Text-based video genre classification using multiple feature categories and categorization methods

Master's Thesis. Definitive version

Chris van der Lee

Department of Communication and Information Sciences

c.vdrlee@tilburguniversity.edu

s4000528

June, 2017

Abstract

The aim of this work is to categorize movies into genres using text-based features. Textual, syntactical and content-specific features are extracted from subtitles in the SUBTIEL corpus. The effectiveness of these three feature types is then compared using five algorithms (AdaBoost, C4.5, Naive Bayes, Random Forest, and SVM) and four methods are tested to combine these features (supervector, add-rule meta-classifier, product-rule meta-classifier, algorithm-based meta-classifier). The experimental results show that of the three feature types, the content-specific features result in the most accurate classifier. Furthermore, it is found that the Random Forest and SVM techniques are the two most accurate algorithms and that combining the textual, syntactical and content-specific features results in a more accurate classifier. However, the effectiveness of combining these three classifiers is largely dependent on the combination method: the algorithm-based meta classifier yields the largest improvement over the individual feature type classifiers, while the supervector approach results in lower classification accuracy compared to the individual content-specific classifier.

1. Introduction

Although genre classification is relevant for many domains, such as prose (e.g. Kessler, Nunberg, & Schütze, 1997), poems (e.g. Mu, 2015), music (e.g. Fell & Sporleder, 2014), and web pages (e.g. Ashegi et al., 2014), the high number of new movie releases every year (approximately 4,500 new professionally produced films published by a commercial organisation; Rasheed, Sheikh, & Shah, 2003) in combination with the surge of video streaming services such as Netflix and YouTube have made automated categorizing of videos by genre increasingly important (Fourati et al., 2014). Genre classification on videos could be done with different modalities (audio, visual and text). In this study we will use a text-only approach to classify videos by genre. This is done with the SUBTIEL corpus, which consists of a large set of Flemish and Dutch subtitles. The corpus contains subtitles that were professionally made as well as subtitles that were made by amateurs (fansubs). The goal of the study is to identify a set of features and a method of analysis that could characterize genres sufficiently well, which allows for the design of an effective genre classifier.

The term 'genre' is an ambiguous term that has been used to refer to different concepts in past classification research. Some studies used the notion of genre to refer to the distinction between letters, prose, poems, etcetera (e.g. Kessler et al., 1997), others used it to refer to the differences in the topic of a medium (geography, history, biology, etcetera; e.g. Dermitaş, 2009), and the notion has also been used in reference to sub-genres of a medium (action, comedy, drama, horror, etcetera). This article will focus on the latter notion of genre. In the interest of simplicity, the term 'genre' will be used instead of talking about 'sub-genres' throughout the article. In addition, as to avoid confusion, the notion to refer to the distinction between e.g. letters, prose, and poems will be called ‘domain’ and the notion to refer to the difference between e.g. geography, history, and biology will be called ‘topic’. The differences between domain, genre and topic go further than level of specificity. The effectiveness of different techniques is different for the three. For instance, lay-out would be an effective feature for domain classifiers, but is unlikely to be as effective for genre or topic classifiers. Similarly, using a WordNet approach to measure the average semantic differences between the text in subtitles and the target words would be effective for topic (e.g. the text of a nature documentary would likely be semantically closer to the word ‘nature’ than ‘music’; Katsiouli, Tsetsos, & Hadjiefthymiades, 2007) and probably less effective for domain (prose is rarely about the topic of prose).

The paper is organized as follows. Section 2 describes relevant literature and specifies the topic of the study. Section 3 describes the SUBTIEL corpus and the applied methods of classification analysis. In Section 4 an analysis is performed to evaluate the different techniques used and how these techniques affect classification performance. Section 5 discusses the significance and limitations of the findings and provides suggestions for future work.

2. Literature

The importance of classifying objects by genre in the domain of computational linguistics has been recognized by various scholars (e.g. Anwar, Salama, Abdelhalim, 2013; Ashegi, Sharoff, & Markert, 2011; Kessler et al., 1997; Brezeale & Cook, 2006; Dermitaş, 2009; Fourati, Jedidi, & Gargouri, 2014; Hong, & Hwang, 2015; Lee, & Myaeng, 2002; Lim, Lee, & Kim, 2004; Wan, Yang, Mo, Liu, Zhang, & Li, 2015; Wortman, 2010). One reason why accurate genre classification could be useful in natural language processing is that the accuracy of parsing, POS tagging and word-sense disambiguation could be increased, since the genre of a text could play a role in some language usages embedded in grammatical constructions and word senses (Giesbrecht, & Evert, 2009; Webber, 2009). Furthermore, genre classification plays an important role in making the information-retrieval process more manageable. It is often difficult for people to find the relevant information in the right genre (Vidulin, Luštrek, & Gams, 2007), and genre classification helps with the indexation of large amounts of data so that the choices can be narrowed down when a user performs a search task. Similarly, genre classification could play an important role in automated systems that recommend media to users based on their preferences.

The linguistic benefits of constructing a genre classifier should not be underestimated as well. Automated classification analyses increase our understanding about, for instance, the semantic and linguistic parameters that vary between different genres. Subtle differences in word choice, stylistic features or grammar, which are often overlooked by humans, could be detected by algorithms used for classification tasks (e.g. Biber, Conrad, Reppen, et al., 2004). Furthermore, the semantic gap, referring to the gap between object features on one hand (e.g., texture, color, text font, lay-out, audio volume, etc.) and semantic concepts on the other (i.e. concepts meaningful to human beings, such as cars, faces, but also genres) is one of the most challenging issues in computer science today (Huang, Fu, & Chen, 2010; Lew, Sebe, Djeraba, & Jain, 2006). Developing genre classifiers could help to obtain a better understanding of the content and structure of several media.

2.1 Different modalities of video genre classification

An interesting aspect about video is that it can incorporate different types of data (i.e. text, audio, and images). Creators of video utilize these textual and audio-visual streams to create semantic meaning. For humans it is often easy to recognize these semantic meanings and information, but to automate this process using information technology is a challenging issue. In automatic genre classification of movies or series episodes, a representation is made of the values of features that may be predictive of a genre. Based on the values that the classifier learned from a training corpus, it can guess the genre of a new movie or series episode. Therefore, it is important to select the features that can make a clear distinction among the genres. There are many different features to choose from with video and there is a lot of variety in the features that researchers utilized for their system. However, it is possible to distinguish three main approaches to video genre classification (VGC): text-based approaches, audio-based approaches, and visual-based approaches. Most authors use a variety of features for their system, sometimes even from more than one modality (Brezeale & Cook, 2008).

2.1.1 Video genre classification using audio and visual features

Most of the approaches to VGC use visual elements, either in combination with text or audio features or visuals-only (Brezeale & Cook, 2008). Audio-only approaches are less popular, although they are still more common than text-only approaches. Furthermore, most of the classifiers using video or audio features focus on the domain level or topic level. For example, Roach, Mason, and Pawlewski (2001) use a visual feature: the motion of foreground objects to classify a video as either a cartoon or a non-cartoon. Truong, Dorai, and Venkatesh (2000) use average shot length, percentage of each type of shot transition (cut, fade, dissolve), camera movement, pixel luminance variance, rate of static scenes, and other visual features to classify videos into a cartoon, commercial, music video, news video or sports video. On the topic-level there are studies by Liu, Wang, and Chen (1998), for example, who observe that sports have a nearly constant level of noise. Therefore, they use the volume standard deviation and volume dynamic range to accurately classify a news video as either sports or non-sports. Kobla, DeMenthon, and Doermann (2000) use instant replay, fraction of frames with motion, standard deviation of the motion, and other visual features on news videos to distinguish sports and non-sports.

The video and audio approach has also been proven effective for genre classification. For instance, Moncrieff, Venkatesh, and Dorai (2003) successfully use changes in sound energy intensity to distinguish horror and non-horror movies and Vasconcelos and Lippman (2000), and Rasheed et al. (2003) use low-level visual features, such as the average shot length, shot motion, lighting and color variance to classify movie previews into different categories of genre, such as comedy, action, drama and horror.

2.1.2 Video genre classification using textual features

Text-based approaches are the least utilized approach for VGC. One of the main reasons for this is that there is a lack of textual information associated with videos (Oger, Rouvier, & Linarès, 2010). Additionally, most of the textual information (e.g. transcripts, verbally anchored text and subtitles) are largely dialog. This means that the textual information does not capture a sizable portion of what is happening in the video. To the advantage of text based methods is that the human language carries more semantic information than visual or audio features, and text processing is computationally less exhaustive compared to video and audio processing (Dermitaş, 2009). There is also a large body of previous research on document text classification that could be utilized for text-based analysis. In other words, despite the difficulties with obtaining useful textual content for videos, there are several benefits to using a text-only strategy for VGC over the other approaches. Furthermore, previous scholars have stated that text information is more reliable and effective for video classification than video and audio information (Wang, Cai, & Yang, 2003).

However, it should be noted that the reliability of textual information depends on the type of text information that is used. For instance, some studies utilize automatic speech recognition or optical character recognition (OCR) of the text on screen to gather text data (Wang et al., 2003; Qi, Gu, Jiang, Chen, & Zhang, 2000). These techniques could be useful, but they still have quite high error rates. When there is background noise, the accuracy of these techniques decreases (Hauptmann, Jin, & Ng, 2002).

