Sentiment Analysis and Topic Detection for Fake News Classifier

Team Members:

Foo Chuan Geng
Loh Xiao Binn
Toh Ying Hui
Moh Qing Loong Darren

TABLE OF CONTENTS

Project Details (an overview)

01

PROJECT TITLE

Sentiment Analysis and Topic Detection for Fake News Classifier

02

OUR PROCESS

Sentiment Analysis & Topic Classifier
-> Fake News Classifier

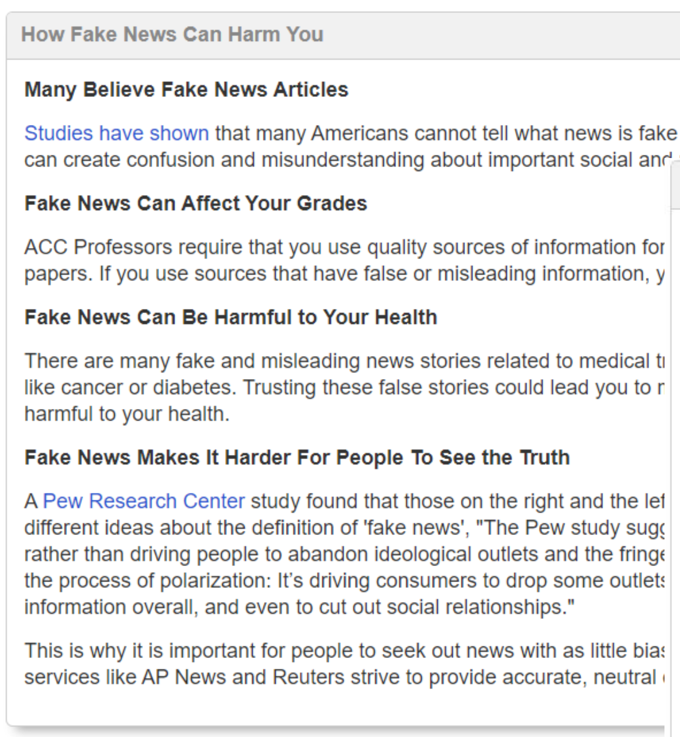
03

TARGET

Identifying fake news & Understanding the trends in Fake News

Problem statement

The plethora of misleading false information in this time where data is highly accessible causes confusion and frustrations to both individuals and businesses.



How Fake News Can Harm You

Many Believe Fake News Articles

Studies have shown that many Americans cannot tell what news is fake and what news is real. This can create confusion and misunderstanding about important social and political issues

Fake News Can Affect Your Grades

ACC Professors require that you use quality sources of information for papers. If you use sources that have false or misleading information, y

