


Fake News Detection using Machine Learning and Natural Language Processing



Group Number : 5
Palak Tiwari (MT20103)
Deepankar Kansal (MT20007)
Vineet Maheshwari (MT20020)

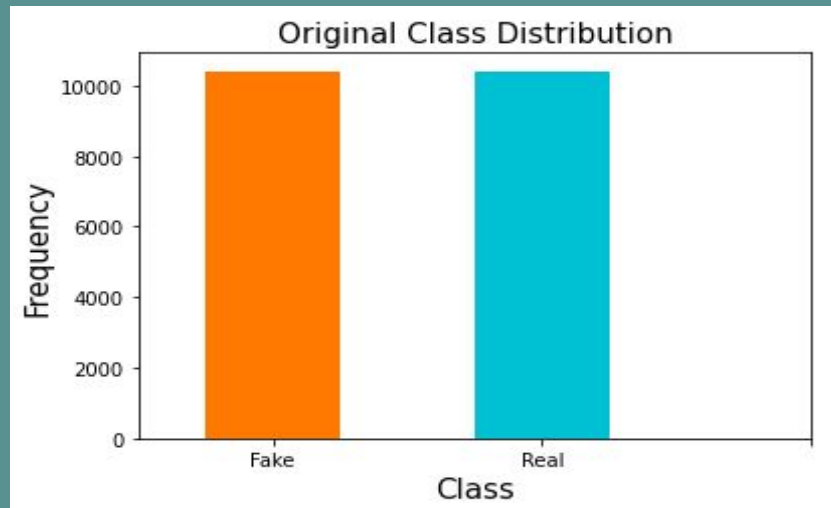


A word cloud centered around the phrase "Fake News". The words are arranged in a circular pattern, with "Fake News" being the largest and most prominent. Other large words include "propaganda", "hoax", "fraud", "bait", "censoring", "fact", "restriction", "channel", "misleading", "television", "business", "headline", "fiction", "corrupt", "theory", "broadcast", "online", "commercial", "communicate". Smaller words include "lie", "banner", "internet", "advertisement", "journalism", "report", "cover", "global", "perception", "story", "media", "manipulation", "checking", "false", "world", "press", "newspaper", "corruption", "technology", "truth", "communication", "reporter", "network", "restrictions", "control", "censorship", "disinformation", "communication", "reporter", "network", "restrictions", "channel", "misleading", "television", "business", "headline", "fiction", "corrupt", "theory", "broadcast", "online", "commercial", "communicate".

Dataset Description

id		title	author	text	label
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
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Aistr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1
5	5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Mason is the Voi...	0
6	6	Life: Life Of Luxury: Elton John's 6 Favorite ...	NaN	Ever wonder how Britain's most iconic pop pian...	1
7	7	Benoît Hamon Wins French Socialist Party's Pre...	Alissa J. Rubin	PARIS — France chose an idealistic, traditi...	0
8	8	Excerpts From a Draft Script for Donald Trump'...	NaN	Donald J. Trump is scheduled to make a highly ...	0
9	9	A Back-Channel Plan for Ukraine and Russia, Co...	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as nat...	0

Unbiased Dataset



As we can see here, the dataset used is Unbiased and therefore it will generate satisfactory results

Main Libraries Used Are:

