

Movie genres classification

Akmalkhon Mukhiddinov

Sahin Rana Betul

Akshma Atreja

Kirill Salokha

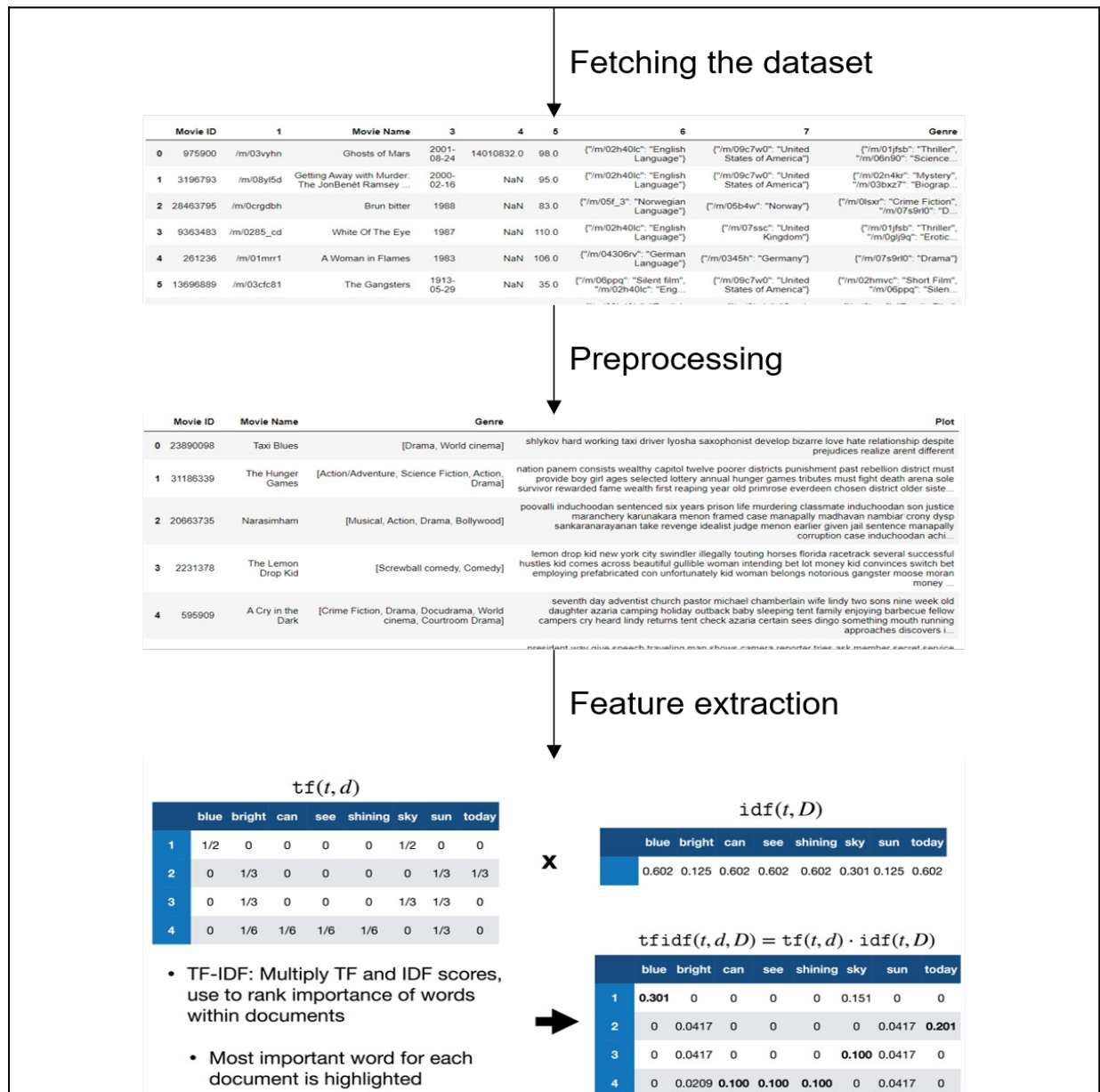
Bohdan Soproniuk

1. Introduction

In the modern day, entertainment is just about everywhere. Videogames, music, and, the topic of this report, all sorts of movies, shows and movie series. The abundance of this material creates a problem for the viewer: what should one choose if one is spoiled for choice this much? Good classification of movies by genre could be the solution. The viewer knows about his preferences and the system serving media to the viewer is aware of the viewer preferences as well, which can help fairly reliably suggest new things that could potentially interest the viewer.

To classify movie genres we used NLP techniques on movie descriptions corresponding to their genres. Logically, descriptions of movies of similar genres are bound to be somewhat similar. Using this assumption, the project introduced in this report was built.

The work process is summarized with the following flowchart:



