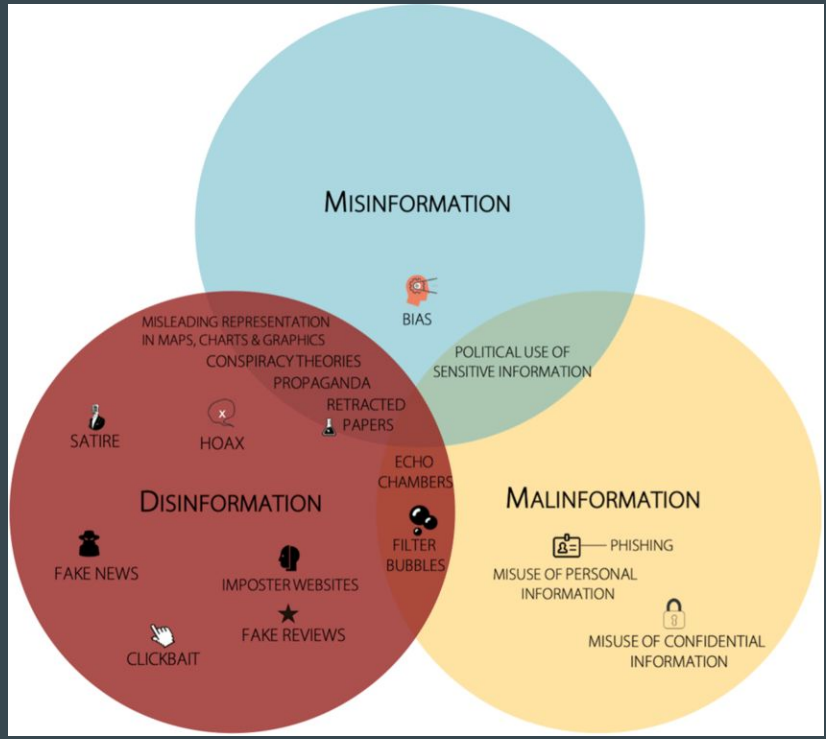
Fake News Detection

Keerthi Gogineni, Rishika Juloori, Ayaan Shaik, Nikita Gupta

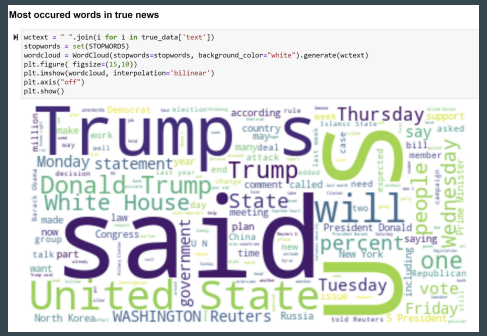
Background and Motivation

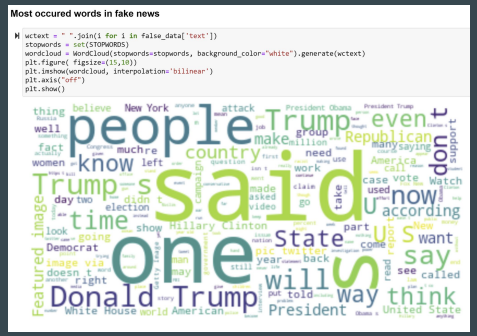
- The dependency on technology has led to the popularization of machine learning and AI
- As digital news becomes popular, it is clear that people are starting to rely more on it.
- The growth of information online is exponential, and it is becoming increasingly difficult to decipher what is true or not.
- Fake news can become indistinguishable from accurate reporting since it spreads so fast.
- Growing up in the technological era, were motivated to tackle this issue as it is a common problem we face in our daily life.