Subtitles or closed captions (textual information for hearing-impaired people) tend to be more accurate. Because of this relative preciseness, subtitles and closed captions are the most common type of text for text-based classification of videos. However, most of the research using subtitles and closed captions focuses on topic-level classification, especially news topics. Several authors (Zhu, Toklu, & Liou, 2001; Nassar, Taha, Nazmy, & Nagaty, 2007; Anwar et al., 2013; Wang et al., 2003; Lin & Hauptmann, 2002) use closed captions to classify news topics such as Politics, Sports, Weather, etcetera. Katsiouli et al. (2007) and Dermitas (2009) use subtitles for topic classification on documentaries by applying NLP techniques on subtitles and using a WordNet textual database and WordNet Domains. They define documentaries in categories such as Geography, History, Politics, etc. and achieve classification accuracy scores above 70% but do not provide a baseline to compare these scores to. Both Katsiouli et al. (2007) and Dermitas (2009) also find that the accuracy of the classifier is highly dependent on what topical labels are considered. They attain high accuracies on topics such as sports, but lower accuracies on topics such as Daily Events, because the categories are unspecific or because there is not enough data available.

Thus, there are multiple studies studying video topic classification using subtitles or closed captions, but studies where subtitles or closed captions are used for VGC are quite rare. However, there is some comparable research where other texts are used for VGC. Blackstock and Spitz's (2008) genre classifier uses a small set of movie scripts, Fourati et al. (2014) use movie synopses to classify movie genres via a WordNet semantic similarity approach and Hong and Hwang (2015) draw social tags from IMDB for genre classification. They achieve accuracy scores of 50% to 65%, but do not mention a baseline to compare the performance to. Additionally, it is interesting to look at similar text-based genre classification for other types of media. Multiple authors (e.g. Fell & Sporleder, 2014; Bou-Rabee, Go, & Mohan, 2012; Walsten & Orth, n.d.; Sadovsky & Chen, 2006) use lyrics to classify songs by their genre (e.g. Rock, Country, Hip Hop), and seem successful in their approach (up to 48% above baseline accuracy in the case of Sadovsky & Chen, 2006). There is also some genre classification research that successfully classifies book genres: Kessler et al. (1997) and Samothrakis and Fasli (2015), for instance, achieve accuracy scores of 20% above baseline.

As stated before, the research on VGC using subtitles or closed captions is quite rare and most of these studies are small-scale studies. Brezeale, and Cook (2006) used closed captions to detect movie genres (e.g. action, adventure, romance, etc.) using 81 movies from the MovieLens dataset. Although they used a relatively small dataset, they achieve a high accuracy score of almost 90%. However, it should be noted that the movies in their dataset could contain several genre labels and the classification was deemed accurate if only one of those labels was correctly identified, plus there was no information provided to determine a baseline. Helmer and Ji (2012) look into the closed captions of film trailers. They use a very small corpus, consisting of only 3,563 words and report a 16% higher accuracy score than the baseline (44% vs. 28% baseline accuracy). Wortman (2010) classifies subtitles of 1,184 movies in several genres. Contrary to the other two studies, the author uses an indirect classification approach: clustering the words into multiple factors first and then using these factors to predict movie genres. This results in an F1-score of around 0.55, but Wortman (2010) fails to provide means to compare this score to a baseline.

2.2 Feature types

2.2.1 Content-free features versus content-specific features

Selecting features that can make a clear distinction among the genres is an important step in automatic genre classification. For text-features, a distinction can be made between two broad categories: content-free features and content-specific features. Content-free features are features that can be regarded as generic features: they are independent of the topics or domains of the text data. Examples of content-free features are average word length, type-token ratio, amount of function words, etc. Content-specific features are features that can represent a specific topic. For instance, when distinguishing football videos from non-football videos, it is useful to identify informative content-specific keywords, such as ‘penalty’, ‘corner’, ‘free kick’, etc. This could also be useful for genre classification, assuming that similar topics arise within the same genre (e.g. the topic of space in science fiction).

When looking at the early history of automatic genre classification (Biber, 1986, 1992, 1995; Karlgren, Bretan, Dewe, Hallberg, & Wolkert, 1998; Karlgren & Cutting, 1994; Kessler et al., 1997; Stamatos, Fakotakis, & Kokkinakis, 2000), there is a focus on automatic genre classification using content-free features. These scholars developed classifiers with various text statistics such as word length and use of exclamation marks. These features have also been considered in more recent text-classification research (Abbasi & Chen, 2008; Abbasi, Chen, & Nunamaker, 2008; Zheng, Li, Chen, & Huang, 2006), but they are rarely incorporated in video classification studies (Huang et al., 2010). Most of the video classification studies incorporated a content-specific bag-of-words approach in their classifier where each term in the feature vector represents a (key)word of a studied text.

2.2.2 Combining features

While using content-specific features has resulted in VGCs with reasonable accuracy (e.g. Brezeale & Cook, 2006; Helmer & Ji, 2012; Wortman, 2010), it seems conceivable that combining these features with content-free features could further improve the classification accuracy. That is because a single feature category often does not provide all the information necessary for accurate classification (Lin & Hauptmann, 2002). It could be difficult, for instance, to distinguish a war movie like *M*A*S*H* from an action movie like *Die Hard* by merely looking at the differences in the lexicon. Combining the content-specific features with content-free features could improve a classifier so that it makes that distinction more accurately.

The assumption that more information useful for categorization could be acquired using different types of features is supported by a vast body of research that has combined different features for their classifier. Most of this research focuses on combining features from different modalities. For example, Lin and Hauptmann (2002) combine classifiers of visual and text features and Wang et al. (2003) use features of the visual, text and audio modalities. However, most multi-modal video classifiers combine audio and visual features (e.g. Huang, Liu, Wang, Chen, & Wong, 1999; Qi, Gu, Jiang, Chen, & Zhang, 2000; Jasinschi & Louie, 2001; Roach, Mason, & Xu, 2002; Rasheed & Shah, 2002). Using different features from one domain is less common. There are some studies where content-specific features and content-free features are combined. Huang et al. (2010) combine these feature types to classify videos based on user-generated text data and report that combining these feature types increases accuracies up to 12%. Research by Dewdney, VanEss-Dykema, & MacMillan (2001) and Finn and Kushmerick (2006) also suggest that classification accuracy could be increased by using multiple types of text-based features. The authors combine the bag-of-words approach with content-free text statistics to make a distinction between different text domains and find that combining both types of features, instead of using only one type, increases accuracy.

In sum, there is a lack of studies that utilize text statistics in combination with bag-of-words features for VGC. However, the existing studies that use this combination of features suggest that it could increase the classification accuracy.

2.3 Classification techniques

While multiple techniques can be applied to a VGC using text features, three techniques have been the most frequently used in text classification studies (e.g., Das & Chen, 2007; Zheng et al., 2006). These techniques are C4.5, Naive Bayes and Support Vector Machines.

C4.5 is an extension of ID3, which was an algorithm that has been proven to have strong predictive power (Chen, Shankaranarayanan, She, & Iyer, 1998; Dietterich, Hild, & Bakiri, 1990). The C4.5 technique is a decision-tree algorithm that classifies mixed objects into categories according to the attribute values of objects (Huang et al., 2010).

The Naïve Bayes classifier is a probabilistic classifier based on Bayes' theorem with strong assumptions of independence of features. The classifier uses the feature values of a new instance to estimate the probability of the instance belonging to each category. Several text-classification studies use this algorithm with good results (e.g. Lewis, 1998; Mccallum & Nigam, 1998; Sahami, 1996).

The idea behind SVMs is to find a linear boundary to separate two or more classes within (a transformation of) the feature space. This technique is powerful (Vapnik, 1998), and efficient for classification tasks (Huang et al., 2010). The SVM is used in text-based authorship classifiers (e.g. De Vel, 2000), but there are also multiple video-classification studies that show the excellent performance of the SVM (Jing, Li, Zhang, & Zhang, 2004; Huang et al., 2010; Lazebnik, Schmid, & Ponce, 2006; Zhang, Marszałek, Lazebnik, & Schmid, 2007).

Additionally, it is worth looking into two more variants of decision trees for the current study: AdaBoost and Random Forest. AdaBoost has been found to produce accurate classification results in several tasks and is well suited to integrate different types of features. Random Forest is a fast and robust classifier. It is sometimes discouraged to use Random Forest with large feature vectors such as bag-of-words, but the technique is not prone to overfitting, making it a useful algorithm for large datasets.

How to combine a large set of different features is a challenging task. Some research groups all these features together in a supervector (e.g. Oger et al., 2010), or base their classification on a list of predetermined rules (Païs, Lambert, Deloule, Beauchêne, & Ionescu, 2012; Silla Jr., Kaestner, & Koerich, 2007). Lin & Hauptmann (2002) posit that classification accuracy is increased when a meta-classification is used. The idea is that there are classification judgements generated by a classifier for each of the feature types. Then, another machine learning classifier, the meta-classifier, treats each generated judgement as a feature. A similar approach is also adopted by, amongst others, Wang et al. (2003), Kennedy and Inkpen (2006), and Giannakopoulos, Makris, Kosmopoulos, Perantonis, and Theodoridis (2010). These scholars report that the machine learning meta-classification approach performs better compared to the supervector approach. Additionally, the performance of rule-based approaches compared to the supervector-approach is found to be slightly better in Silla Jr. et al.'s (2007) study. However, the performance of the machine learning meta-classifier compared to the rule-based meta-classifier has rarely been investigated for genre classification research.