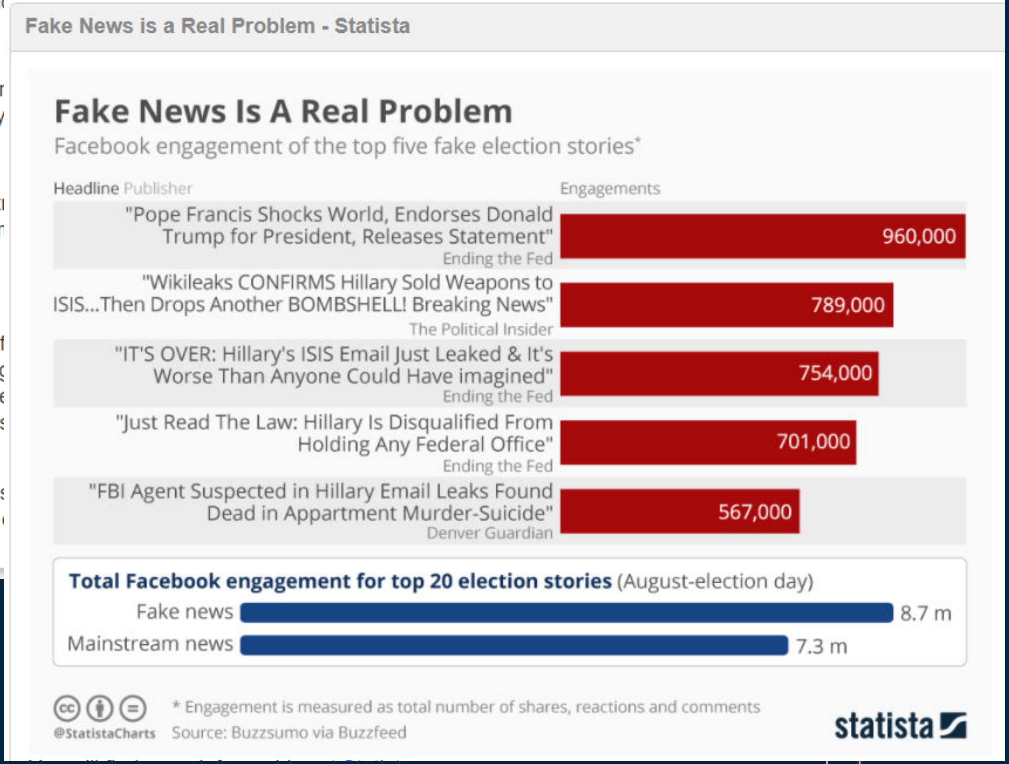
Fake News Can Be Harmful to Your Health

There are many fake and misleading news stories related to medical t like cancer or diabetes. Trusting these false stories could lead you to r harmful to your health.

Fake News Makes It Harder For People To See the Truth

A Pew Research Center study found that those on the right and the lef different ideas about the definition of 'fake news', "The Pew study sugg rather than driving people to abandon ideological outlets and the fringe the process of polarization: It's driving consumers to drop some outlets information overall, and even to cut out social relationships."

This is why it is important for people to seek out news with as little bias services like AP News and Reuters strive to provide accurate, neutral



Fake News is a Real Problem - Statista

Fake News Is A Real Problem

Facebook engagement of the top five fake election stories*

* Engagement is measured as total number of shares, reactions and comments

Source: Buzzsumo via Buzzfeed

Motivation – UNDERSTANDING THE PROBLEM

Fake News

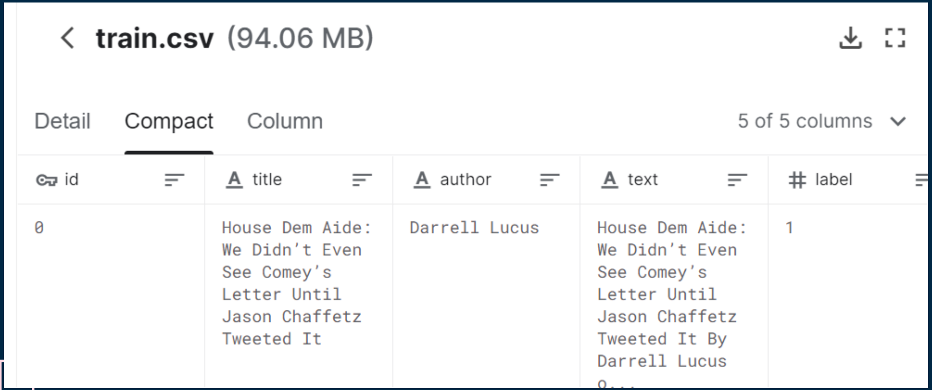
We want to **identify fake news** from a large dataset and **understand trends in fake news**.

Real News

We must learn to identify true information and it will be easier if we understood the trends in false news. For example, a hypothesis could be that political news is more likely to be false.

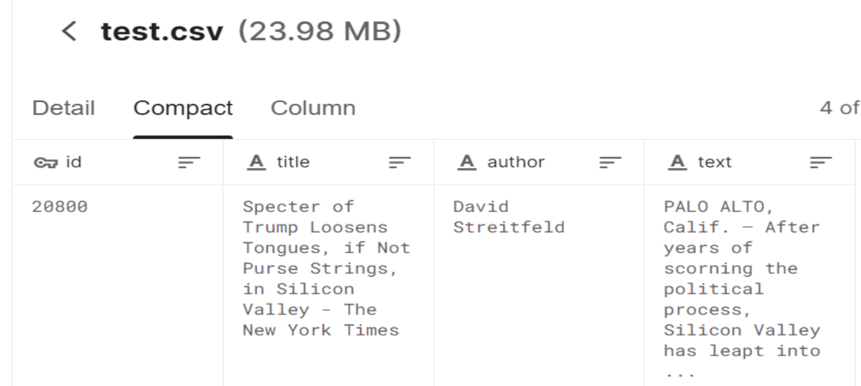
Datasets

- Obtained from Kaggle (https://www.kaggle.com/c/fake-news/data).
- Contains a total of 20,800 records and 5 variables: id', 'title', 'author', 'text' and 'label'.
- For test data, there are a total of 5,200 rows.



