Should You Buy This Game??

Team: Daniel Mata, YingHsuan Lo

Project Repository

Increment Video Link

Final Project Demo Link

Final Project Presentation Link

Fig. extracted from our EDA in text classification model notebook

Motivation

With video games being in abundance and game console technology getting better and better, video game journalists are constantly writing game reviews about the latest video games at any given moment. Sometimes, readers may just want a quick opinion on whether or not to buy a video game. They may want to get the gist of how a video game journalist has described a certain video game. Additionally, some readers may be visually impaired or have some other form of disability that requires some audio component when browsing the web. Or some may just simply want to pick the game when they are doing chores, driving, or other routines-- treat a purchase choice more like an experience of listening to an audio book.

As such, we choose to focus on text summarization, text classification, and speech synthesis in order to help facilitate video selection for buyers as well as to elevate market penetration for sellers. The main aspect of our project is the text classification, we would like to see if we can improve upon past work using the Steam Reviews data (more information provided in sections below). We want to provide a quick way for users to decide on whether or not to buy a video game.

A transparent and automatic analyzing and classifying system is needed, yet few seriously consider it in academia. For future market analysis research and for more people to accurately pick the right video games, we propose a system which performs video game review text collection from IGN video game website, text summarization and classification on the review, and finally performs text-to-speech. The result will be incorporated showing a game's summary, visualized recommendation, and speech

synthesis.

Significance

Video games have become more and more influential in every endeavor. From school educational platforms to museum programs, video games ignite interest in many. An interesting and appropriate video game can be utilized in various ways ranging from relaxation(and de-stressing), making new friends, and even advanced topics such as reinforcement learning. As a result, a functional system that can be used to quickly assess a user's decision on whether or not to buy a video game can be very helpful. In this design, we especially add speech synthesis as an additional component to visualization elements in order to better assist those who suffer from any visually-impairing disability.

Background

Research in Recent Years Using Steam Dataset:

Steamvox(2019) is a platform built to scrape using steam reviews, clean data using NLTK, SpaCy, gensim, Syntok,do topic modelling using Latent Dirichlet Allocation(LDA) and analyze reviews using VADER SentimentIntensityAnalyser from Steam to identify topics and its sentiment for each topic. It is currently focusing on the Total war game: Three kingdoms. Besides it capacity that only be able to focus on analysing one game for the time being, it also observed other problems like: spam reviews are common, often hold no meaning and usually be posted as "spam bomb game" among reviewers; most reviews are too short to do further analysis; vedar analyzer usually failed on discern sarcasms; some reviewers do not use punctuation, and this can affect tokenization process.

The paper *A Study on Video Game Review Summarization*(2019) examines aspect-based summarization and sentiment analysis applied on the game reviews and also offers an evaluation process on the performance of the summarization task aiming to minimize supervision that constantly performs in the previous research. The preprocessing includes converting reviews into tf-idf vectors, tokenization, stopword removal and lemmatization. The experiment set up most frequent words in 5 clusters and applied k-means clustering to perform aspect extraction and Aspect labelling. Then, they use VADER analyzer to combine lexical features to sentiment scores with a set of five heuristics. It did not fully resolve the pitfalls stated previously.

In *Summarizing Game Reviews: First Contact* (2020), its summarization pipeline includes : Pre-processing and parsing, Topic Modeling, Sentiment Analysis, summarization. The first two are based on keyword detection and clustering, the second step performs sentimental analysis and the third step is bi-directional BERT model. The algorithm then uses the generated embedding to train binary Ridge Logistic Regression classifiers in each aspect. Each candidate sentence gains a confidence score. Sentences with a high prediction confidence score will be selected as summarization candidates. This model also displays aspect extraction and Aspect labelling as the previous paper, but it includes sentiment analysis as a feature for summarization. The limitation of the pipelines lies in the mixture of sentiments from users about various features, and it makes labelling and extraction processes harder to get a clear result.

In *Recommender System: Rating predictions of Steam Games Based on Genre and Topic Modelling* (2020), the research focuses on implementing a genre-based and topic modeling recommender system to predict rating of games. The process includes, data cleaning, convert playtime data to rating, topic extraction by LDA (as previous paper), implementing K-NN algorithm to determine the game rating target, computation of user and item similarity, using RMSE for model evaluation. However, the evaluation shows non-outstanding performance using genre and topic modelling for games recommender systems.

From these papers, we can observe some similarities in the project design. For example, VADAR analyzer is used in sentiment analysis, LDA used to do Topic Modelling, performing clustering according to each setup clustering tables. (e.g. most frequent words clusters table) Sentiment analysis, topic modeling or genre classification are used as features

while implementing text summarization, sentiment analysis, or recommender.

In our project, we scrutinized the review texts and decided to focus on text analysis because we noticed that these papers did not focus on expanding contractions, replacing video game slang. So it will be our focus to make some breakthrough from the previous research progress, and we will also implement ensemble learning for our project design. And from this point, for a better performance, we want to expand the project architecture to scrap the review content, clean and preprocessing in order to perform summarization which may largely condense the essence of the review. Then conduct our classification. The result of summarization will be read by our Text-to-Speech model, as well as be inputted in our classification task.

Our Model:

Architecture Diagram:

Note: We implemented 3 different models:

A. Summarize the review - T5 Base Model
B. Classify the video game as "recommended" or "not recommended" - Scikit Learn ML model
C. Perform speech synthesis - MelGAN / TransformerTTS

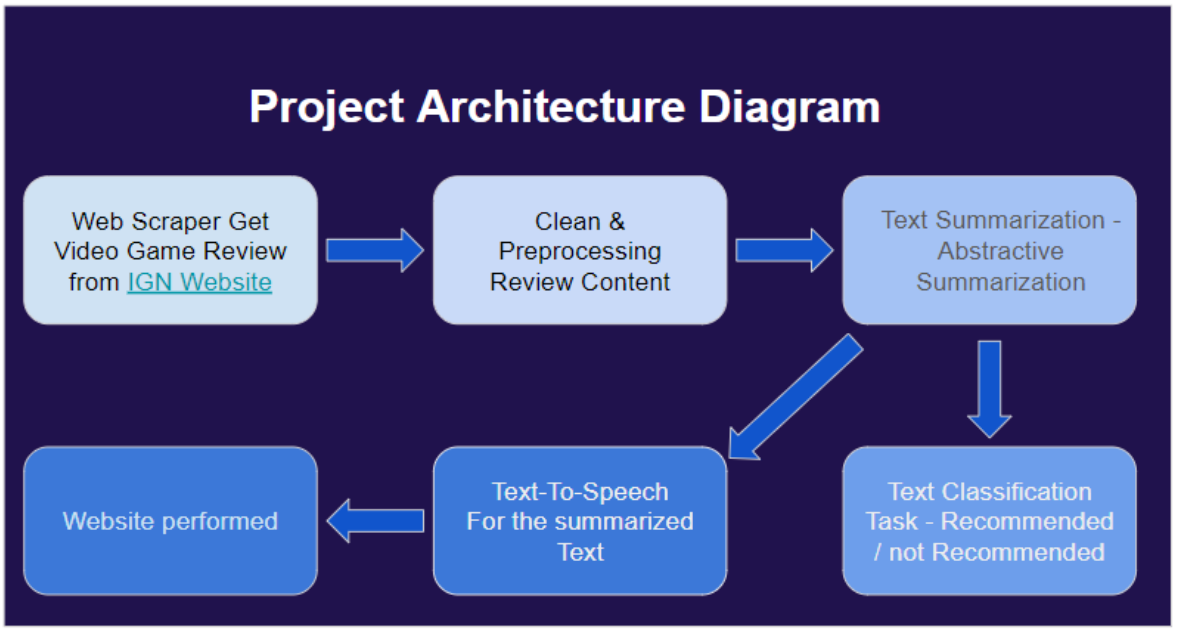


Figure 1: Our project design diagram

We will incorporate the models to perform the task where a user submits a link to the Steam video game review: From here, we web scrape the review. There will be three different models used to do classification of recommended or not recommended, review summarization, and text-to-speech summarization tasks.

Workflow Diagram:

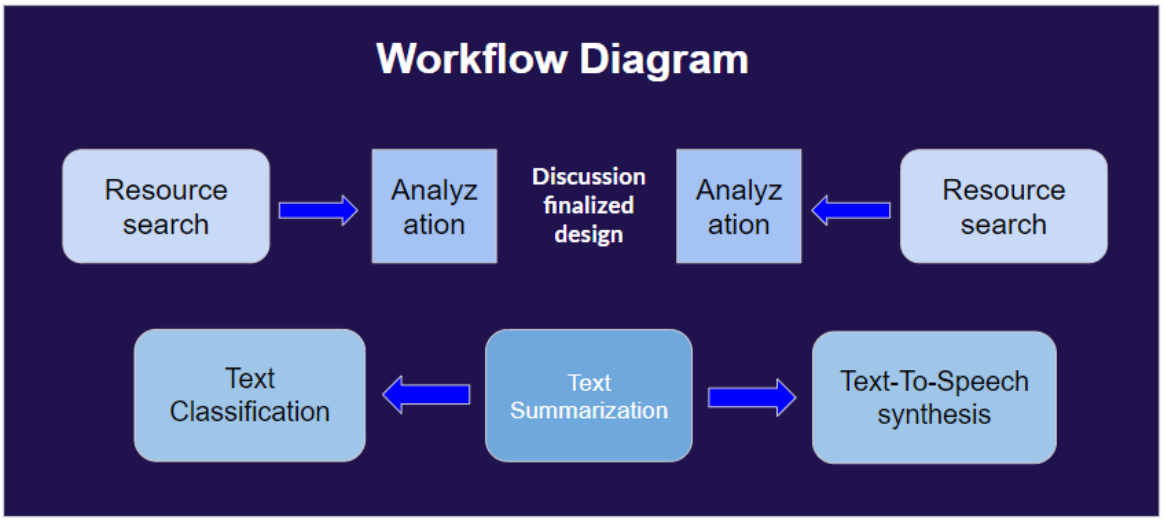


Figure 2: Project workflow diagram

In Our project, we focus on the Classification task and its dataset--Steam review. Therefore, we focus on searching the resources separately, through repeatedly discussing the possibilities of the project development, and reach a consensus to innovate this classification from a semantic perspective, then implement T5 model because the pretrained model uses text-to-text method, emphasizing providing a comprehensive framework for a task specific architecture. And then perform Text-To-Speech function for a more convenient user experience. The Steam Dataset, WikiHow Dataset, and LJSpeech Dataset are datasets for training purposes only. Our Goal is to use Web Scraper to get video game reviews from IGN video game review website to perform text summarization, text classification, and Text-To-Speech tasks.

Dataset:

Stream Website and Review Dataset

Dataset 1 : Steam Reviews

The final aim for the project is to piece all models together and create a process that summarizes, classifies, and performs TTS. However, our main focus was geared towards improving upon the text classification, followed by applying what we've learned throughout this semester in the form of the additional two models. Our innovations are described under the ***Models*** *section right below our EDA on the Steam Reviews Dataset.*

Steam is a gaming platform that acts as a third-party medium to sell games online and download them. Users are generally aged between 18 to 30. Consequently, the culture around Steam is built on sarcasm, memes, and wit. This community-driven content often informs other users' purchases and is also monitored by developers and publishers in order

to glean opinions on specific aspects of the game which can be patched or improved in updates to the game.

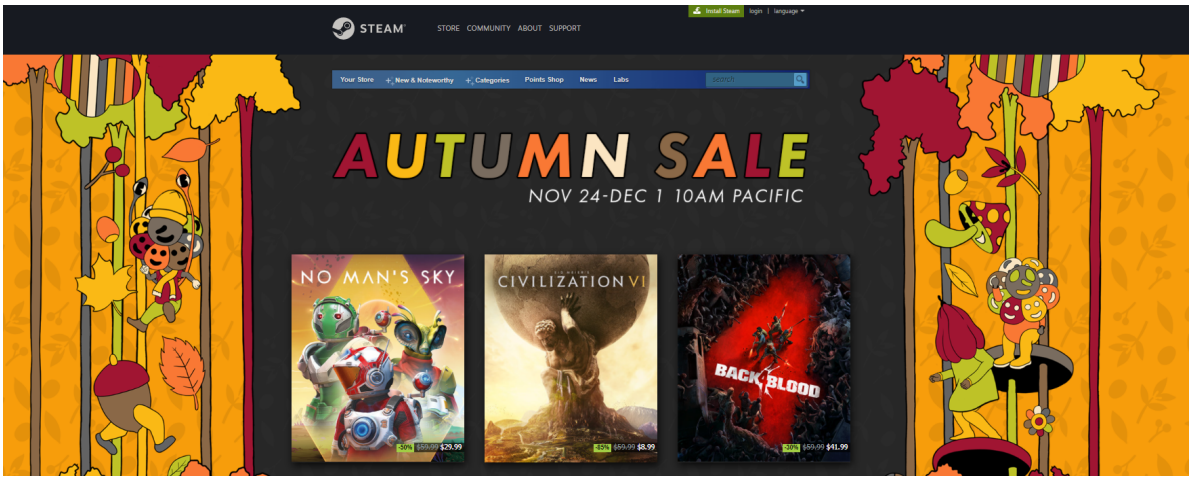


Figure 3: Steam website

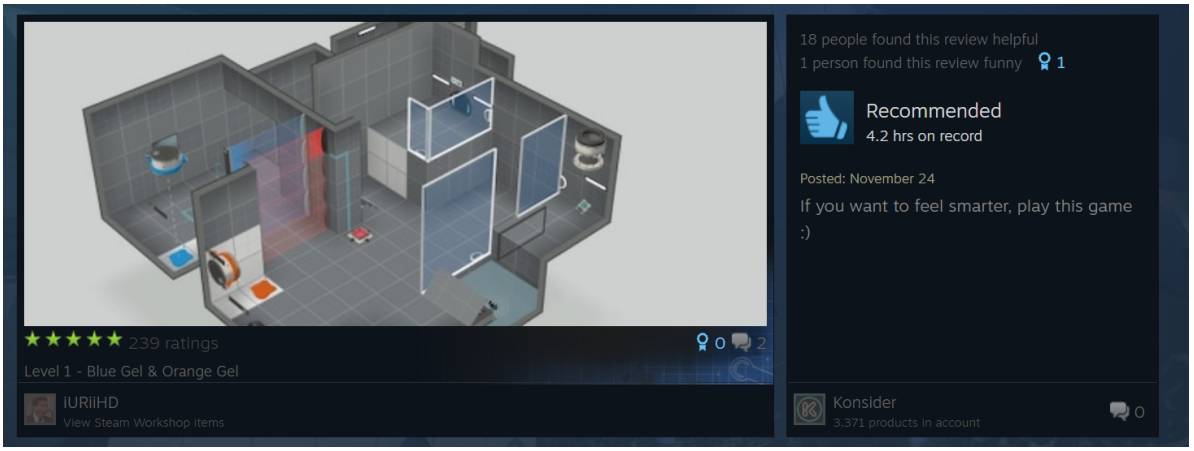


Figure 4: Steam Website Review

Steam Community allows users to post reviews of games once they have played them. Instead of using a 5-star rating system, players are asked to provide their feeling about the game as Recommended , or Not Recommended. The number of playing hours of the reviewed game, the number of games played, and the number of previously posted reviews by the reviewer at this moment are shown aside from reviews. The positive review rate is displayed on the Steam Store page of the game, to advise potential customers.