2.4 Research gaps

Based on the review of the literature, several important research gaps have been exposed. For one, there is a scarcity of text-based video classifiers. The existing text-based video classifiers rarely look into genre-level classification and even more rarely use closed captions or subtitles. This lack of literature on text based VGC is unfortunate, especially since classification based on subtitle and closed caption features are generally more accurate compared to audiovisual features, classification using text-features impose the fewer computational requirements compared to audiovisual features, and a need for accurate genre-level classifiers is expressed by multiple scholars (Brezeale & Cook, 2008; Rasheed et al., 2003; Fourati et al., 2014). In addition, it should be noted that the existing literature on text-based VGC worked with small corpora, and that the features that most of these video genre classifiers use are solely content-specific features, while classifiers in other domains have proven that content-free text statistics are likely to increase classification accuracy (Huang et al.,2010; Dewdney et al., 2001; Finn and Kushmerick, 2006).

The present research will make an attempt to address the aforementioned research gaps. First of all, by using the SUBTIEL corpus: a corpus that is considerably larger than the corpus used in

Table 1: *Numbers of movie and TV series subtitle documents per genres used in this study*

	Documents				Tokens			
Genre	**NL BTI**	**BE BTI**	**OS**	**Total**	**NL BTI**	**BE BTI**	**OS**	**Total**
Drama	19.515	8526	26.385	54.426	28.557.895	13.579.893	63.612.075	105.749.863
Comedy	13.371	6100	14.690	34.161	16.609.521	7.387.871	37.300.652	61.298.044
Crime	6563	3191	11.752	21.506	10.355.245	5.197.874	28.899.319	44.452.438
Action	5482	1876	12.991	20.349	8.000.793	2.692.911	30.083.145	40.776.849
Mystery	4984	1681	7086	13.751	7.586.760	2.943.631	15.439.003	25.969.394
Adventure	3173	1130	8191	12.494	4.455.613	1.570.201	20.118.353	26.144.167
Romance	5133	951	5377	11.461	6.504.396	1.511.698	15.090.559	23.106.653
Thriller	1424	514	8433	10.371	2.574.450	968.382	19.654.866	23.197.698
Sci-Fi	2012	704	4918	7634	2.882.174	884.426	10.916.066	14.682.666
Fantasy	1918	565	5002	7485	2.692.100	842.192	11.000.977	14.535.269
Family	2626	1244	2486	6356	3.115.843	1.451.949	6.446.973	11.014.765
Horror	671	184	4946	5801	1.046.589	273.138	9.441.596	10.761.323
Total	34.101	14.695	48.018	96.814	46.358.189	20.853.462	114.980.363	182.192.014

Note: each document can be involved in multiple genres: if the document, for instance, contains the labels *drama* and *comedy*, the document and its tokens are added to the totals of drama and comedy. For the *total* category, this document and its tokens would only be added once.

previous subtitle-based VGC literature. Additionally, the present study will combine the content-specific bag-of-words approach that is popular in current classifiers with content-free text statistics that has lost popularity over time.

3. Methodology

3.1 Data

For the current research a new corpus was compiled, called the SUBTIEL corpus. This corpus contains over 500,000 subtitles in Dutch and English for a broad variety of movies and television shows. They are either obtained from OpenSubtitles.org, a website that offers an extensive collection of fan-made subtitles (or: fansubs), or from the Dutch and Belgian branch of BTI Studios, a company that provides professional subtitling for broadcasters, film studios, VoD platforms and distributors. These subtitles were converted to the FoLiA XML format, which is a rich format for linguistic annotation (Van Gompel & Reynaert, 2013). In addition, the Internet Movie Database (http://www.imdb.com; IMDb), a website that provides official information and content for professionally produced television shows and movies, was mined for additional information about the film using an automated mining tool, including genre labels that conform to IMDb’s own genre labeling system. This tool located the genre labels of roughly 50% of movies and series. The metadata that was retrieved with this method was then saved in the CMDI XML format, which is a format used for the description of linguistic resources (Dima, Hoppermann, Hinrichs, Trippel, & Zinn, 2012). The movies and television shows that remained without genre labels were subsequently removed from the dataset for this study. The English subtitles were also not used for this study and thus discarded from the dataset. Additionally, the subtitles for videos of another domain than movies or TV shows (i.e. News, Reality-TV, Shorts, Talk Shows, and Documentaries) were not included in the final dataset of this study. This leaves a total of 96,814 movies and TV shows; see Table 1 for more detailed statistics on genre labels. Each of the

Table 2: *Co-occurrence frequency among the genres of this research*

	Drama	Comedy	Crime	Action	Mystery	Adventure	Romance	Thriller	Sci-Fi	Fantasy	Horror
Comedy	9325										
Crime	15330	3643									
Action	8428	2991	6675								
Mystery	10837	751	6792	1433							
Adventure	3682	2300	518	6641	602						
Romance	7249	5324	281	344	386	593					
Thriller	5042	262	3495	2761	1864	629	231				
Sci-Fi	2646	971	188	3055	1524	1786	100	1174			
Fantasy	4014	1268	257	1703	845	2445	395	185	485		
Horror	1824	552	274	718	1039	180	112	1736	952	1459	
Family	1732	2676	29	284	27	1466	323	0	250	801	4

films and TV shows in the corpus contain between 1 and 6 genre tags, with an average of 2.31. There are a total of 22 distinct genre labels in the dataset. However, to provide ample data to the classifier, only the genres that applied to more than 5% of movies were considered, which limits the dataset to 12 genres. There is a wide array of combinations between the different genres; detailed co-occurrence statistics are provided in Table 2.

3.2 Features

The current study combines the content-specific bag-of-words features of most recent text-based VGC research with the more uncommon content-free text statistics features. For the content-free text statistics, a selection of features is chosen from research on text domain classification and modified to be more suitable for genre classification of subtitles. This means that some text statistics that are not applicable to the current study were discarded (e.g. the number of paragraphs, since subtitles do not contain any paragraphs). Furthermore, text statistics such as the total number of words or characters are useful for classifying the domain of written texts (prose is generally longer than poetry), but for genre classification it might be more useful to instead use the average amount of words or characters per minute. Average amount of words or characters per minute could indicate how dialogue heavy a movie or TV show is, which possibly helps discriminate between genres (an action movie would probably contain less dialogue than a comedy).

Therefore, a selection of features is adapted from previous literature, which is shown in Table 3. A distinction is made between textual, syntactic, and content-specific text statistics, based on Abbasi & Chen (2008) and Huang et al. (2010). Textual features are statistical measures on the word or character level, such as sentence/word length and word length distributions. Syntactic features are indications of the syntactical patterns present in sentences. These patterns can be found by identifying function words, punctuation and part-of-speech tag n-grams. Content-specific features

Table 3: *Text features adopted in the present research*

Group	Category	Description	Feature counts
Textual	Average words per minute		1
	Average characters per minute		1
	Average word length		1
	Average sentence length in terms of words		1
	Average sentence length in terms of characters		1
	Type/token ratio	Ratio of different words to the total number of words	1
	Hapax legomena ratio	Ratio of once-occurring words to the total number of words	1
	Dis legomena ratio	Ratio of twice-occurring words to the total number of words	1
	Short words ratio	Words < 4 characters to the total number of words	1
	Long words ratio	Words > 6 characters to the total number of words	1
	Word-length distribution	Ratio of words in length of 1-20	20
Syntactic	Function words ratio	Ratio of function words (e.g. 'dat', 'de', 'ik') to the total number of words	1
	Descriptive words to nominal words ratio	Adjectives and adverbs to the total number of nouns	1
	Personal pronouns ratio	Ratio of personal pronouns (e.g. 'ik', 'jou', 'mij') to the total number of words	1
	Question words ratio	Proportion of wh-determiners, wh-pronouns, and wh-adverbs (e.g. 'wie', 'wat', waar') to the total number of words	1
	Question mark ratio	Proportion of question marks to the total number of end of sentence punctuation	1
	Exclamation mark ratio	Proportion of exclamation marks to the total number of end of sentence punctuation	1
	Part-of-speech tag n-grams	Part-of-speech tag n-grams (e.g. 'NP', 'VP')	Varies
Content-specific	Word n-grams	Bag-of-word n-grams (e.g. 'lat', 'erg hoog')	Varies

are words or phrases that are an indication of certain topics.

3.3 Text preprocessing

Some preprocessing steps have already been performed in the SUBTIEL corpus. For instance, the sentences of the subtitle file were split from each other, all the words in a sentence were tokenized (i.e. punctuation markers were split from word strings), and the beginning and end times of every subtitle line were logged as metadata. In addition to these preprocessing steps, some additional preprocessing was undertaken for the present study. The syntactical features required part-of-speech information on the sentence level. Therefore, the Dutch POS-tagger *Pattern* (De Smedt, & Daelemans, 2012) was applied to the subtitles, which determined the word class of every word in a sentence.

The unprocessed content-specific feature set was initially too large and required strategic data reduction to make classification with these features possible. Therefore, stop words (common words such as 'and' and 'the), number strings, and punctuation were removed. These elements substantially increase the computational requirements, while it is argued that they contain relatively little information about the category of the document (e.g. Lin & Hauptmann, 2002). Lemmatization was also applied and all words were converted to lowercase. Lemmatization is normalizing word variations by removing the prefixes and suffixes of words so the affix-free lemma of the word can be used for further processing. This ensured that different realizations of the same word were pooled together, which could increase classification performance. These preprocessing steps for the content-specific feature category decreased the computational requirements, while retaining most of the useful information.