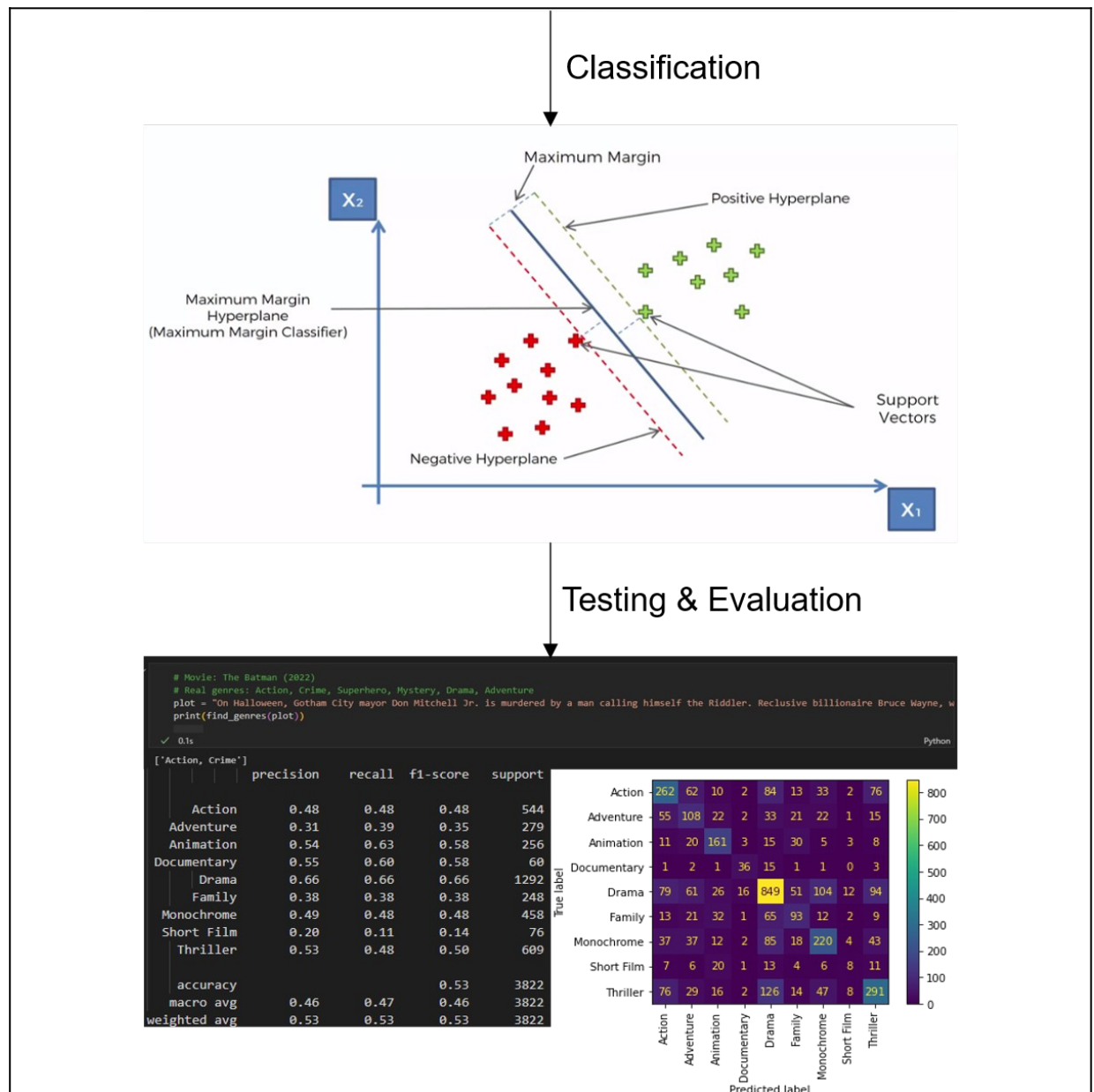


Figure 1: Flowchart of the project

2. Data analysis

2.1. Genres

Overall, our dataset contained 363 different genres.

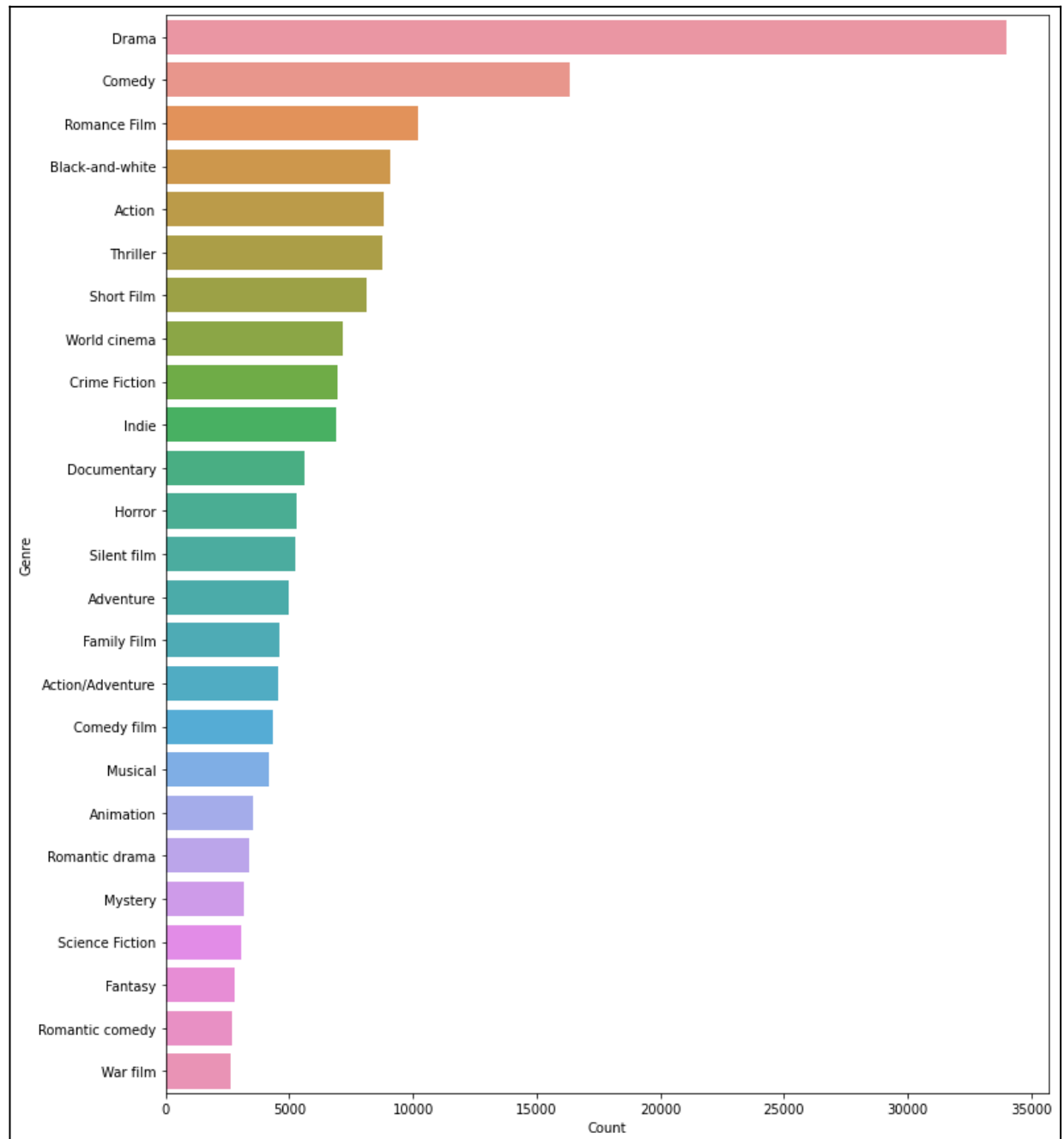


Figure 2: Bar plot of 25 most popular genres

Obviously, we did not need 363 genres. Hence, we reduced the number of genres to nineteen. We selected those nineteen genres from the twenty-five most popular genres above.

Another point to consider, is that movies often have multiple genres. For instance, movie can be “Musical drama biography” or “Family action comedy”. There exist multi-label classification models, which would map one movie into multiple genres, however, single-label classifiers are usually more accurate than multi-label ones, hence, we decided to build single-label classifiers. But, it is impossible to describe a movie by just one genre from one single-label classifier. Therefore, we split our nineteen genres into two classes: general and specific, and built two single-label classifiers.

