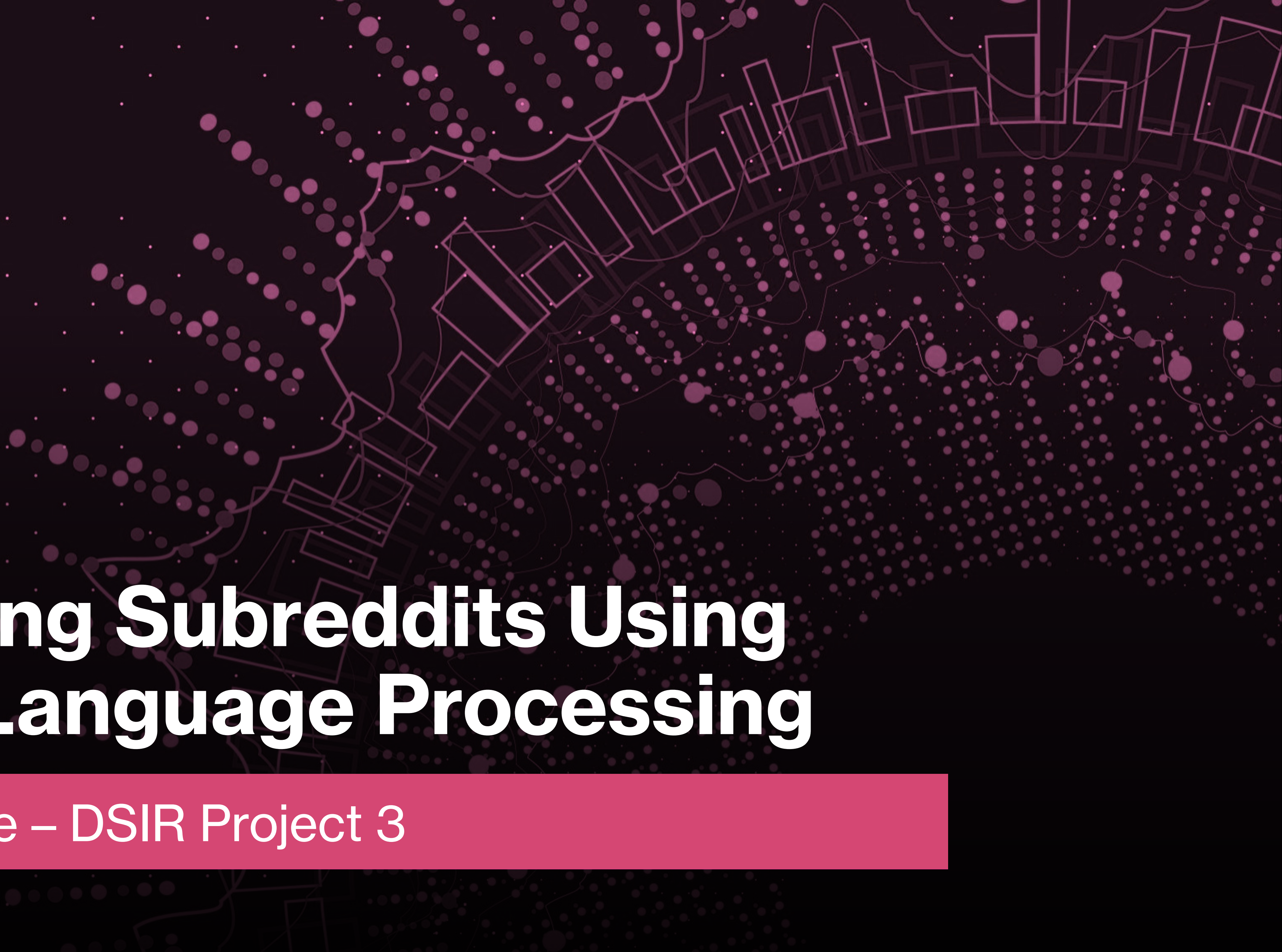


Classifying Subreddits Using Natural Language Processing

Andrea Pascale – DSIR Project 3

Problem Statement

- Explore the unstructured text data in two different video game subreddits in order to build a classification algorithm that can distinguish between the two categories
- Analyze trends in word usage to gain insight on type of gamer (PC vs. Console), as well as reasons for posting online (mods/DLCs, bugs/glitches, fan-fiction, etc.)

