



# Classifying Reddit Post Titles

## /r/WorldNews vs /r/NotTheOnion

Manlai Amarsaikhan

DSI-919

## /r/worldnews



Russia/Ukraine Russia pumping millions into US-based propaganda outlets [rawstory.com/russia...](https://rawstory.com/russia...)

Posted by u/AstronautWombat 12 hours ago



Russia/Ukraine Russia loses 480 military personnel, 2 helicopters and 1 plane in past 24 hours [pravda.com.ua/eng/ne...](https://pravda.com.ua/eng/ne...)

Posted by u/Albatross9121 4 hours ago



Russia/Ukraine Russia 'Miscalculated its Strength' and 'Can't Win,' State TV Admits [newsweek.com/russia...](https://newsweek.com/russia...)

Posted by u/Core2score 12 hours ago



Russia/Ukraine Putin says Russia is losing 10 times fewer troops than Ukraine [news.yahoo.com/putin-...](https://news.yahoo.com/putin-...)

Posted by u/Sxzym 9 hours ago

## /r/nottheonion



Woman sues an Arizona city over her arrest for feeding homeless people [infidelpro.com/woman-...](https://infidelpro.com/woman-...)

Posted by u/Voiceamerica 14 hours ago 🚩



Afro hair: School bans probably illegal, says watchdog [bbc.com/news/e...](https://bbc.com/news/e...)

Posted by u/ProFoxxxx 17 hours ago



Painting by Mondrian has been hanging upside down in German museum for decades [paudal.com/2022/1...](https://paudal.com/2022/1...)

Posted by u/BrainStew\_HS 5 hours ago



Real life Overwatch 2 charm costs less than it does in game [pcgamesn.com/overwa...](https://pcgamesn.com/overwa...)

Posted by u/Unlucky\_Lifeguard\_81 2 hours ago



# Background and Problem Statement



- /r/WorldNews (30.1 million subscribers):
  - A place for major news from around the world, excluding US-internal news.
- /r/NotTheOnion (21.9 million subscribers):
  - For true stories that are so mind-blowingly ridiculous that you could have sworn they were from The Onion.
- Reddit moderators routinely remove inappropriate posts
- Can they use machine learning to more efficient (and perhaps more accurate) filtering?



# Data

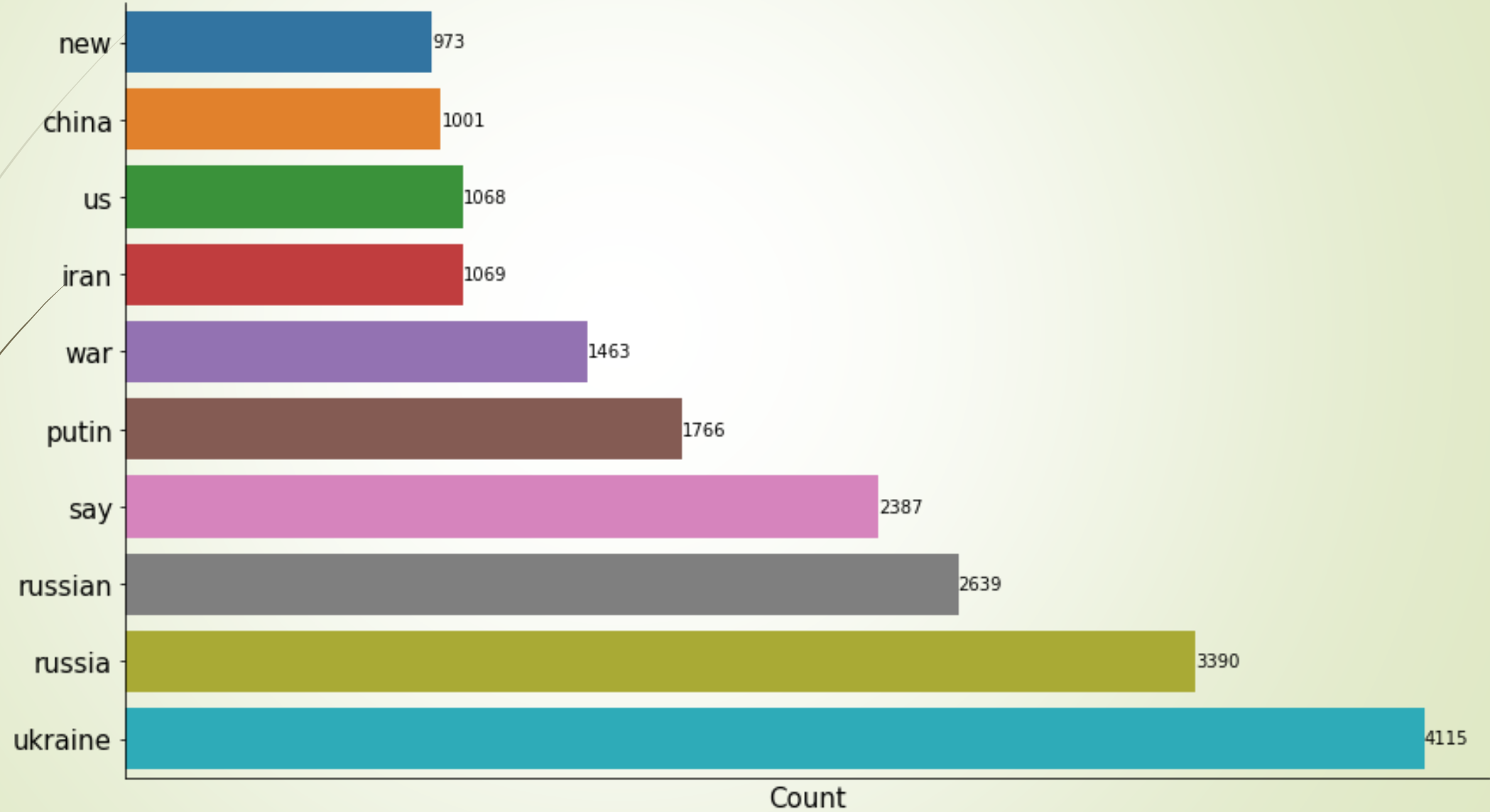
- Pushshift API
- **time**, **title** and **subreddit** of each post were downloaded.
- 24,956 /r/WorldNews posts
- 24,978 /r/NotTheOnion posts
- The fate of a post is never tracked



