DSO 560 Final Project

Jewon Ju, Matthew Lee, Krish Vora

TABLE OF CONTENTS

1. Goals & Objectives
2. Data Cleaning & Text Preprocessing
3. Topic Modeling
4. Sentiment Analysis
5. Recommendations
6. Evaluation/Improvements

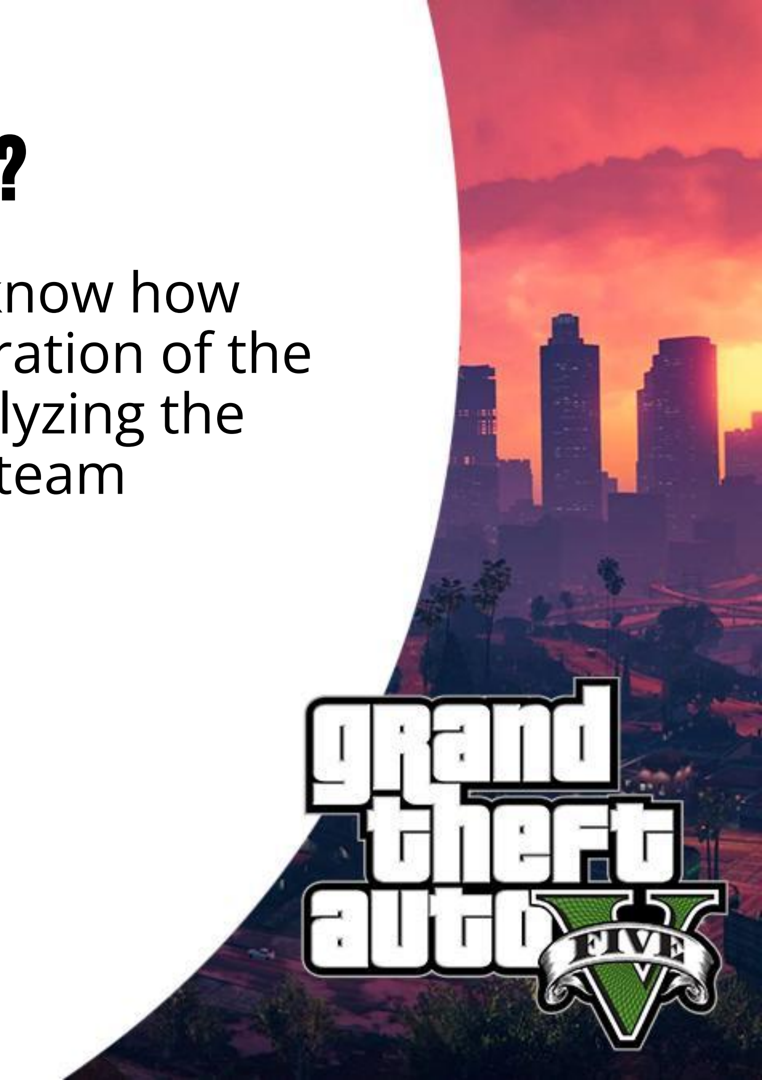
1. GOALS & OBJECTIVES

WHAT DO WE WANT TO ACHIEVE?

Rockstar Games would like to know how they can improve their next iteration of the Grand Theft Auto series by analyzing the sentiment of user reviews on Steam

Topic Modelling

Sentiment Analysis



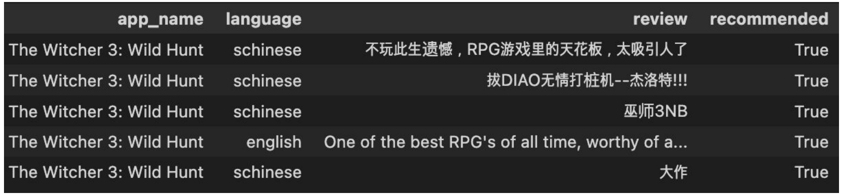
2. DATA CLEANING & TEXT PREPROCESSING

DATA SET

Data Source: Steam Reviews Dataset 2021 from Kaggle.com

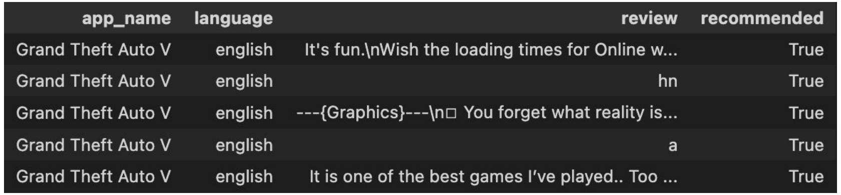
We filtered only for reviews in **English** about **Grand Theft Auto V**