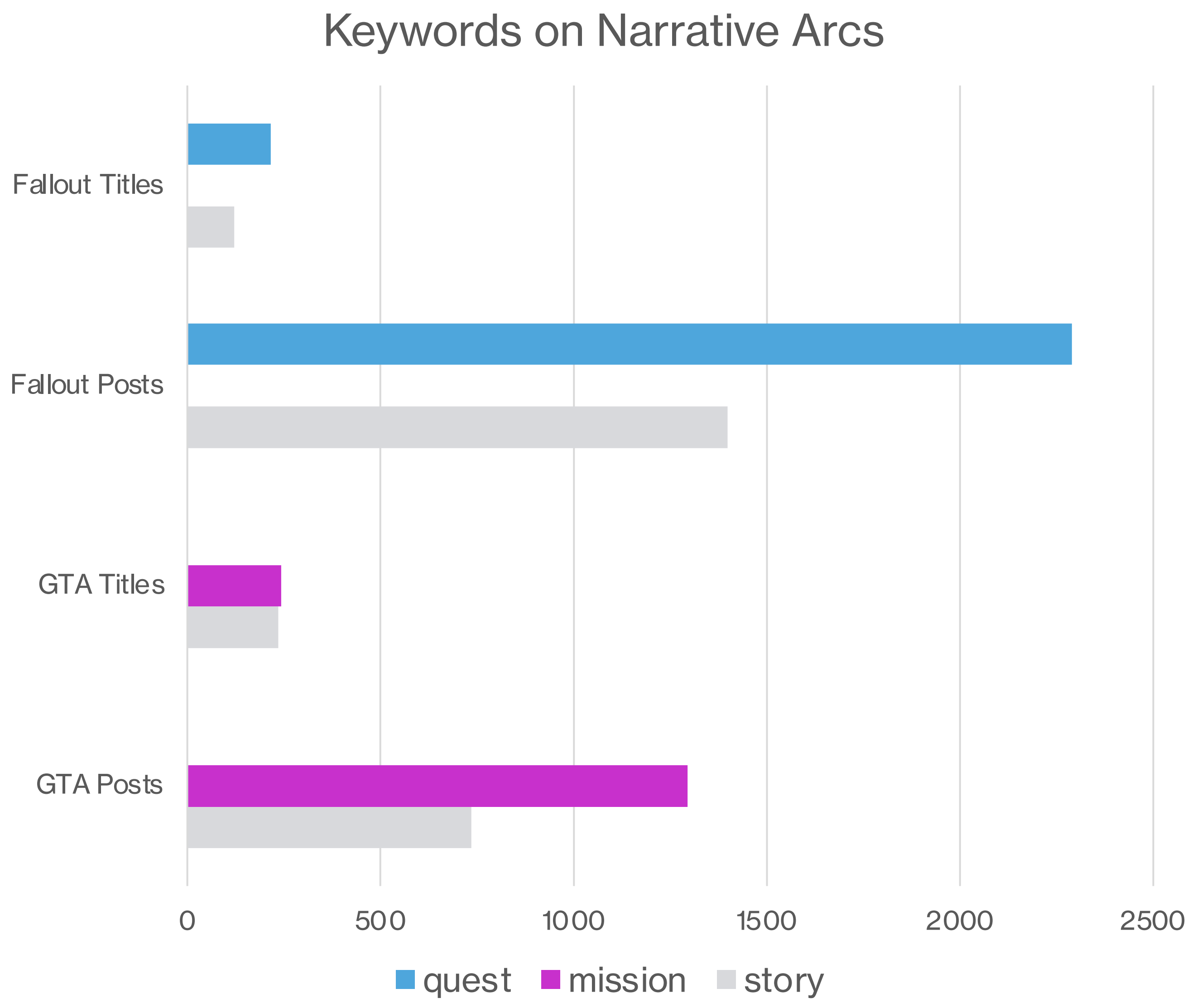
The Games

GrandTheftAutoV and Fallout

Data Collection

- Scraped a sample of 20,000 posts from each subreddit
- Dropped duplicates and missing values
- Data cleaning reduced number of observations by 30-60%
- Total number of observations remaining:
    - Grand Theft Auto - 8,500
    - Fallout - 13,500