	date_posted	funny	helpful	hour_played	is_early_access_review	recommendation	review	title
114132	2017-06-15	0	0	290	False	Not Recommended	Going after single player mods? Do you want a ...	Grand Theft Auto V
293152	2017-09-24	1	1	72	True	Recommended	got a headshot with a crossbow 10/10	PLAYERUNKNOWN'S BATTLEGROUNDS
115139	2018-03-24	1	3	4647	False	Not Recommended	After 3466 hours of enjoyable gameplay and hit...	Grand Theft Auto V
205435	2016-11-24	0	0	38	False	Recommended	is nice	Grand Theft Auto V
416865	2017-10-27	0	0	82	True	Recommended	really fun game but you will need friends to h...	Rust

Figure 5: Steam review dataset csv table

Steam Review Dataset is a binary sentiment classification dataset that extracts contents from the steam website containing over hundreds of millions reviews in multiple languages labeled by Steam community members. In our paper, we downloaded steam reviews from kaggle. The dataset contains eight features: date_posted, funny, helpful, hour_played, is_early access review, recommendation, review, title as shown in Figure 5.

```
**************************************************************************************************************************
RECOMMENDATION:
Recommended
--------------------------------------------------------------------------------------------------------------------------
REVIEW:
This game is pretty if you ask me the kids make the game so much worse to play I honestly want to die when I play this game 10/10 would play again
**************************************************************************************************************************
RECOMMENDATION:
Not Recommended
--------------------------------------------------------------------------------------------------------------------------
REVIEW:
THIS GAME BLOWS IN MANY WAYS )IF YOU GET MY DRIFT.
**************************************************************************************************************************
```

Figure 6: Recommended and Not recommended reviews

```
**************************************************************************************************************************
RECOMMENDATION:
Not Recommended
--------------------------------------------------------------------------------------------------------------------------
REVIEW:
I do not recommend this game and I've played for almost 1000hrs now. I play both killer and Survivor to clarify. This game used to be very fun and with friends it was amazing. The Devs have
grown to only want money and do not care about their fanbase. They either nerf killers and call it balanced or nerf survivors and say the same. The game is not balanced and unless you're a
sadist or masochist you will not enjoy this game. They only break the game more with each update and only care about their new game DeathGarden which positively sucks There are also more
hackers now because the Devs do nothing to stop it. They did a "Tournement" which had so many hackers it wasn't funny. They had an announcer who knew absolutely nothing about the game and
refused to listen when people callled out hackers. My friend was banned from a Dev's stream when telling some issues with a killer he was having. There is no longer a way to contact the
devs and the game is slowly going dowhill quickly. Some suggestions if I may.1. Instead of putting more cosmetics in the game actually fix the bugs that are present. Cosmetics mean nothing
if you're permenently stuck in one place due to a bug. 2. reach out to the players. You're making people wanna stop playing with the ranking system. A new killer should not be with rank 10
survivors and vise versa. Also with the hackers it makes others who are veterans not want to play this game. 3. fix the balancing. You removed pallets but people still get face camped and
looped. The toxcicity of the players are from them. So with balancing fix the exhaustion perks debuff. If you're running balanced landing you should be able to recover from exhaustion when
you jump down. You will eventually get cornered and killed. Also fix the whole body blocking scenario. If a killer can block you from moving that's a major flaw. this goes for survivors
being able to trap killers in "fatshame" places. 4. Fix the EAC. Make it so it actually stops hackers. Simple5. Updates should focus on fixing the game not ruining it. With every update as
I've said it breaks the game. It tips it to one side each time. Killers should not be able to face camp. Maybe fix that in your very important COSMETICS Update. Cause actually fixing the
game is totally unnecessary right?There's more top fix but Devs don't care about their players so they won't bother reading this
**************************************************************************************************************************
RECOMMENDATION:
Recommended
```

Figure 7: Long Review from Steam (Outliers)

Feature Design Diagram

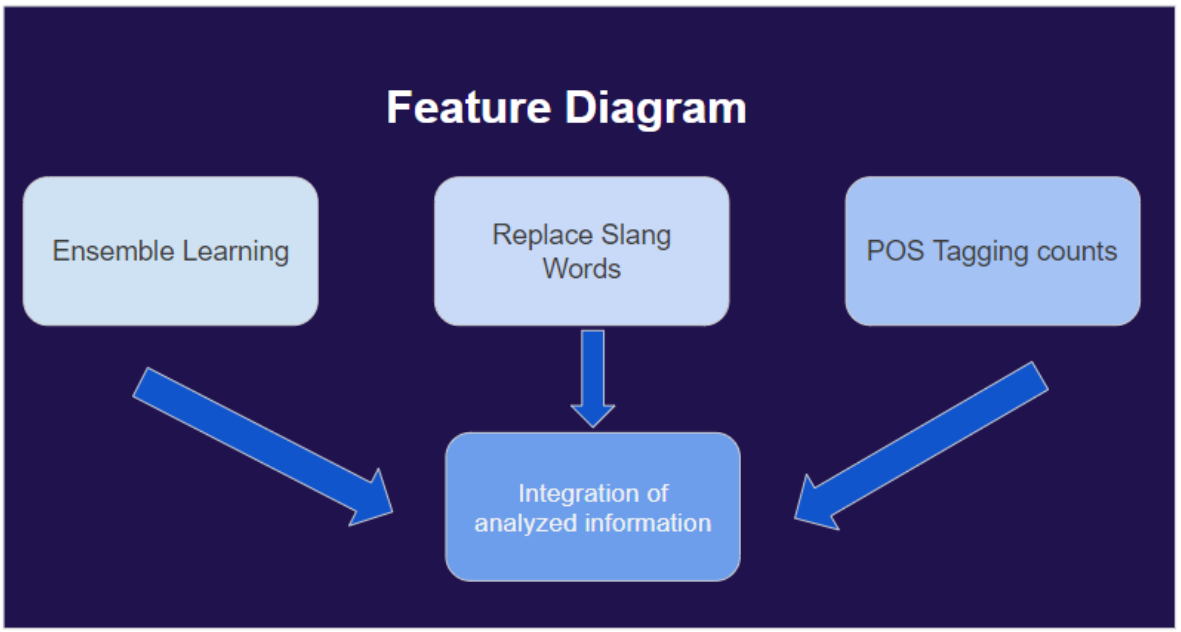


Figure 8: Feature diagram

Analysis of Data

1. Distribution of recommended/not recommended instances

```
Recommended          7044
Not Recommended      2956
Name: recommendation, dtype: int64
```

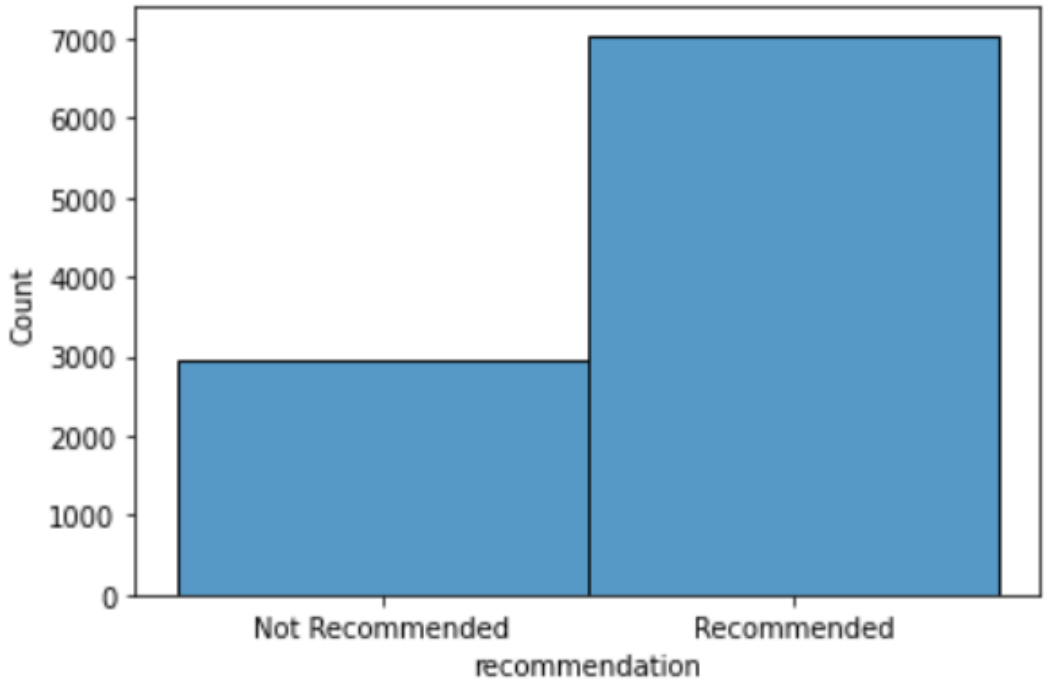


Figure 9: distribution of Recommended and Not Recommended instances

2. Minimum and Maximum words per review per label

```
Max number of words per review in "Not Recommended": 800
Min number of words per review in "Not Recommended": 0
Mean number of words per review in "Not Recommended": 29
------------------------------------------------------------
----------
Max number of words per review in "Recommended": 781
Max number of words per review in "Recommended": 0
Mean number of words per review in "Recommended": 19
```

Figure 10: Minimum and Maximum words in Recommended and Not Recommended reviews

Based on the Max and Mean number of words in each class, we can see that

our data appears to be skewed and we could have "outliers" (see Figure 7)- that is, reviews with an unusually large number of words. These reviews are typically long because they go in depth with the specific game they are describing. This can become unnecessary information as describing what actually happens in the game does not necessarily dictate whether the reviewer actually enjoyed the game or not. Let's take a look at what our n_words columns look like on a box and whisker plot. From there, we will take a closer look at these video game reviews.

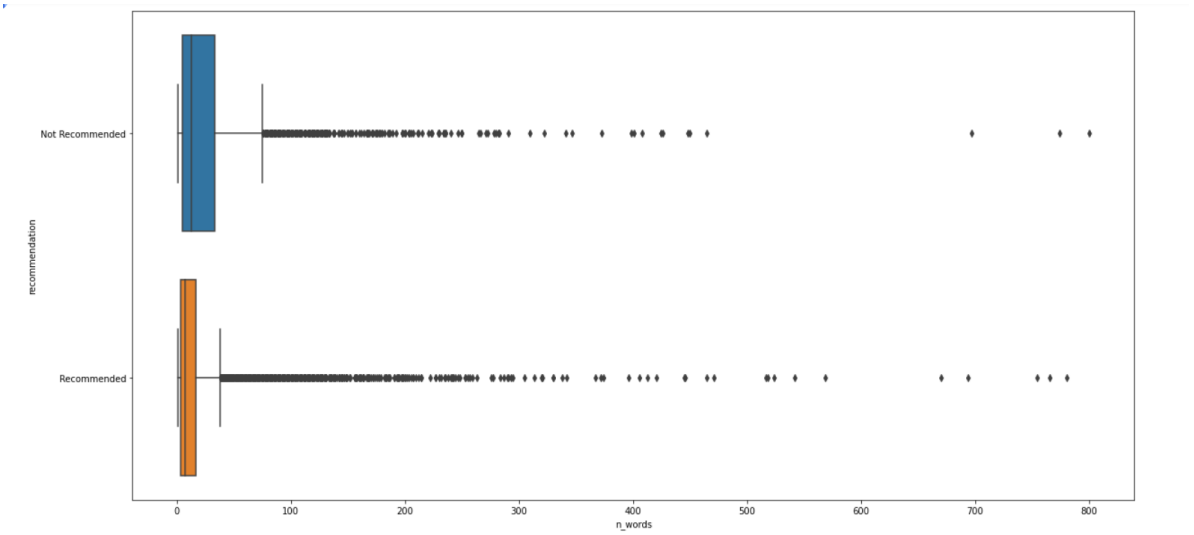


Figure 11: Number of words of each reviews

After taking a closer look, we chose to keep the reviews because although they contain a lot of fluff describing the game, they do so by using words that we believe would help a model decide whether or not a game is recommended. For example, "It's even worse for me because the game doesn't load at all for me and I have to close it from task manager, start it again and reconnect.All of the above issues have been in the game for MONTHS and nothing has been done to fix or even address them. Don't waste your time and nerves on this game".

