1 Problem statement

Use Network Analysis and NLP techniques to investigate whether we can make a model predict a movies genre given the description of this movie. In addition, we would like to investigate whether it is possible to predict the choice of similar movies given this description.

2 Data acquisition

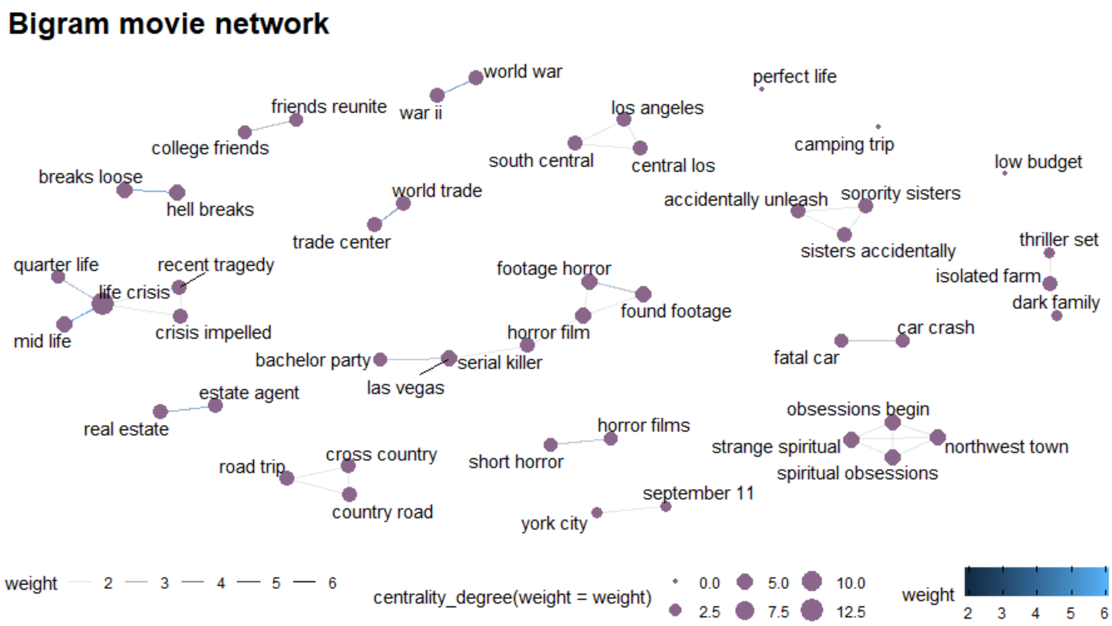
Given the problem statement we choose and obtain datasets which we consider interesting and appropriate for this analysis. Through Kaggle we were able to find a relevant dataset. The dataset contains 22 columnsfeatures, which describes different characteristics for selected movies ranked and reviewed on IMDB eg. genre and title.

Based on these, we describe the variables that are used the most in relation to our problem statement. This includes four character strings. The first variable **imdb_title_id** assign each movie with an individual identification name. The second variable **title** refers to the title of each movie. The third variable **genre** describes which genre or several genres a given movie belongs to. In order to make the analysis more applicable only the four most common genres (Drama, Thriller, Comedy and Horror) are used and only observations from the year 2000 to the present. The fourth variable **description** contains a text that gives a brief description of the action of each movie.

3 Results

3.1 Network analysis

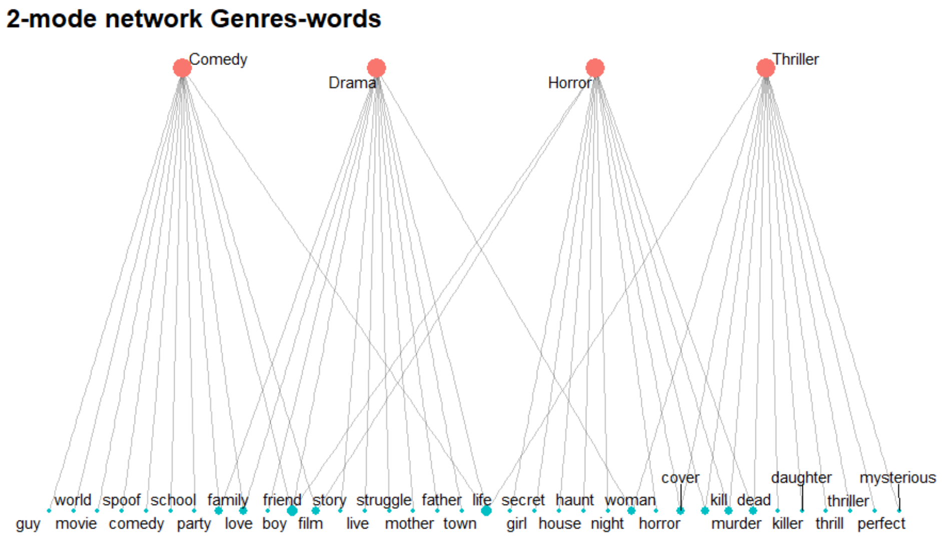
During the project multiple networks have been created to visualize different forms of attachments between movies and words in the used dataset. The following is a network of bigrams, which are a sequence of adjacent words from movie description with the genre Drama, Thriller, Comedy or Horror from the year 2000 to the present.