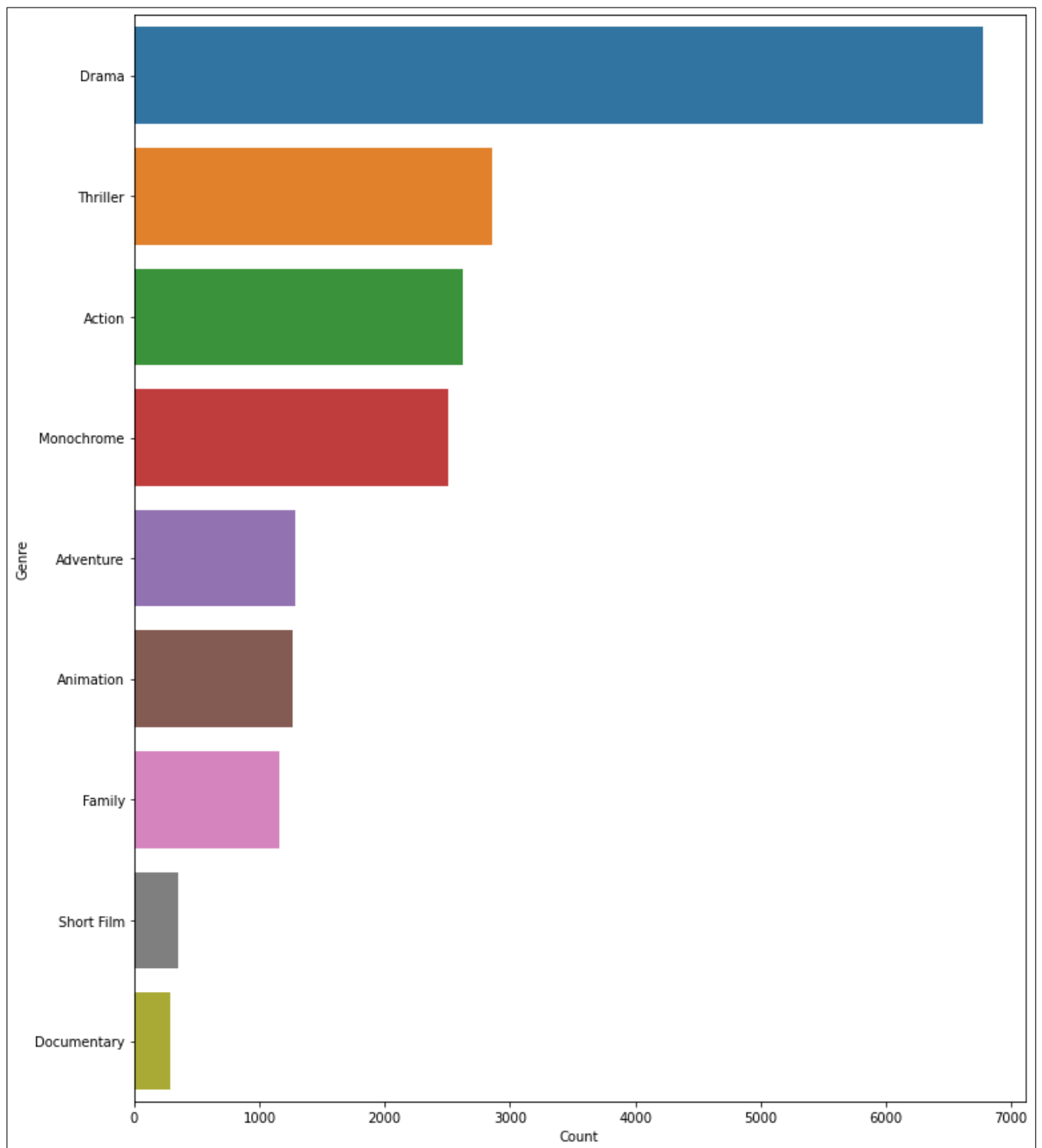


Figure 3: General genres bar plot

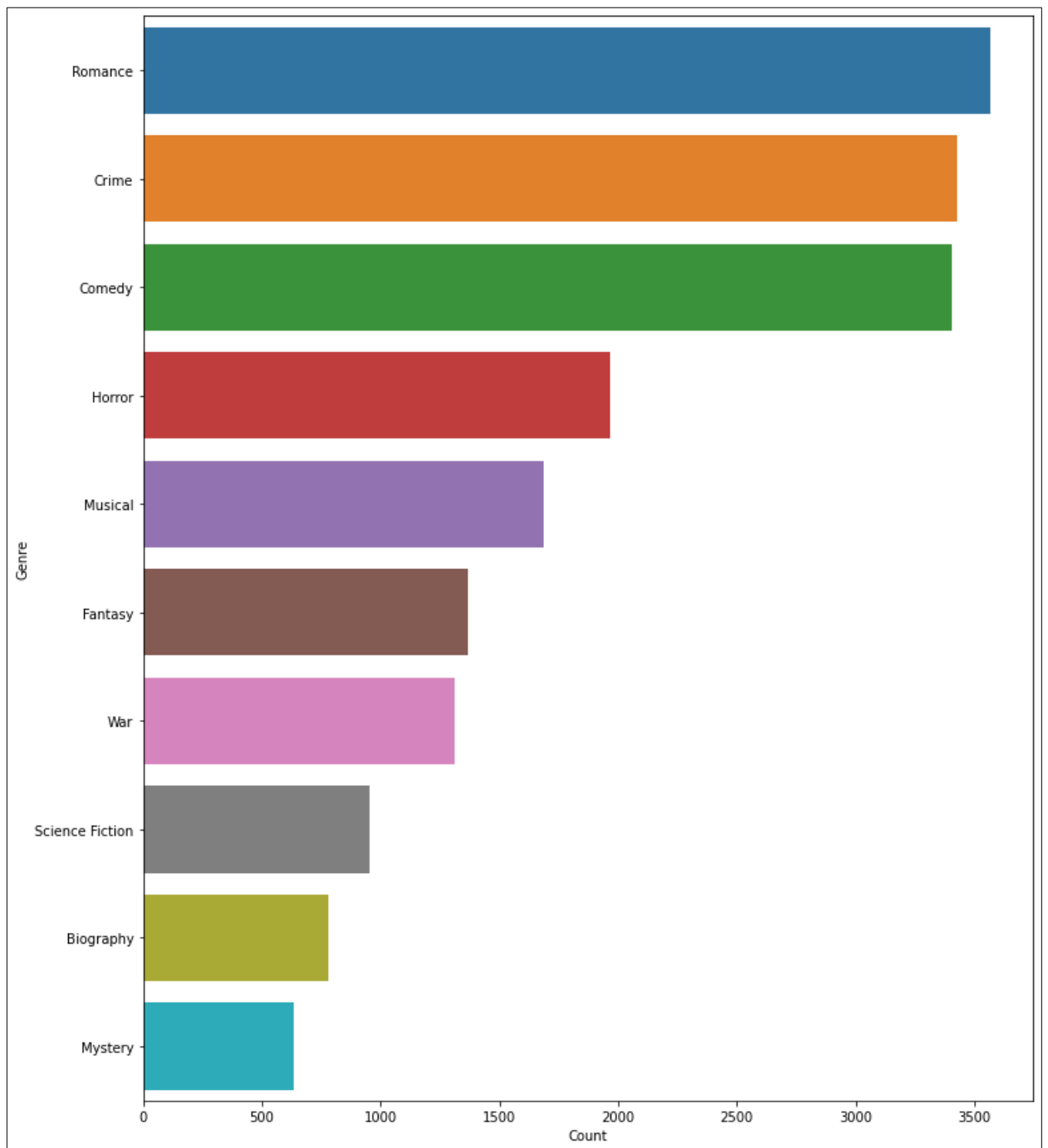


Figure 4: Specific genres bar plot

3. Data fetching and preprocessing

In this step, the dataset was preprocessed with the help of NLTK. Firstly, the files which, provide information about movie names, genres, and plot descriptions, were loaded.