3. Word frequency recommended and not recommended review

```
Unique Words in Recommended: 8901
Unique Words in Recommended: 7041
```

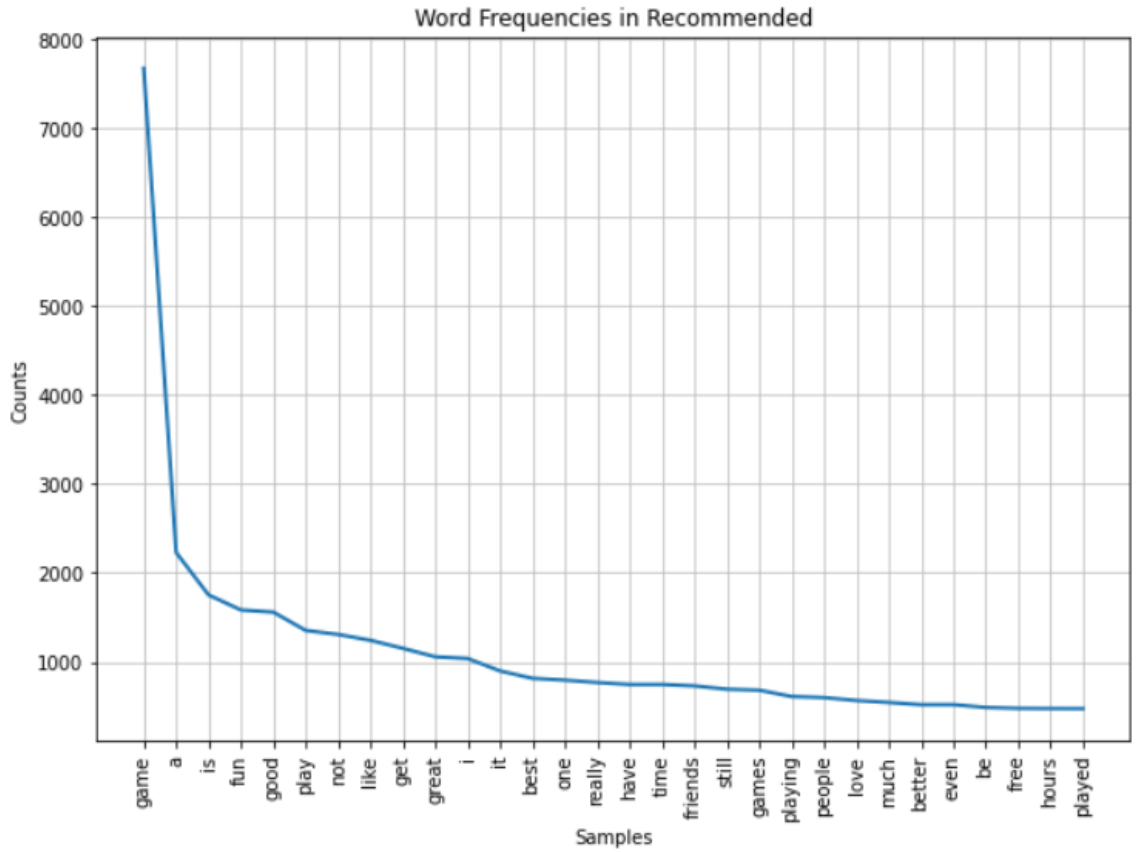


Figure 12: Word frequency in Recommended review

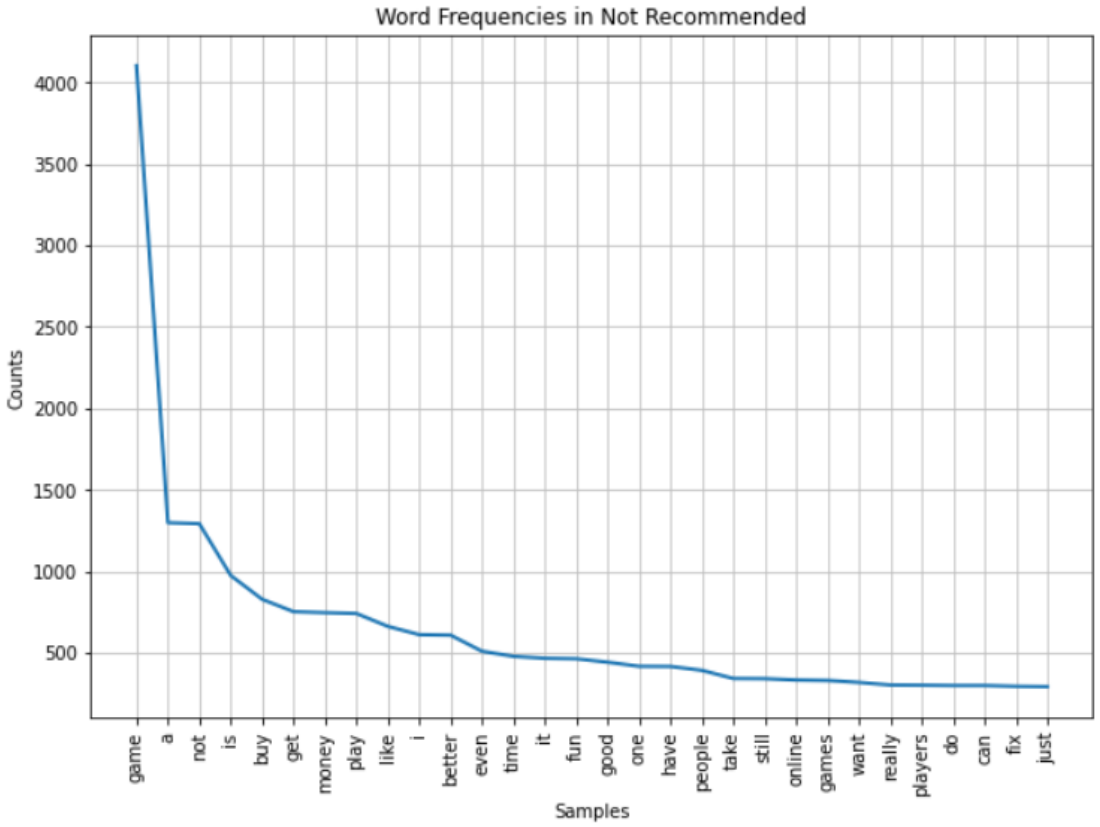


Figure 13: Word frequency on Not Recommended reviews

After taking a look at our word frequencies, there is an issue we need to address. The word "game" is mentioned the most in both reviews. Given that we are analyzing game reviews, this is obviously expected. However, we will opt out of including the word since "game" is mentioned in both classes the most.

4. Most common words per label

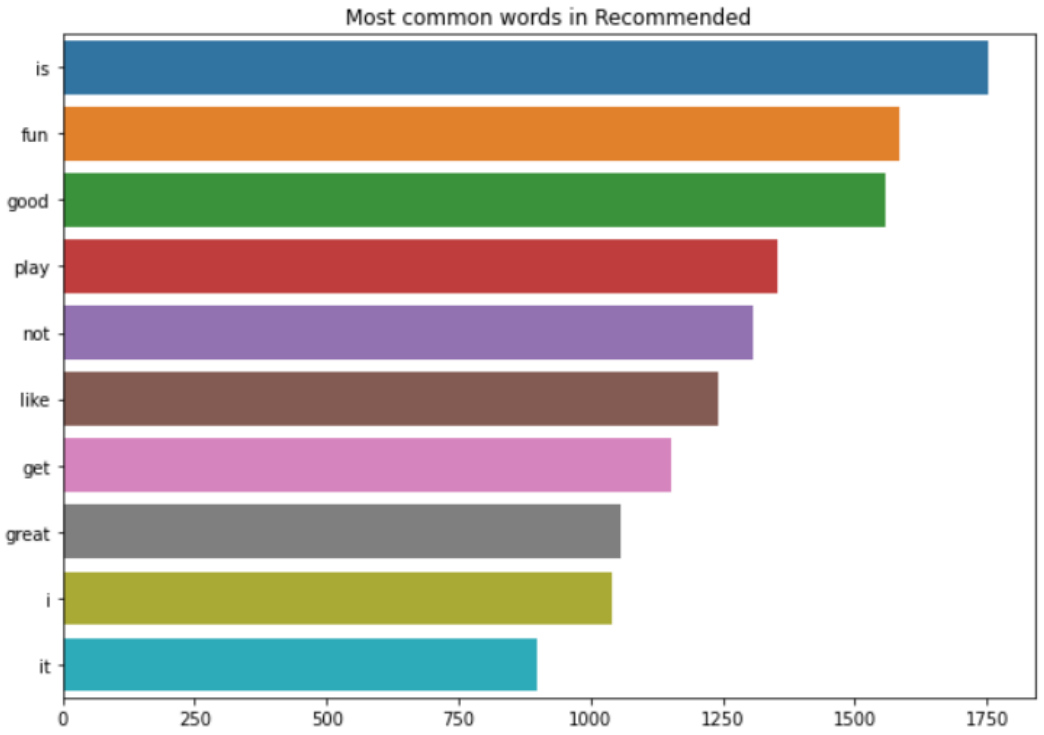


Figure 14: Most frequent words in Recommended reviews.

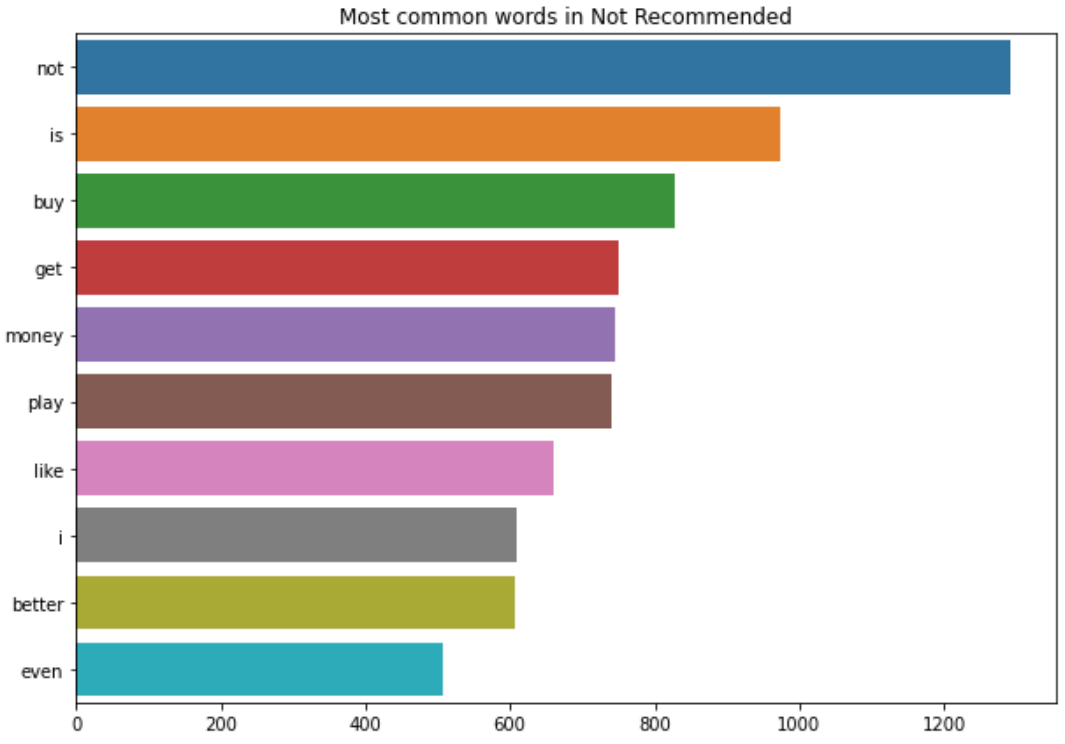


Figure 15: Most frequent words in Not Recommended reviews.

Model

Project Innovation: Given that our project consisted of only extracting features from text alone, we could not make a direct comparison of the models that were mentioned in the paper that used the Steam Reviews Dataset. As a result, we tried our best to make a comparison by using the same models and only the TF-IDF features used in the paper (other features from the dataset such as “helpful” and “funny” features were used in the paper, but because an IGN review does not have either of these features, we simply stuck with TF-IDF).

For our innovation, we tried to replace modern video game slang words with their standard vocab pair (e.g. “gud” - “good”, “poggers” - “awesome”). We felt that by switching our these words would help improve the importance of given words based on our dataset. Additionally, we tried to include the POS Tagging counts as supplementary features in hopes that this would boost performance. Finally, we included three additional models that were not included in the paper: Random Forest, Ensemble Learning with Soft Voting, and Ensemble Learning with Hard Voting.

In total, we used 2 model variations for all 6 models (Naive Bayes, Linear SVC, Log Regression, Random Forest, Ensemble Learning Soft/Hard voting), the first 6 were models used slang replacement and POS Tagging count as a feature in addition to TF-IDF. The last 6 simply used the TF-IDF as a feature.