21747371 Rows



app_name	language	review	recommended
The Witcher 3: Wild Hunt	schinese	不玩此生遗憾，RPG游戏里的天花板，太吸引人了	True
The Witcher 3: Wild Hunt	schinese	拔DIAO无情打桩机--杰洛特!!!	True
The Witcher 3: Wild Hunt	schinese	巫师3NB	True
The Witcher 3: Wild Hunt	english	One of the best RPG's of all time, worthy of a...	True
The Witcher 3: Wild Hunt	schinese	大作	True

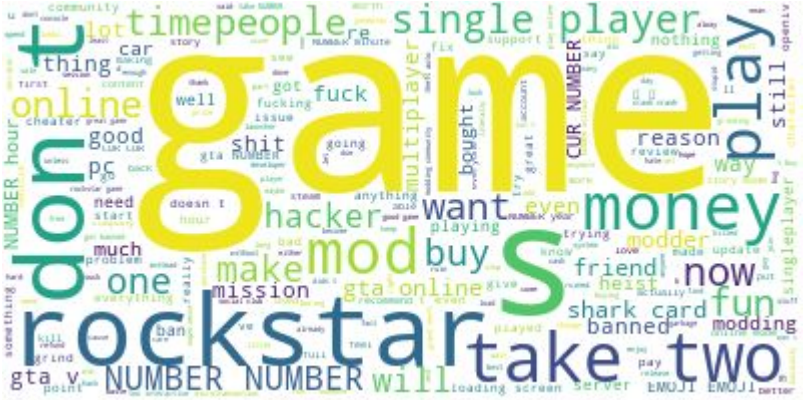
319751 Rows



app_name	language	review	recommended
Grand Theft Auto V	english	It's fun.\nWish the loading times for Online w...	True
Grand Theft Auto V	english	hn	True
Grand Theft Auto V	english	---{Graphics}---\n□ You forget what reality is...	True
Grand Theft Auto V	english	a	True
Grand Theft Auto V	english	It is one of the best games I've played.. Too ...	True

EXPLORATORY DATA ANALYSIS

We used a word cloud in order to see what other stop words we want to remove from our data set