train.csv (94.06 MB)

id	title	author	text	label
0	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It By Darrell Lucus o...	1



test.csv (23.98 MB)

id	title	author	text
20800	Specter of Trump Loosens Tongues, if Not Purse Strings, in Silicon Valley - The New York Times	David Streitfeld	PALO ALTO, Calif. — After years of scorning the political process, Silicon Valley has leapt into ...

Training & Test dataset (fake news classifier)

Field Name	Data Type	Description
id	int64	Unique ID for a news article
title	object	The title of a news article
author	object	Author of a news article
text	object	The text content of the article (may be incomplete)
topic	string	Category of news
sentiment	integer	Neutrality from title
label	int64	A label that marks the article as potentially unreliable •1: unreliable •0: reliable

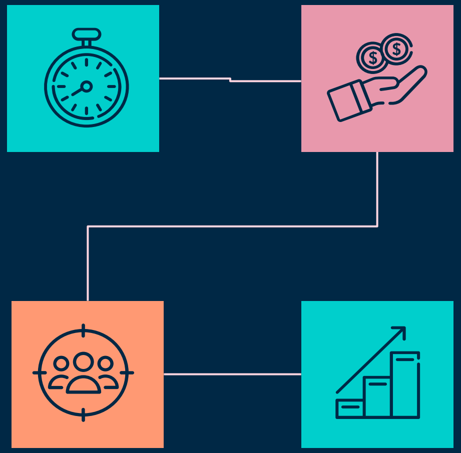
From topic classifier model and sentiment classifier model →

* Test data is not labelled →

Literature review

Fake News Classifiers (Supervised)	Sentiment Analysis	Topic Detection (Unsupervised)
KNN(K-Nearest Neighbour)	SVM(Support Vector Machine)	SVM, Personalized
SVM(Support vector Machine)	KNN(K-Nearest Neighbour)	Elasticsearch (based on business rules, not ML)
Random Forest	Rule based lexicon (NRC, SentiWordNet)	Logistic Regression
Naïve bayes	Random Forest Classifiers	CatBoost with default VoW embeddings
GWO(Grey Wolf Optimisation), SSO(Salp Swa rm Optimisation)	Ensemble of label powerset classifiers	CatBoost on TF-IDF features
Logistics regression	Vader, Textblob	Naïve Bayes , Hidden Markov Model

METHODOLOGY



Sentiment Analysis

Positive, Negative, Neutral

Topic Modelling

e.g. Politics, Sports, Health, Crime, etc.

Fake News Classifier

Binary (0,1)

Analysis

Trends and characteristics of fake news

We will be using the same dataset for all 3 models & the output of the sentiment analysis and topic classifier will be used to train the fake news classifier model.

Tools and Resources:

General

NLTK, Pandas, numpy, sckit-learn

Sentiment analysis [tools]

Textblob & NLTK-Vader package

Topic modelling [unsupervised]

LDA [from Gensim], K-means, Density based [DBSCAN]