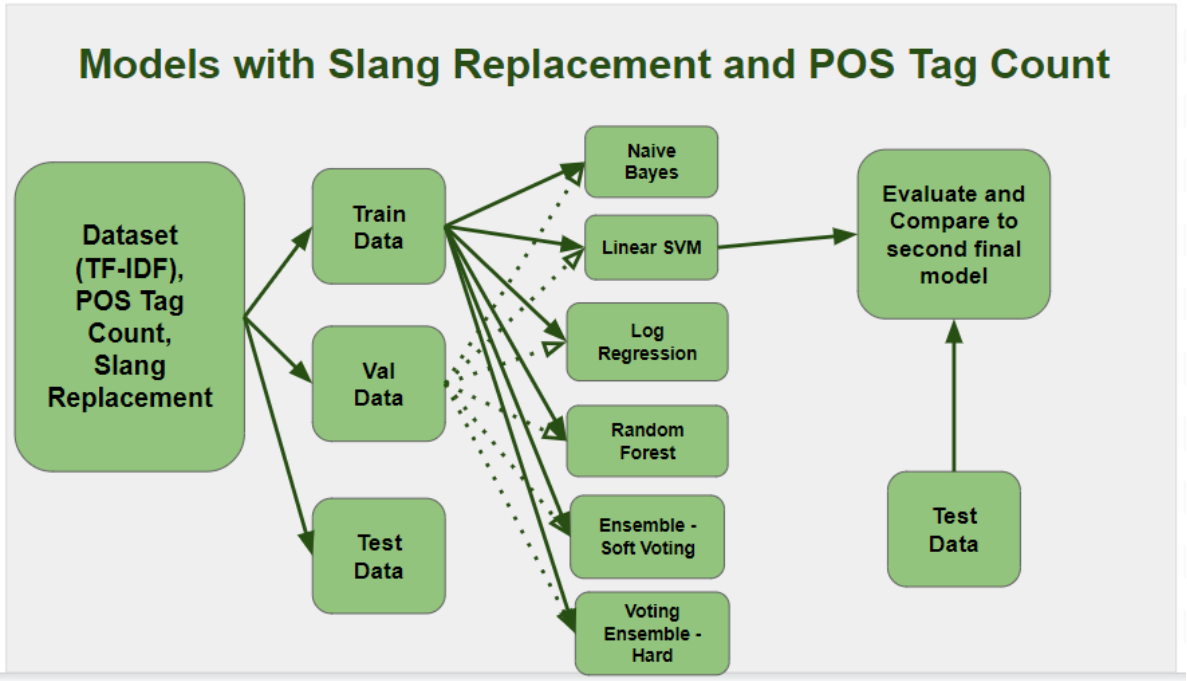


Figure 16: Variation 1 Model design

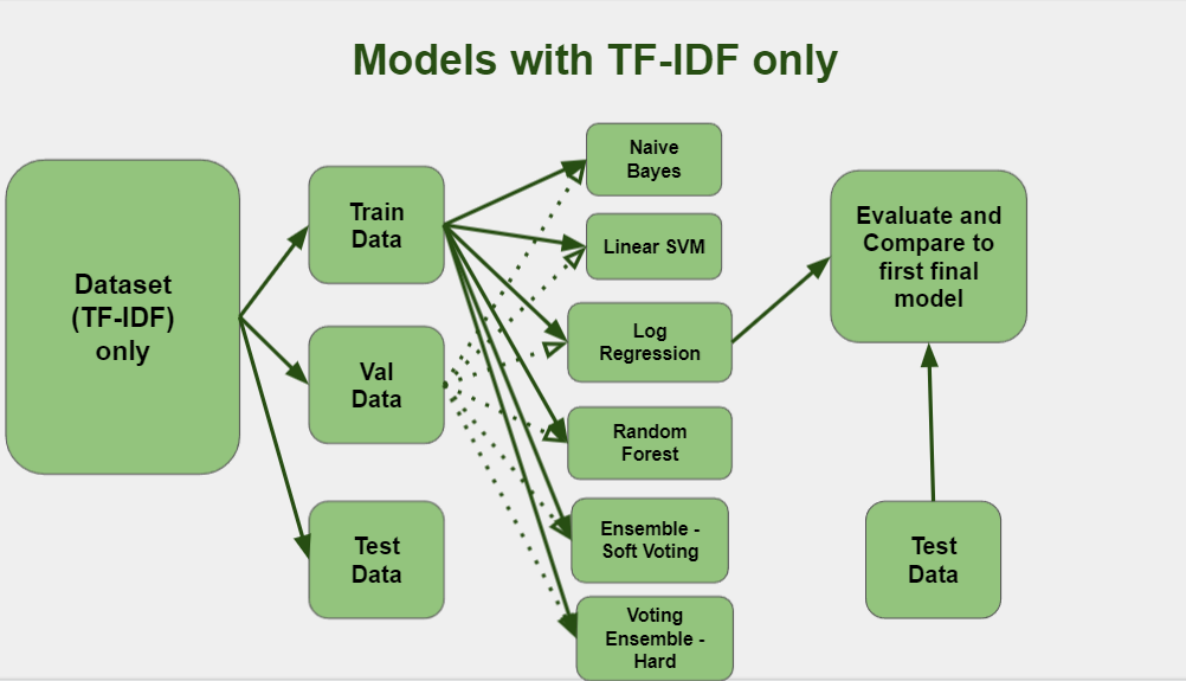


Figure 18: Variation 2 Model design (TF-IDF only)

We include the original models that were mentioned and used in the previous research paper (Naive Bayes, SVC, Logistic Regression). In our project design, we decided to include some additional models:

- Random Forest
- Ensemble Learning - Hard Voting (SVC, Log Reg, and Random Forest)
- Ensemble Learning - Soft Voting (Naive Bayes, Log Reg, and Random Forest)

Dataset 2: [WikiHow]

WikiHow contains more than 230,000 article and summary pairs extracted from the [wikihow online knowledge base] written by various authors, therefore each article is constructed with a wide range of topics and styles from existing news. The advantage of the dataset is its large-scale data and its diversity and complications. Each article contains multiple paragraphs starting with a sentence summarizing it, the summary thus formed by outlining paragraphs.

Figure 19: WikiHow website

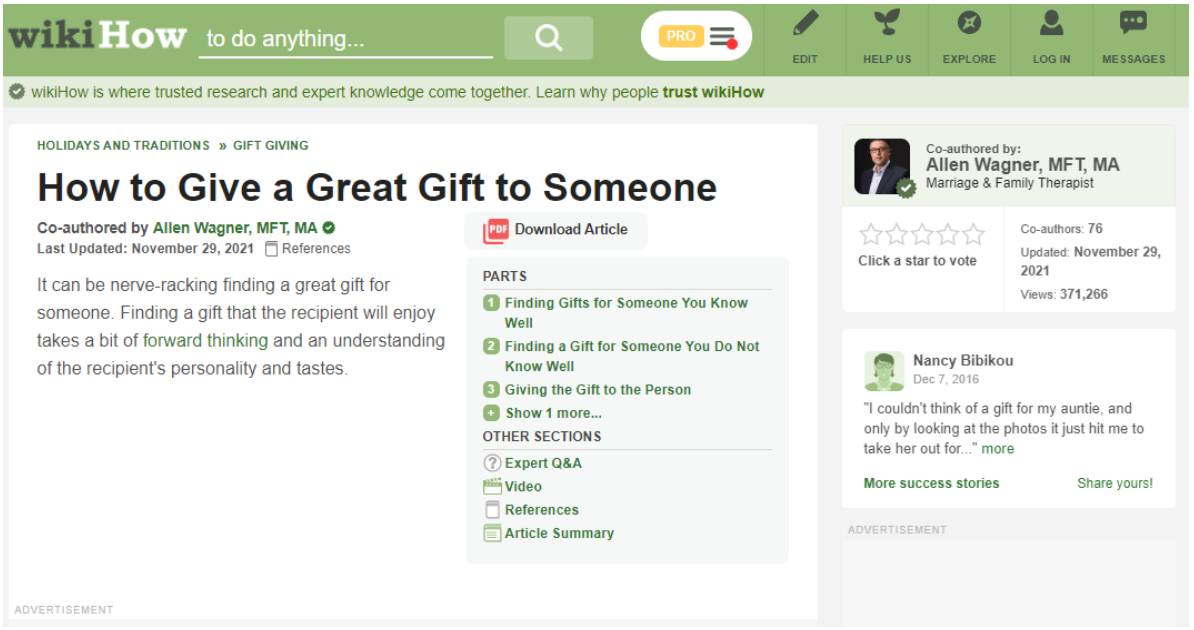


Figure 20: An article summary, its author, and the whole article to be downloaded

The title of articles start from “How to”, and some articles describe single method tasks in steps, whereas some describe different methods in steps for a specific task.

	headline	title	text
0	\nKeep related supplies in the same area.,\nMa...	How to Be an Organized Artist1	If you're a photographer, keep all the necess...
1	\nCreate a sketch in the NeoPopRealist manner ...	How to Create a Neopoprealist Art Work	See the image for how this drawing develops s...
2	\nGet a bachelor’s degree.,\nEnroll in a studi...	How to Be a Visual Effects Artist1	It is possible to become a VFX artist without...
3	\nStart with some experience or interest in ar...	How to Become an Art Investor	The best art investors do their research on t...
4	\nKeep your reference materials, sketches, art...	How to Be an Organized Artist2	As you start planning for a project or work, ...

Figure 21: The dataset detail, headline indicates summary, title indicates the content title, and text indicates the article content.

text from splitted test file “test.iloc[188][0]”	Choose a sunflower with bright, undamaged petals and pluck these petals off one by one using your fingers., Keep a layer or two of blotting paper above and below the petals. Place these sheets of paper, with the sunflower petals still inside, in between to pieces of stiff cardboard. If you do not have blotting paper, you could use tracing paper, parchment paper, or clean paper towels, instead. None of these options will work quite as well as blotting paper, but they are decent alternatives.

	Make sure that the petals do not touch one another. If they touch or overlap, they may end up sticking together, and you could cause the petals to break when you attempt to peel them apart. Place the petals-filled pieces of cardboard underneath a stack of books or other heavy objects. Alternatively, you can create an even more secure flower press with a few layers of cardboard and wooden press boards.If you do so, you will not need to place the petals beneath a heavy stack of books, and you can move the structure around as needed while the petals dry out. Place cardboard or padded board below the blotting paper sandwich securing the petals and a second padded board above it. Place a third padded board on top of the others. Sandwich everything between two wooden press boards and hold the pressing structure together with rubber bands. Press the sunflower petals for a few weeks, disturbing them as little as possible during the process. When you do check the petals after two or three weeks, remove the cardboard and blotting paper carefully and pick the petals up gently. If the petals still feel moist, lay out new blotting paper and continue pressing them for another few days to another week before checking again.
Its headline(summary) test.iloc[188][1]	Collect the petals., Place the petals in between sheets of blotting paper., Put pressure on the flower petals., Let dry for several weeks.

Figure 22: A splitted test sample of text(article) and its summary

Analysis of Data

1. Visualize some random text and its corresponding headline(reference summary)

```
Text:
 It's the black, apple-shaped icon in the upper-left corner of your screen., It's in the second section of the drop-down menu.

, It's on the left side of the window.

, It's in the lower-left corner of the window.

All of your iCloud data, including calendar entries and iCloud photos, will be removed from the Mac.


Headline:

Click on the Apple menu.,
Click on System Preferences.,
Click on iCloud.,
Click on Sign Out.
```

Figure 23: A splitted test sample of text(article) and its summary

2. Sentence Count — Total number of sentences in the text

	headline	title	text	headline_ns_count	text_ns_count
0	\nKeep related supplies in the same area.,\nMa...	How to Be an Organized Artist1	If you're a photographer, keep all the necess...	1	31
1	\nCreate a sketch in the NeoPopRealist manner ...	How to Create a Neopoprealist Art Work	See the image for how this drawing develops s...	1	31
2	\nGet a bachelor's degree.,\nEnroll in a studi...	How to Be a Visual Effects Artist1	It is possible to become a VFX artist without...	1	21
3	\nStart with some experience or interest in ar...	How to Become an Art Investor	The best art investors do their research on t...	1	46
4	\nKeep your reference materials, sketches, art...	How to Be an Organized Artist2	As you start planning for a project or work, ...	1	23
5	\nKeep all of your past work organized and acc...	How to Be an Organized Artist3	When you finish a project, whether it sells o...	1	26
6	\nCreate a compelling reel or portfolio.,\nLan...	How to Be a Visual Effects Artist2	This should be a short video showcasing the b...	1	18
7	\nJoin a professional society.,\nEnjoy working...	How to Be a Visual Effects Artist3	Networking is a great way to find new opportu...	1	18
8	\nMake sure you know what is expected of you.,...	How to Be Good at Improvisation	Some entire movies are improvised, some plays...	1	33
9	\nMake a list of what your friends watch, read...	How to Always Catch Pop Culture References1	Use your friends' conversations to figure out...	1	31

Figure 24: The total number of sentence in each text can be seen in column 'text_ns_count'

3. Word Count — Total number of words in the text

	headline	title	text	headline_ns_count	text_ns_count	headline_word_count	text_word_count
0	\nKeep related supplies in the same area.,\nMa...	How to Be an Organized Artist1	If you're a photographer, keep all the necess...	1	31	98	696
1	\nCreate a sketch in the NeoPopRealist manner ...	How to Create a Neopoprealist Art Work	See the image for how this drawing develops s...	1	31	153	702
2	\nGet a bachelor's degree.,\nEnroll in a studi...	How to Be a Visual Effects Artist1	It is possible to become a VFX artist without...	1	21	48	512
3	\nStart with some experience or interest in ar...	How to Become an Art Investor	The best art investors do their research on t...	1	46	138	1023
4	\nKeep your reference materials, sketches, art...	How to Be an Organized Artist2	As you start planning for a project or work, ...	1	23	75	505

Figure 25: The total number of words in each text can be seen in column 'text_word_count'

4. Find maximum of headline number of words and maximum of text number of words

```
maximum number of words in headline is:5465
maximum number of words in text is:13837
```

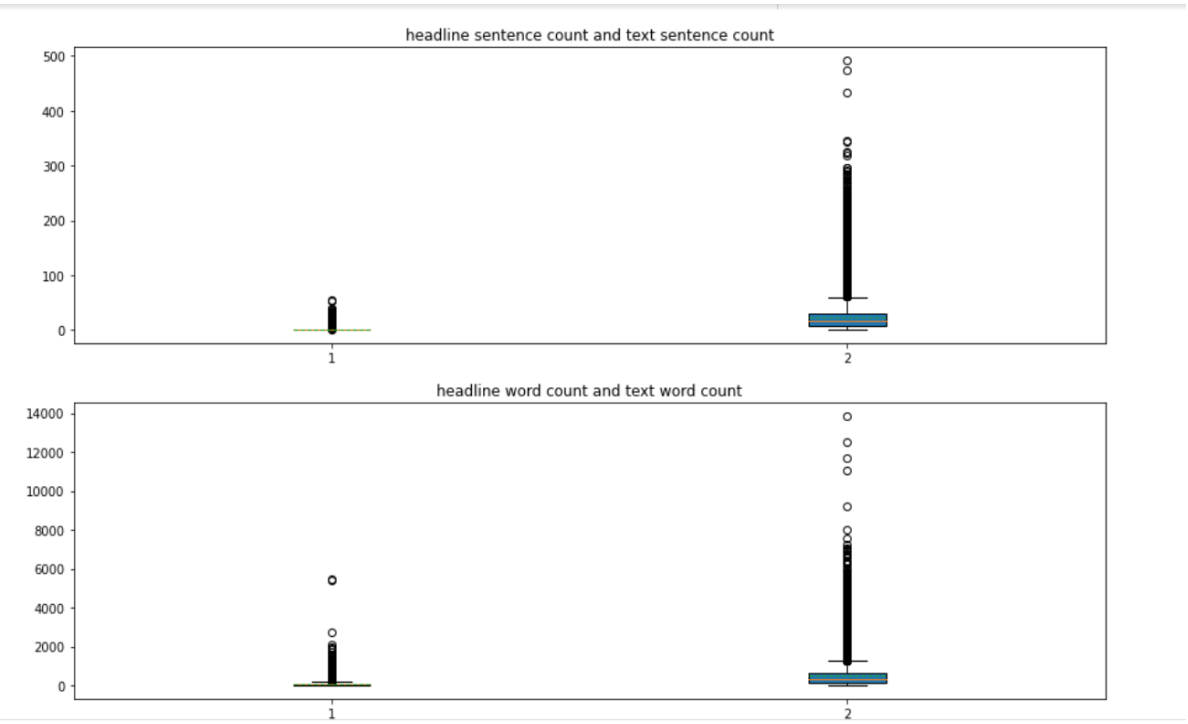
Figure 26: Maximum number of words in headline and the maximum number of words in text