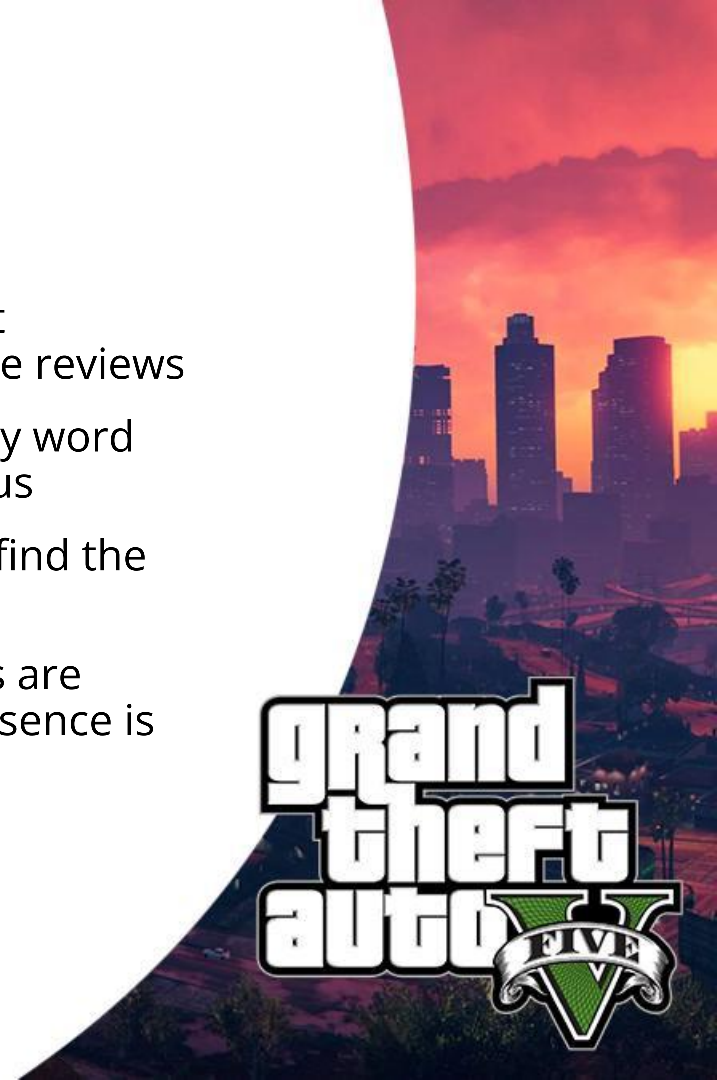
TEXT PREPROCESSING

- Lower Case
- Replace Common Entity (urls, hashtags, numbers, currency symbols, emojis, emails, numbers)
- Remove punctuations
- Remove stopwords (NLTK + Custom words)
- Lemmatization

3. TOPIC MODELING

METHODOLOGY

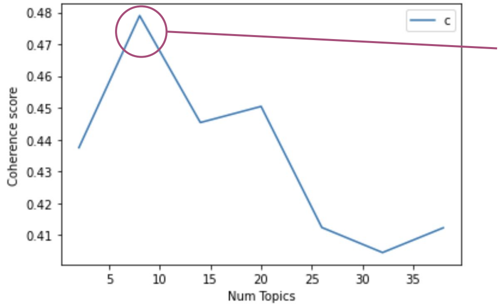
- Filter for reviews that indicated that they would not recommend GTA V and look for main topics in those reviews
- Use **N-Gram (Bigrams and Trigrams)** to isolate key word pairings/groups that appear frequently in the corpus
- Build **Latent Dirichlet allocation (LDA) model** to find the keywords for each topic
  - LDA is used to do topic modeling where words are collected into documents and each word's presence is attributed to one of the document's topics