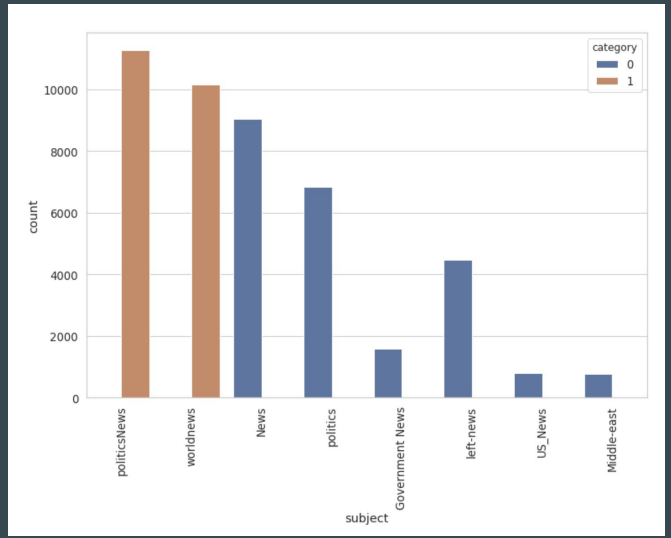


Initial visualizations

We created few visualizations hoping they might help us understand the data better. We used wordcloud and matplotlib. This helped us learn more about which subject was most prominent in fake and real new and other features that distinguish real news from fake news.







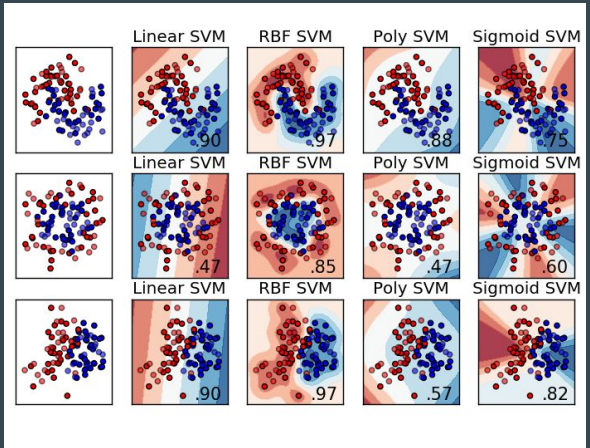
Data

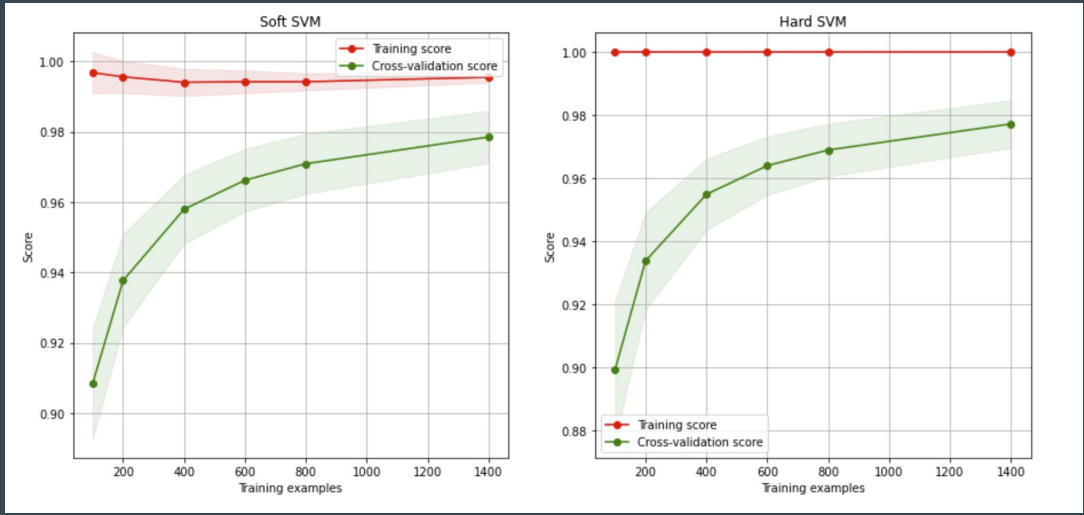
- The data was obtained from Kaggle as two different csv files.
- The data contained the subject, title, date and text related to the article. We added another column (binary), signifying if the article is true or fake. We then combined the data into one set
- Cleaned the data (removing unnecessary columns, words)
- Changed the data into vector and encoder to use it in models (TfidfVectorizer and LabelEncoder)

	title	text	subject	date	fake
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	0
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	0
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	1
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	1
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	1
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	1