5. Sentence density — Number of sentences relative to the number of words

	headline	title	text	headline_ns_count	text_ns_count	headline_word_count	text_word_count	headline_sen_density	text_sen_density
0	\nKeep related supplies in the same area.,\nMa...	How to Be an Organized Artist1	If you're a photographer, keep all the necess...	1	31	98	696	0.010101	0.044476
1	\nCreate a sketch in the NeoPopRealist manner ...	How to Create a Neopoprealist Art Work	See the image for how this drawing develops s...	1	31	153	702	0.006494	0.044097
2	\nGet a bachelor's degree.,\nEnroll in a studi...	How to Be a Visual Effects Artist1	It is possible to become a VFX artist without...	1	21	48	512	0.020408	0.040936
3	\nStart with some experience or interest in ar...	How to Become an Art Investor	The best art investors do their research on t...	1	46	138	1023	0.007194	0.044922
4	\nKeep your reference materials, sketches, art...	How to Be an Organized Artist2	As you start planning for a project or work, ...	1	23	75	505	0.013158	0.045455

Figure 27: Number of sentences relative to the number of words can be seen in column 'text_sen_density'

Visualize our observation in graphs:



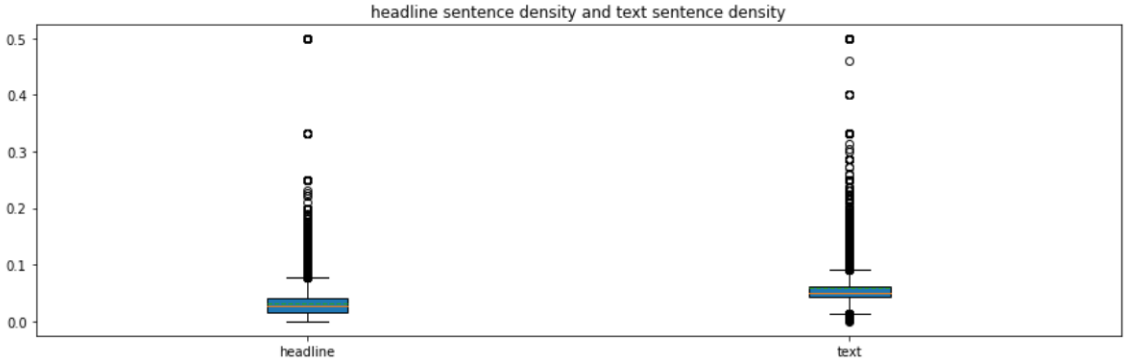


Figure 28: Visualization of previous statistic numbers

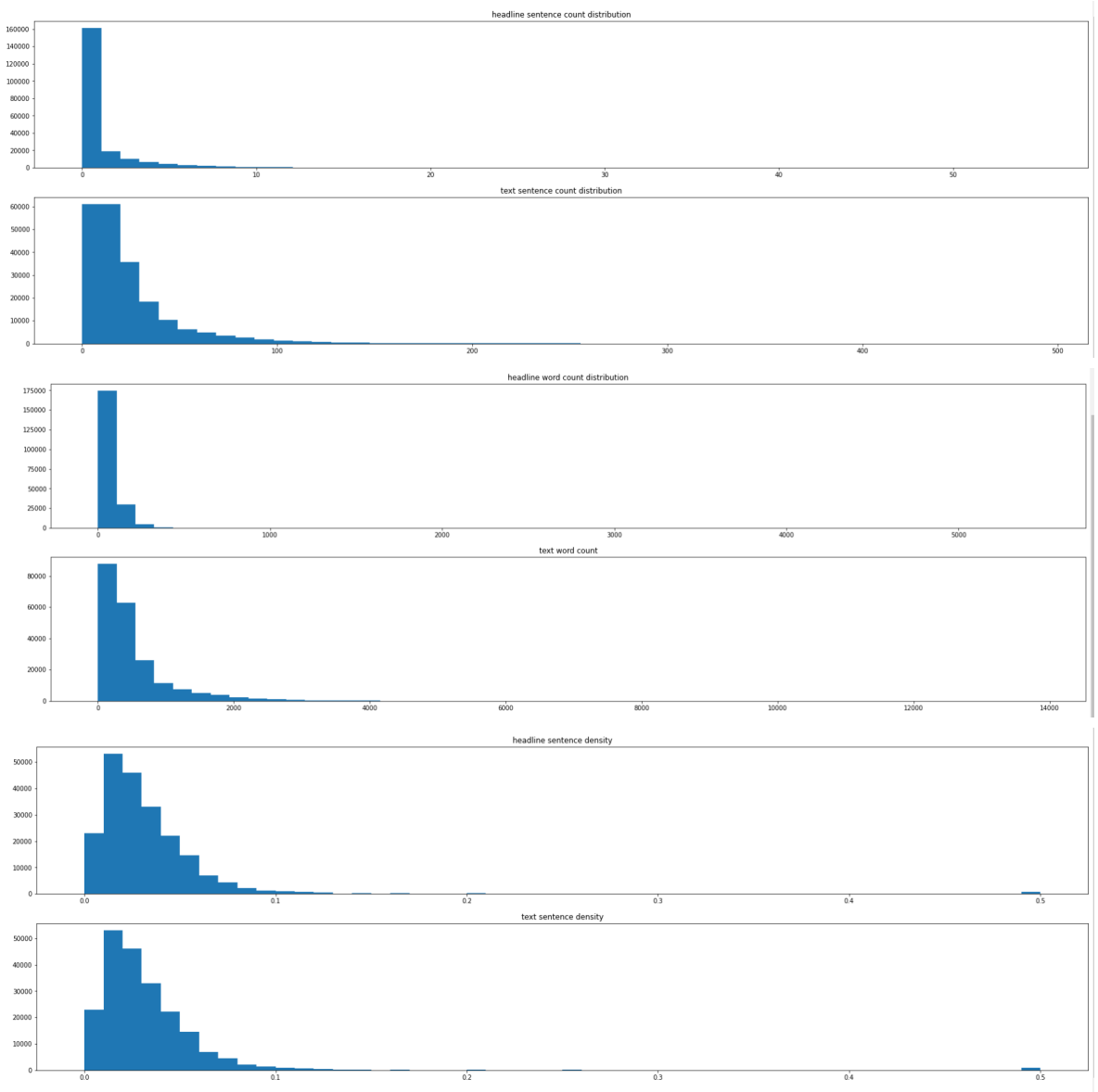


Figure 29: Visualization of previous statistic numbers

Model

We fine-tuned the T5-Text-to-Text-Transfer-Transformer model to perform our text summarization task. The model proposes reframing all NLP tasks into an unified text-to-text format where both its input and output are text strings. This format makes the model to perform multiple tasks including our desired summarization. We then utilize the advantage of this model setting, specifically train this model to a specific format summarization style to fit our needs.

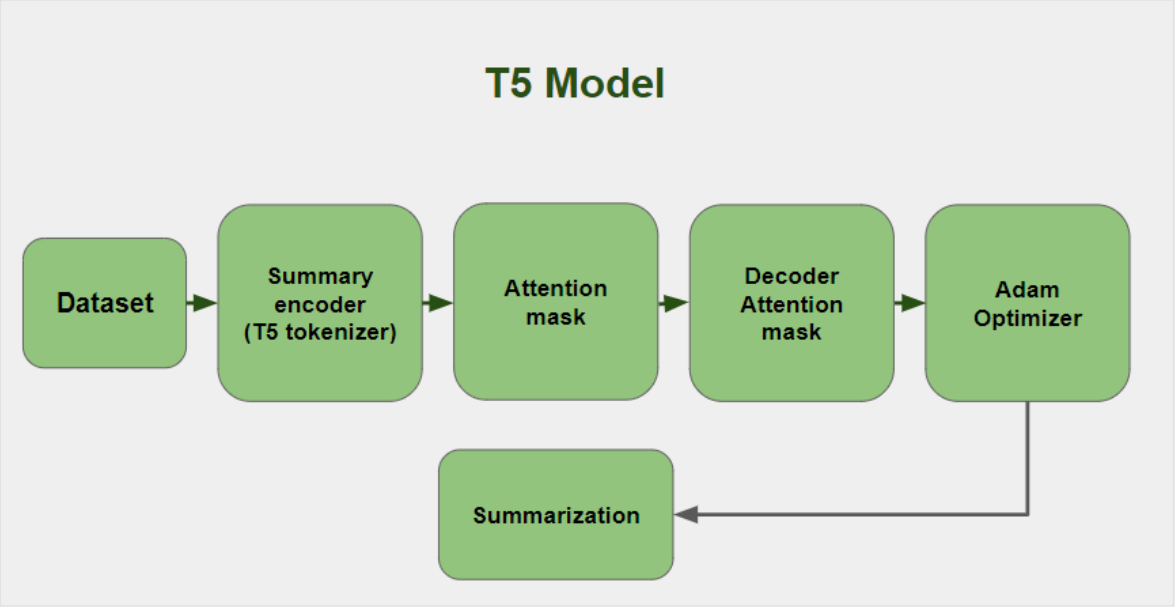


Figure 30: T5 transformer model in our task

Dataset 3: LJ Speech Dataset

LJ Speech dataset contains 13100 short audio clips of a single speaker reading paragraphs from seven non fiction books. The texts were public domain and recorded in 2016 to 2017 by the LibriVox project. The data samples are audios, so we cannot insert in this documentation therefore, please click the link to listen to the sample. [sample link]

As we can see in Figure 29 below, each clip is provided with a transcription. The transcript is stored in metadata and the sound track in the wav clips. Each clip varies in

length from 1 to 10 seconds and adds up to a total length of approximately 24 hours.

Analysis of Data

This is a audio dataset, so first we look at its metadata.csv

LJ001-000	in the onl	differs frc	in the onl	differs from most if not from all the arts and crafts represented in the Exhibition					
LJ001-0002	in being comparatively modern.	in being comparatively modern.							
LJ001-000	by a simil	by a similar process							
LJ001-000	which we	produced	which were the immediate predecessors of the true printed book						
LJ001-0005	the invention of movable metal letters in the middle of the fifteenth century may justly be considered as the invention of the art of printing.	the invention c							
LJ001-000	as an exa	And it is v	as an example of fine typography						
LJ001-000	the Guter	or "forty-	the earlie	the Guter	or "forty-two line Bible" of about fourteen fifty-five				
LJ001-0008	has never been surpassed.	has never been surpassed.							
LJ001-000	then	for our pu	may be cc	then	for our pu	may be considered as the art of making books by means of movable types.			
LJ001-001	as all boo	Now	as all books not primarily intended as picture-books consist principally of types composed to form letterpress						
LJ001-0011	it is of the first importance that the letter used should be fine in form;	it is of the first importance that the letter used should be fine in form;							
LJ001-001	or cost in	in casting	setting	or printin	or cost in	in casting	setting	or printing beautiful letters	
LJ001-0013	than in the same operations with ugly ones.	than in the same operations with ugly ones.							
LJ001-001	when the	And it wa	when the craftsmen took care that beautiful form should always be a part of their productions whatever they were						
LJ001-001	and that t	and that their arrangement on the page should be reasonable and a help to the shapeliness of the letters themselves.							
LJ001-001	and it wa	and it was natural therefore							
LJ001-001	and they	and they followed them very closely							

Figure 31 LJSpeech dataset metadata csv content

```
There are 1 directories and 2 files in 'LJSpeech-1.1/'.
There are 0 directories and 13100 files in 'LJSpeech-1.1/wavs'.
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8750 entries, 0 to 8749
Data columns (total 5 columns):
 #   Column                                                                                Non-Null Count  Dtype
---  ------                                                                                --------------  -----
 0   LJ001-0001|Printing                                                                   8750 non-null   object
 1    in the only sense with which we are at present concerned                             4972 non-null   object
 2    differs from most if not from all the arts and crafts represented in the Exhibition|Printing  4080 non-null   object
 3    in the only sense with which we are at present concerned.1                           1933 non-null   object
 4    differs from most if not from all the arts and crafts represented in the Exhibition  1151 non-null   object
dtypes: object(5)
memory usage: 341.9+ KB
```

Figure 32 LJSpeech dataset folder content

Because the constituent of the dataset and the group method of the metadata file, let's visualize of a random audio from the dataset:

1. Plot the audio array and display the waveplot
   We use librosa to plot the 'LJ037-0171.wav' file, and zoom in with Zero-Crossing Rate.