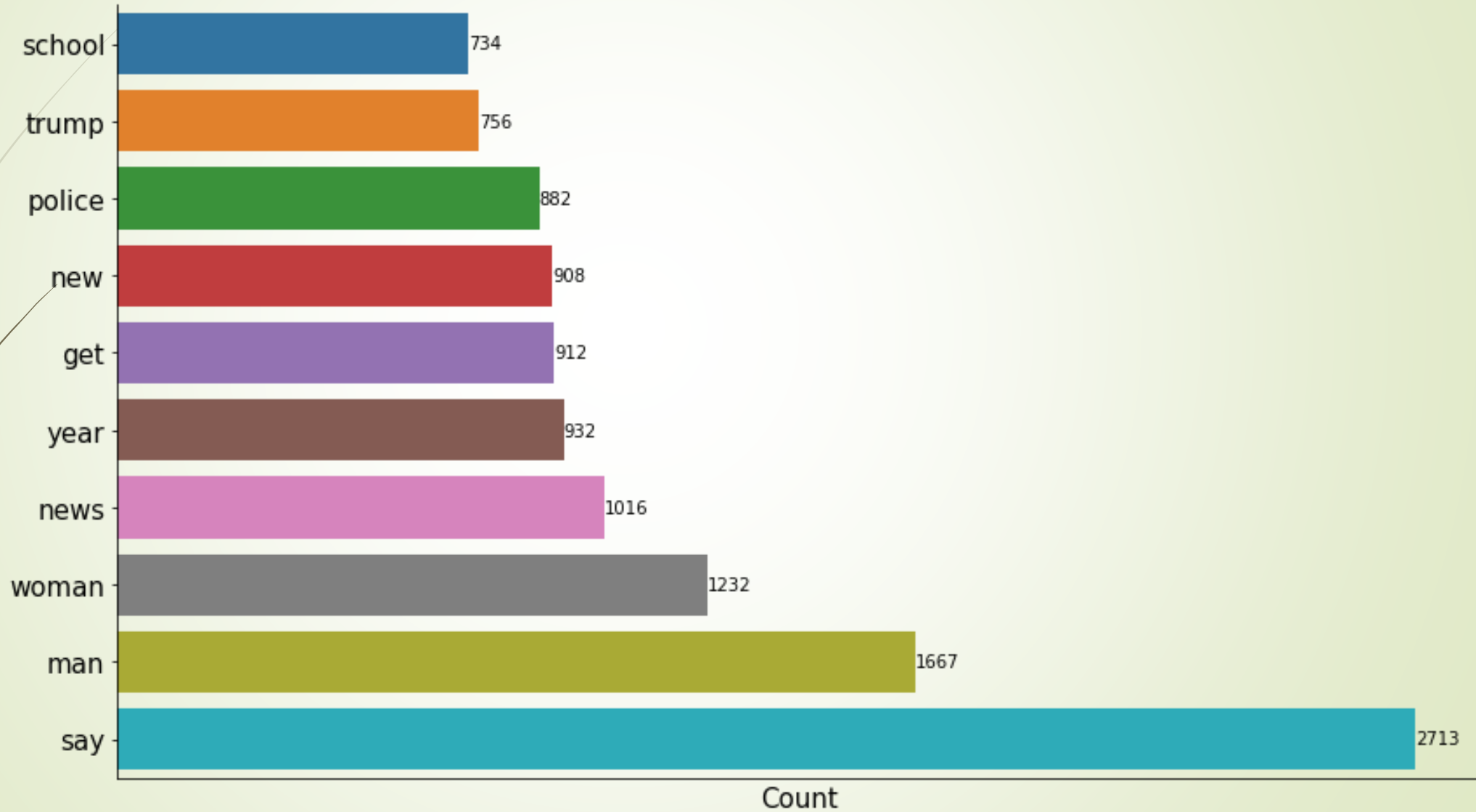
# Data Cleaning

- English stop words were removed.
- `NLTK.tokenize.RegexpTokenizer` that matches
  - alphanumeric characters,
  - dollar amounts
  - non-white space characters
- `NLTK.pos_tag` to identify parts of speech of each word
- `NLTK.stem.WordLemmatizer` was used to group word inflections as a single word
- `sklearn.feature_extraction.text.CountVectorizer` to convert strings to numeric features
- Training data: 37444 x 25075
- Testing data: 12482 x 25075
- The subreddits share about half of their unique tokens

