Movie Genre Classification from Subtitles

Enes Akdogan, Omer Akgul, Ali Burak Erdogan, Mesut Gurlek, Mustafa Yildiz

How can we learn movie genre from a subtitle content?

OVERVIEW

- DATASET
- METHODS
  - HEARING IMPAIRED DESCRIPTIONS
  - DPM AND WPM
  - FULL TEXT OF SUBTITLES
- MODEL COMBINATION
- OVERALL RESULTS & EXPERIENCE
- CONCLUSION

Dataset

Before;
1000+ Subtitles

- 9000+ Subtitles
- 6000+ hearing impaired
- Opensubtitles API
- 8 genres from IMDB
- Action, Comedy, Crime, Horror, Musical, Romance, War, Western

Data Separation

- %20 Test Set
- %80 Training Set

Cross-Validation

- Training Data:
  - %20 Validation Set
  - %80 Training Set
- Tune Parameters with several iteration
- In each iteration shuffle training data and divide again

Hearing Impaired Descriptions

Motivation and Methodology

- Distinguishing hearing impaired descriptions
- Preprocessing to separate descriptions
- NLTK: stopwords, stemming
- TF-IDF vectorizer
- Multinomial Naive Bayes

```
00:13:48,766 --> 00:13:49,966
[MUTTERING]

183
00:13:51,466 --> 00:13:52,566
SEND--

184
00:13:52,633 --> 00:13:56,232
[WHISPERING]
They're sending a bus.
```

```
00:01:18,805 --> 00:01:20,371
(PLAYING)

35
00:01:22,943 --> 00:01:24,509
(CLAPPING RHYTHMICALLY)

36
00:01:31,084 --> 00:01:32,333
- (GUNSHOT)
- (HORSES NEIGHING)

37
00:01:32,334 --> 00:01:34,752
- (GUNSHOTS)
- (GROANING)

38
00:01:39,592 --> 00:01:41,725
(INDISTINCT SCREAMING)
```

Results and Experience

- Distinguishing Feature for specific genres
- Analysing f1 scores not average accuracy score
- Horror, Musical, War, Western gives promising scores
- Leads to model combination

	precision	recall	f1-score	support
Action	0.36	0.24	0.29	75
Comedy	0.29	0.30	0.30	69
Crime	0.34	0.35	0.35	68
Horror	0.51	0.57	0.54	58
Musical	0.62	0.54	0.58	80
Romance	0.26	0.31	0.28	65
War	0.64	0.66	0.65	65
Western	0.69	0.80	0.74	60
avg / total	0.46	0.46	0.46	540

Dialog per Minute & Word Per Minute

Motivation and Methodology

- Movie dialog frequencies can have

  a pattern

- Knn with Dialog per Minute, Word

  per Minute

Action	50.92884278822343
Adventure	50.29885300837908
Comedy	70.20427639111396
Crime	59.13997907503586
Horror	46.04569823281735
Musical	74.99845382251573
Romance	62.0360795765368
War	55.79225354108949
Western	55.404143056896224

Table: Word per minute

Action	11.752364947756044
Adventure	12.398248315034923
Comedy	15.543227618801078
Crime	12.997386458500177
Horror	11.029250406923195
Musical	14.046774517883888
Romance	13.973299009645306
War	11.637057205037284
Western	11.394383759610237

Table: Dialog per minute

Results and Experience

- Crime, Comedy, Romance scores are promising
- Can be supporting model for some categories

	precision	recall	f1-score	support
Action	0.11	0.14	0.12	66
Comedy	0.31	0.24	0.27	80
Crime	0.39	0.37	0.38	97
Horror	0.24	0.27	0.25	83
Musical	0.30	0.31	0.30	75
Romance	0.47	0.54	0.50	114
War	0.29	0.24	0.27	103
Western	0.15	0.16	0.16	68
avg / total	0.30	0.30	0.30	686

Table: Score for dpm, wpm

Whole Subtitle Text

Methodology & Results

- Multiclass SVM model
- Single label & Multi label
- Word2vec

	precision	recall	f1-score	support
Action	0.72	0.64	0.68	89
Comedy	0.70	0.70	0.70	89
Crime	0.74	0.73	0.73	89
Horror	0.87	0.85	0.86	89
Musical	0.85	0.90	0.87	89
Romance	0.77	0.75	0.76	89
War	0.86	0.93	0.89	89
Western	0.93	0.97	0.95	88
avg / total	0.81	0.81	0.81	711

Table: Only full text classification with single-label statistics

Number of labels	Accuracy for the number
1	80%
2	90%
3	95%

Table: Multi Label Results

Model Combination

Motivation and Methodology

- 3 different model and each have strength and weakness in different genres
- Promising f1 scores for some categories in each model
- Decided to use f1 scores to combine models

Combination Logic

Overall Results and Experience

- **Since we have tried different features, models and model combination we could not get perfects results.**

- **Pure Full Text classification with SVM gave the best results for our project.**

- **Data Collection, Preprocess and training are time consuming processes**

- **Learning with small dataset results might be deceitful**

- **Larger Dataset, better and reliables results**

- **Multi label classification makes difficult to combine models**

- **Intuitive features might give trivial results**

Conclusion

- 8 categories
- The best model was whole text svm model (~79%)
- Combination of models system was adequate (~75%)
- Tension modeling was not very successful.
- Hearing impaired data and the whole text was useful