Figur 1: Bigram movie network

The network shows that a lot of bigrams interact in groups or in small communities. This is due to bigrams often being identical in the same genre, but bigrams like serial killer often aint mentioned in genre like drama and comedy.

Below a 2-mode network have been created from the 12 words with the highest td_idf score from the four genres in the used dataset. The 2-mode network insures that genre nodes cant connect with other genre nodes and the same for the word nodes.

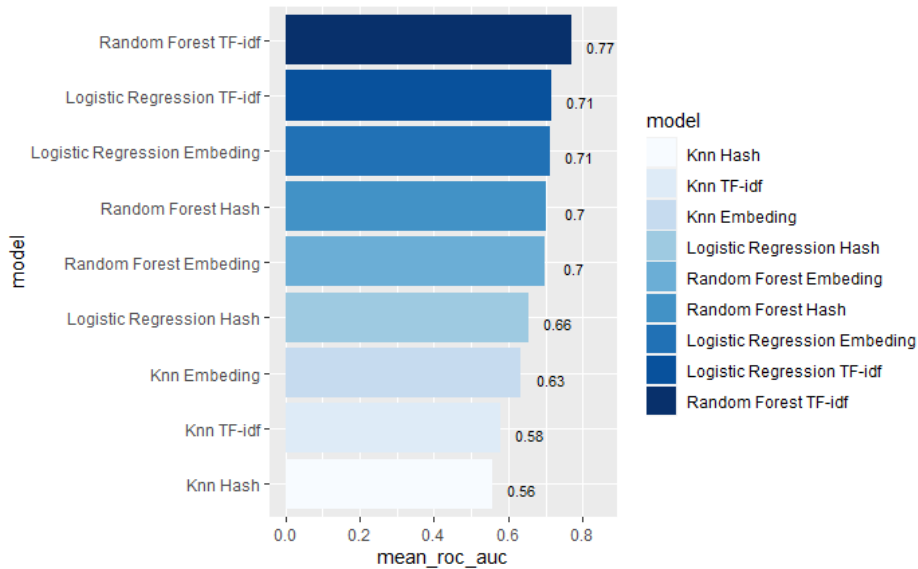


Figur 2: 2-mode network, genres-words

This 2-mode network visually shows how certain popular words are used in multiple genres, which makes them cross-genre popular. This applies to words like "friend" and "life" which are commonly used in three out of the four genres.

3.2 Supervised machine learning

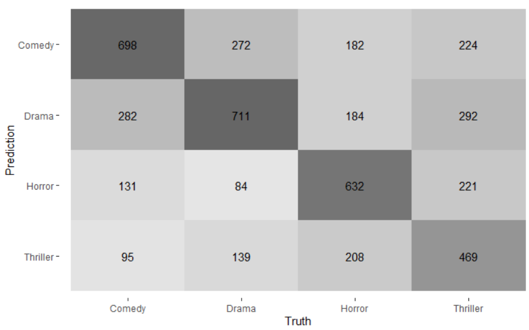
The goal for the supervised machine learning model is to be able to predict the genre of the movie based on its description. First the data is split into training and test data, Then the training data is downsampled to get an equal representation of each genre used. The data is preprocesed using



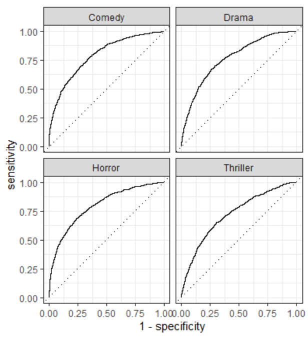
Figur 3: Model comparison

3 different methods: tf-idf, word-embedding and text-hash. Next three models are defined including a Logistic model, K-Nearest neighboor model and last a Random forest model. After hypertuning, the three models are fitted using resampled data, and different measures are collected for each of the 9 models, they are plotted for comparison. In figure 3 where the models are ranked by area under the curve.

After looking at other measures the Random forest model and logistic regression model performs the best, so the confusion matrix and roc-curve is made for these models shown in figure 4. The more observations in the diagonal the better for the confusion matrix, which means more true predictions. For a multiclass model it seems to do an okay job at predicting the genres. Also the area under the curve seems to give an okay result showing the False-positive fraction agains the True-positive fraction.