## 10 most common words in WorldNews



## 10 most common words in NotTheOnion





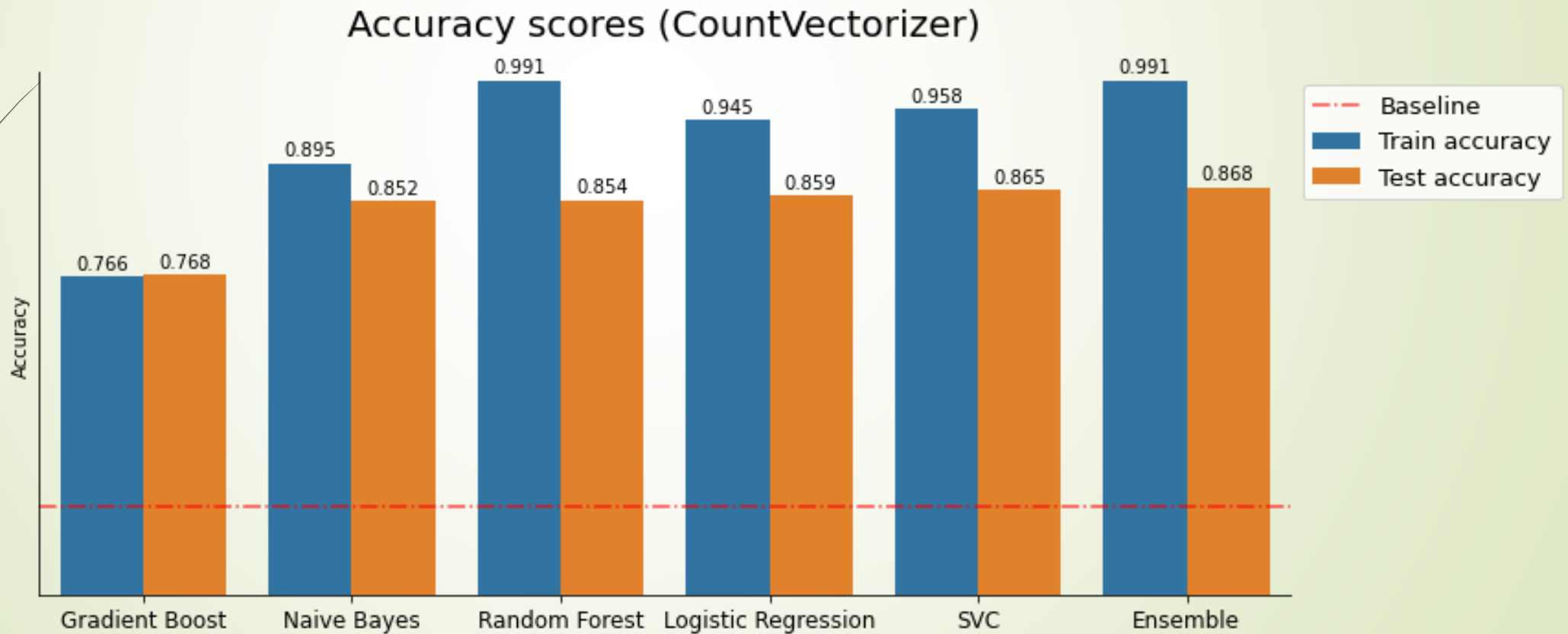
# Models



- Logistic Regression
- Multinomial Naive Bayes
- Random Forest
- Support Vector Machine
- Gradient Boosting
- Ensemble of
  - Logistic Regression,
  - Multinomial Naive Bayes,
  - Random Forest,
  - Decision Tree,
  - Support Vector Machine