	Movie ID	1	Movie Name	3	4	5	6	7	Genre
0	975900	/m/03vyhn	Ghosts of Mars	2001-08-24	14010832.0	98.0	{"/m/02h40lc": "English Language"}	{"/m/09c7w0": "United States of America"}	{"/m/01jfsb": "Thriller", "/m/06n90": "Science..."}
1	3196793	/m/08yl5d	Getting Away with Murder: The JonBenét Ramsey ...	2000-02-16	NaN	95.0	{"/m/02h40lc": "English Language"}	{"/m/09c7w0": "United States of America"}	{"/m/02n4kr": "Mystery", "/m/03bxz7": "Biograp..."}
2	28463795	/m/0crgdbh	Brun bitter	1988	NaN	83.0	{"/m/05f_3": "Norwegian Language"}	{"/m/05b4w": "Norway"}	{"/m/0lsxr": "Crime Fiction", "/m/07s9rl0": "D..."}
3	9363483	/m/0285_cd	White Of The Eye	1987	NaN	110.0	{"/m/02h40lc": "English Language"}	{"/m/07ssc": "United States of America"}	{"/m/01jfsb": "Thriller", "/m/0glj9q": "Erotic..."}
4	261236	/m/01mrr1	A Woman in Flames	1983	NaN	106.0	{"/m/04306rv": "German Language"}	{"/m/0345h": "Germany"}	{"/m/07s9rl0": "Drama"}

Figure 5: Raw movies genres dataset

	Movie ID	Plot
0	23890098	Shlykov, a hard-working taxi driver and Lyosha, a saxophonist, develop a bizarre love-hate relationship, and despite their prejudices, realize they aren't so different after all.
1	31186339	The nation of Panem consists of a wealthy Capitol and twelve poorer districts. As punishment for a past rebellion, each district must provide a boy and girl between the ages of 12 and 18 selected by lottery for the annual Hunger Games. The tributes must fight to the death in an arena; the sole...
2	20663735	Poovalli Induchoodan is sentenced for six years prison life for murdering his classmate. Induchoodan, the only son of Justice Maranchery Karunakara Menon was framed in the case by Manapally Madhavan Nambiar and his crony DYSP Sankaranarayanan to take revenge on idealist judge Menon who had e...
3	2231378	The Lemon Drop Kid , a New York City swindler, is illegally touting horses at a Florida racetrack. After several successful hustles, the Kid comes across a beautiful, but gullible, woman intending to bet a lot of money. The Kid convinces her to switch her bet, employing a prefabricated con. Unfo...

Figure 6: Raw movies plots dataset

From the dataset, a movie could possess multiple genres which were strings in our case. Those strings had to be converted to a dictionary type so that they can be utilized in further steps.

	Genre	Genres Updated
0	{"/m/01jfsb": "Thriller", "/m/06n90": "Science Fiction", "/m/03nnpn": "Horror", "/m/03k9f": "Adventure", "/m/0fdjb": "Supernatural", "/m/02kdv5l": "Action", "/m/09zvmj": "Space western"}	[Thriller, Science Fiction, Horror, Adventure, Supernatural, Action, Space western]
1	{"/m/02n4kr": "Mystery", "/m/03bxz7": "Biographical film", "/m/07s9rl0": "Drama", "/m/0hj3n01": "Crime Drama"}	[Mystery, Biographical film, Drama, Crime Drama]
2	{"/m/0lsxr": "Crime Fiction", "/m/07s9rl0": "Drama"}	[Crime Fiction, Drama]
3	{"/m/01jfsb": "Thriller", "/m/0glj9q": "Erotic thriller", "/m/09blyk": "Psychological thriller"}	[Thriller, Erotic thriller, Psychological thriller]
4	{"/m/07s9rl0": "Drama"}	[Drama]
5	{"/m/02hmv": "Short Film", "/m/06ppq": "Silent film", "/m/0219x_": "Indie", "/m/01g6gs": "Black-and-white", "/m/01z4y": "Comedy"}	[Short Film, Silent film, Indie, Black-and-white, Comedy]
6	{"/m/0hqxf": "Family Film", "/m/01hmn": "Fantasy", "/m/03k9f": "Adventure", "/m/03q4nz": "World cinema"}	[Family Film, Fantasy, Adventure, World cinema]
7	{"/m/04t36": "Musical", "/m/01z4y": "Comedy", "/m/01g6gs": "Black-and-white"}	[Musical, Comedy, Black-and-white]

Figure 7: Clean genres

After that, the table of movie names and their genres were merged with the corresponding plot descriptions based on their IDs.

	Movie ID	Movie Name	Genre	Plot
0	23890098	Taxi Blues	[Drama, World cinema]	Shlykov, a hard-working taxi driver and Lyosha, a saxophonist, develop a bizarre love-hate relationship, and despite their prejudices, realize they aren't so different after all.
1	31186339	The Hunger Games	[Action/Adventure, Science Fiction, Action, Drama]	The nation of Panem consists of a wealthy Capitol and twelve poorer districts. As punishment for a past rebellion, each district must provide a boy and girl between the ages of 12 and 18 selected by lottery for the annual Hunger Games. The tributes must fight to the death in an arena; the sole...
2	20663735	Narasimham	[Musical, Action, Drama, Bollywood]	Poovalli Induchoodan is sentenced for six years prison life for murdering his classmate. Induchoodan, the only son of Justice Maranchery Karunakara Menon was framed in the case by Manapally Madhavan Nambiar and his crony DYSP Sankaranarayanan to take revenge on idealist judge Menon who had e...
3	2231378	The Lemon Drop Kid	[Screwball comedy, Comedy]	The Lemon Drop Kid , a New York City swindler, is illegally touting horses at a Florida racetrack. After several successful hustles, the Kid comes across a beautiful, but gullible, woman intending to bet a lot of money. The Kid convinces her to switch her bet, employing a prefabricated con. Unfo...
4	595909	A Cry in the Dark	[Crime Fiction, Drama, Docudrama, World cinema, Courtroom Drama]	Seventh-day Adventist Church pastor Michael Chamberlain, his wife Lindy, their two sons, and their nine-week-old daughter Azaria are on a camping holiday in the Outback. With the baby sleeping in their tent, the family is enjoying a barbecue with their fellow campers when a cry is heard. Lindy r...
5	5272176	End Game	[Thriller, Action/Adventure, Action, Drama]	The president is on his way to give a speech. While he is traveling there a man shows up with a camera. A reporter tries to ask a member of the secret service a question. When the president enters he is shot by the man with the camera. The president's main bodyguard, Alex Thomas , is grazed by t...

Figure 8: Movies genres and plots merged

Moreover, the unnecessary symbols and stopwords were removed from the plot description using the NLTK library.

Movie ID	Movie Name	Genre	Plot	Plot Updated	
0	23890098	Taxi Blues	[Drama, World cinema]	Shlykov, a hard-working taxi driver and Lyosha, a saxophonist, develop a bizarre love-hate relationship, and despite their prejudices, realize they aren't so different after all.	shlykov a hard working taxi driver and lyosha a saxophonist develop a bizarre love hate relationship and despite their prejudices realize they arent so different after all
1	31186339	The Hunger Games	[Action/Adventure, Science Fiction, Action, Drama]	The nation of Panem consists of a wealthy Capitol and twelve poorer districts. As punishment for a past rebellion, each district must provide a boy and girl between the ages of 12 and 18 selected by lottery for the annual Hunger Games. The tributes must fight to the death in an arena; the sole...	the nation of panem consists of a wealthy capitol and twelve poorer districts as punishment for a past rebellion each district must provide a boy and girl between the ages of and selected by lottery for the annual hunger games the tributes must fight to the death in an arena the sole survivor is...
2	20663735	Narasimham	[Musical, Action, Drama, Bollywood]	Poovalli Induchoodan is sentenced for six years prison life for murdering his classmate. Induchoodan, the only son of Justice Maranchery Karunakara Menon was framed in the case by Manapally Madhavan Nambiar and his crony DYSP Sankaranarayanan to take revenge on idealist judge Menon who had e...	poovalli induchoodan is sentenced for six years prison life for murdering his classmate induchoodan the only son of justice maranchery karunakara menon was framed in the case by manapally madhavan nambiar and his crony dysp sankaranarayanan to take revenge on idealist judge menon who had earlier...