A second round of feature reduction was undertaken for the content-specific and syntactic features, since the size of these feature sets were still deemed too large. To further optimize performance, word and part-of-speech n-grams that occurred less than 10 times in the corpus were removed. Additionally, feature selection was performed for all feature categories. Using chi-square, the remaining features were ordered by importance. Then, when performing the classification task, features were added gradually by their importance until no more increases for accuracy were observed. Using this technique, no more than 10% of the part-of-speech and word features were found to be necessary for maximum classification accuracy.

3.4 Classification

The features described above are used as input for the machine learning classifiers in different ways, in order to test which type of feature combination method results in the most accurate classification. This means that the features are either combined into a supervector, a rule-based meta-classifier, or a machine-learning meta-classifier. For the supervector approach all groups of features are combined as the input vector. Using ten-fold cross-validation, the classifier is trained on the training input after which it will be tested with the remaining test data. The genre predictions are subsequently compared to the IMDb-tags to evaluate the classifier (see Figure 1 for a schematic workflow). The classifiers tested and evaluated in this study are AdaBoost, C4.5, Naive Bayes, Random Forest, and SVM algorithms (Quinlan, 1993; Freund, & Schapire, 1995; Breiman, 2001; Wu, Lin, & Weng, 2004; Zhang, 2004; Fan, Chang, Hsieh, Wang, & Lin, 2008; Zhu, Zou, Rosset, & Hastie, 2009;).

For the meta-classifier approach, the textual, syntactic and content-specific features are also extracted, and the training and test data is also separated by ten-fold cross-validation, just like with the supervector approach. However, whereas the supervector merged all the different groups of data to train and test the classifier on, the meta-classifier approach will keep these feature groups separate for the first part of classification. Each of the feature groups are converted into an input vector on which a classifier is trained. AdaBoost, C4.5, Naïve Bayes, Random Forest or SVM are then applied to the test-data, which results in prediction accuracy scores for each genre. For the machine learning meta-classifier these prediction accuracy scores are then used as features to train and test another C4.5, Naïve Bayes, or SVM classifier on. For the rule-based meta classifier a combination rule is applied to these prediction accuracy scores to come to a final classification. Two rules are used: for the sum rule all prediction accuracy scores are *summed up* for each genre; for the *product* rule all prediction accuracy scores are *multiplied* for each genre. The genres with a value above a certain

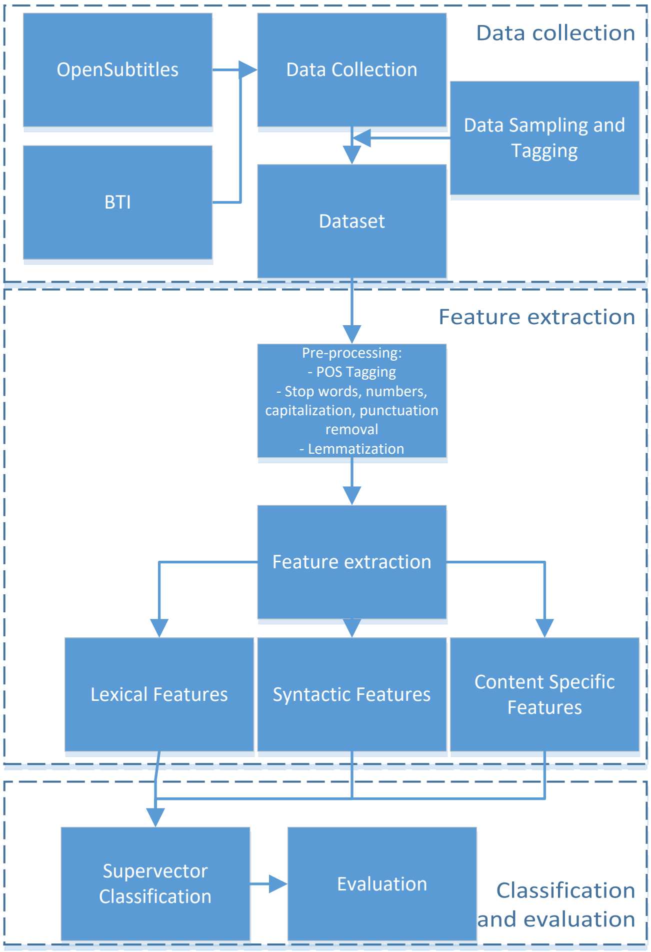


Figure 1: Video classification system using a supervector

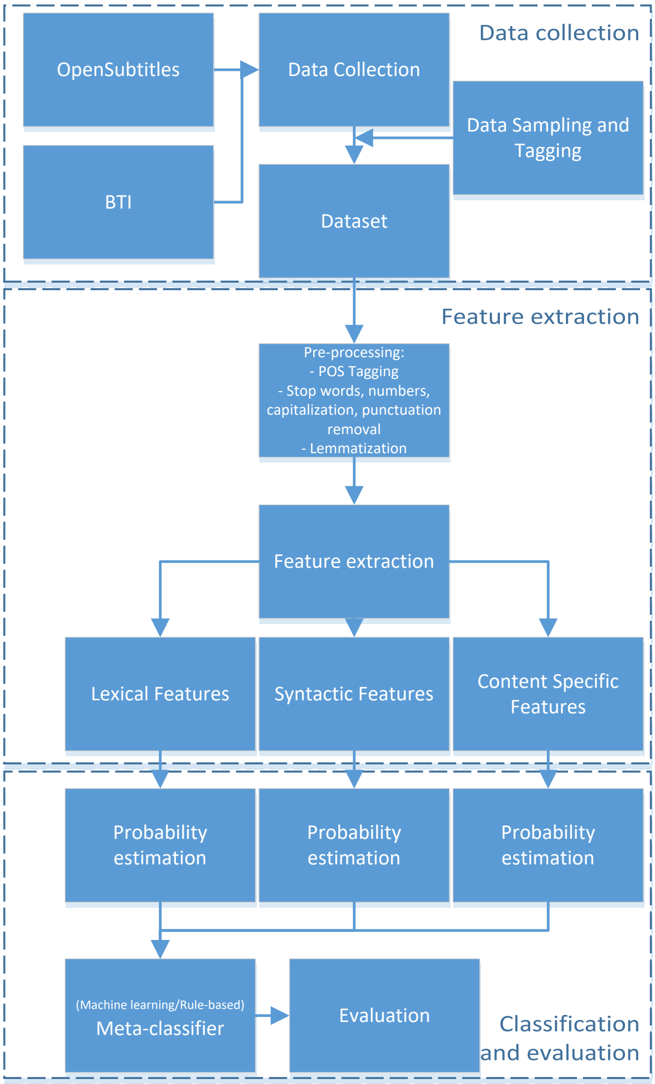


Figure 2: Video classification system using a meta-classifier

threshold are then chosen in the final classification. The optimal threshold for this is calculated with 30-step Bayesian optimization. After choosing the genres, an evaluation is done by comparing the predicted genres to the actual genres according to IMDB, as visualized schematically in Figure 2.

To optimize the performance of all classifiers, hyperparameter optimization is applied to all the machine learning algorithms. The best parameters for the algorithm were determined using 30-step Bayesian optimization on 10% of the corpus.

3.5 Accuracy measure

To measure whether one version of the classifier outperforms another, it is necessary to evaluate the outcomes of the classifiers in an objective way. To accomplish this, two standard information retrieval performance measurements were used: *precision* (P) and *recall* (R). *Precision* is the percentage of total items labeled as genre *g* that are actually members of *g*. *Recall* is the proportion

Table 4: *Classification accuracy per feature category and combination approach*

Method	Algorithm	# of features	Precision	Recall	F1-score
Baseline	-	-	0.56	0.26	0.36
Textual only	Random Forest Classifier	30	0.90	0.63	0.74
Syntactical only	Random Forest Classifier	180	0.91	0.69	0.79
Content-specific only	Linear SVM	188.500	0.91	0.77	0.84
Textual/Syntactical	Random Forest Classifier	180	0.92	0.71	0.80
Textual/Content-specific	Linear SVM	188.500	0.91	0.82	0.86
Syntactical/Content-specific	Random Forest Classifier	1886	**0.93**	0.75	0.83
Supervector	Random Forest Classifier	1886	**0.93**	0.75	0.83
Meta classifier (added rule)	-	-	0.87	**0.84**	0.85
Meta classifier (product rule)	-	-	0.90	0.80	0.85
Meta classifier (machine learning)	Linear SVM	36	0.92	0.83	**0.88**

Note, bold scores indicate the highest scores per column

of total items that are actually genre *g*, that are also correctly labeled as *g*. The classifier performance was based on the harmonic mean of *precision* and *recall*, which is called the *F1-score*. Similar to the formula of Van Rijsbergen (1979), a beta of 1 is used.

4. Results

4.1 Classifier evaluation

The first analysis assessed how well the different types of classifiers performed. For every classifier, the precision, recall and F1-score were calculated at the micro-level, which uses the total number of true positives, false negatives and false positives. These results were compared to a baseline that always predicted the most frequent labels in the training set. Table 4 shows the classification results for every feature category and combination of categories. The results were obtained using the optimal number of features in combination with the best performing algorithm.

The Random Forest and SVM classifiers were found to be the strongest algorithms amongst all the classifiers. Both algorithms are known to be effective on larger datasets. Additionally, while the accuracy of Random Forest has been known to deteriorate when more features are used, this effect was probably prevented by using the method of gradually introducing features until a saturation point was reached.