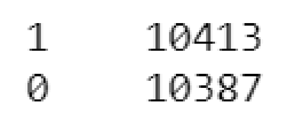
Fake news Classifier

Naïve Bayes, SVM, Logistic Regression, Random forest

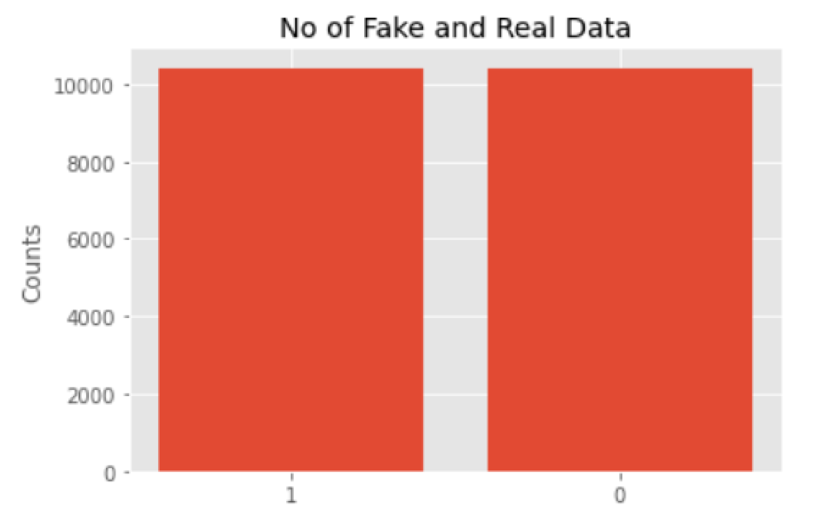
Metrics

Cross Validation [K-folds], Confusion Matrix

Preliminary results of Baseline model (EDA and Text Processing)



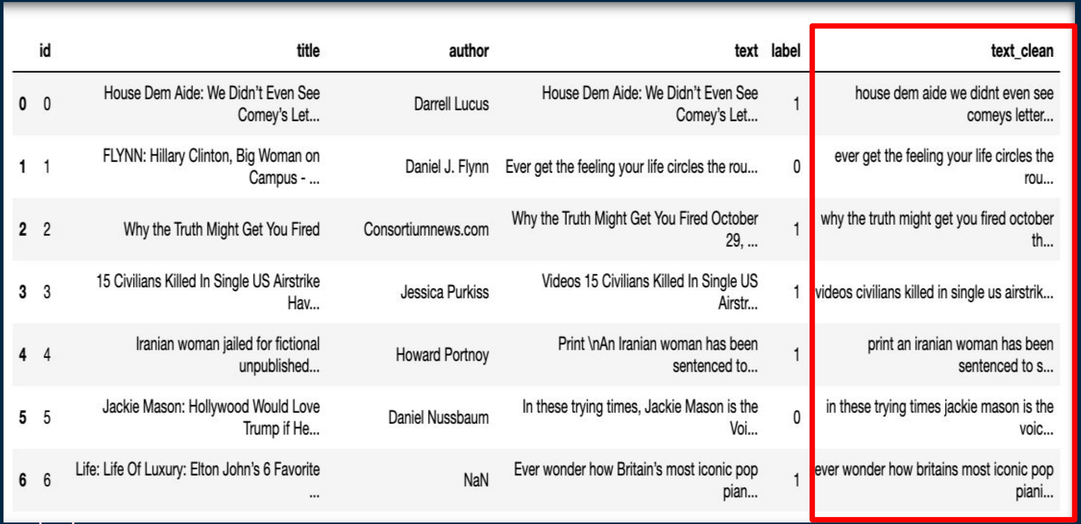
```
1      10413
0      10387
```



No of Fake and Real dataset labels

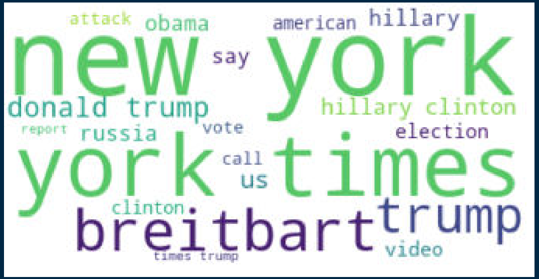
Preliminary results of Baseline model (EDA and Text Processing)

Performed text processing to remove stopwords, tokenize the words and stemmer to prepare as input to the baseline models

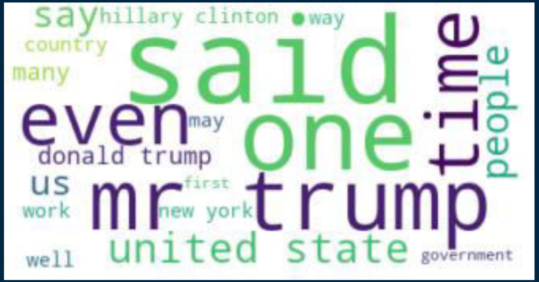


	id	title	author	text	label	text_clean
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1	house dem aide we didnt even see comeys letter...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	ever get the feeling your life circles the rou...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	why the truth might get you fired october th...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1	videos civilians killed in single us airstrik...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1	print an iranian woman has been sentenced to s...
5	5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Mason is the Voi...	0	in these trying times jackie mason is the voic...
6	6	Life: Life Of Luxury: Elton John's 6 Favorite ...	NaN	Ever wonder how Britain's most iconic pop pian...	1	ever wonder how britains most iconic pop piani...