Figure 9: Removing unnecessary symbols and making lowercase the dataset

After all, this is the final version of our dataset.

Movie ID	Movie Name	Genre	Plot	
0	23890098	Taxi Blues	[Drama, World cinema]	shlykov hard working taxi driver lyosha saxophonist develop bizarre love hate relationship despite prejudices realize arent different
1	31186339	The Hunger Games	[Action/Adventure, Science Fiction, Action, Drama]	nation panem consists wealthy capitol twelve poorer districts punishment past rebellion district must provide boy girl ages selected lottery annual hunger games tributes must fight death arena sole survivor rewarded fame wealth first reaping year old primrose everdeen chosen district older siste...
2	20663735	Narasimham	[Musical, Action, Drama, Bollywood]	poovalli induchoodan sentenced six years prison life murdering classmate induchoodan son justice maranchery karunakara menon framed case manapally madhavan nambiar crony dysp sankaranarayanan take revenge idealist judge menon earlier given jail sentence manapally corruption case induchoodan achi...
3	2231378	The Lemon Drop Kid	[Screwball comedy, Comedy]	lemon drop kid new york city swindler illegally touting horses florida racetrack several successful hustles kid comes across beautiful gullible woman intending bet lot money kid convinces switch bet employing prefabricated con unfortunately kid woman belongs notorious gangster moose moran money ...
4	595909	A Cry in the Dark	[Crime Fiction, Drama, Docudrama, World cinema, Courtroom Drama]	seventh day adventist church pastor michael chamberlain wife lindy two sons nine week old daughter azaria camping holiday outback baby sleeping tent family enjoying barbecue fellow campers cry heard lindy returns tent check azaria certain sees dingo something mouth running approaches discovers i...
5	5272176	End Game	[Thriller, Action/Adventure, Action, Drama]	president way give speech traveling man shows camera reporter tries ask member secret service question president enters shot man camera presidents main bodyguard alex thomas grazed bullet hits president shooter gunned alex secret service agents president dies hospital kate crawford investigative...

Figure 10: The final dataset

4. Features extraction

To extract features from the text we used the TF-IDF (term frequency - inverse document frequency) method. Every word in every description is assigned a value of the following product:

$$TFIDF(word, description) = TF(word, description) \times IDF(word)$$

TF function takes two arguments: current word and current description. Then, it calculates the word frequency in the current document and divide it by the number of total words in the current document.

$$TF(word, description) = \frac{\text{number of 'word' } \in \text{description}}{\text{total number of words } \in \text{description}}$$

IDF function takes only one argument: current word. Then, it calculates the logarithm base of the number of descriptions containing the current word of the number of total descriptions in the corpus.

$$IDF(word) = \log_a b$$

a – number of descriptions, which contain ‘word’

b – total number of descriptions in the corpus

We configured our vectorizer to calculate TF-IDF value for both unigrams and bigrams. We also defined the maximum document frequency as 0.8, which removed all corpus-specific stop words. Although, since the movie descriptions are usually very diverse, we do not think that the maximum document frequency had a real impact on our models. Finally, our vectorizer selected top 15000 features, sorted by word frequency, across the whole corpus.

Of course, due to the overfitting problem, we extracted features from train and test sets separately and independently.

5. Model building

Two models have been built for this project: the model for genres that are more general, i.e., genres like “Adventure” or “Family” or perhaps “Action”, which do not tell us all that much about the contents of the movie, and the model for more descriptive and specific genres, like perhaps “Horror” or “War”. Any one movie is expected to have both a “general” genre and a more descriptive one (as seen on examples of 2.1).

To actually build the model, hyperparameter optimization was used, focusing on creating the “best” sets of words that describe certain genres. The model is then later evaluated through cross validation. It is also important to once again note that two models are used in this project. Despite their similarities in what they have to achieve, they are nonetheless rather different “under the hood”, so their hyperparameters differ.

```
else:
    # use hardcoded models (parameters were obtained by hyperoptimization)
    general_svc = LinearSVC(C=0.8, class_weight='balanced', tol=0.5, random_state=42)
    specific_svc = LinearSVC(C=0.2, tol=0.5, random_state=42)
```

Figure 11: Hyperoptimized models. Note, that parameters of two models are quite different

The model building is also done with the use of Support Vector Machines for classification of the genres. Every movie description is being “fed” to an SVM, which then determines the class of this movie description, giving its probabilities of this description belonging to any one class. After this process is finished, the model is ready.

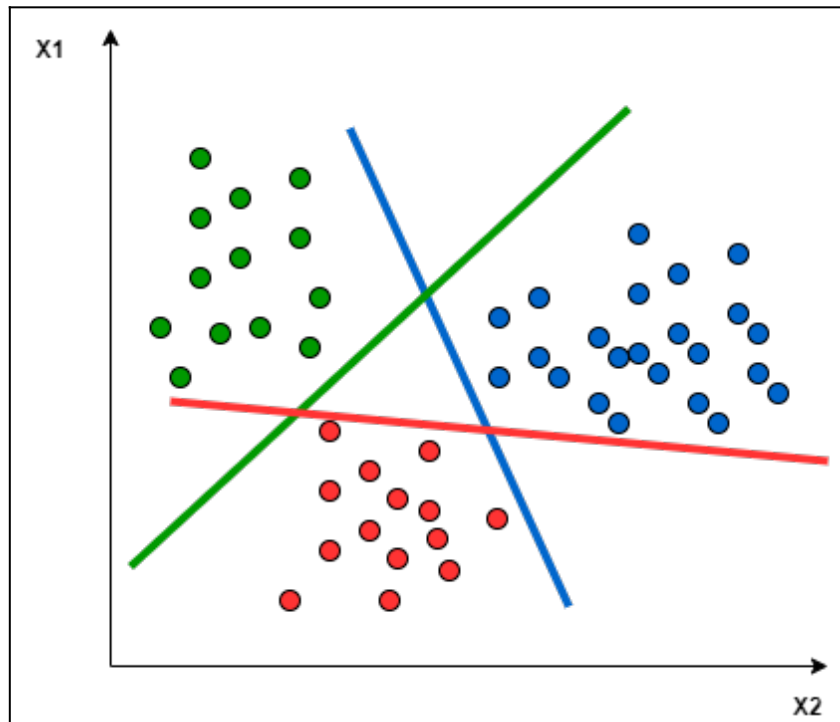


Figure 12: Support Vector Machine

6. Parts-of-speech tagging, named entity recognition

For the next part of this project, we turned our attention to a different tool for natural language processing – spaCy. It is a feature-rich library written in Python and Cython that provides parts-of-speech tagging, dependency parsing, named entity recognition, text classification for more than 60 languages.

After tokenizing raw text, we classified words by their part of speech and counted their occurrence in text for further analysis.

For example, we found an association between the name of a genre and the noun that occurs most frequently in the descriptions of movies belonging to that genre.