Inspiration: Topic Modelling with Gensim

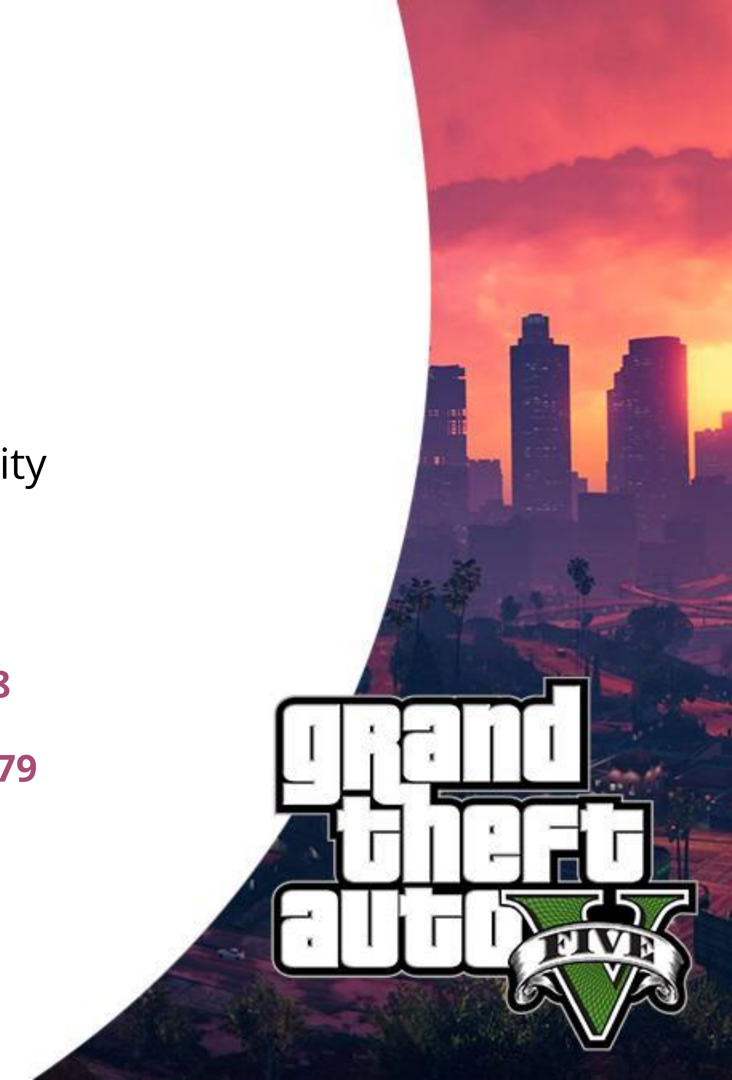
OPTIMIZATION

- Use iterative approach to determine the optimal number of topics using topic coherence as the evaluation metric
- **Topic Coherence:** the degree of semantic similarity between high scoring words in the topic

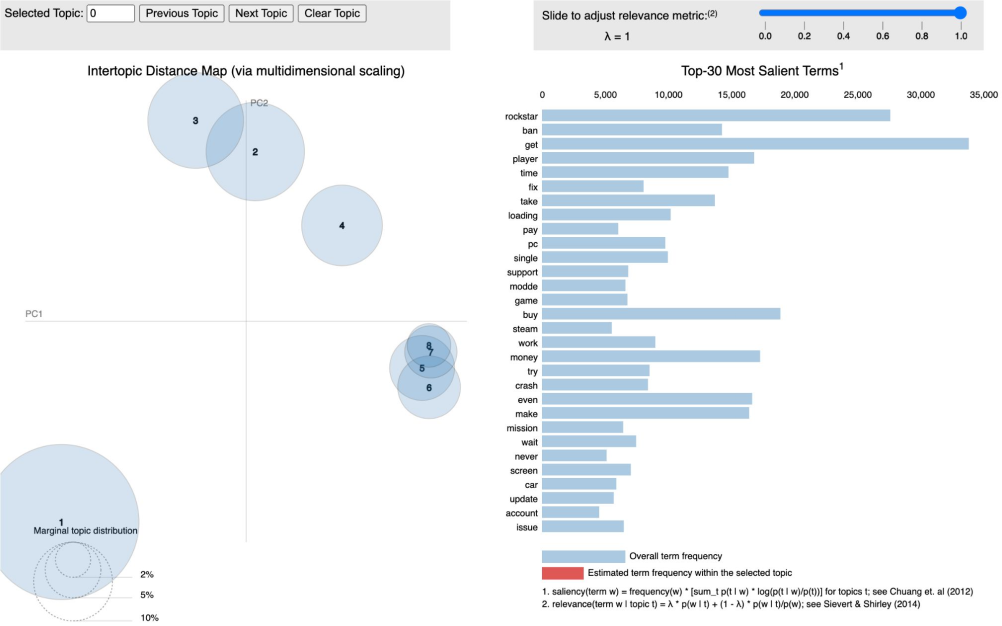


Optimal Number of Topics: 8

Topic Coherence Value = 0.479



RESULTS



Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

λ = 1

🟦 Overall term frequency

🟥 Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

TOP 8 TOPICS

Topic 0: Player, single, money, buy, multiplayer, mod, take, fun, want, pay → **Value for Money**

Topic 1: loading, screen, load, wait, take, time, crash, minute, suck, long → **Long Loading Times**

Topic 2: fix, crash, hacker, server, issue, many, get, problem, session, bug → **Bugs and Crashes**

Topic 3: rockstar, buy, work, even, steam, support, try, game, launch, account → **Issues with Steam**

Topic 4: hacker, good, mod, singleplayer, ruin, multiplayer, great, rip, rockstar, kill → **Hackers**

Topic 5: pc, community, console, version, hate, port, garbage, run, gaming, time → **Issues with PC Version**

Topic 6: get, money, make, time, car, people, even, go, buy, fun → **Long time required to achieve things in the game**

Topic 7: ban, get, take, modde, rockstar, mod, buy, reason, support, people → **Getting Banned**