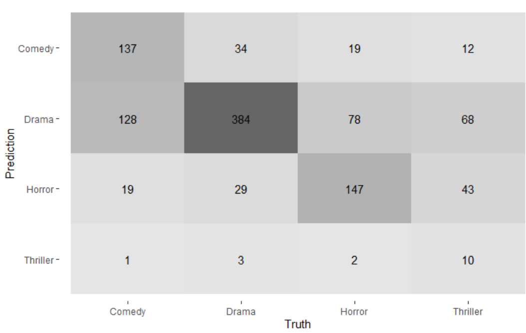
(a) Confusion matrix score



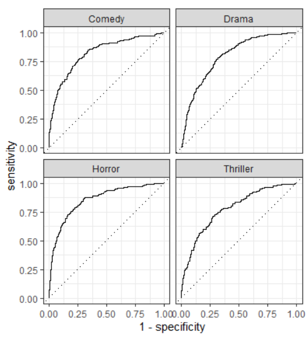
(b) Area under the curve names

Figur 4: Model Comparison plot

Now the two models are used on the test data, if the models are not over fitted we should get almost the same results.



(a) Confusion matrix on test data score



(b) Roc-auc on test data names

Figur 5: Model Comparison test data plot

The testing data is not downsampled like the training data, which also leads to the model beeing exposed to a lot more Drama movies. Still looking at figure 5 the model does an okay job at predicting Comedy, Drama and Horror movies, but has some problems predicting Thrillers.

3.3 Similarity prediction

In this part the main objective is to create a model using "doc2vec", that can act as a search predictor. This will predict similar words or movies based on either a keyword, a sentence or another movie. For example one can enter ones favorite movie to find a similar movie. Multiple models have been made. In the first model the input is a keyword and the output is similar words. These words are ranked based on a similarity score.

term1	term2	similarity	rank
<chr>	<chr>	<dbl>	<int>
vampire	vampires	0.6475565	1
vampire	blood	0.5586129	2
vampire	werewolf	0.5436665	3
vampire	bitten	0.5167841	4
vampire	clans	0.5080011	5

Figur 6: Similarity prediction: Word to word

In figure 6 the results from the keyword "vampire" can be seen. All words seems to be related to vampire in some way. Another model has been used to predict a movie based on a sentence of keywords:

term1	term2	similarity	rank
<chr>	<chr>	<dbl>	<int>
sent1	tt1656179	0.6045629	1
sent1	tt0050530	0.5979193	2
sent1	tt7200946	0.5957882	3
sent1	tt1228987	0.5945967	4
sent1	tt3898776	0.5797982	5

(a) Similarity scores

imdb_title_id	title	original_title	year	date_published	genre	duration	country	language
<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<chr>	<chr>
tt0050530	I Was a Teenage Werewolf	I Was a Teenage Werewolf	1957	1957-06-19	Drama, Fantasy, Horror	76	USA	English
tt1228987	Blood Story	Let Me In	2010	2011-09-30	Drama, Fantasy, Horror	116	UK, USA	English
tt1656179	I Kissed a Vampire	I Kissed a Vampire	2010	2012-03-30	Musical	91	USA	English
tt3898776	Aaron's Blood	Aaron's Blood	2016	2017-06-02	Drama, Horror, Mystery	80	USA	English
tt7200946	Oh, Ramona!	Oh, Ramona!	2019	2019-06-01	Comedy, Romance	109	Romania	English

(b) Movie names

Figur 7: Similarity prediction: Sentence to movie

The input in figure 7 is the sentence "vampire", "werewolf", "teenager". The results show that the most similar movies are *I Was a Teenage Werewolf*, *Blood Story* and *I kissed a vampire*. It should be noted that not every movie ever made is part of the dataset.

Finally a model predicting movies from the similarity score of another movie was made as can be seen below.

term1	term2	similarity	rank
<chr>	<chr>	<dbl>	<int>
tt0004873	tt0021599	0.6652263	1
tt0004873	tt1577811	0.6132298	2
tt0004873	tt0068190	0.5994261	3
tt0004873	tt0035771	0.5745484	4
tt0004873	tt0465407	0.5403202	5

(a) Similarity scores

imdb_title_id	title	original_title	year	date_published
<chr>	<chr>	<chr>	<dbl>	<chr>
tt0021599	Alice in Wonderland	Alice in Wonderland	1931	1931-09-30
tt0043719	Nuda ma non troppo...	Lady Godiva Rides Again	1951	1951-10-25
tt0068190	Le avventure di Alice nel paese delle meraviglie	Alice's Adventures in Wonderland	1972	1973-04-22
tt0111276	Nella trappola	Stalked	1994	1994-10-01
tt1577811	Fun in Balloon Land	Fun in Balloon Land	1965	2009

(b) Movie names

Figur 8: Similarity prediction: Movie to movie

The input in figure 8 is the movie id of the movie *Alice in Wonderland*. The results show that the most similar movies are another version of *Alice in Wonderland*, *Lady Godiva Rides Again* and *Alice's Adventures in Wonderland*.

Additional code, models and elaborations can be found in the attached pdf file.