SVM

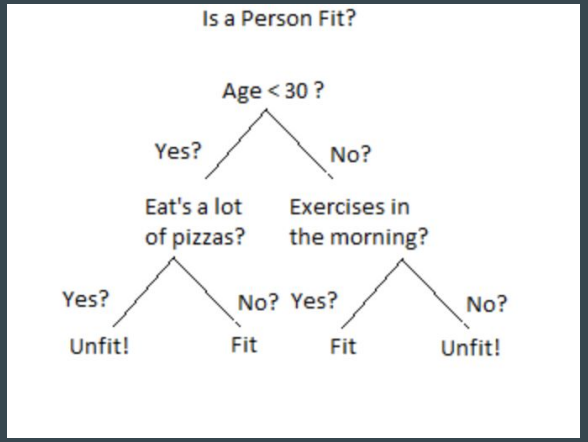
- We used the SVC model from sklearn's model_selection.svm library for the project.
- We trained both Soft and Hard margin SVM by changing the hyperparameter C.
- With testing, we realized that the best results were when the kernel was set as rbf (radial basis function).
- The hard SVM did overfit and resulted in 100% recall and accuracy on the training set.
- The SVM gave us the best results in our project.

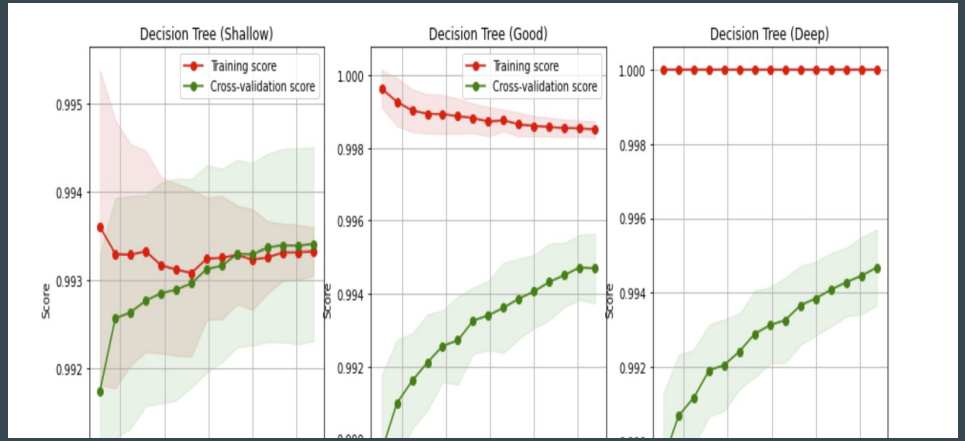




Decision Trees

- We used the DecisionTreeClassifier model from sklearn.tree library to implement decision tree model.
- The test accuracy depended on the depth of the tree.
- We observed that the tree gave great test and train accuracies when the depth was set to 10. A number significantly below or above this number either underfit or overfit the model.
- There is a parameter called splitter, which when set to random significantly decreased the accuracy for the model.

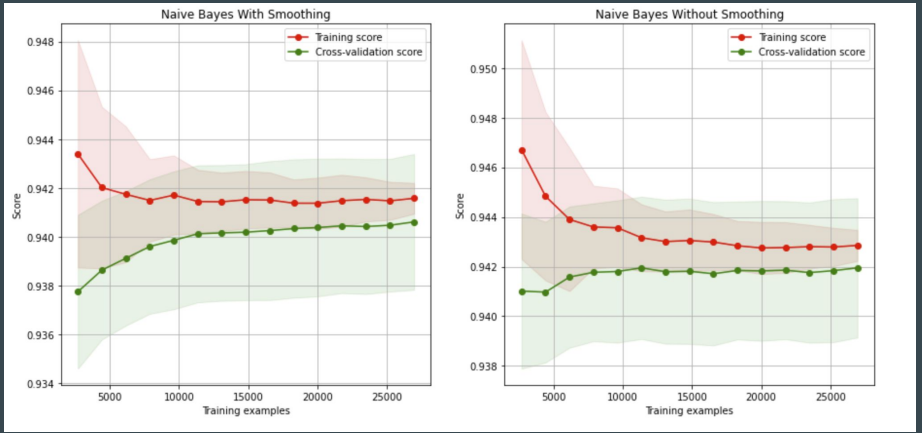




Naive-Bayes

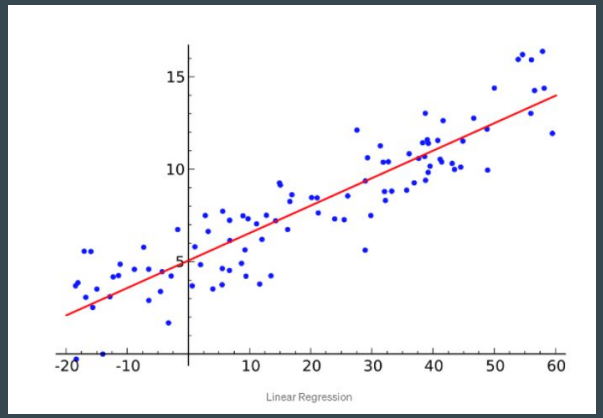
- We used the MultinomialNB model from sklearn's model_selection.naive_bayes library for the project.
- We compared the results with and without Laplacian smoothing. (handles 0 probabilities)
- The model performed slightly better with smoothing.
- It accounts for cases where the word is absent and its 0 probability is taken of by smoothing.
- However, this model did not do well compared to other models with an accuracy of 94.4.