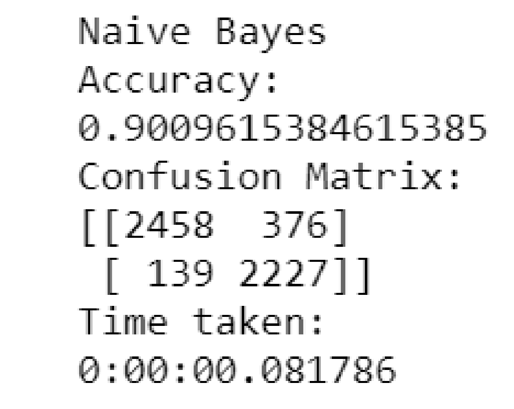
Top 20 most common words in title



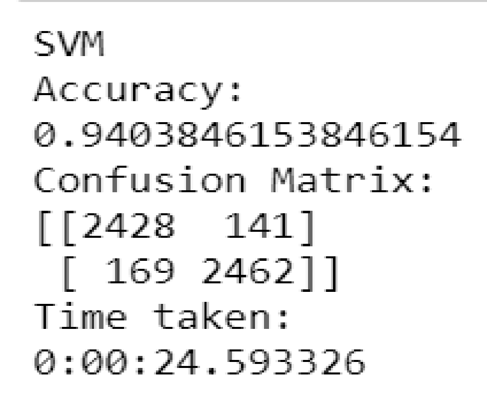
Top 20 most common words in text



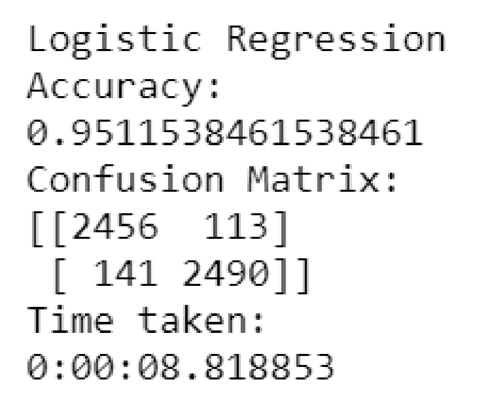
Preliminary results and evaluation (Baseline Model)



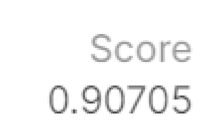
```
Naive Bayes
Accuracy:
0.9009615384615385
Confusion Matrix:
[[2458  376]
 [ 139 2227]]
Time taken:
0:00:00.081786
```



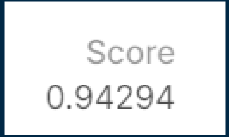
```
SVM
Accuracy:
0.9403846153846154
Confusion Matrix:
[[2428  141]
 [ 169 2462]]
Time taken:
0:00:24.593326
```



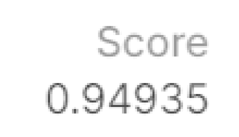
```
Logistic Regression
Accuracy:
0.9511538461538461
Confusion Matrix:
[[2456  113]
 [ 141 2490]]
Time taken:
0:00:08.818853
```