Most importantly, all the classifiers performed well above baseline score: classification resulted in a 0.38 increase over the baseline in the worst case up to a 0.52 increase in the best case, which indicates that all the attempted classification techniques were successful. When looking at the classifiers that used only one feature category, the best results were obtained with the content-specific classifier: 0.84 F1-score. This is expected because word n-grams also have some textual and syntactical information stored in them. However, the performances of the textual and syntactical classifiers were also quite robust, with F1-scores of 0.74 and 0.79 respectively. Although these scores are lower than the content-specific classifier, they are still well above the baseline score of 0.36. The findings also show that combining the different feature categories could result in a performance

Table 5: *Classification accuracy per genre (ordered by occurrence in the corpus) for the machine learning-based meta-classifier*

Genre	Precision	Recall	F1-score	% in dataset
Drama	0.92	**0.92**	**0.92**	26.45%
Comedy	**0.94**	0.90	**0.92**	16.60%
Crime	0.92	0.85	0.89	10.45%
Action	0.92	0.83	0.87	9.89%
Mystery	0.89	0.76	0.82	6.68%
Adventure	0.93	0.81	0.86	6.07%
Romance	0.90	0.64	0.75	5.57%
Thriller	0.93	0.72	0.81	5.04%
Sci-Fi	0.92	0.74	0.82	3.71%
Fantasy	0.93	0.76	0.83	3.64%
Family	0.89	0.60	0.72	3.09%
Horror	**0.94**	0.81	0.87	2.82%

Note, bold scores indicate the highest scores per column

increase over using one feature category only. However, whether increased performance is achieved or not depends on the feature categories that are combined and the combination technique that is used. Combining textual and syntactical features (F1-score: 0.80), and combining textual and content-specific features (F1-score: 0.86), resulted in higher F1-scores than the F1-scores obtained when using merely one of these feature categories. On the other hand, the results for the syntactical and content-specific combination (F1-score: 0.83) and the supervector results (F1-score: 0.83) were less accurate than the content-specific classifier. The most reliable increases were achieved with the meta-classifier approach. While the rule-based meta-classifiers achieve a slight increase over the content-specific classifier (F1-score: 0.85), the best result of all the classifiers was achieved with the machine learning-based meta-classifier (F1-score: 0.88). Interestingly, the machine learning-based meta-classifier did not perform the best in terms of precision or recall out of all classifiers, but rather had the most consistently high precision and recall.

While overall F1-scores were high, there were noticeable differences between genres in regard to classification accuracy. Table 5 shows these differences for the machine learning-based meta-classifier. The size of the training data seemed to have some effect on the classification accuracy: five of the six most frequently occurring genres had an F1-score of above 0.85. However, the F1-score for horror, the least occurring genre in the corpus, was also relatively high. This suggests that other factors also affect the performance. One of the factors that could have played a role is the distinctness of the genre and co-occurrence with other genres: horror is quite a specific genre that does not frequently co-occur with other genres, as is shown in Table 2. Thus, it would not often be confused with other genres. Romance and family on the other hand, the lowest performing genres in the current study, are less distinct genres that frequently co-occur with other genres. Romance frequently co-occurs with drama and comedy and family, and family with drama, comedy and adventure. This frequent co-occurrence with other genres could have made it more difficult for the classifier to correctly assign these genres.

Table 6: *Top 10 most important features per feature category*

Textual	Syntactical	Content-specific
Type/token ratio	.	agent
Words per minute	Descriptive/nominal ratio	wapen
Characters per minute	. DT	vampier
Hapax legomena ratio	Question mark ratio	moord
Dis legomena ratio	PRP	vermoorden
Short words ratio	VBD	zaak
Ratio of 4-letter words	DT NNP	pistool
Ratio of 1-letter words	VBN .	gedood
Ratio of 10-letter words	JJ	vernietigen
Ratio of 3-letter words	. .	gaan

4.2 Important features

The most important features for every feature category can be observed in Table 6, Appendix A shows the differences between the genres for these features. This information can also provide insight into the differences between genres. The information from Table 6 and Appendix A will be analyzed in this paragraph. It should be noted that these are speculative interpretations of the data and should not be confused with proper linguistic research, but it could help to instigate more fundamental linguistic research on language variety differences.

It could be derived from Appendix A and Table 6 that the biggest differences between genres are at the content-specific level, and that the most important features at this level are unigrams. Furthermore, most of the content-specific features in Table 6 were words frequently occurring for action, thriller, mystery, and/or crime documents. Appendix A shows that the word ‘agent’ (police officer) occurred most frequently for action, thriller, mystery and crime documents; ‘wapen’ (weapon) occurred most for action, thriller, crime and sci-fi documents; ‘moord’ (murder) most for mystery and crime; ‘vermoorden’ (to kill) and ‘zaak’ (case) most for thriller, mystery and crime; and ‘pistool’ (pistol) most for action, thriller and crime documents. Therefore, it seems that there is a jargon for action, thriller, mystery and crime movies that helped the classifier to accurately classify these genres. Furthermore, it can be observed in appendix A that ‘vampier’ (vampire) is a very specific word for fantasy and horror, which seems intuitively logical. ‘Vernietigen’ (to destroy) occurs most frequently in the adventure and sci-fi genre, and also occurs relatively frequently for fantasy and action. Defeating a villain is a common plot for these genres, which could explain this feature. The feature ‘gaan’ (to go) could be observed the most in subtitles with a romance, action, adventure, horror, or thriller label. This feature is somewhat more difficult to explain, since it could indicate movement (‘ga daarheen’/‘go there’), and a future plan (‘ik ga rennen’/’I am going to run’). Therefore, the higher instances of ‘gaan’ could indicate that there is more movement or travel in movies or TV shows of the aforementioned genres, or more conversations about future plans.

At the syntactical level, the period seems to be important. Not only is the period unigram the most important feature, the period-period, period-determiner (DT), and past particle verb (VBN)-

period bigrams were found to be important features as well. The period and period-period features were most frequent among thrillers. Thrillers would possibly contain more ellipses (…) in the subtitles, which are often used as a tool to raise suspense. Not only were the part-of-speech n-grams important features, the descriptive/nominal ratio and question mark ratio were also found to be effective tools for genre classification. Romantic movies and television shows had the highest amount of adjectives and adverbs per noun, while action movies and television shows had the lowest. Similarly, romance documents had the highest rate of adjectives (JJ). This could suggest that romance movies are more verbose and contain more extensive descriptions in conversations, while movies with a low descriptive/nominal ratio and low rate of adjectives (e.g. action) contain more direct and straight-to-the-point dialogue. For the question mark ratio, the mystery and crime documents contained the highest number of question marks, which seems to conform to intuition, since solving a problem (a crime) is generally the main topic in these genres.

At the textual level, the ratio of the unique words played an important part: type/token ratio, hapax legomena ratio and dis legomena ratio were all three important features. These three features showed that comedy and family movies and television shows contained the highest amount of unique words, which could indicate more complex dialogues and more topic variety. The dialogue-heaviness of a movie or television show was also found to be important, as the importance of words per minute and characters per minute illustrate. The most dialogue-heavy documents were mystery, comedy, thriller and crime documents.

5. Conclusion

In this study, a framework was proposed for text-based video genre classification. Three categories of text features (textual, syntactical and content-specific features) were used for classification, five classification algorithms (AdaBoost, C4.5, Naive Bayes, Random Forest, and SVM) were compared, and four methods of combining the text features (supervector, add-rule meta-classifier, product-rule meta-classifier, and machine learning-based meta-classifier) were tested. Feature selection was also adopted to further improve accuracy and identify key features. Experiments were conducted on subtitles from the SUBTIEL corpus, and showed a very good performance of the proposed framework.

Based on the findings of the current study, several conclusions can be drawn. The main finding of the current study is that accurate classification of video genres is possible using text features only. The potential of text-based video genre classifiers has been stated previously (Wang et al., 2003), but it was not often utilized as an approach for video genre classification, mostly due to a lack of textual information associated with videos (Oger et al., 2010). The SUBTIEL corpus, containing a large amount of annotated subtitle data, had solved this problem for the current study. The expansiveness of the SUBTIEL corpus could be attributed as one of the main reasons for the good performance of the classifiers. The best F1-score obtained in the current study (0.88) surpassed the baseline score (0.36) by a sizable margin. Furthermore, it was higher than the F1-scores obtained in similar text-based multiclass video genre classification (e.g. Blackstock & Spitz, 2008: 0.47; Fourati et al., 2014: 0.48; Hong & Hwang, 2015: 0.65; Wortman, 2010: 0.55). However, it is difficult to compare these studies to the current one. The labels for all these studies are based on the IMDb labels, but the number of labels in most of these studies is either less (Hong & Hwang, 2015), more (Blackstock & Spitz, 2008), or unknown (Fourati et al., 2014 talk about 'all cinematic genres'). Only Wortman (2010) has a similar set of labels. However, there is not enough information to determine whether the classifier described in this study performs better because of the dataset, the framework, or both.

Another finding of the current work is that the different feature categories in the current study perform well separately, but the features could also complement each other when they are combined. Classifiers using only one of the three feature categories did already produce accurate predictions. Of these categories, the content-specific features resulted in the best performance, followed by the syntactical features. The textual features classifier performed the worst of the three separate classifiers, although it still produced results well above baseline. This supports the notion that the features adopted from Abbasi & Chen (2008) and Huang et al. (2010) are also useful (in modified form) for genre classification.

