction.text import CountVectorizer
- The next step is to normalize the count matrix using tf-idf representation.
- from sklearn.feature_extraction.text import TfidfTransformer
- I used the pipeline function to do all steps together.

Bidirectional Representation for Transformers (BERT)

- BERT is a powerful technique for natural language processing that can improve how well computers comprehend human language. The foundation of BERT is the idea of exploiting bidirectional context to acquire complex and insightful word and phrase representations. By simultaneously examining both sides of a word's context, BERT can capture a word's whole meaning in its context, in contrast to earlier models that only considered the left or right context of a word.
- This enables BERT to deal with ambiguous and complex linguistic phenomena including polysemy, co-reference, and long-distance relationships.
- For that, the paper also proposed the architecture of different tasks. In this
 post, we will be using BERT architecture for Sentiment classification tasks
 specifically the architecture used for the CoLA (Corpus of Linguistic
 Acceptability) binary classification task.

SINGLE SENTENCE CLASSIFICATION TASK



Step 1: Import the necessary libraries

- · PROGRAM:
- import os
- · import shutil
- · import tarfile
- · import tensorflow as tf
- · from transformers import BertTokenizer, TFBertForSequenceClassification
- · import pandas as pd
- · from bs4 import BeautifulSoup
- · import re
- · import matplotlib.pyplot as plt
- · import plotly.express as px
- · import plotly.offline as pyo
- · import plotly.graph_objects as go
- · from wordcloud import WordCloud, STOPWORDS
- · from sklearn.model_selection import train_test_split
- · from sklearn.metrics import classification_report

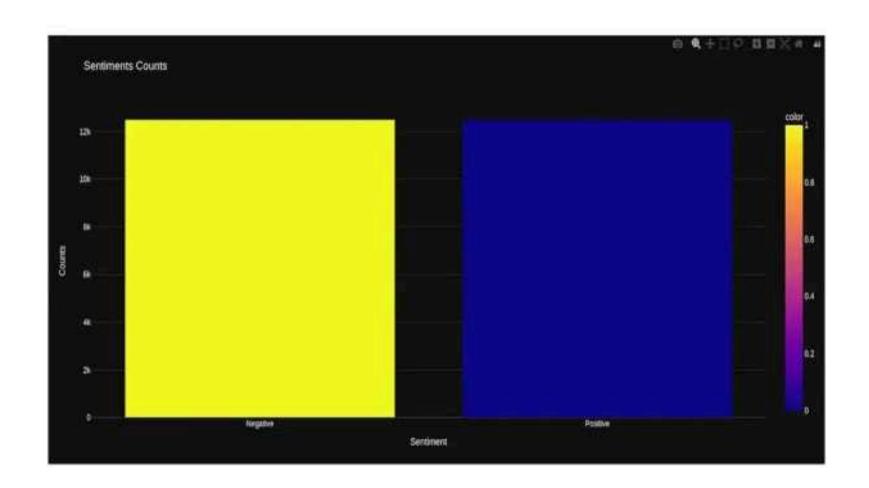
Step 2: Load the dataset

- # Get the current working directory
- current_folder = os.getcwd()
- dataset = tf.keras.utils.get_file(
- fname ="aclImdb.tar.gz",
- origin
 ="http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz",
- cache dir=current folder,
- extract = True)

Output:

• ['aclImdb.tar.gz', 'aclImdb']

Step 3: Preprocessing



CONCLUSION:

Sentiment analysis deals with the classification of texts based on the sentiments they contain. This article focuses on a typical sentiment analysis model consisting of three core steps, namely datapreparation, review analysis and sentiment classification, and describes representative techniques involved in those steps.