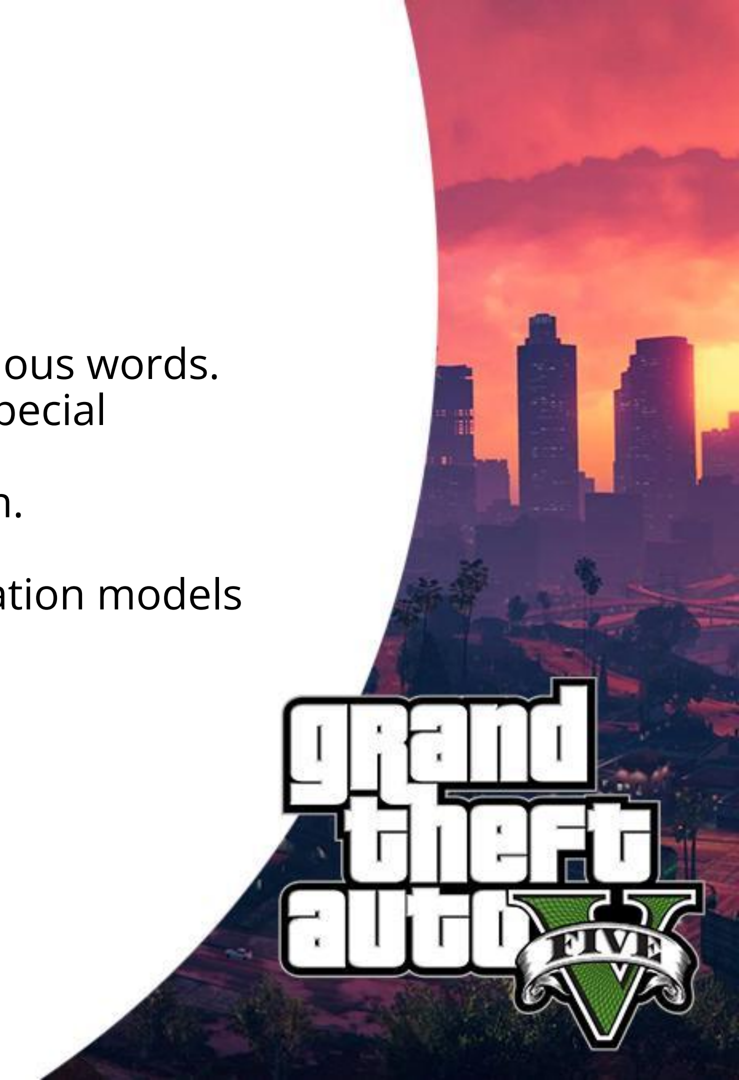
4. SENTIMENT ANALYSIS

METHODOLOGY

- Stopword removal using a custom stopword list.
- Regex grouping for uniform representation of various words.
- Regex cleaning to remove punctuations, emojis, special characters.
- Using CountVectorizer and TF-IDF for vectorization.
- Splitting the data in a 70:30 ratio.
- Employing Supervised Machine Learning Classification models to classify reviews.

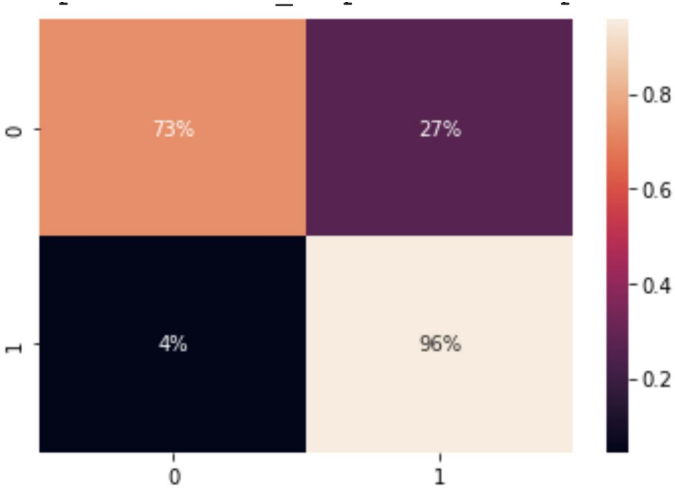
Models deployed:

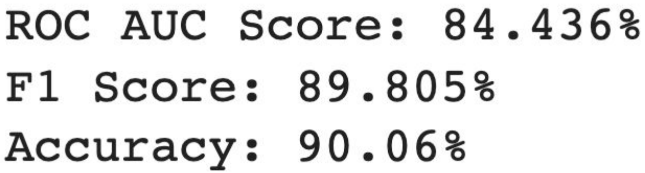
- Logistic Regression
- Random Forest
- Recurrent Neural Network



LOGISTIC REGRESSION

Logistic regression is a statistical model that in its basic form uses a logistic function to (usually) model a binary dependent variable (i.e., fraud label).

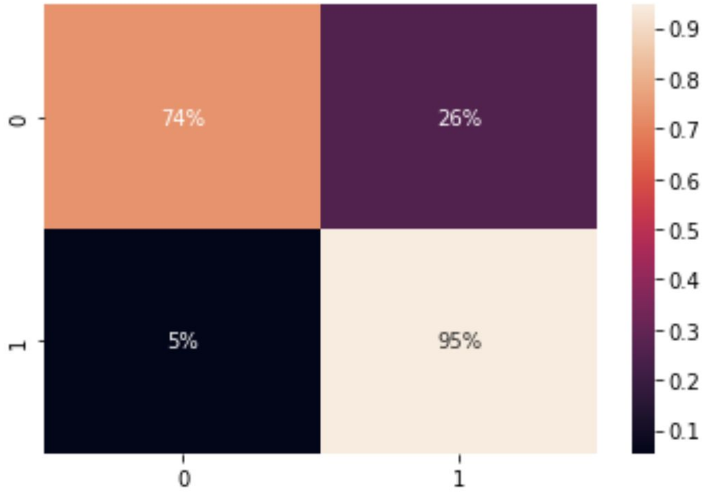


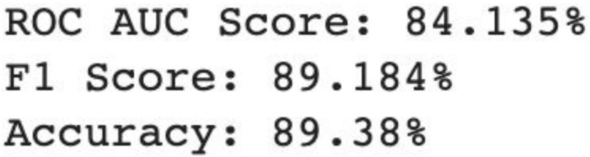


```
ROC AUC Score: 84.436%
F1 Score: 89.805%
Accuracy: 90.06%
```

RANDOM FOREST

Random forest is an ensemble learning method for classification that operates by constructing multitude of decision trees.

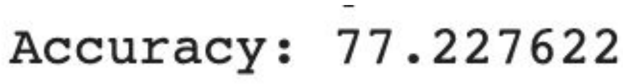




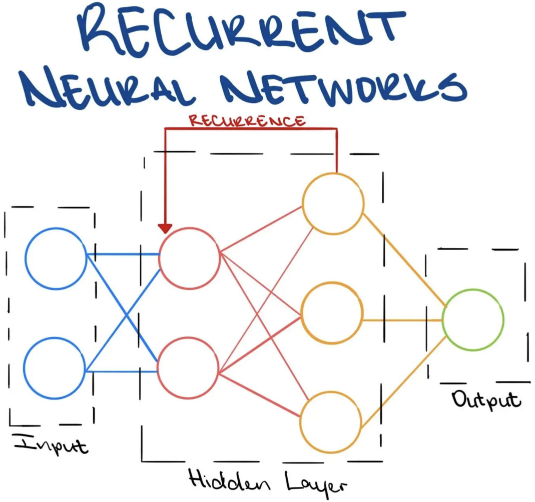
```
ROC AUC Score: 84.135%
F1 Score: 89.184%
Accuracy: 89.38%
```