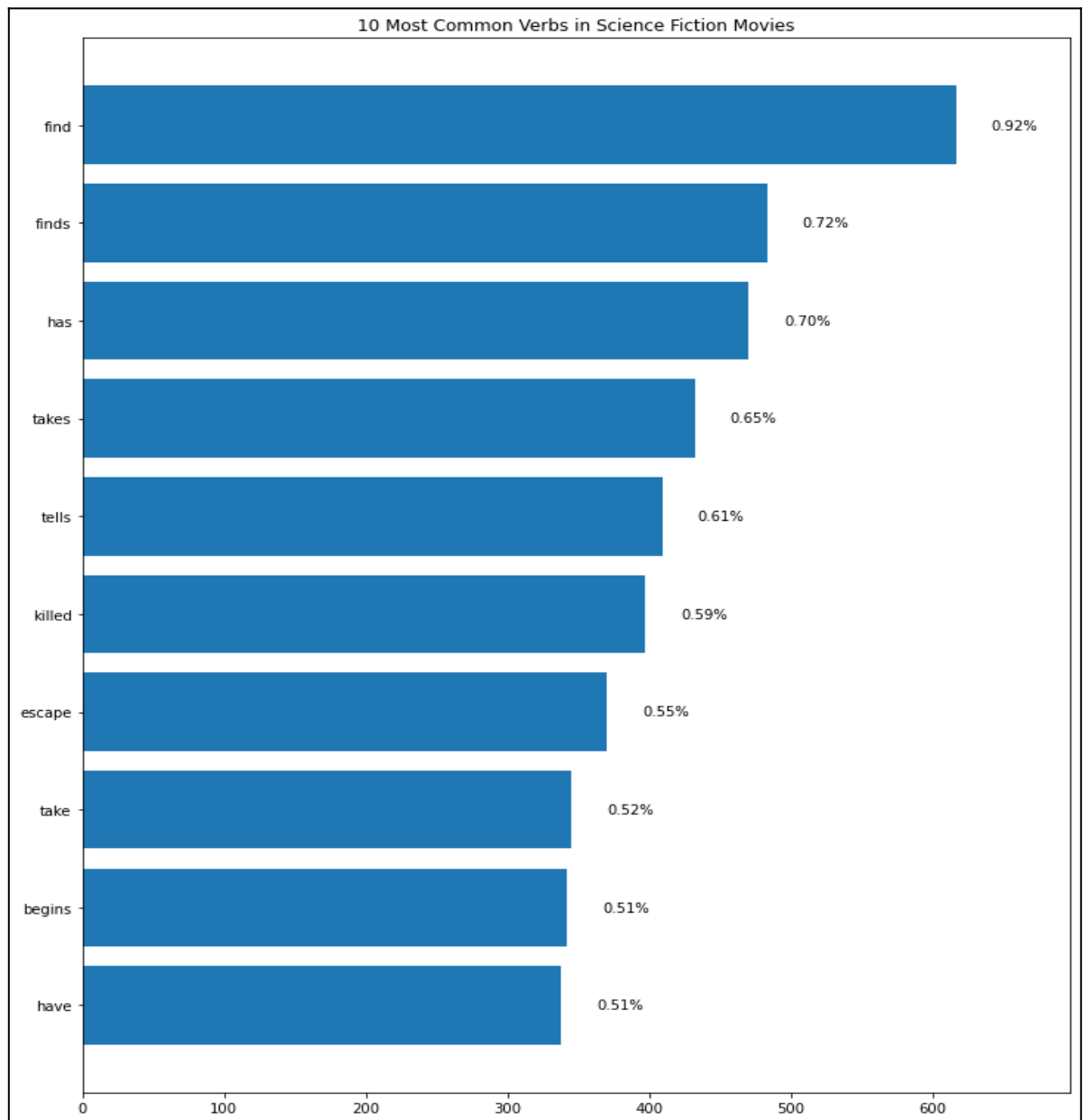


Figure 13: Most common verbs in science fiction

General genre	Most common noun	Specific genre	Most common noun
Action	police	Biography	film
Adventure	father	Comedy	father
Animation	time	Crime	police
Documentary	film	Fantasy	time
Drama	love	Horror	house
Family	father	Musical	love
Monochrome	man	Mystery	police
Short Film	film	Romance	love
Thriller	house	Science Fiction	time
		War	war

Figure 14: Most common nouns in all genres

We also found out that most common adjectives do not differ much, with the word “other” being the top one for every general and specific genre with the only exception being war movies, where the word “german” occurs most frequently.

For the sake of another interesting investigation, we decided to evaluate the following coefficient. We collected a list of all named entities for every plot of a specific genre, and a set of unique entities among them. By dividing the length of the first collection by the length of the second collection, we obtain a number indicating how often a certain person, location, organization, product etc. is mentioned in the descriptions. Not surprisingly, documentary films are very informative, and, on average, each such object is repeated less than twice in the same text. In contrast, descriptions of horror films are mostly focused on the same entities, and they repeat in the same text 4.57 times on average.

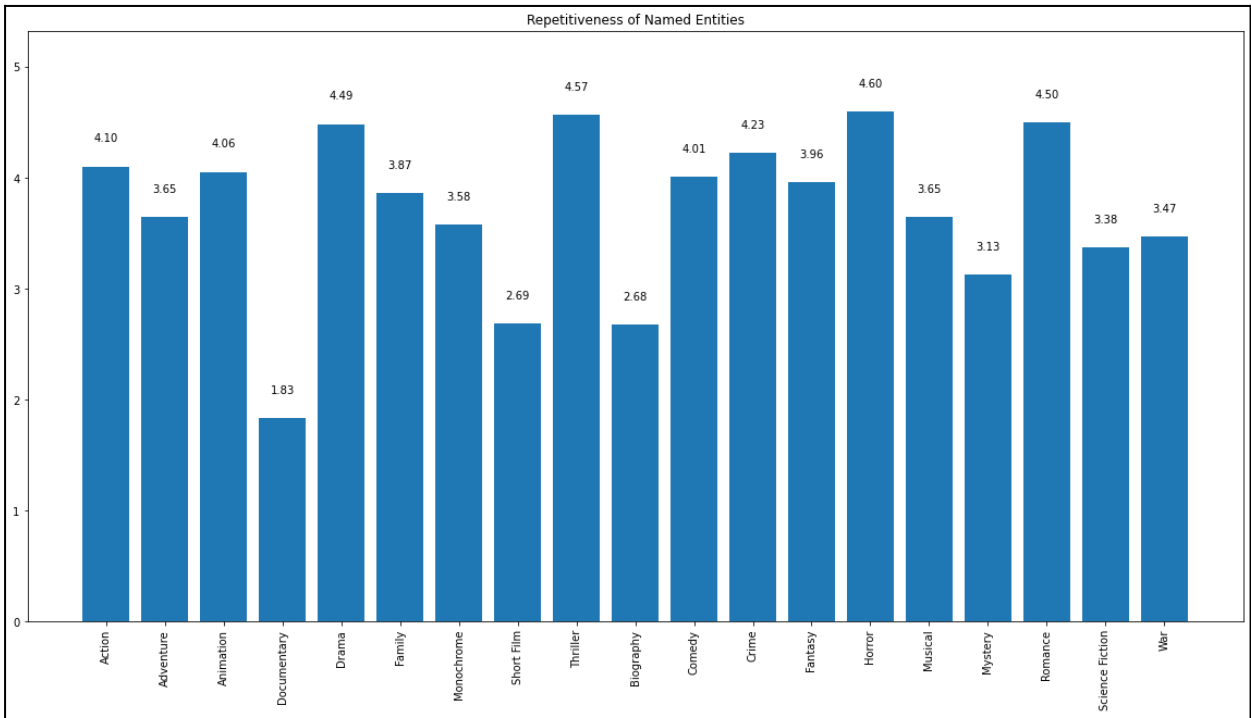


Figure 15: Repetitiveness coefficients of all genres

7. Model evaluation and testing

For the model evaluation, the first parameter to find was Accuracy which implies the success of the model we made. The accuracy was 53% for general genres and 57% for specific genres. Since the naive random model would have the accuracy around 10% (100% divided by the number of genres), we conclude, that our models accuracy are very good. Also, there were precision, f1-score and recall calculated for each genre as seen below:

Genre	Precision	Recall	f1-score	Support
Action	0.48	0.48	0.48	544
Adventure	0.31	0.39	0.35	279
Animation	0.54	0.63	0.58	256
Documentary	0.55	0.60	0.58	60
Drama	0.66	0.66	0.66	1292

Family	0.38	0.38	0.38	248
Monochrome	0.49	0.48	0.48	458
Short Film	0.20	0.11	0.14	76
Thriller	0.53	0.48	0.50	609
-	-	-	-	-
Accuracy			0.53	3822
Macro average	0.46	0.47	0.46	3822
Weighted average	0.53	0.53	0.53	3822

Figure 16: General genres metrics

Genre	Precision	Recall	f1-score	Support
Biograpy	0.52	0.21	0.30	149
Comedy	0.55	0.55	0.55	672
Crime	0.59	0.73	0.65	690
Fantasy	0.58	0.44	0.50	268
Horror	0.66	0.64	0.65	426
Musical	0.49	0.33	0.40	327
Mystery	0.22	0.01	0.03	139
Romance	0.50	0.68	0.58	659
Science Fiction	0.64	0.54	0.59	193
War	0.72	0.78	0.75	299
-	-	-	-	-
Accuracy			0.57	3822
Macro average	0.55	0.49	0.50	3822
Weighted average	0.56	0.57	0.55	3822

Figure 17: Specific genres metrics

Next we constructed the confusion matrices. Confusion matrix is a table with predicted and actual values from the model.