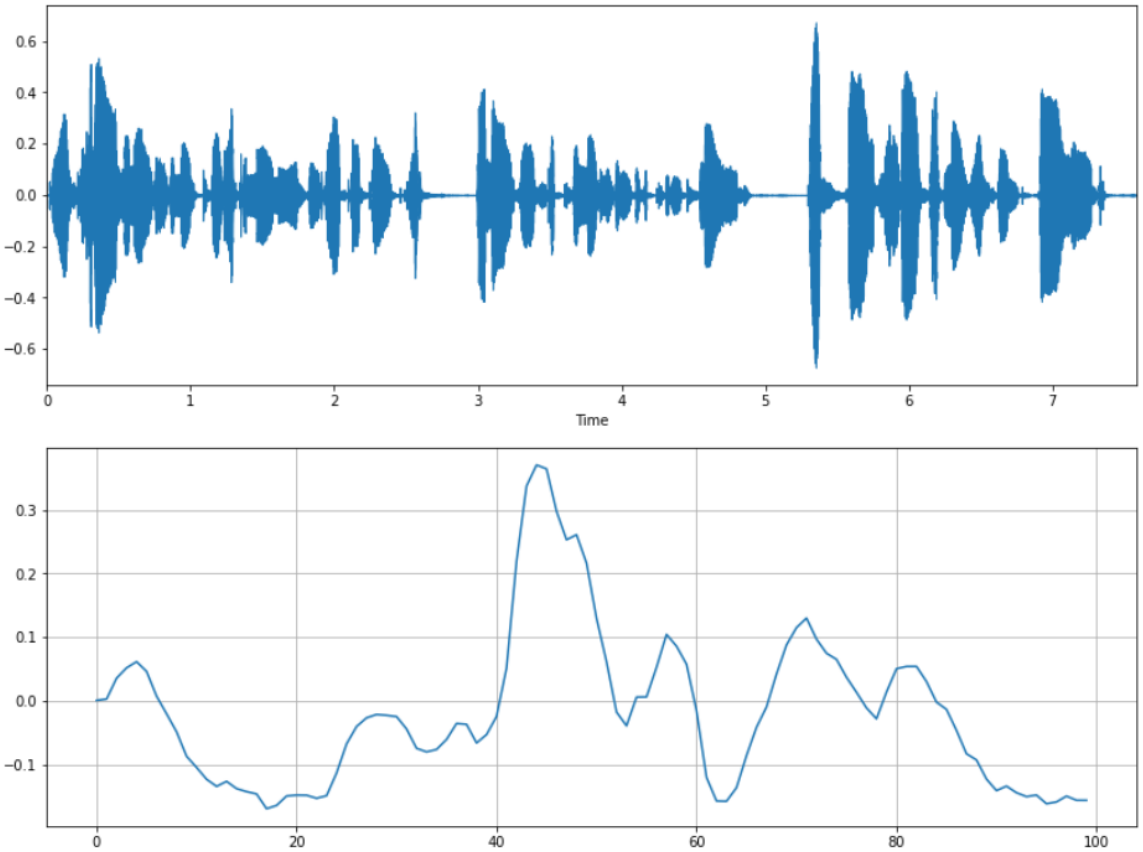


Figure 33: speech signal wave and its Zero-Crossing Rate plot

2. Present the speech signal in a spectrogram to visualize the strength of a signal over time at various frequencies.

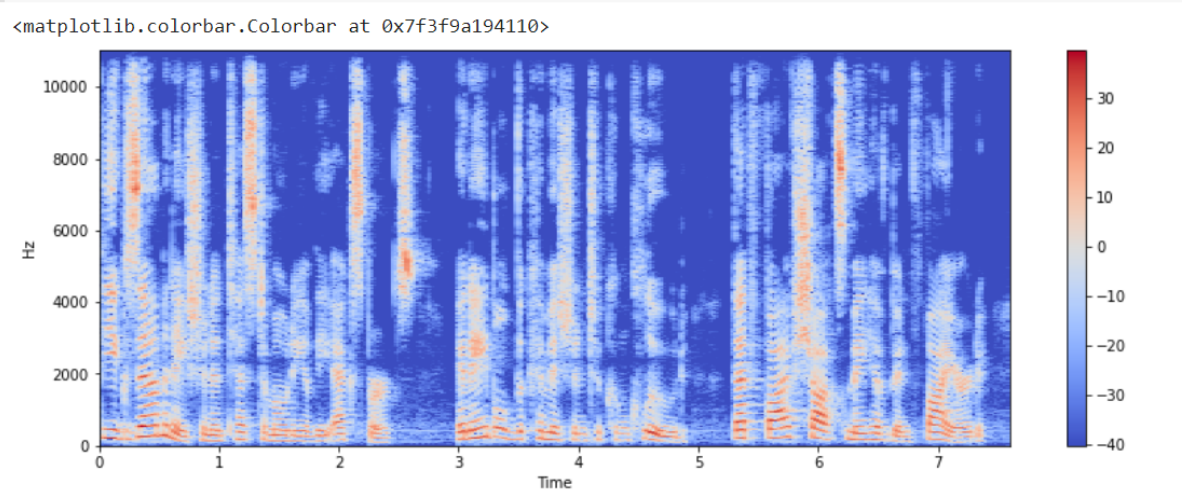


Figure 34: A spectrogram of the speech signal

3. The vertical axis shows the frequencies and the horizontal axis shows the time of the speech signal.

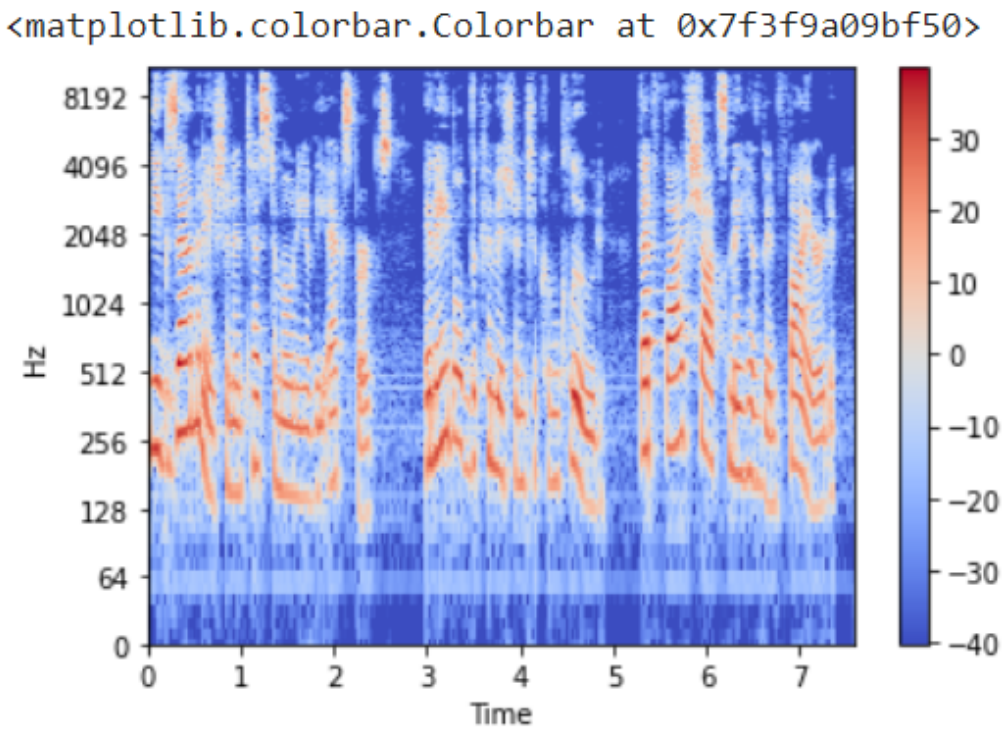


Figure 35: A spectrogram of the speech signal

4. Spectral Centroid plot of the speech signal

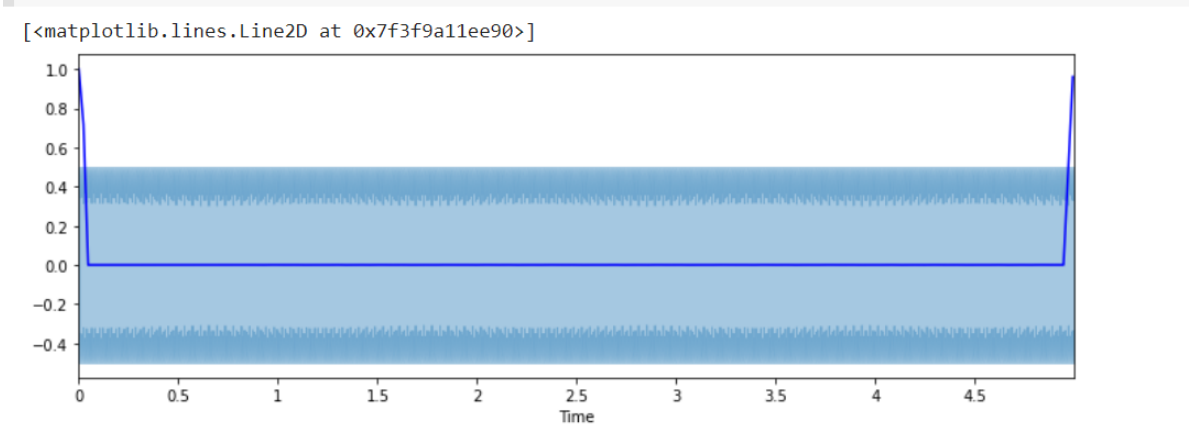


Figure 36: Spectral centroid for each frame in the speech signal

Model

For the Text-to-Speech model, we install many libraries, such as phonemizer, webrtcvad, pyworld to perform sound synthesis. Utilize the pretrained Transformer TTS model and fine tune it with LJSpeech dataset for a crispy and calm speech tone to meet our need for our

system.

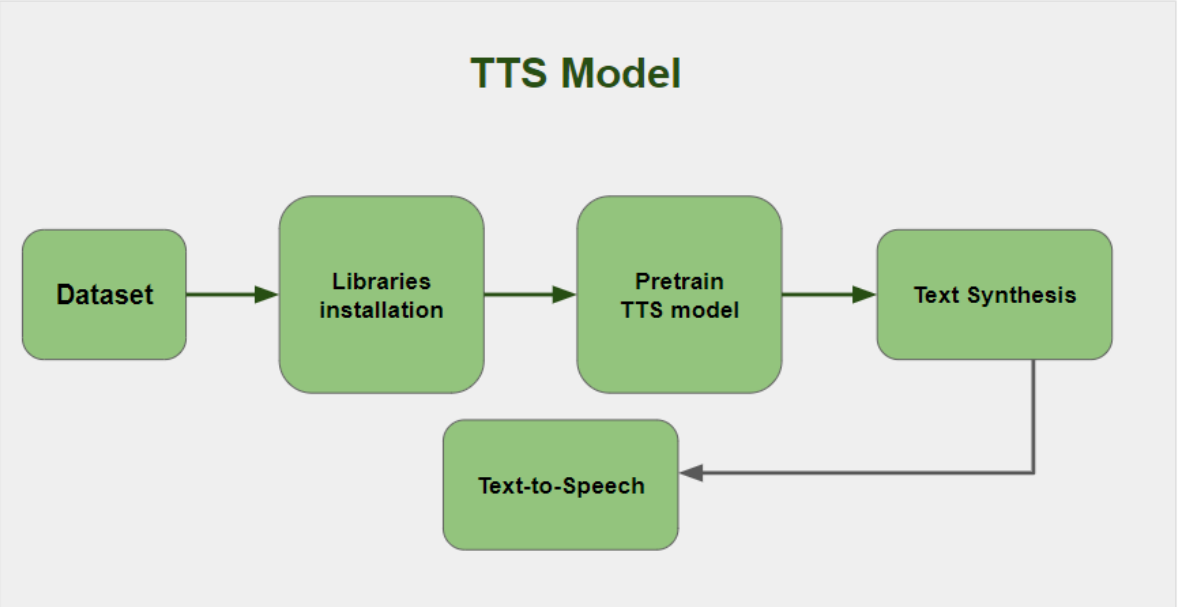


Figure 37: Text-to-Speech Model

Implementation

Pseudocode:

1. Write a Web scraper
2. Clean text and preprocessing the text for Text Summarization Model analyzation
    a. T5 model implementation
    b. Fine tune with WikiHow dataset
3. Get Summary and input into Text Classification model
    a. The first model section: all models using Slang replacement and POS Tagging as a feature
        i. Additional Random Forest Model

ii. This also includes two different ensemble models for soft and hard voting

b. The second model section: using the TF_IDF feature (Baseline)

i. Additional Random Forest model included

ii. The two different ensemble models are used as well

4. Perform Text-to-Speech on the Summary get from Text Summarization Model
    a. implement Transformer TTS model
    b. Fine tune with LJ-Speech dataset
5. Save and integrate models above and perform the IGN website review analysis.

Explanation of Implementation:

For our text classification, we trained six different models with 2 variations (TF-IDF only feature vs. slang replacement and POS Tag Count) and performed cross validation for each model with hyper-parameter tuning. After choosing our best model for each variation, we performed one final comparison against each variation.

Here are the models we work with:

1. Naive Bayes
2. Linear SVM
3. Logistic Regression
4. Random Forest
5. Ensemble Learning
    1. Linear SVM, Logistic Regression, Random Forest - Hard Voting
    2. Naive Bayes, Logistic Regression, Random Forest - Soft Voting

Result and Result Confusion Metrics:

Text Classification Results:

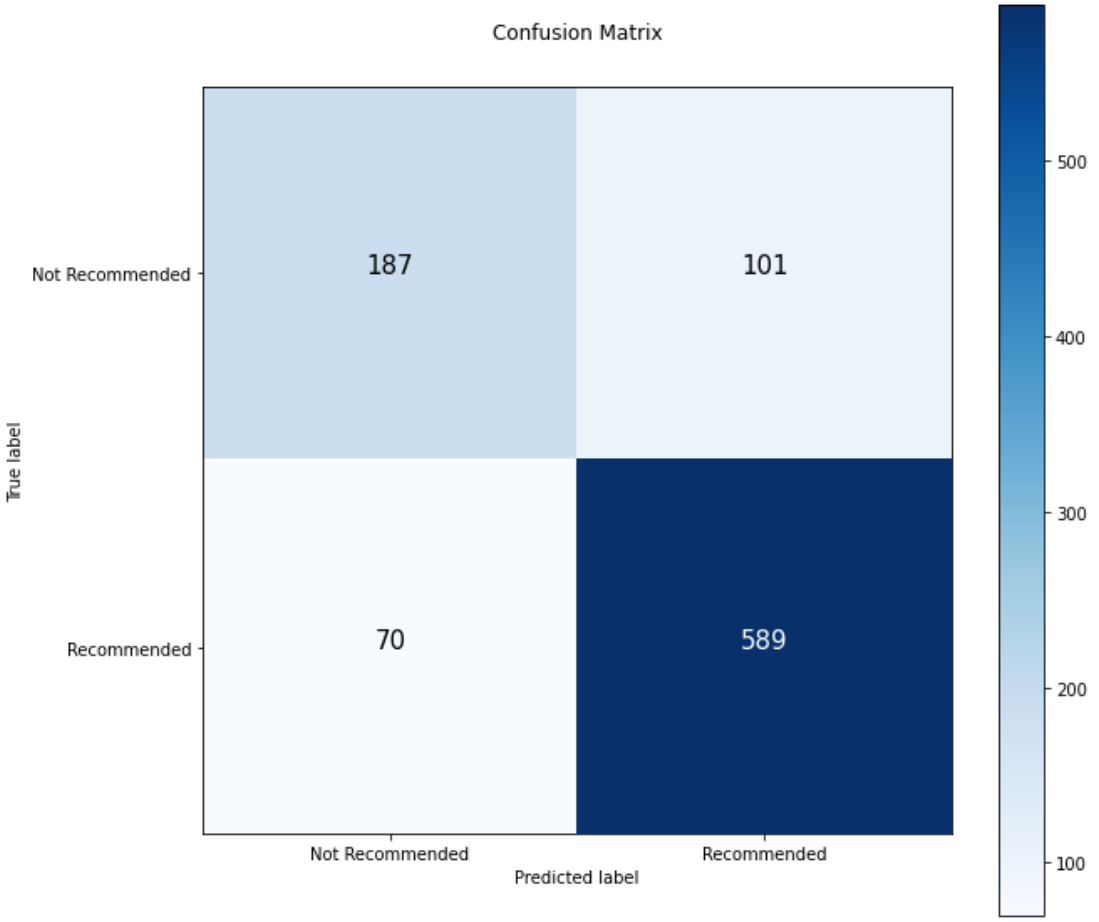
After comparing the model performance metrics, it appears that our replacement of slang in the video game reviews and inclusion of POS Tagging did not help improve performance. Ensemble learning, although performing on par with the rest of the models in both cases, did not make a good enough improvement to compensate for the use of multiple models.