Score
0.90705



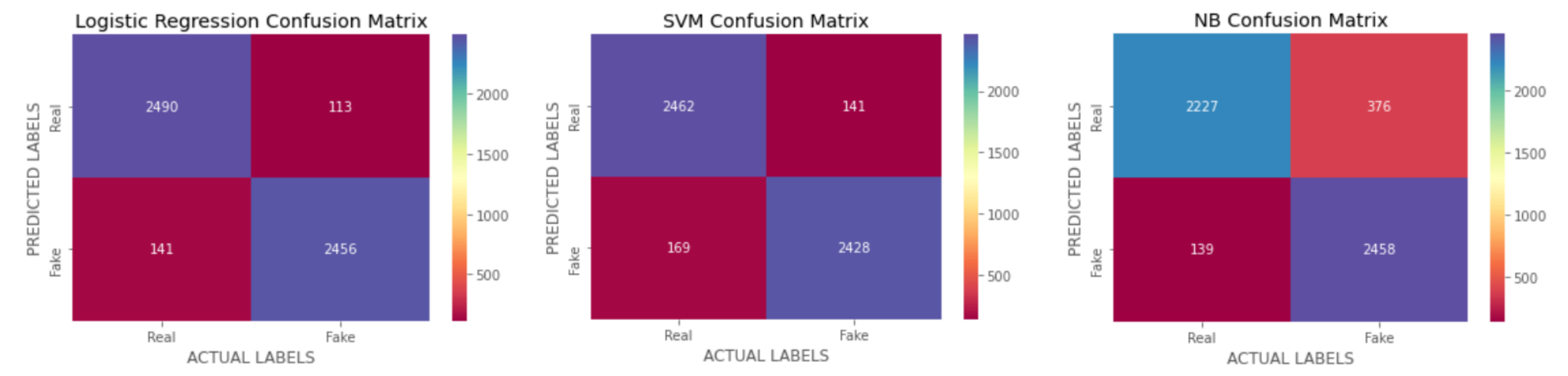
Score
0.94294



Score
0.94935

Respective output for preliminary models for **Fake news classification** before adding topic/sentiment analysis in increasing accuracy rate (from left to right).
Logistic regression gives the best accuracy of 95.1

Preliminary results and evaluation (Baseline Model)