Keywords : Missions vs. Quests



Bag of Words

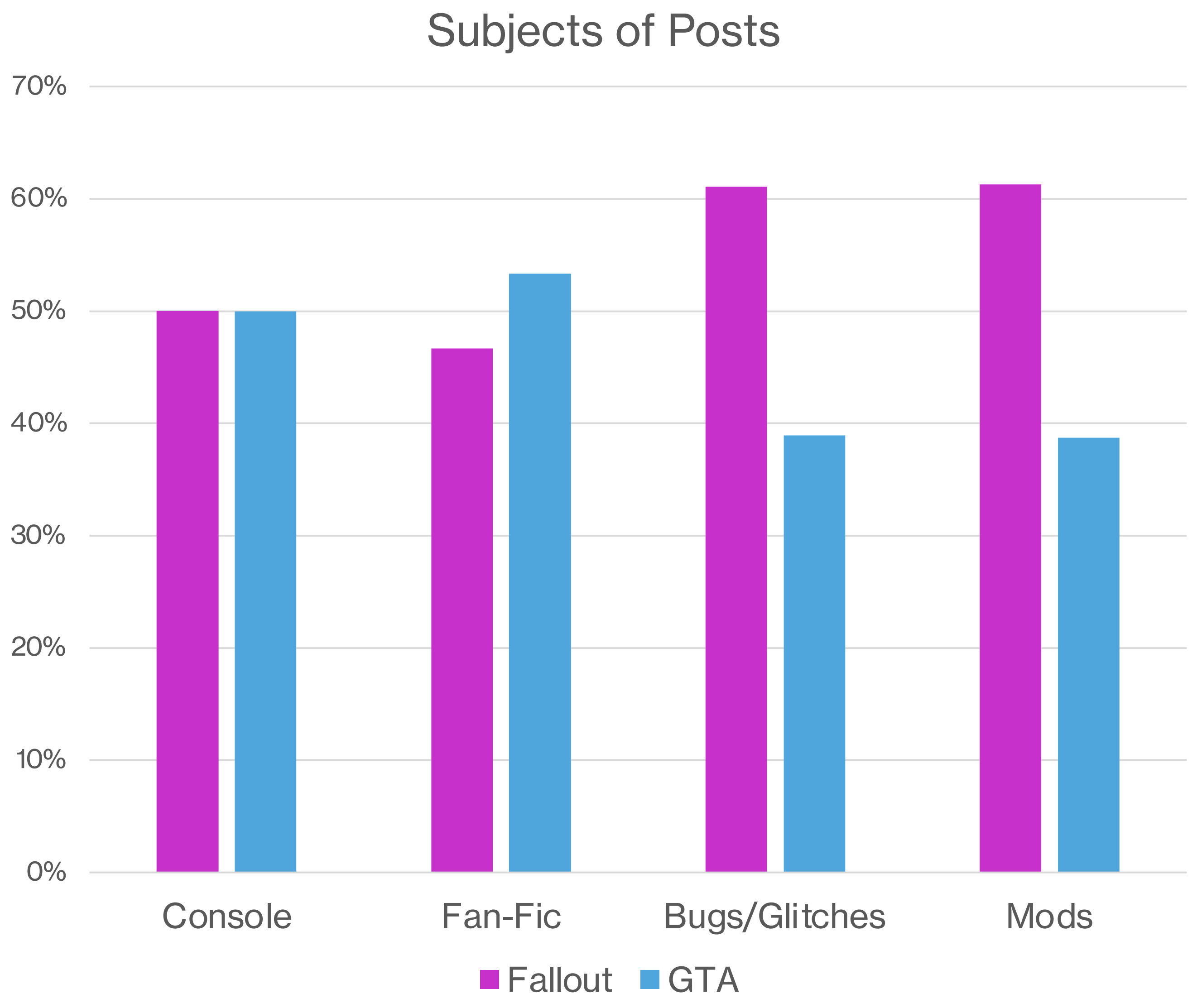
Fallout

- Vault
- World
- Armor
- Character
- Brotherhood
- Wasteland
- Faction

Grand Theft Auto

- Car
- Rockstar
- Money
- Epic
- Club
- Casino
- Crew

Topic Clustering



Building the Model

- Pre-processed raw text using Count Vectorizer
    - Set maximum number of words collected to 14,000 per subreddit
    - Words must appear in at least two posts
    - Words cannot appear in more than 80% of all posts
- Built pipelines and grid searched over multiple estimators
    - Multinomial Naïve Bayes
    - Random Forest
    - Logistic Regression