In the end, the Logistic Regression (one of the original models from the papers) performed the best overall. Keep in mind, the overall performance of each model hovered around the same scores for roughly all metrics (Precision, Recall, F1-score)

Model Confusion Matrix Table:

Models with TF-IDF only

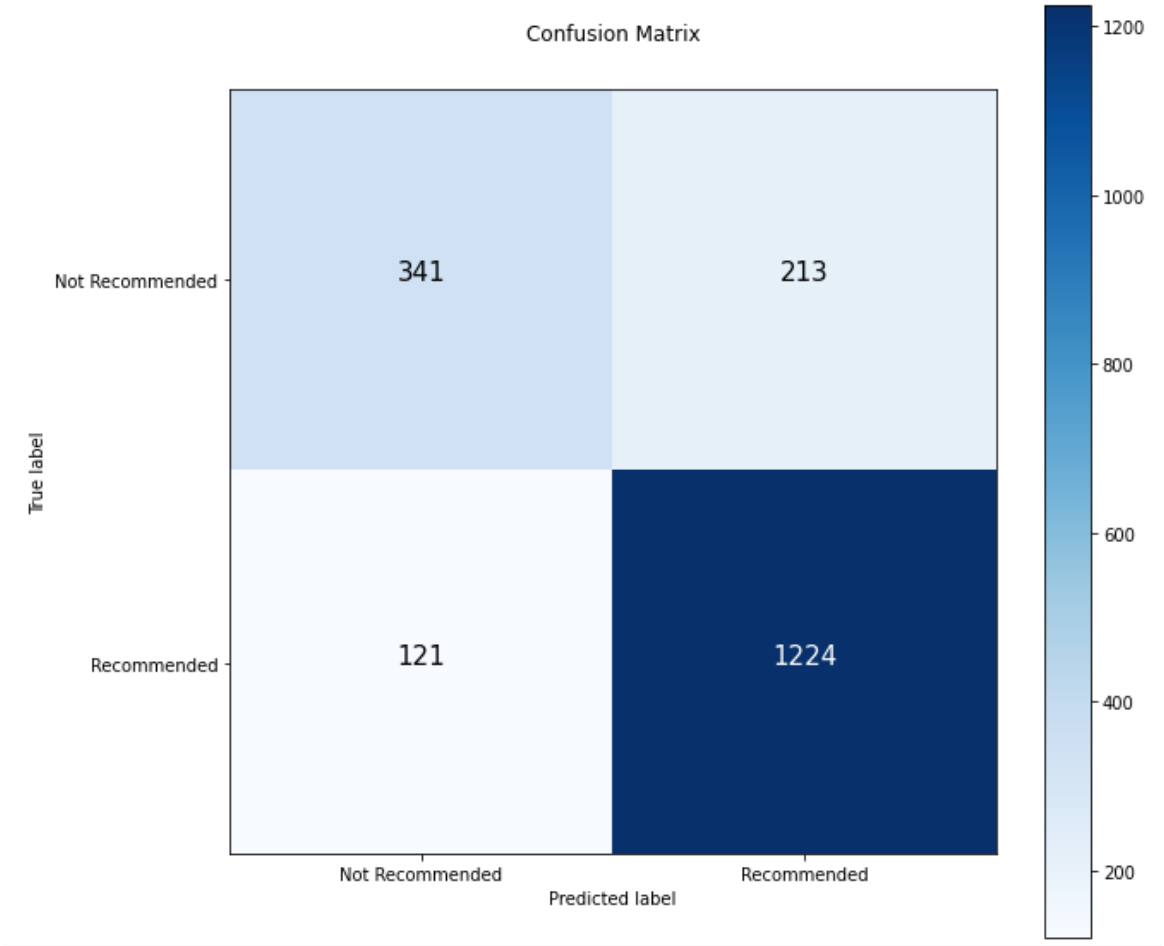
	precision	recall	f1-score	support
Not Recommended	0.73	0.65	0.69	288
Recommended	0.85	0.89	0.87	659
accuracy			0.82	947
macro avg	0.79	0.77	0.78	947
weighted avg	0.82	0.82	0.82	947



Logistic Regression Our Final Model Evaluation:

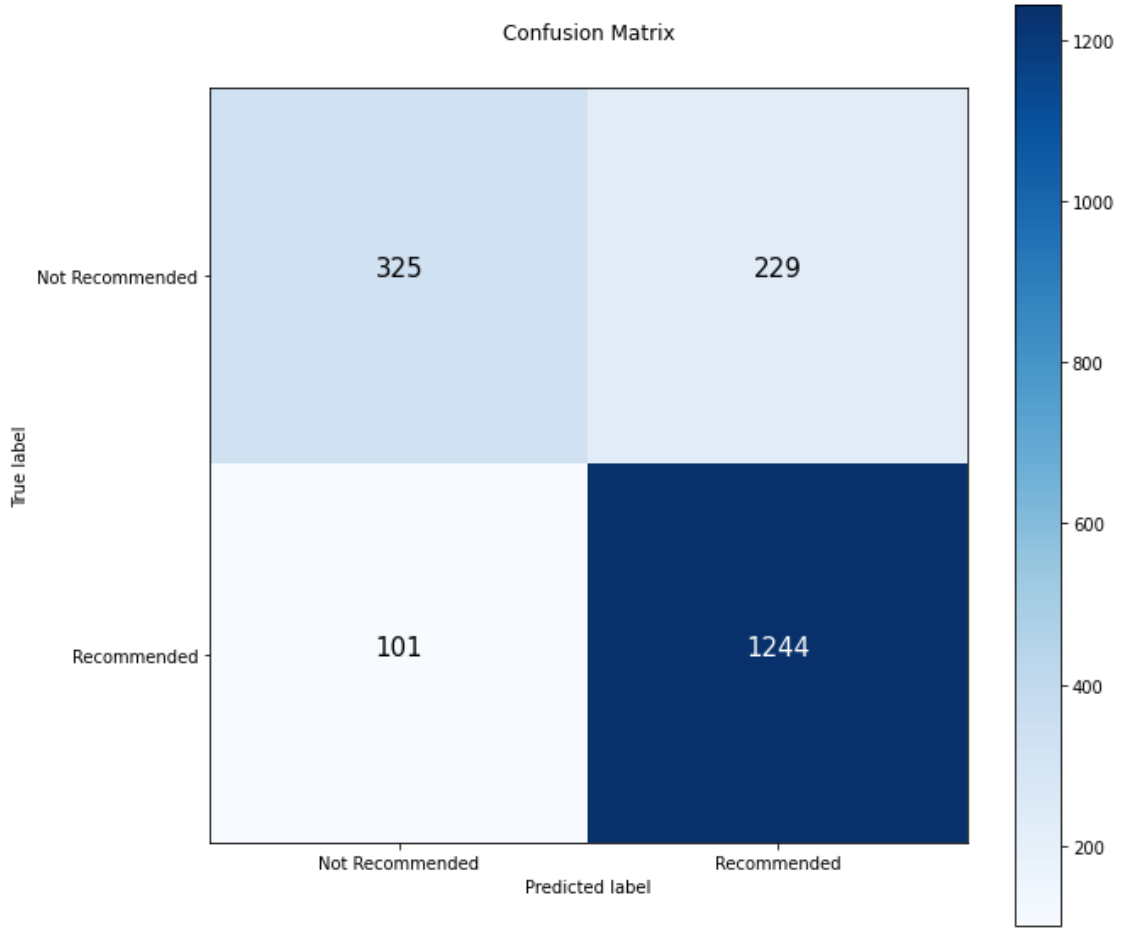
Our final model did not perform as well as we were hoping in the final model evaluation. Still, our scores were very similar to the original Linear SVC selected in the paper.

	precision	recall	f1-score	support
Not Recommended	0.74	0.62	0.67	554
Recommended	0.85	0.91	0.88	1345
accuracy			0.82	1899
macro avg	0.79	0.76	0.78	1899
weighted avg	0.82	0.82	0.82	1899



Ensemble Learning (Linear SVC, Log Reg, Random Forest) - Hard Voting (baseline model)

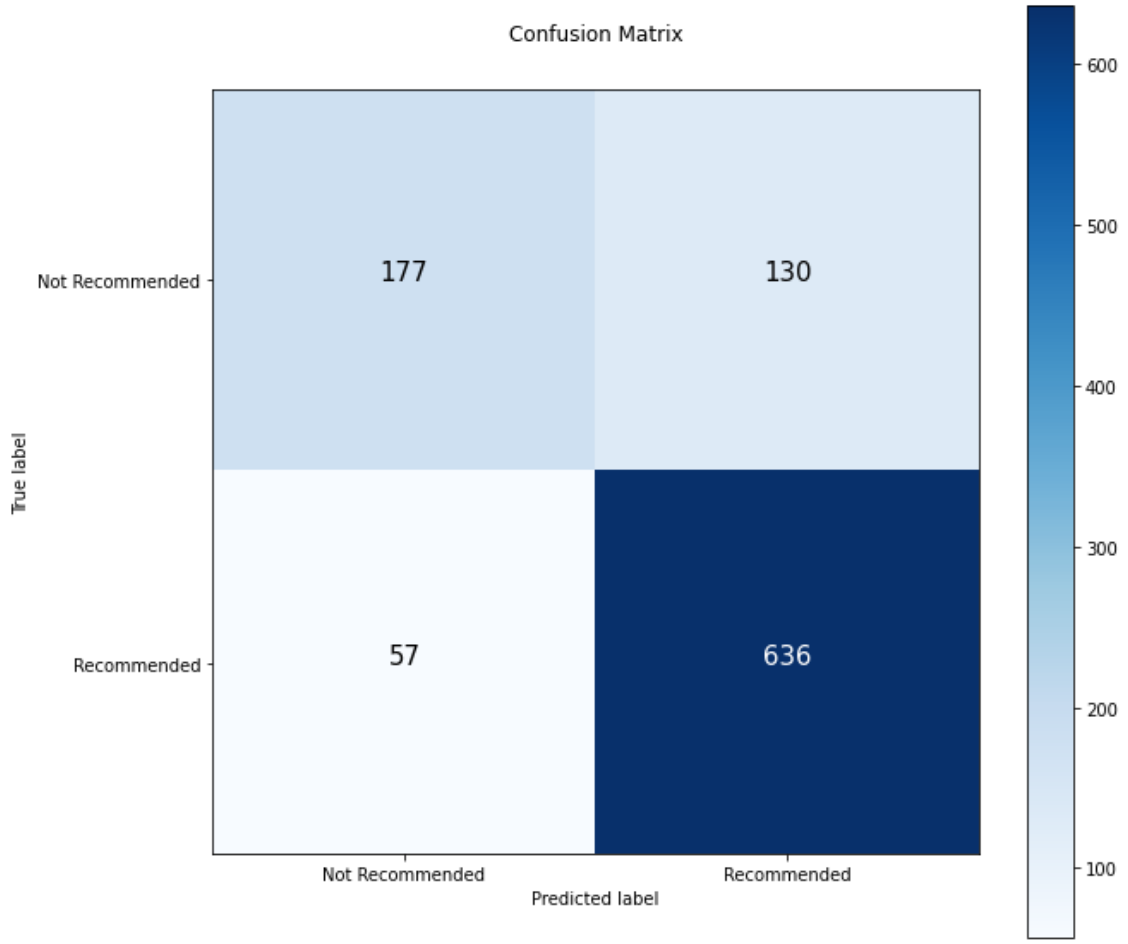
	precision	recall	f1-score	support
Not Recommended	0.76	0.59	0.66	554
Recommended	0.84	0.92	0.88	1345
accuracy			0.83	1899
macro avg	0.80	0.76	0.77	1899
weighted avg	0.82	0.83	0.82	1899



Ensemble Learning (Naive Bayes, Log Reg, Random Forest) - Soft Voting(baseline model)