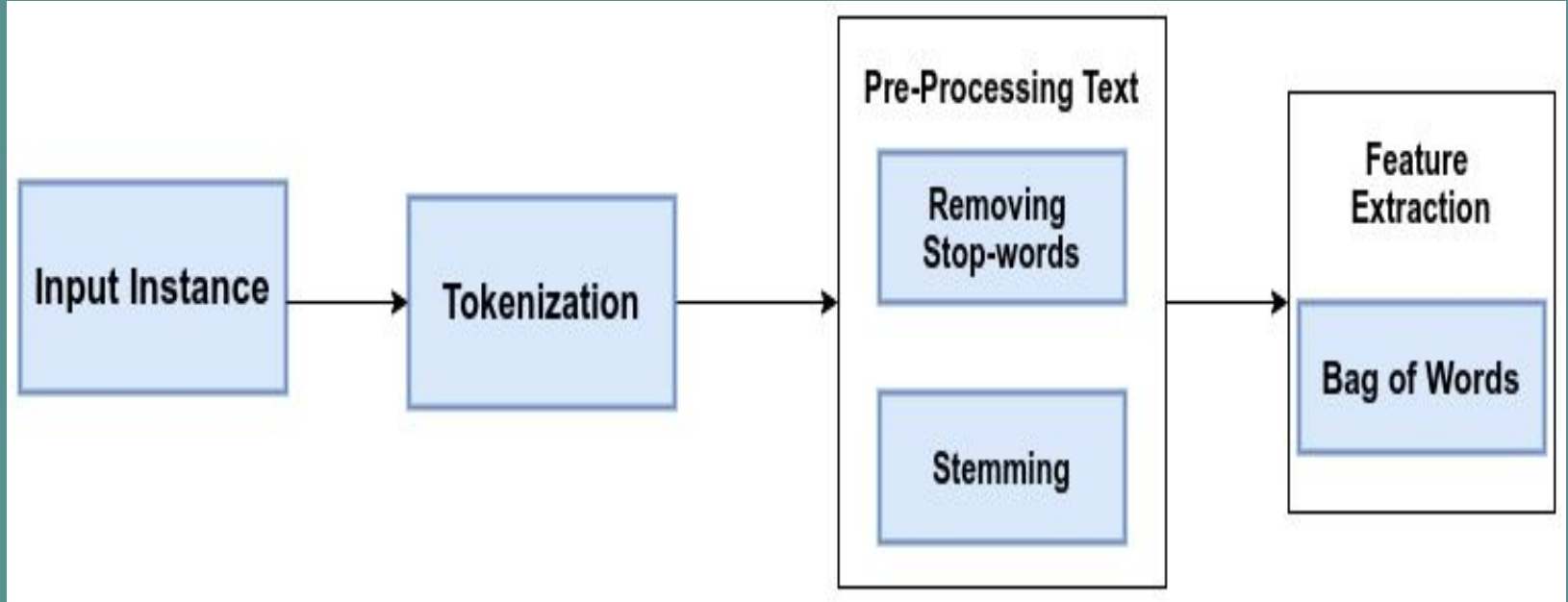
For text preprocessing and feature extraction:

1. nltk (Natural Language Toolkit):- Like Corpus, PorterStemmer
2. re (Regular Expression)
3. CountVectorizer class from SciKit-learn

Classification Models Applied: From SciKit-learn

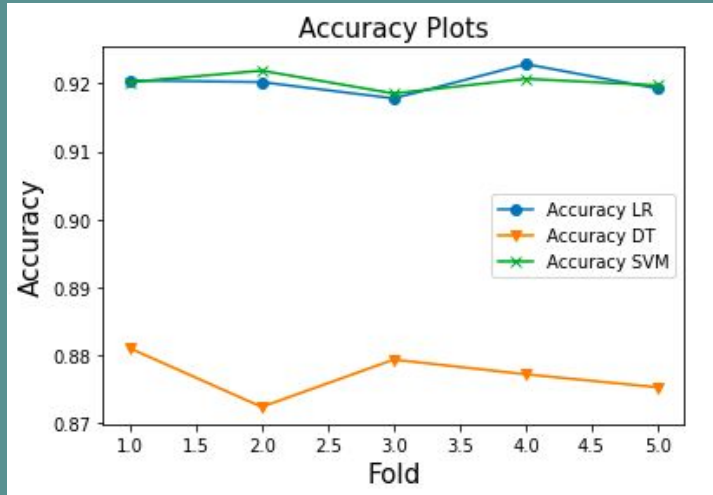
Logistic Regression (LR), Decision Tree (DT), Support Vector Machines (SVM)

Text Preprocessing

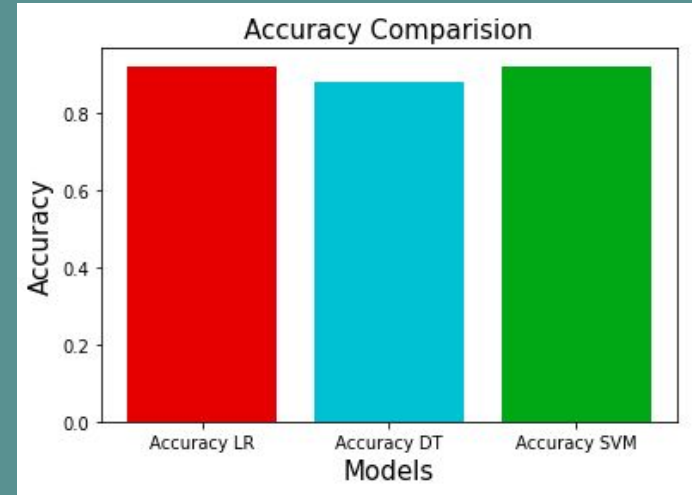


Using libraries nltk, re, and CountVectorizer (class)

Evaluating Models



Accuracies plot at different folds



Accuracies comparison of three classification models

Future Work

- Implementing TF-IDF on bi-grams and n-grams as one of the feature extraction method
- Training data on various Machine learning models/algorithms like Naive Bayes, kNN etc.
- Applying various ensembling machine learning techniques like Bagging and Boosting
- Thinking of implementing some deep learning techniques for improving the models performance using word embedding.
- Evaluating model using the headline, body and author of the article and flagging it as fake or real.