Depending on the combination approach that is used for the feature categories, the classification performance could increase, but also decrease, compared to the single feature type classifiers. Combining the different feature types into one single vector (supervector-approach) was not a reliable method to achieve accuracy increases. While a textual-syntactical and textual-content-specific combination performed better than their separate counterparts, the syntactical-content-specific combination and the combination of all three feature categories resulted in worse accuracy compared to a classifier using merely content-specific features. Meta-classifier approaches were more reliable. The add-rule and product-rule meta-classifiers achieved a higher accuracy than the classifier using only content-specific features did. However, the predictions of the rule-based classifiers were not as good as those of the machine learning-based meta-classifier. The best results were obtained with the latter approach. This result was expected, since previous meta-classification approaches in genre classification using machine learning algorithms also substantially increased classification accuracy (Lin & Hauptmann, 2002; Wang et al., 2003; Kennedy & Inkpen, 2006; Giannakopoulos et al., 2010). However, this approach had not been investigated using one video modality only, and had also not been compared to rule-based approaches in previous video classification studies.

Furthermore, the feature selection process in conjunction with some of the tested algorithms would likely have had a positive impact on classification accuracy. The SVM was found to be one of the best performing algorithms in all classification tasks, being the most accurate for the content-specific classifier, textual-content-specific classifier and machine learning-based meta-classifier. This is not unexpected, as the SVM algorithm has often been found to be a good algorithm for text classification tasks (e.g. Jing, Li, Zhang, & Zhang, 2004; Huang et al., 2010; Lazebnik, Schmid, & Ponce, 2006; Zhang, Marszałek, Lazebnik, & Schmid, 2007). However, the Random Forest algorithm was the most accurate algorithm for all the other classifiers. The strength of the algorithm for the textual and syntactical classifier was not unexpected, since it is a classifier known to perform well on larger datasets. However, previous scholars have noted that the performance of the Random Forest algorithm drops for tasks with a large amount of features, such as bag-of-words classification. The strength of the Random Forest algorithm in these situations is likely due to the chi-square based feature selection.

5.1 Future work

While the current study offers a framework that was found to be effective for video genre classification, it also offers various paths for further exploration. The accuracy results per genre showed substantial differences between the classification accuracy for various genres. Some genres (e.g. romance, family) lagged somewhat behind in terms of accuracy, partly due to their mutual confusability. Therefore, further improvements could be made to the preciseness of the predictions, for which several approaches are possible. One approach would be to add more features to the classifier. The focus was on word-level features in the current work, though there are also several character-level features that could be added to future classifiers (character n-grams, character ratios, etc.) and possibly help increase accuracy. Another approach would be to not only use text-based features, but to also add features from other modalities (i.e. visual, auditory). In the SUBTIEL corpus, details are available of the movie or television show a subtitle belongs to. It is therefore possible to connect these subtitles to trailers and use visual and auditory features from these trailers in conjunction with the text features of the subtitles.

Besides accuracy-improving changes, it would also be useful to explore the generalizability of the current results. The current research has focused on Dutch-language subtitles for movies and television shows. The research could therefore be extended to subtitles in other languages to see whether the current framework is an effective approach independent of language. This would also provide insight into whether the genre-revealing features found in the current study can be generalized across multiple languages. While an attempt was made to look into genre-revealing features in the current work, it would take follow-up work to be able to tell the importance of these features. A more accurate idea of which features are important for genre classification would provide us with more insight into the notion of genre and provide a more accurate way of objectively determining the genre of a movie or TV show.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, *26*(2), 1-29.

Abbasi, A., Chen, H., & Nunamaker, J. F. (2008). Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, *25*(1), 49-78.

Anwar, A., Salama, G. I., & Abdelhalim, M. B. (2013). Video classification and retrieval using Arabic closed caption. In *ICIT 2013 The 6th International Conference on Information Technology VIDEO*.

Asheghi, N. R., Sharoff, S., & Markert, K. (2014). Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 1339-1346).

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*(2), 384-414.

Biber, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities, 26*(5-6), 331-345.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.* Princeton, NJ: Educational Testing Service.

Blackstock, A., & Spitz, M. (2008). *Classifying Movie Scripts by Genre with a MEMM using NLP-Based Features*. Student Report Stanford University.

Bou-Rabee, A., Go, K., & Mohan, K. (2012). *Classifying the Subjective: Determining Genre of Music From Lyrics*. Student Report Stanford University.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Brezeale, D., & Cook, D. J. (2006, August). Using closed captions and visual features to classify movies by genre. In *Poster session of the Seventh International Workshop on Multimedia Data Mining (MDM/KDD2006)*.

Brezeale, D., & Cook, D. J. (2008). Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(3), 416-430.

Chen, H. Shankaranarayanan, G. She, L., Iyer, A. (1998). A machine learning approach to inductive query by examples: An experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing. *Journal of the American Society for Information Science and Technology*, *49*(8), 693-705.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375-1388.

Dermitaş, K. (2009). *Automatic Video Categorization and Summarization*. Master's Thesis Middle East Technical University.

Dewdney, N., VanEss-Dykema, C., & MacMillan, R. (2001). The form is the substance: Classification of genres in text. In *Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001* (pp. 7-16).

Dietterich, T. G., Hild, H., & Bakiri, G. (1990). A comparative study of ID3 and backpropagation for English text-to-speech mapping. *Machine Learning*, *50*, 55-60.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research 9*, 1871-1874.

Fell, M., & Sporleder, C. (2014). Lyrics-based Analysis and Classification of Music. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 620-631).

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, *57*(11), 1506-1518.

Fourati, M., Jedidi, A., & Gargouri, F. (2014). Automatic audiovisual documents genre description. In *6th International Joint Conference on Knowledge Discovery and Information Retrieval (KDIR 2014)* (pp. 21-24).

Freund, Y., & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In Proceedings of the Second *European Conference on Computational Learning Theory* (pp. 23-37).

Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., & Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos. In *Hellenic Conference on Artificial Intelligence* (pp. 91-100).

Giesbrecht, E., & Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In *Proceedings of the Fifth Web as Corpus Workshop* (pp. 27-35).

Gliozzo, A., & Strapparava, C. (2006). Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 553-560).

Hauptmann, A. G., Jin, R., & Ng, T. D. (2002). Multi-modal information retrieval from broadcast video using OCR and speech recognition. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries* (pp. 160-161).

Helmer, E., & Ji, Q. (2012). *Film classification by trailer features.* Student Report Stanford University.

Hong, H. Z., & Hwang, J. I. G. (2015). Multimodal PLSA for movie genre classification. In *International Workshop on Multiple Classifier Systems* (pp. 159-167).

Huang, C., Fu, T., & Chen, H. (2010). Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, *61*(5), 891-906.

Huang, J., Liu, Z., Wang, Y., Chen, Y., & Wong, E. K. (1999). Integration of multimodal features for video scene classification based on HMM. In *IEEE 3rd Workshop on Multimedia Signal Processing,* (pp. 53-58).

Jasinschi, R. S., & Louie, J. (2001). Automatic tv program genre classification based on audio patterns. In *Proceedings of the 27th Euromicro Conference,* (pp. 370-375).

Jing, F., Li, M., Zhang, H. J., & Zhang, B. (2004). An efficient and effective region-based image retrieval framework. *IEEE Transactions on Image Processing*, *13*(5), 699-709.

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics* (Vol. 2, pp. 1071-1075).

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., & Wolkert, N. (1998). Iterative information retrieval using fast clustering and usage-specific genres. In *Eight DELOS Workshop on User Interfaces in Digital Libraries* (pp. 85-92).

Katsiouli, P., Tsetsos, V., & Hadjiefthymiades, S. (2007). Semantic video classification based on subtitles and domain terminologies. In *Knowledge Acquisition from Multimedia Content, 57*.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110-125.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 32-38).

Kobla, V., DeMenthon, D., & Doermann, D. S. (2000). Identifying sports videos using replay, text, and camera motion features. In *SPIE conference on Storage and Retrieval for Media Databases* (pp. 332-343).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 2169-2178).

Lee, Y. B., & Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 145-150).

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *2*(1), 1-19.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European Conference on Machine Learning* (pp. 4-15).

Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*, *41*(5), 1263-1276.

Lin, W. H., & Hauptmann, A. (2002). News video classification using SVM-based multimodal classifiers and combination strategies. In *Proceedings of the Tenth ACM International Conference on Multimedia* (pp. 323-326).

Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, *20*(1-2), 61-79.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization* (Vol. 752, pp. 41-48).

Moncrieff, S., Venkatesh, S., & Dorai, C. (2003). Horror film genre typing and scene labeling via audio analysis. In *Proceedings of the 2003 International Conference on Multimedia and Expo* (Vol. 1, pp. 193-196).

Mu, Y. (2015). Using keyword features to automatically classify genre of Song Ci poem. In *Workshop on Chinese Lexical Semantics* (pp. 478-485).

Nassar, H., Taha, A., Nazmy, T., & Nagaty, K. (2007). Classification of video scenes using Arabic Closed-Caption. In *Proceedings of the Third International Conference on Intelligent Computing and Information Systems* (pp. 186-192).

Oger, S., Rouvier, M., & Linares, G. (2010). Transcription-based video genre classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5114-5117).

Païs, G., Lambert, P., Beauchêne, D., Deloule, F., & Ionescu, B. (2012). Animated movie genre detection using symbolic fusion of text and image descriptors. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1-6).

Petrenz, P. (2012). Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 11-21).

Petrenz, P., & Webber, B. (2012). Robust cross-lingual genre classification through comparable corpora. In *The 5th Workshop on Building and Using Comparable Corpora* (pp. 1-9).

Qi, W., Gu, L., Jiang, H., Chen, X. R., & Zhang, H. J. (2000). Integrating visual, audio and text analysis for news video. In *Proceedings of the International Conference on Image Processing* (Vol. 3, pp. 520-523).

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning* (Vol. 1). Burlington, MA: Morgan Kaufmann.