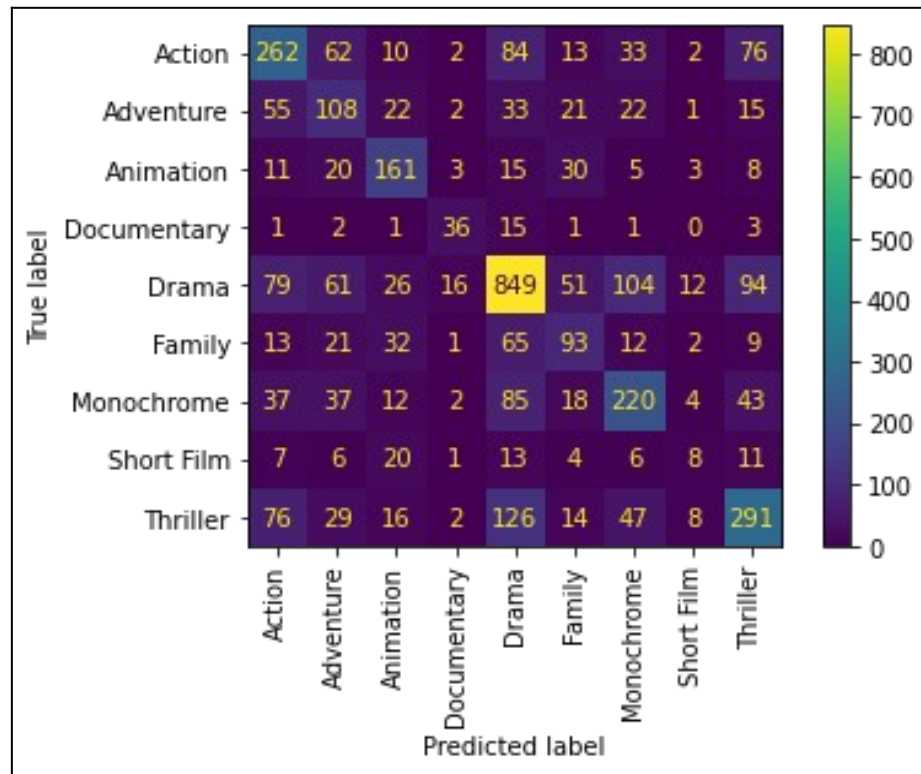


Figure 18: General genres confusion matrix

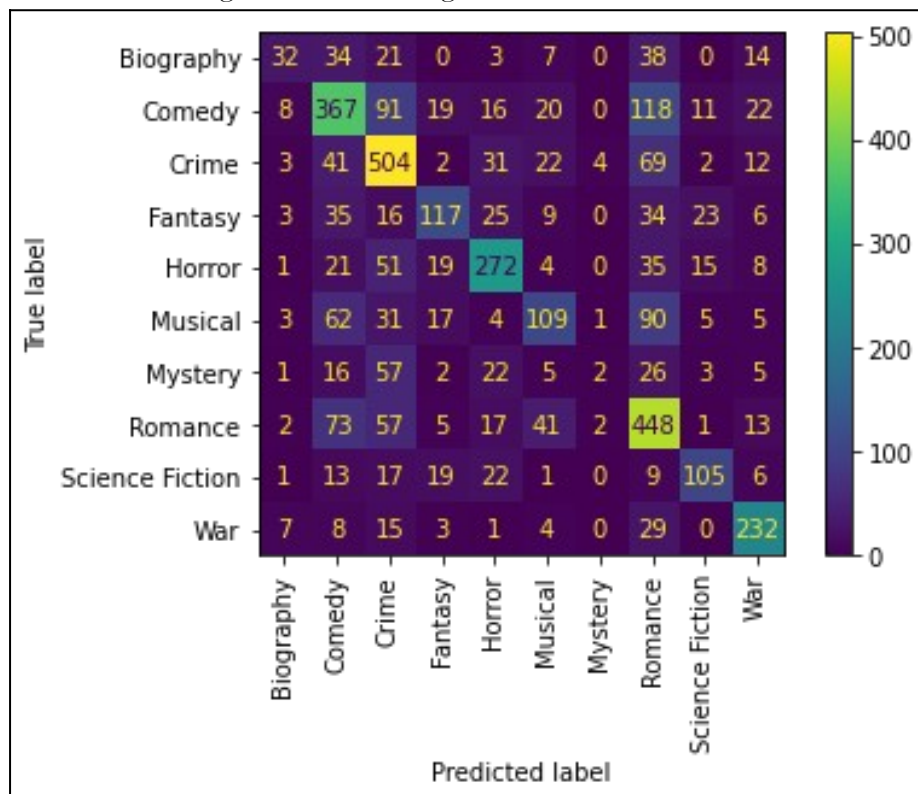


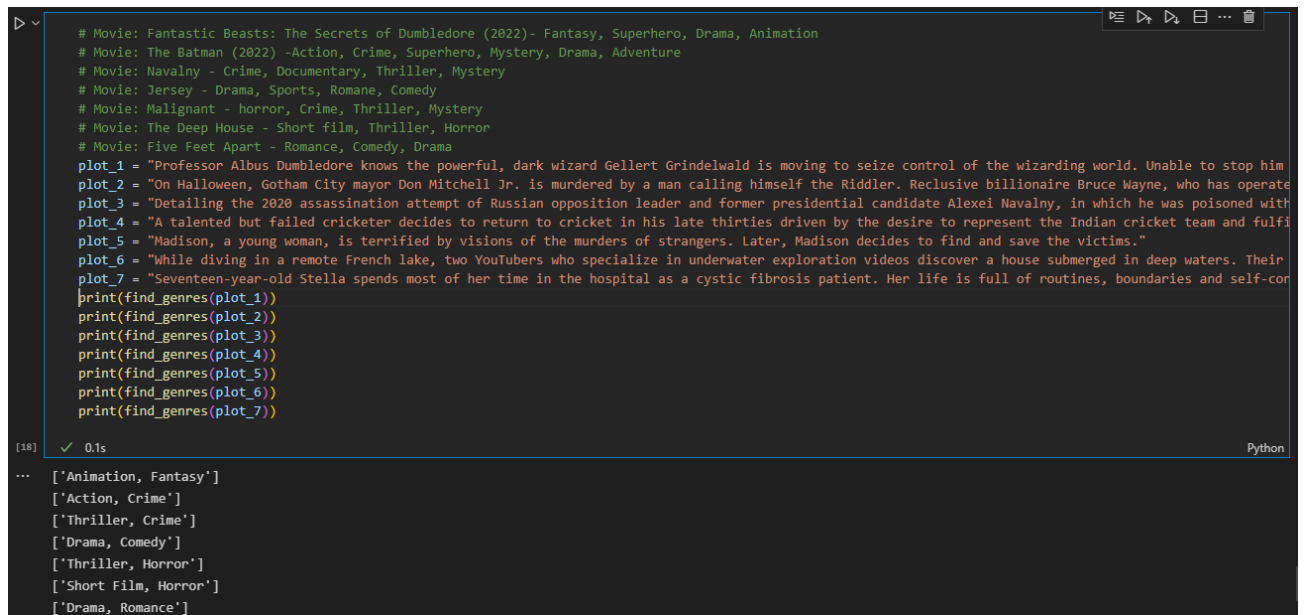
Figure 19: Specific genres confusion matrix

Looking at the confusion matrix above, the diagonal values are relatively more, which can be considered as a success of our model, as it implies that the actual genres are correctly matched with predicted one.

Nonetheless, there are few anomalies. In general genres, we noticed that 126 thriller movies were classified as drama by our model. That may be, because these genres are the most popular in our dataset, and because drama movies in the reality can be quite often also interpreted as thrillers.

The worst predicted genres are “short films” from general subset and “mystery” from specific subset. These two genres contain very small amount of movies, and this is definitely one of the reasons of their bad classification. However, it is not the only reason. As we can observe, “documentary” from the general subset has the least amount of movies, but was classified pretty decently. Hence, another problem may be, that genres “short films” and “mystery” are by nature not quite distinguishable. In fact, it makes sense, that a movie hardly can be recognized as “short film” based only on it’s description.

Next, we manually tested our models. We wrote a function, that takes the string description, clean it by using methods from the chapter 3, then extracts features by using methods from chapter 4, and, finally, fits the numerical array into the models from chapter 5. Our empirical tests were following:



```
# Movie: Fantastic Beasts: The Secrets of Dumbledore (2022)- Fantasy, Superhero, Drama, Animation
# Movie: The Batman (2022) -Action, Crime, Superhero, Mystery, Drama, Adventure
# Movie: Navalny - Crime, Documentary, Thriller, Mystery
# Movie: Jersey - Drama, Sports, Romance, Comedy
# Movie: Malignant - horror, Crime, Thriller, Mystery
# Movie: The Deep House - Short film, Thriller, Horror
# Movie: Five Feet Apart - Romance, Comedy, Drama
plot_1 = "Professor Albus Dumbledore knows the powerful, dark wizard Gellert Grindelwald is moving to seize control of the wizarding world. Unable to stop him
plot_2 = "On Halloween, Gotham City mayor Don Mitchell Jr. is murdered by a man calling himself the Riddler. Reclusive billionaire Bruce Wayne, who has operate
plot_3 = "Detailing the 2020 assassination attempt of Russian opposition leader and former presidential candidate Alexei Navalny, in which he was poisoned with
plot_4 = "A talented but failed cricketer decides to return to cricket in his late thirties driven by the desire to represent the Indian cricket team and fulfil
plot_5 = "Madison, a young woman, is terrified by visions of the murders of strangers. Later, Madison decides to find and save the victims."
plot_6 = "While diving in a remote French lake, two YouTubers who specialize in underwater exploration videos discover a house submerged in deep waters. Their
plot_7 = "Seventeen-year-old Stella spends most of her time in the hospital as a cystic fibrosis patient. Her life is full of routines, boundaries and self-cor
print(find_genres(plot_1))
print(find_genres(plot_2))
print(find_genres(plot_3))
print(find_genres(plot_4))
print(find_genres(plot_5))
print(find_genres(plot_6))
print(find_genres(plot_7))
[18] ✓ 0.1s
Python
... ['Animation, Fantasy']
['Action, Crime']
['Thriller, Crime']
['Drama, Comedy']
['Thriller, Horror']
['Short Film, Horror']
['Drama, Romance']
```

Figure 20: Empirical tests. First two tests are overfitting examples. Other “real” tests passed extremely well: all predicted genres are almost the same as actual ones.

There were a few pitfalls in the empirical tests:

1. We should not test on movies earlier than 2015, as they were used in model training or testing. Since our dataset was constructed in 2014, it consists of movies released before 2014, so we can just do empirical tests on movies released later than 2015-2016 and avoid overfitting.
2. Nowadays, there are a lot of franchises, sequels, reboots, remakes, spin-offs and so on. For instance, first two tests are Batman and Fantastic Beasts, which are overfitting, since there were dozens of other movies about Batman and Harry Potter, and they are all included in our dataset. Hence, model just recognizes specific names, like "Batman", "Joker", "Gotham", or "Dumbledore", "Hogwarts", "Voldemort", and gives the same genres from the old movies, that it used during the training.

As the result, we should avoid using old movies and popular franchises, sequels and use new unique movies, like Navalny.

8. Conclusion

The goal of our project was to classify movie genres based on its Wikipedia description. After the data fetching, preprocessing, and features extraction, we constructed two models for the classification, namely Linear Support Vector Machines. By looking at the metrics, we can conclude, that both models performed almost equally well.

Afterwards, we manually tested our models on some descriptions taken from Wikipedia.

9. Contributions

Akkmalkhon Mukhiddinov – Feature extraction, model building code

Sahin Rana Betul – Data fetching, preprocessing

Akshma Atreja – Model evaluation, testing

Kirill Salokha – Data analysis, model building report

Bohdan Soproniuk – Data analysis, parts-of-speech tagging and named entity recognition