Models with Slang Replacement and POS Tagging

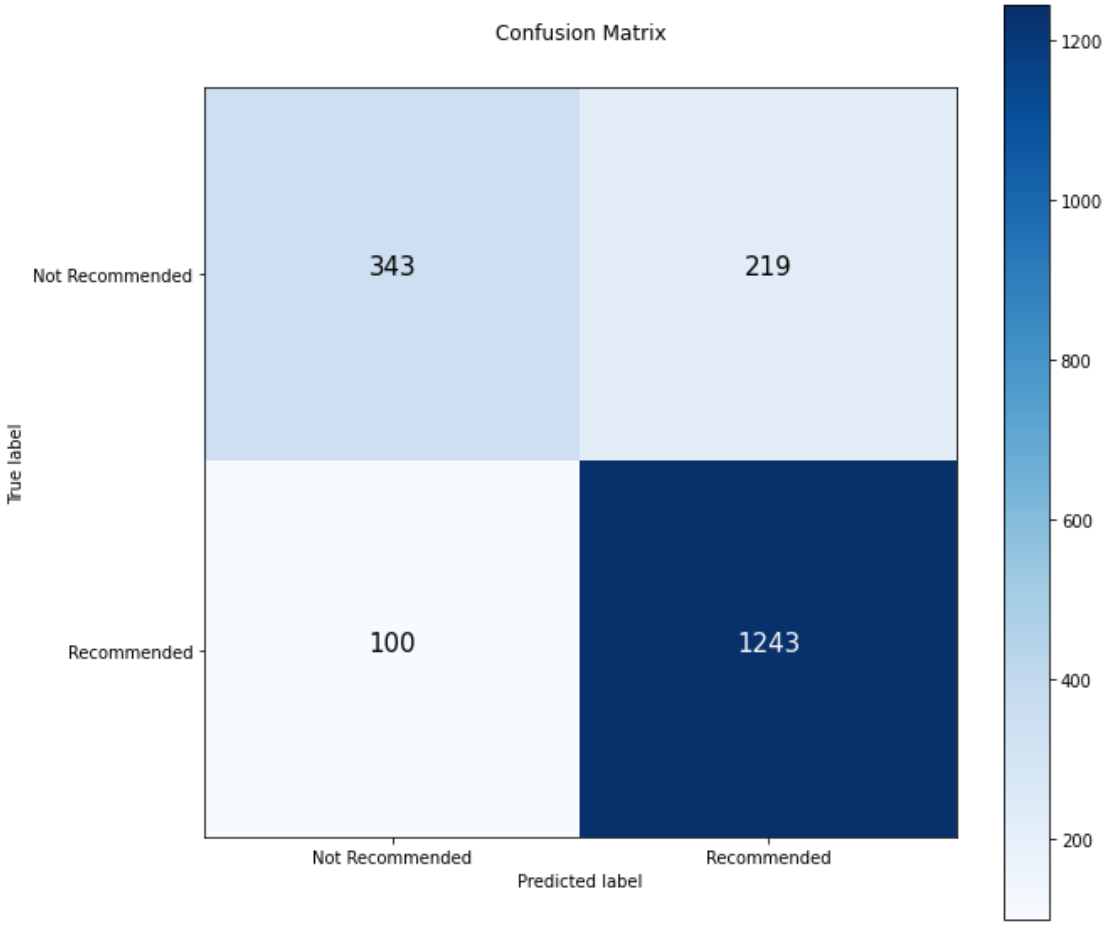
	precision	recall	f1-score	support
Not Recommended	0.76	0.58	0.65	307
Recommended	0.83	0.92	0.87	693
accuracy			0.81	1000
macro avg	0.79	0.75	0.76	1000
weighted avg	0.81	0.81	0.81	1000



Linear SVM Final Model Evaluation with Videogame Slang replacement and POS Tagging

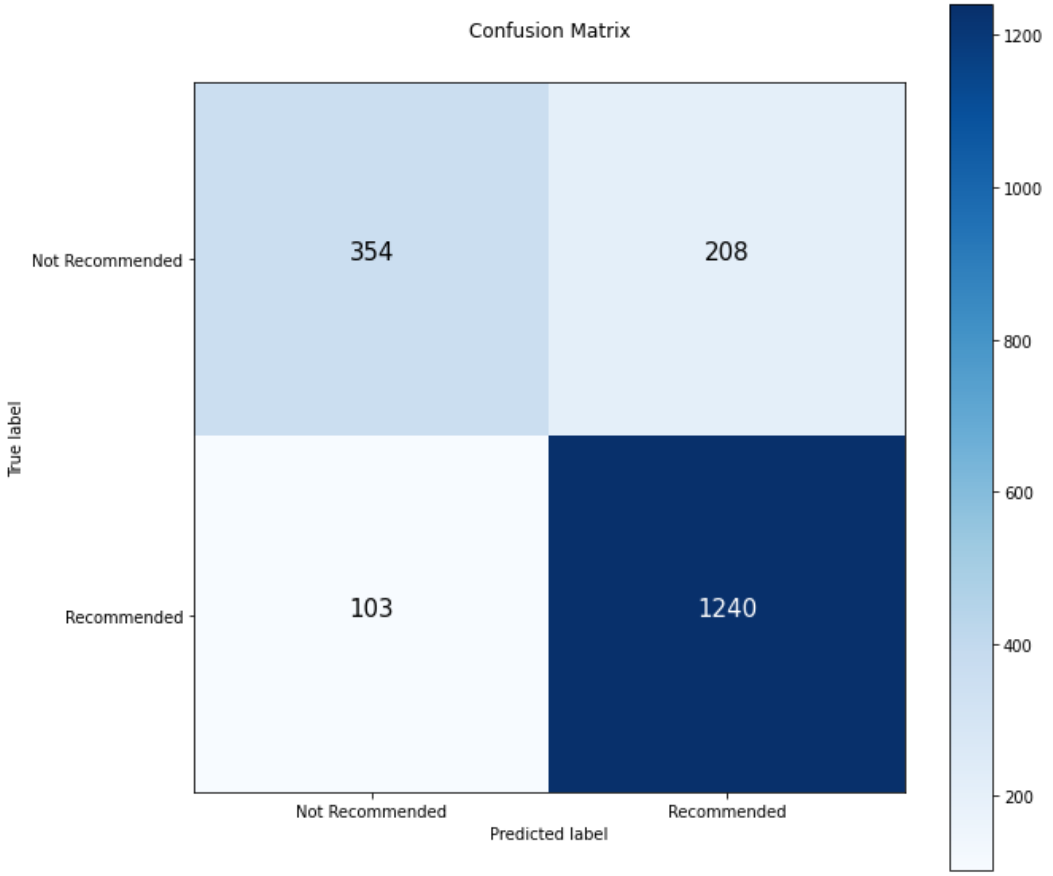
It seems like our ensemble model didn't give as much of a boost as we had hoped for. Our second model did not perform as well given we switched out our Linear SVM for Naive Bayes in order to use soft voting. After looking at our F1 scores, Accuracy, Average, and Weighted Average, our Linear SVC came out on top ever so slightly.

	precision	recall	f1-score	support
Not Recommended	0.77	0.61	0.68	562
Recommended	0.85	0.93	0.89	1343
accuracy			0.83	1905
macro avg	0.81	0.77	0.78	1905
weighted avg	0.83	0.83	0.83	1905



Ensemble Learning (Naive Bayes, Log Reg, Random Forest) - Soft Voting(experiment model)

	precision	recall	f1-score	support
Not Recommended	0.77	0.63	0.69	562
Recommended	0.86	0.92	0.89	1343
accuracy			0.84	1905
macro avg	0.82	0.78	0.79	1905
weighted avg	0.83	0.84	0.83	1905



Ensemble Learning (Linear SVC, Log Reg, Random Forest) - Hard Voting(experiment model)

Figure 38: Confusion matrices of all models

Our Text Summarization Result:

We can compare our model's summarization result and the test set headline(summarization) conducted by the WikiHow website, and easily find our model performs better than the test set in a random selection of text. For example, in Text-1, Our Model details the instruction.

Our model's summary sentence: "Wear a scarf around your neck"; WikiHow's summary sentence: "Add a scarf.". The rest of the sentence displays the same difference. For Text-3, Our model's summary expresses more natural than the WikiHow summary does. Our Model summary sentence:"Learn how to not sweat the small stuff., Take a break from homeschooling"; Wikihow summary:"Don't take yourself and your homeschool too seriously."

Text -1
A scarf around your neck is a good way to gain some warmth. It helps add heat to your body, namely your neck and back. Plus, you can choose a high-end scarf to still look professional. Try keeping one in your desk to take out as needed., Another option for the office is a wrap that you can use around the top half of your body. If you pick one in a nice material (such as silk or cashmere), it will still look professional as it keeps you warm.If you pick a wrap made from a thin, warm material, you can easily keep it neatly folded in your desk. Athletic cold-weather gear is made to fit close to the skin. In fact, most of the time, it will work under work clothes. Add a layer or two underneath your regular clothes to stay warm at work.For instance, you could wear thin jogger's leggings made from polypropylene or capilene under your work pants or a long-sleeved shirt made from the same material under a dress shirt.You can also try long underwear in silk., A cardigan can help keep you warm, but it only helps your top half. However, if you choose a long cardigan, you'll keep more of yourself warm. In fact, you can keep a long cardigan at work specifically for this purpose., You may need to switch to more sensible shoes to help keep your feet warm. When you do, you can add a pair of wool socks, which will keep your feet nice and toasty in the cold office., Cotton or polyester are common fabrics for professional clothes and sweaters. However, if you want more warmth out of your fabrics, choose wool or cashmere, which will insulate you more efficiently.
Our Model Summarization -1
Wear a scarf around your neck., Choose a wrap for the top half of your body., Add layers to your regular clothes., Use a long cardigan., Switch to sensible shoes., Choose fabrics that will keep you warm.
Test Set headline(summarization) -1
\nAdd a scarf.,\nTry a wrap.,\nUse cold-weather gear.,\nDrape on a long cardigan.,\nKeep your feet warm.,\nUpgrade your fabrics
Text -2
Don't forget to get the same information for the recipient's provider as well!; If their mailbox is limited to 10mb, for example, 20mb of attachments won't fit and will be "returned to sender". If the storage limit is too small, suggest creating an account from one of the several free providers that offer greater storage space.

HJSplit does not need to be installed to run, and versions are available for both PC and Mac platforms.

If either of the email providers involved have a limit for file sizes sent or received, the "parts" should be sized a bit smaller than the allowed limit.

Be sure to note in the message how many parts are to be expected. If the recipient does not have the file splitting program, send it as well (or a link to where it can be obtained).

Our Model Summarization -2

Get a list of email providers., Check the storage limit., Install HJSplit., Make sure that the "parts" are smaller than the allowed limit., Note in the message how many parts will be needed., Send an email to the recipient., Remind the recipient to send more attachments.

Test Set headline(summarization) -2

Determine the largest file size allowed by your email provider.,

Determine the amount of storage available to your recipient.,

Obtain a "file splitter" program, the best of which is HJSplit- completely free and very easy to use.,

Using the file splitting program, break the large file into several smaller files.,

Send the parts to the recipient as attachments to individual emails.,

Once the recipient has received all of the parts, they need only use the file split program to reassemble the parts into the original file.

Text -3

Yes, you have taken on a huge responsibility, but you need to learn how to not sweat the small stuff. Find the humor in your days. Kids are constantly saying hilarious things and if we are not careful we can be all too busy to catch them.

, Some days things are just going so far south that it’s time to simply throw in the towel for the day and head to the park or turn on Netflix or just send the kids outside to play while mom takes some time to scream into a pillow., Go on a date with your spouse or partner, hang out with friends, or simply go shopping or get your hair done. You need this time to get away from your children and be in adult company. It's so important to foster relationships that matter to you, and assure your family and friends that they don't always take a back seat to homeschooling., Sometimes the burnout may be due to using a curriculum that is not best suited to your needs. You may want to take a good look at what you are using and reevaluate its usefulness for you. Do your kids balk at using it? Is it difficult for you to understand how to use it? These may be indications that this curriculum is not a good fit for your family., Nearly every state and county has a local homeschool support group. You can most likely find one in your area by googling “homeschool support group”. There are also online support groups. You don’t have to do this alone. This can go a long way to helping avoid burnout., Homeschool parents have a tendency to neglect themselves. Having a hobby that you enjoy addresses this issue and makes sure that you are spending some time doing something that you like to do, at least a few hours a week. It can be anything, from scrapbooking to running. Even just finding time to read a book you've been wanting to read., This can apply no matter what age they are. Of course, the way you play with a teen looks a lot different than how you play with 1st and 2nd graders. However, you will never regret this one. After

all, isn’t this one of the major reasons you decided to homeschool in the first place?
Our Model Summarization -3
Learn how to not sweat the small stuff., Take a break from homeschooling., Make time for yourself., Find a support group., Have a hobby that you enjoy., Play with your kids.
Test Set headline(summarization)-3
Don’t take yourself and your homeschool too seriously., Know when to take a break., Make time for yourself and your loved ones., Adjust your curriculum., Join a homeschool support group., If you don’t have a hobby find one., Make time to play with your kids.
Text-4
Turn on your broiler and get it heated to about 550ºF (290ºC). Also position the top rack 5 inches (12 cm) from the top of the broiler. Unless your grill pan has a non-stick surface, coat it with non-stick cooking spray or oil to prevent the meat from sticking to it. Place your steak in the grill pan and place the pan on the top rack of the preheat broiler. For rare, close the door and allow it to cook for 4 minutes, before opening the door and flipping it to cook for another 4 minutes on the other side. For medium rare, add 1-3 minutes and for medium rare, add 3-5 minutes, to the total cooking time. Remove the steak from the broiler if it is charred to your liking. Use a small sharp knife, to make a small cut in the middle of the steak. If it looks done, serve immediately; if not, return it to the oven and broil it for another minute before removing and serving it. Serve whole or sliced.
Our Model Summarization -4
Heat your broiler., Spray the grill pan with non-stick cooking spray or oil., Cook for 4 minutes., Check if the steak is done.
Test Set headline(summarization) -4
Preheat your broiler., Prepare your pan.,

Broil your steak.,

Check and serve your steak.

Our Text-to-Speech Model result:

Because samples are all audio files, we will present them in our video report, and for the details, please check our github repository. We did not put as much emphasis on our Text-to-Speech so this can be considered for future work.

Figure 39: A snapshot of our result audio with a pretty nice natural tone. Please check this in our video presentation and code notebook.

Project Management

Implementation Status Report:

- Work completed:
    - Responsibilities -
        - Daniel Mata
            - Text Classification Models & EDA
            - Text Summarization Model
            - Video Game Implementation
            - Documentation
        - Lori YingHsuan Lo

        - Text Summarization EDA
        - TTS Model & EDA
        - Video Game Implementation
        - Documentation
    - Contributions - 50% - 50%
        - We both feel that we split the work evenly.
    - Issues / Concerns -
        - Unfortunately, we did not have enough time to focus on the Text-to-Speech model, so this can be considered for future work.
        - We were not able to implement all 3 models into a website (although we did try) although we did finish a final implementation.

References

- Resource for Text Classification
- Resource for Text Summarization
- Paper describing WikiHow dataset
- Speech synthesis resource / CSS10 / Dataset paper
- Resource guide for speech synthesis model
- IGN Review website for user integration
- Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier (2018)
- Steam Review Dataset - new, large scale sentiment dataset (2016)
- Steamvox: SteamVox is built to obtain the “Voice of the Player” from players’

(2019)

- A Study on Video Game Review Summarization (2019)
- Summarizing Game Reviews: First Contact (2020)
- An Empirical Study of Game Reviews on the Steam Platform (2019)
- Recommender System: Rating predictions of Steam Games Based on Genre and Topic Modelling (2020)
- T5 Transformer Tutorial
- Coqui-Ai/TTS
- Wikihow Online Knowledge Base