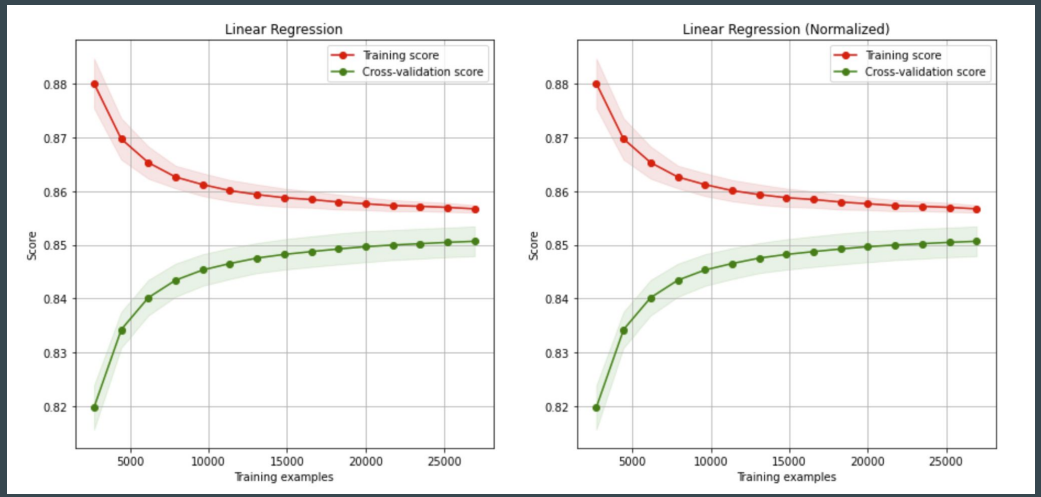
$$P(w'|positive) = \frac{\text{number of reviews with w' and y = positive } + \ \alpha}{\text{N } + \ \alpha * K}$$



Linear Regression

- We used the LinearRegression() model from sklearn's linear_model library for the project.
- We trained the model with and without normalization by changing the parameter LinearRegression(normalize=True).
- We created a function that rounds the values predicted by the Linear Regression Model to produce an accurate classification.
- After testing, we observed that the model gave the same accuracies in both the conditions (normalized and without normalized).





ACCURACIES

Method	Accuracy of Training Set	Accuracy of Testing Set
Naive Bayes with Smoothing	94.1436759421495	94.36971046770601
Naive Bayes without Smoothing	94.25355626169335	94.43207126948775
SVM (soft SVM underfit)	99.40011285005791	99.38530066815144
SVM (Hard SVM Overfit)	100.0	99.679287305122
Linear Regression without Normalizing	98.74677040952692	98.83296213808464
Linear Regression Normalized	98.74677040952692	98.83296213808464
Decision Trees with Depth = 10 (Good)	99.8574525584296	99.53674832962139
Decision Trees with Depth = 5 (Shallow) - Underfit	99.33477860600482	99.43875278396436
Decision Trees Splitting is random for shallow	99.05265346123007	99.06458797327394
Passive-Aggressive Classifier with Maximum Step Size = 1	99.68223799483266	99.38530066815144
Passive-Aggressive Classifier with Maximum Step Size = 100	99.52781159979806	99.0913140311804
Passive-Aggressive Classifier with Maximum Step Size = 10000	99.15659430404182	98.75278396436525

CONCLUSION AND RESULTS

- We’ve managed to create models that differentiate between True and Fake news with almost 99.7 accuracy.
- We’ve developed different models with multiple cases like different depths for decision trees, different step sizes for the Passive-Aggressive Classifier, Hard and Soft SVM, and Naive Bayes.
- Out of all the models, Naive Bayes theory with smoothing have the least accuracy of 94.3697%.
- Hard SVM model gives the best accuracy out of all the models with accuracy of 99.6793%.