Rasheed, Z., & Shah, M. (2002). Movie genre classification by exploiting audio-visual features of previews. In *Proceedings of the 16th International Conference on Pattern Recognition* (Vol. 2, pp. 1086-1089).

Rasheed, Z., Sheikh, Y., & Shah, M. (2003). Semantic film preview classification using low-level computable features. In *3rd International Workshop on Multimedia Data and Document Engineering (MDDE-2003)*.

Rijsbergen, C. J. van (1979). *Information Retrieval.* London, UK: Butterworths.

Rippey, S., & Zimmerman, E. (n.d.). *Genre Classification with Video Game Transcripts and Descriptions*. Student Report Stanford University.

Roach, M. J., Mason, J. D., & Pawlewski, M. (2001). Video genre classification using dynamics. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001 (ICASSP'01)* (Vol. 3, pp. 1557-1560).

Roach, M., Mason, J., & Xu, L. Q. (2002). Video genre verification using both acoustic and visual modes. In *IEEE Workshop on Multimedia Signal Processing* (pp. 157-160).

Sadovsky, A., & Chen, X. (2006). *Song genre and artist classification via supervised learning from lyrics*. Student Report Stanford University.

Sahami, M. (1996). Learning Limited Dependence Bayesian Classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (pp. 335-338).

Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop* (pp. 83-94).

Sharoff, S., Wu, Z., & Markert, K. (2010). The Web Library of Babel: evaluating genre collections. In *Proceedings of the Seventh Conference on International Language Resources*.

Silla Jr, C. N., Kaestner, C. A., & Koerich, A. L. (2007). Automatic music genre classification using ensemble of classifiers. In *2007 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1687-1692).

Samothrakis, S., & Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS One*, *10*(11).

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471-495.

Truong, B. T., Dorai, C., & Venkatesh, S. (2000). Automatic genre identification for content-based video categorization. In *Proceedings of the 15th International Conference on Pattern Recognition* (Vol. 4, pp. 230-233).

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Vasconcelos, N., & Lippman, A. (2000). Statistical models of video structure for content analysis and characterization. *IEEE Transactions on Image Processing*, *9*(1), 3-19.

Vel, O. de (2000). Mining e-mail authorship. In *Proceedings Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000)*.

Vidulin, V., Luštrek, M., & Gams, M. (2007). Using genres to improve search engines. In *Proceedings of the International Workshop Towards Genre-Enabled Search Engines* (pp. 45-51).

Walsten, D. & Orth, D. (n.d.). *Song Genre Classification through Quantitative Analysis of Lyrics.* Unpublished manuscript.

Wan, Y., Yang, X., Mo, W., Liu, Z., Zhang, J., & Li, Z. (2015). Study on genre and its role in discourse information processing. *Open Cybernetics & Systemics Journal*, *9*, 1478-1484.

Wang, P., Cai, R., & Yang, S. Q. (2003). A hybrid approach to news video classification multimodal features. In *Proceedings of the Fourth International Conference on Information, Communications and Signal Processing* (Vol. 2, pp. 787-791).

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Vol. 2, pp. 674-682).

Wortman, J. (2010). *Film classification using subtitles and automatically generated language factors*. Master's Thesis Israel Institute of Technology.

Zhang, H. (2004). The optimality of Naive Bayes. In *Proceedings of the 17th International FLAIRS Conference*.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, *73*(2), 213-238.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378-393.

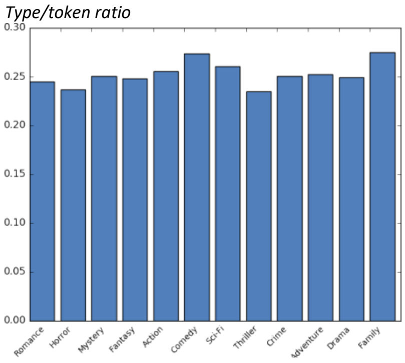
Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class adaboost. *Statistics and Its Interface*, *2*(3), 349-360.

Zhu, W., Toklu, C., & Liou, S. P. (2001, August). Automatic news video segmentation and categorization based on closed-captioned text. In *IEEE International Conference on Multimedia and Expo (ICME 2001)* (pp. 829-832).

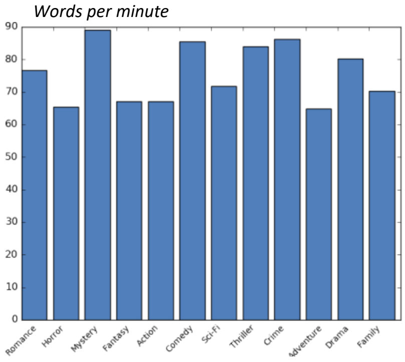
Appendix A: Graphs showing the genre differences for the 10 most important features per feature category

Textual

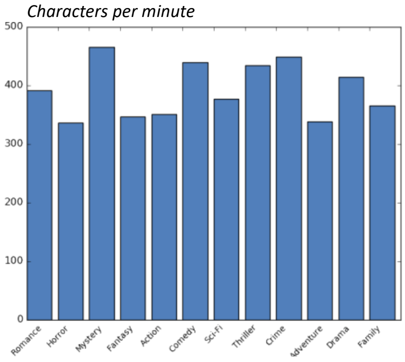
Type/token ratio



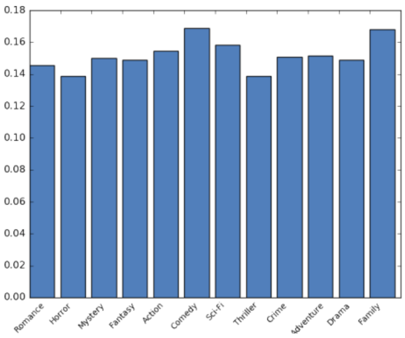
Words per minute



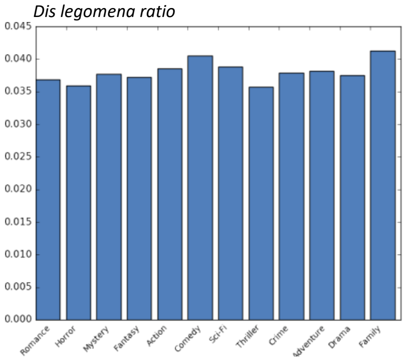
Characters per minute



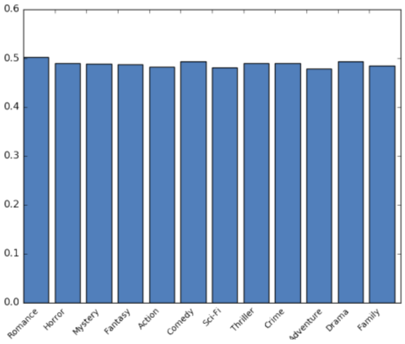
Hapax legomena ratio



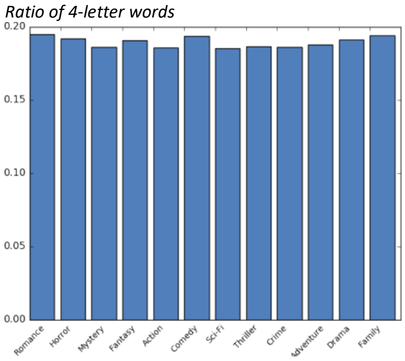
Dis legomena ratio



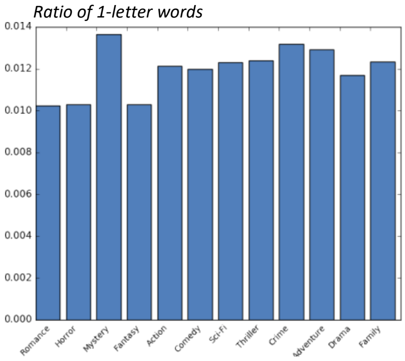
Short words ratio



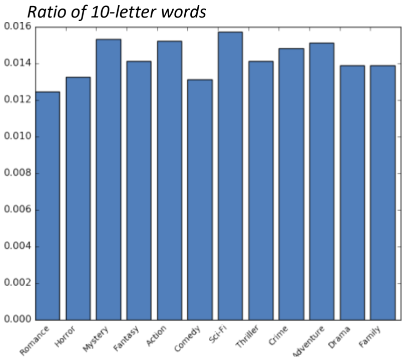
Ratio of 4-letter words



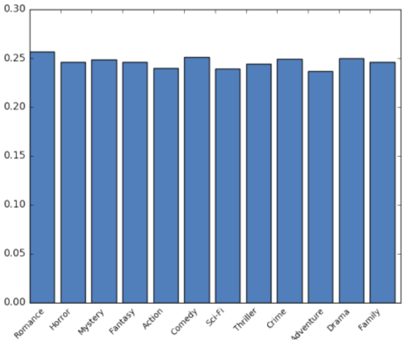
Ratio of 1-letter words



Ratio of 10-letter words

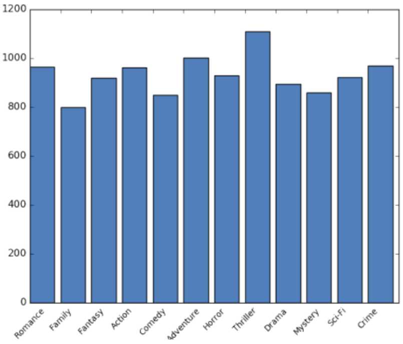


Ratio of 3-letter words

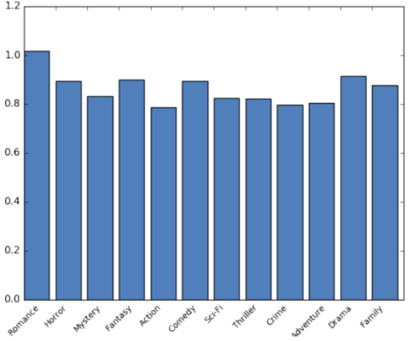


Syntactical

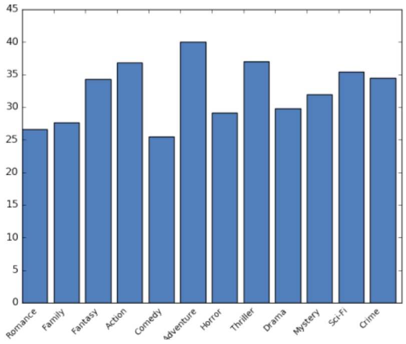
.