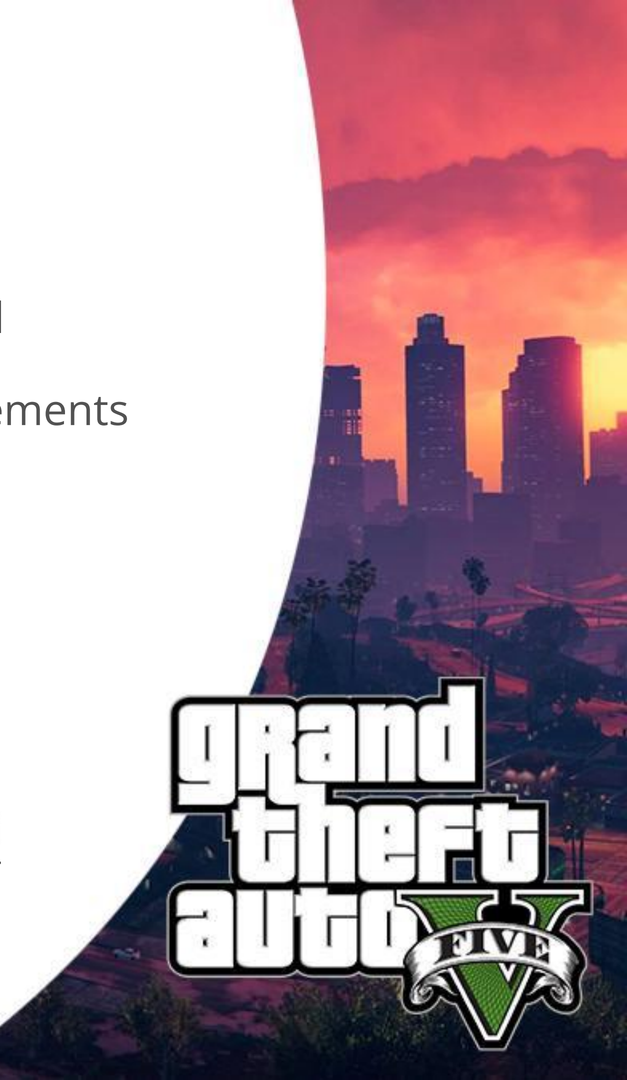
RECURRENT NEURAL NETWORKS

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. The output of recurrent neural networks depend on the prior elements within the sequence.



```
Accuracy: 77.227622
```





MODEL RESULTS

Model	**Accuracy**
Logistic Regression	90.06%
Random Forest	89.38%
Recurrent Neural Network	77.23%

Logistic Regression has the best accuracy amongst all models. Random Forest also performs similarly without any hyperparameter tuning, upon further tuning, it could potentially be a better performer.