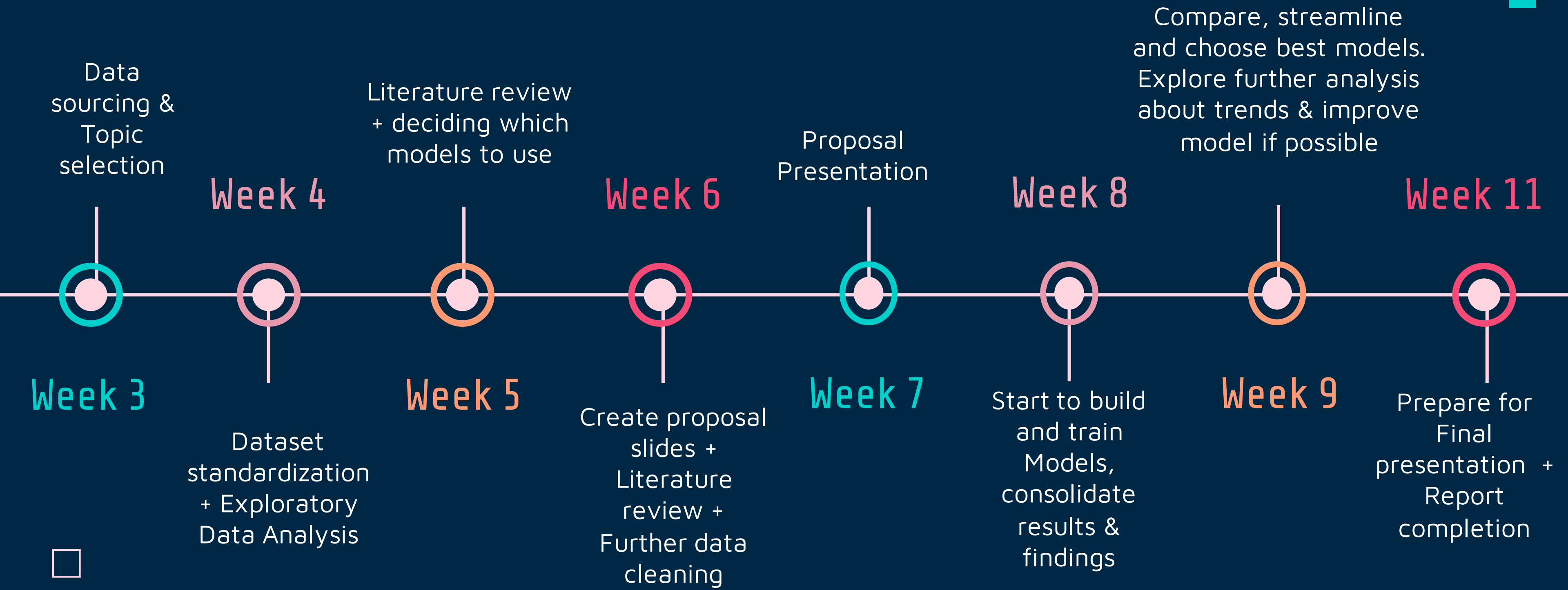
Logistic Regression Confusion Matrix

SVM Confusion Matrix

NB Confusion Matrix

https://github.com/chuangeng555/fakenewseda/blob/master/fake_news.ipynb

Project Milestones / Tentative Schedule



Week 3
Data sourcing & Topic selection

Week 4
Dataset standardization + Exploratory Data Analysis

Week 5
Literature review + deciding which models to use

Week 6
Create proposal slides + Literature review + Further data cleaning

Week 7
Proposal Presentation

Week 8
Start to build and train Models, consolidate results & findings

Week 9
Compare, streamline and choose best models. Explore further analysis about trends & improve model if possible

Week 11
Prepare for Final presentation + Report completion

References – APA (continued)

Ozbay, F., & Alatas, B. (n.d.). A Novel Approach for Detection of Fake News on Social Media Using Metaheuristic Optimization Algorithms. Retrieved September 27, 2020, from https://eejournal.ktu.lt/index.php/elt/article/view/23972

Austin Community College, Library Services. (2020, September 24). Fake News and Alternative Facts: Finding Accurate News. Retrieved September 27, 2020, from https://researchguides.austincc.edu/c.php?g=612891

Mohammed, T. J. (2019, May 12). NLU: Topic Discovery [Web log post]. Retrieved September 25, 2020, from https://medium.com/@b.terryjack/nlu-topic-discovery-85b492c4beb7

Thushan, G. (2018, August 23). Intuitive Guide to Latent Dirichlet Allocation [Web log post]. Retrieved September 25, 2020, from https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158

White, B. (2020, May 27). Sentiment Analysis: VADER or TextBlob? [Web log post]. Retrieved September 25, 2020, from https://towardsdatascience.com/sentiment-analysis-vader-or-textblob-ff25514ac540

Koch, K. (2020, March 26). A Friendly Introduction to Text Clustering [Web log post]. Retrieved September 28, 2020, from https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04

Foley, D. (2018, October 21). Building an ETL Pipeline in Python [Web log post]. Retrieved September 25, 2020, from https://towardsdatascience.com/building-an-etl-pipeline-in-python-f96845089635

Garg, H., Goyal, A., & Joshi, A. (n.d.). Techniques for fake news detection. Retrieved from http://ijcmes.com/upload_file/issue_files/2IJCMES-APR-2020-1-Techniques.pdf

References – APA (continued)

Pandey, P. (2018, September 23). Simplifying Sentiment Analysis using VADER in Python (on Social Media Text) [Web log post]. Retrieved September 24, 2020, from https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

Genç, Ö. (2019, April 16). The basics of NLP and real time sentiment analysis with open source tools [Web log post]. Retrieved September 23, 2020, from https://towardsdatascience.com/real-time-sentiment-analysis-on-social-media-with-open-source-tools-f864ca239afe

Foley, D. (2019, February 8). K-Means Clustering [Web log post]. Retrieved September 24, 2020, from https://towardsdatascience.com/k-means-clustering-8e1e64c1561c

Curcuma_, DhruvPathak, Adonis, &amp; Alec_djinn. (2017, June 28). Is there a way to improve performance of nltk.sentiment.vader Sentiment analyser? [Web log review]. Retrieved September 28, 2020, from https://stackoverflow.com/questions/45296897/is-there-a-way-to-improve-performance-of-nltk-sentiment-vader-sentiment-analyser

Hutto, C.J.; Eric, G. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (pp. 1-10, Rep.). Atlanta: Association for the Advancement of Artificial Intelligence. Retrieved September 28, 2020, from http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Yadollahi, A., Shahraki, A., & Zaiane, O. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys (CSUR)*, *50*(2), 1–33. https://doi.org/10.1145/3057270

Przybyla, M. (2020, May 15). Developing a Data Science Model to Predict Fake News. Retrieved September 28, 2020, from https://towardsdatascience.com/developing-a-data-science-model-to-predict-fake-news-184c25a13cb8

Q & A