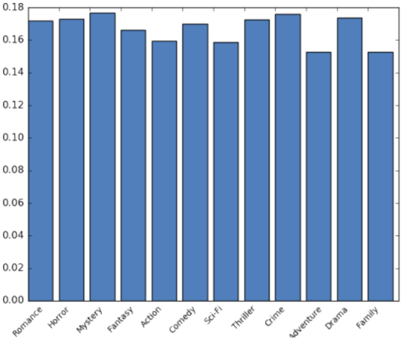
Descriptive/nominal ratio



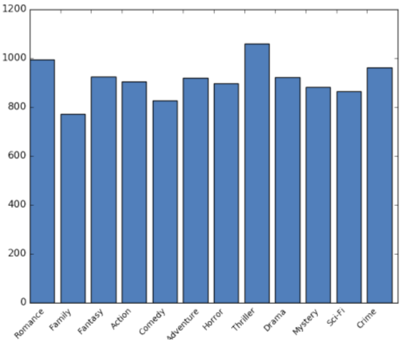
. DT



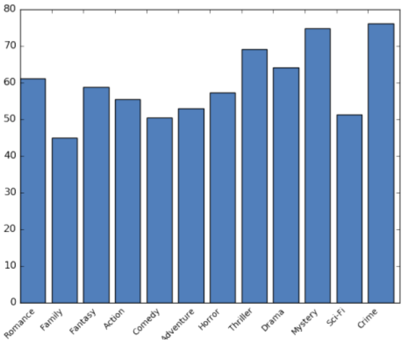
Question mark ratio



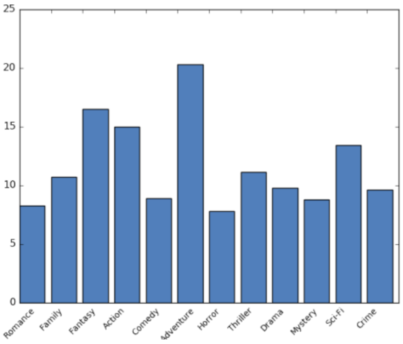
PRP



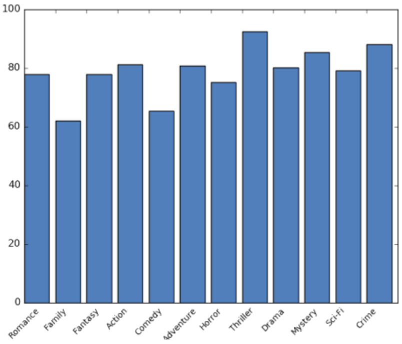
VBD



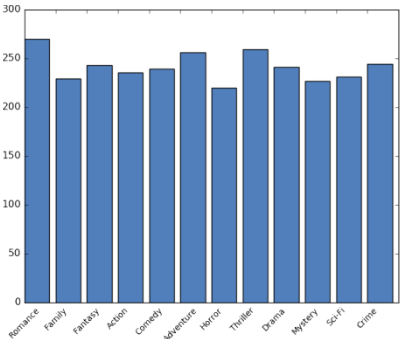
DT NNP



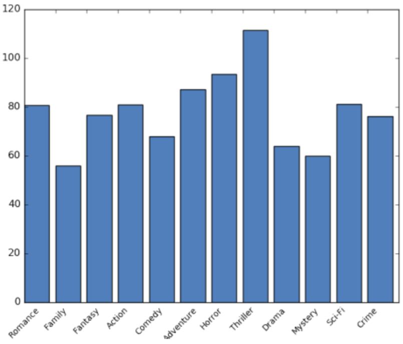
VBN .



JJ

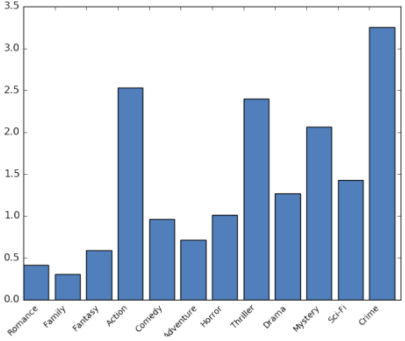


. .

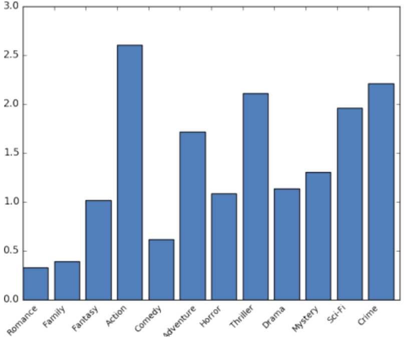


Content-specific

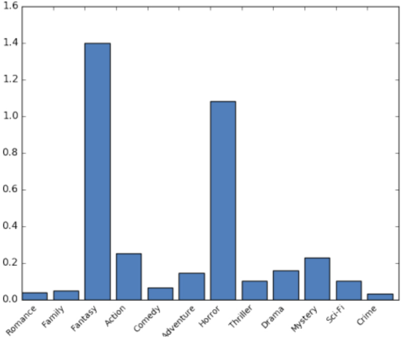
Agent



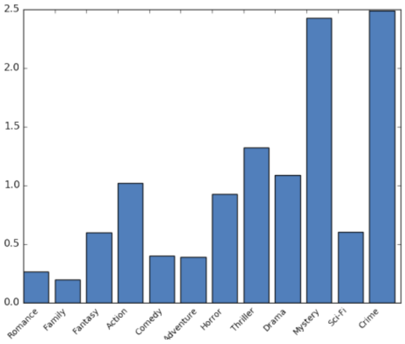
Wapen



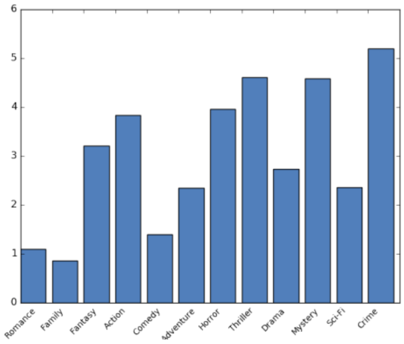
Vampier



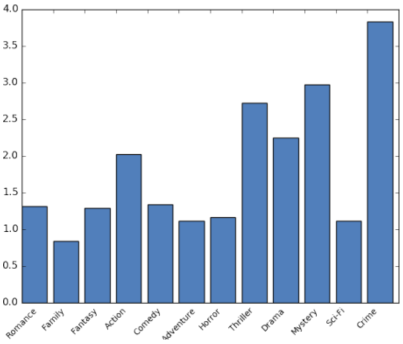
Moord



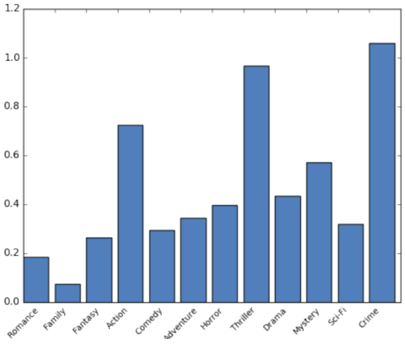
Vermoorden



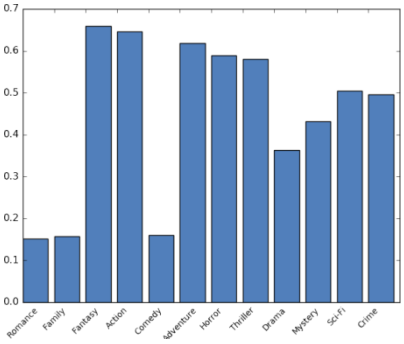
Zaak



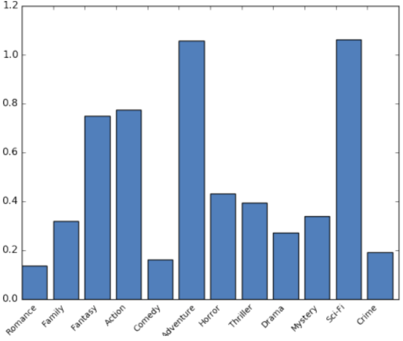
Pistool



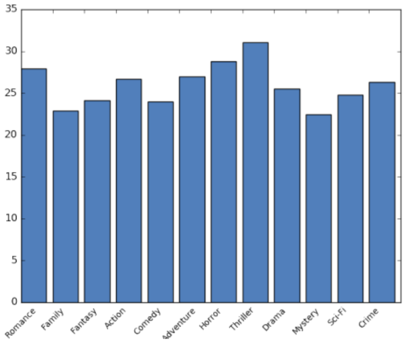
Gedood



Vernietigen



Gaan


Romance
Family
Fantasy
Action
Comedy
Adventure
Horror
Thriller
Drama
Mystery
Sci-Fi
Crime