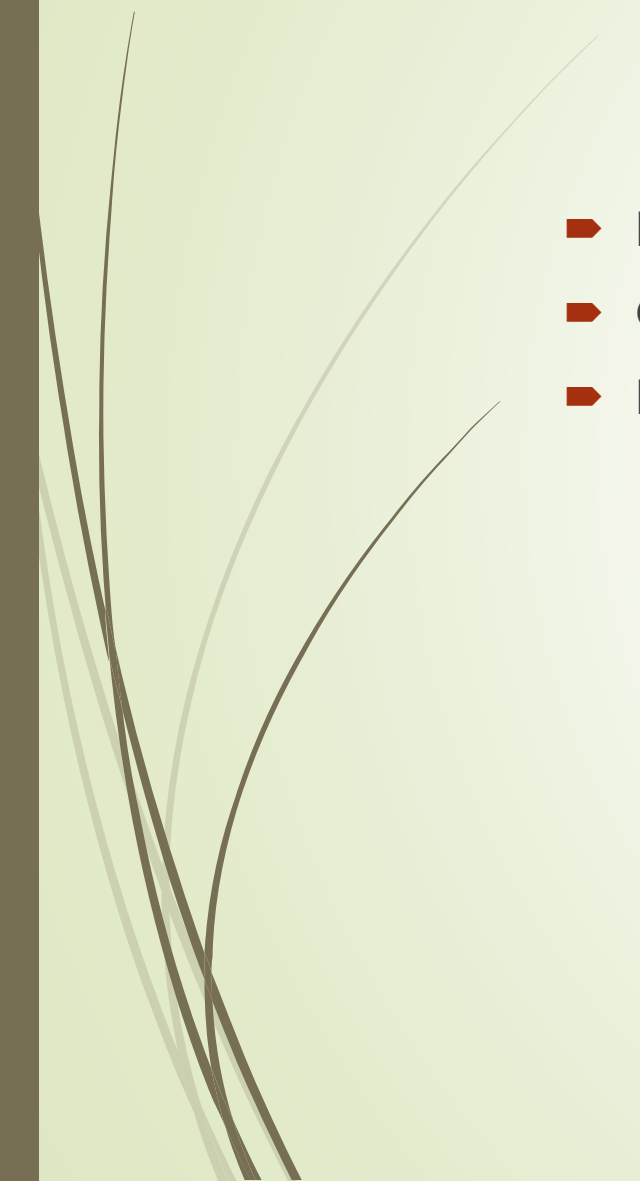


# Model evaluation



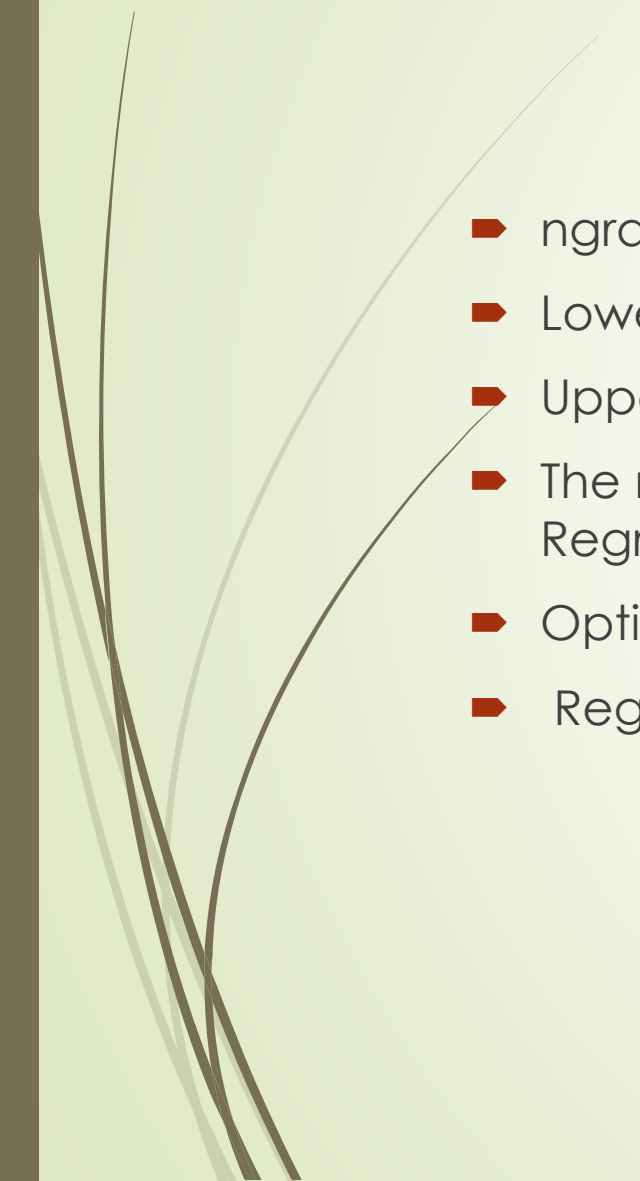


# Hyper Parameter Tuning: Logistic Regression

- High-ish accuracy
  - Overfit
  - Fast
- 



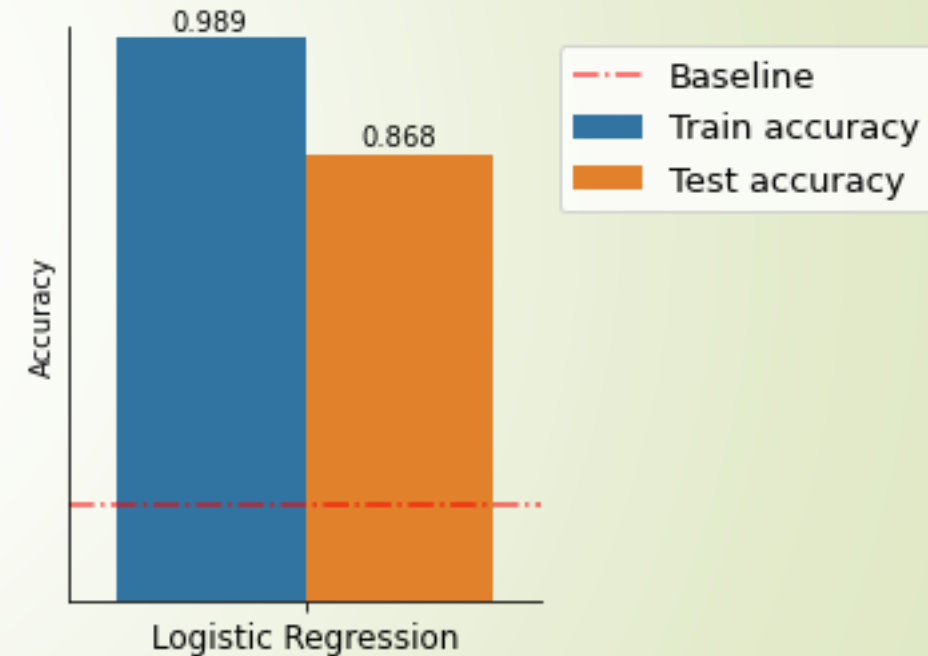
# Hyper Parameter Tuning: Logistic Regression

- 
- ▀ ngram range
  - ▀ Lower cutoff point for document frequency
  - ▀ Upper cutoff point for document frequency
  - ▀ The norm of the penalty term of the Logistic Regression
  - ▀ Optimization algorithm
  - ▀ Regularization strength

# Hyper Parameter Tuning: Logistic Regression

- ngram range
- Lower cutoff point for document frequency
- Upper cutoff point for document frequency
- The norm of the penalty term of the Logistic Regression
- Optimization algorithm
- Regularization strength

Accuracy scores (Lemmatized Tokens)





# Conclusion

- ▶ Performs much better than baseline
  - ▶ Still overfit
  - ▶ Accurately predicts 87%.
- 
- ▶ Use lower cutoff for document frequency
  - ▶ Modify the tokenizer method to filter out more words
  - ▶ Feature engineering