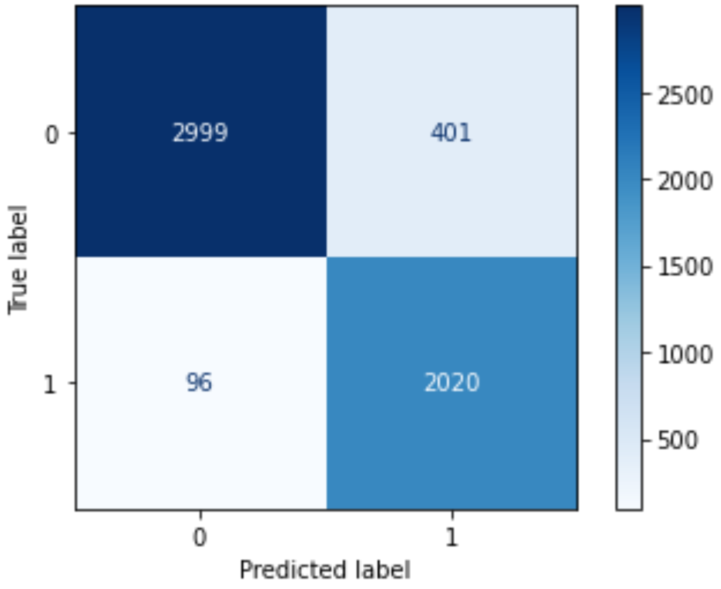
Scoring the Model

	Naïve Bayes	Random Forest	Logistic Regression
Training Group	0.97	0.90	0.97
Testing Group	0.96	0.89	0.92

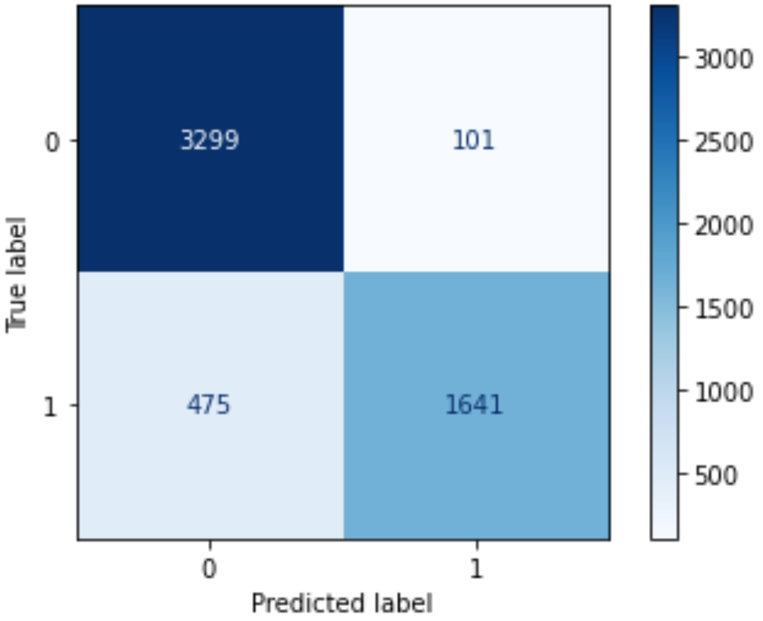
Baseline model ~ .61

Evaluating the Model

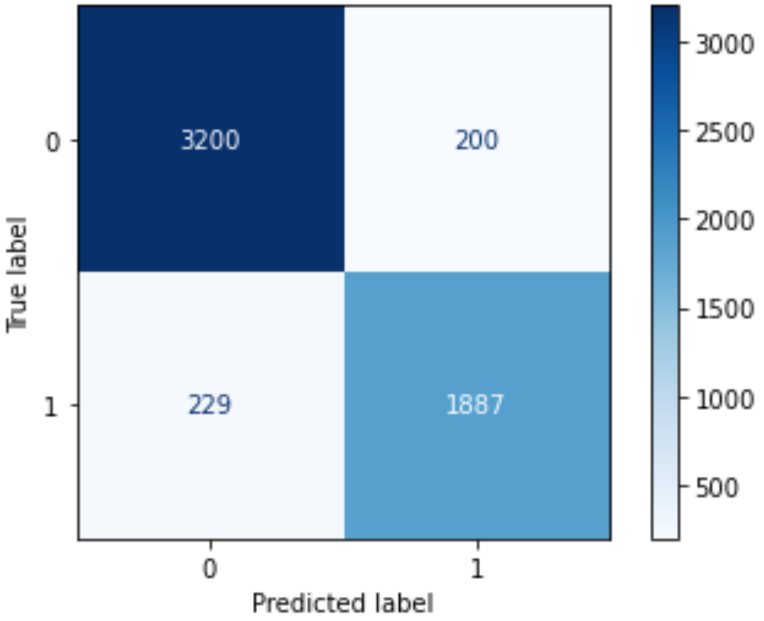
Naïve Bayes



Random Forest



Logistic Regression



Improving the Model

- Better lemmatization/stemming
- More comprehensive tagging of words
- Sentiment Analysis on most frequently recurring word pairs
- Testing additional estimators such as Support Vector Machines