5. RECOMMENDATIONS

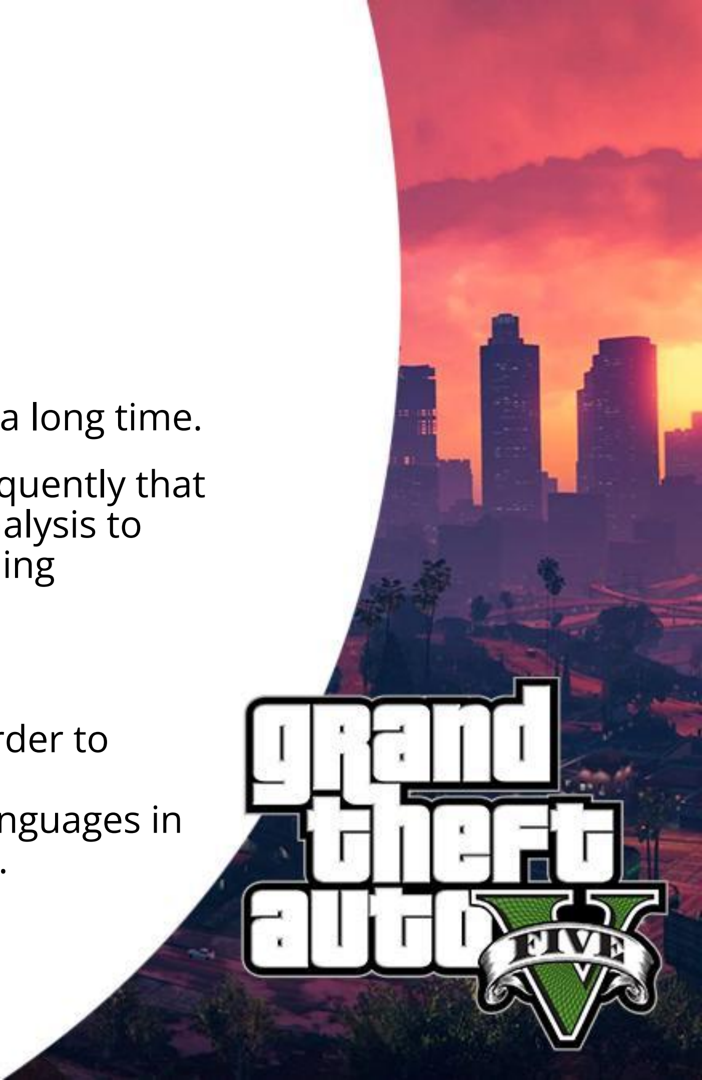
RECOMMENDATIONS

Topic Modelings

- Shorten loading screen time by having things loaded in the background so the users don’t feel like they are waiting for a long time.
- Limit bugs and number of crashes by releasing updates frequently that address these issues. Use topic modeling and sentiment analysis to keep checking on what specific bugs or crashes are happening frequently.

Sentiment Analysis

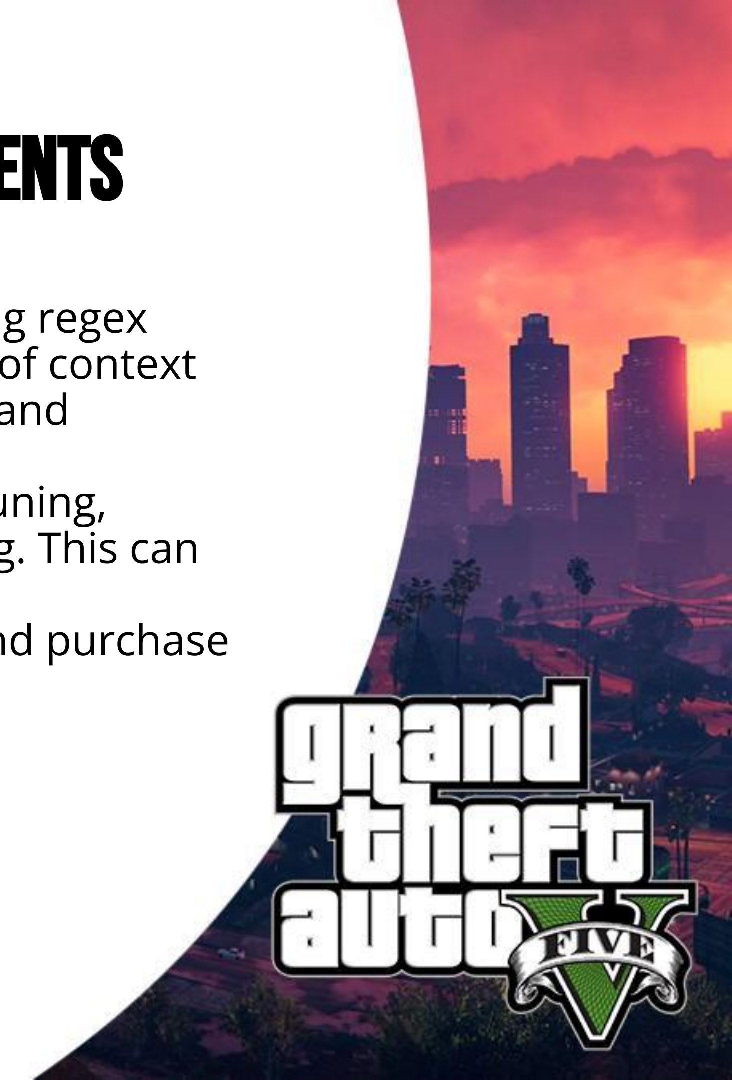
- Build a recommendation system based on the reviews in order to target their audience better.
- Increase the scope of the analysis to various regions and languages in order to get a better understanding of the overall audience.



6. EVALUATION/IMPROVEMENTS

EVALUATION & POTENTIAL IMPROVEMENTS

- Further text preprocessing for topic modeling using regex
- Using transformers like BERT to take into account of context
- Use more advanced models like Neural Networks and Transformers.
- Improve the current models by hyperparameter tuning, feature engineering and further text preprocessing. This can be achieved by understanding the data better and augmenting it with other data points like region and purchase history.



THANK YOU