UNIVERSITY
OF WOLLONGONG
AUSTRALIA

ISIT219 – Assignment 2

Categorising Videos Based on Their Description

Lab Group	:	Lab1_7	
Last Amended	:	29 May 2020	
Group Members	:	Minh Huy Tran	6330617
		Trang Nguyen	6166994
		Josh Duncan	5206315
		Le Anne Ng	5943243

Table of Contents

1.0 Introduction

With YouTube being one of the most popular online video sharing platforms, there comes a lot of knowledge that it can accumulate. With information such as a user's personal preference on a specific genre can be a huge influence when it comes to advertising. Hence, causing the existence of many studies on analysing and categorising YouTube's metadata. Aside from streamlining advertising schemes, there are still many other studies that focus on various types of topics that are done to not only prevent false advertising but also to prevent conflicts in areas such as those regarding politics and mental health.

Nonetheless, amongst the many types of analysis done, one of the prominent ones that our study will focus on is using knowledge to categorise videos mainly to validate the legitimacy of video content based on their genre. From the given specification of this assignment, we were provided with a list of categorises and respective "category_id" that YouTube uses to categorise videos. In addition to an owner of a channel selecting the genre of their video content, using a given knowledge to validate against that would help reduce the sharing of content that might cause an adverse impact to both YouTube and the community.

There are many types of information that can be grouped into different categories as spoken in [1] therefore, for our study, we look at categorising videos based on their description using Association Rules during our data extraction and Naïve Bayes during our analysis to see if it is possible to find a relationship between YouTube's "category_id" and a video's description.

2.0 Acquiring Knowledge

The knowledge acquisition stage of our study started by exploring possible methods with which to approach the problem of comment-text based classification of videos on the YouTube website. At first, we sought to find out if any academics had attempted to solve a problem from this or a similar domain in the past. Madden, Ruthven and McMenemy [19] were able to create a classification schema for YouTube videos by analysing 66,637 comments, and sorting said comments into 55 categories. This study employed qualitative content analysis, and deviant case analysis. The categorisation technique employed in [19] involved grouping comments into categories based on similarities, such as whether or not they referenced video timestamps, or another comment made earlier. This led us to investigate the possibility of applying Association Rules to the same problem as a knowledge-creation technique.

For our project, which was on the classification of videos by description, the grouping was to be based on the genre with given "category_id" of the video, as opposed to the kind of category the comments themselves fell into, as in the case of [19]. Because of this, we began to explore ways to discern the likelihood that any comment was from a video of any category. Because Naive Bayes as mentioned in [20] is "One of the most efficient inductive learning algorithms, NB classifier is often used as a baseline in text classification because it is fast and easy to implement." we decided to consider using Naive Bayes as one of our knowledge-creation techniques.

Naive Bayes makes judgements statistically based on the data that has been given to it as input. As such, it is important that the data that is given is both representative of the desired result and also meaningful. For Data Preparation, the Apriori algorithm is known to be the most powerful Association

Rule miner [21]. Additionally, we found it to be well-suited to the data available; our frequent datasets were already identified as YouTube video categories. The predefined categories meant that if we were to use Apriori, all that was left to do was to remove outliers in the sample data. Data Preparation can often take from 50-70% of total project time and is crucial for achieving desired results [22]. As such, this was an important factor in the decision to use Apriori and Naive Bayes.

We researched academic journals and studies that used Association Rules and Naive Bayes for similar purposes, specifically for classification of text, or better yet the classification of an item based on associated text. Padmanabhan and Kummamuru in their 2007 paper propose a procedure using Apriori along with K-means clustering to extract from call-centres information that represents procedures from other kinds of conversation or information being conveyed. Agnihotri, Verma, and Tripathi in their 2017 paper [23] discuss methods for automatic classification of text using both Apriori and Bayes. They propose improving the Data-Mining method by considering the presence of multiple correlated key words in succession, for example "bacterial infection" as being of a higher value than "bacterial" and "infection" would be individually or spaced out over multiple paragraphs. This was determined to be beyond our abilities but was considered.

3.0 Knowledge Creation

3.1 JUSTIFICATION OF CHOSEN TECHNIQUES

3.1.1 Association Rules

To identify frequent terms, we apply a text mining technique capable of extracting rules and relationships in the description section of each video category. The Association Rule is our selection for several reasons. The technique has been applied successfully in many areas to find frequent patterns [11-15]. The method can avoid the effects of noises [14]. Moreover, the outcomes are simple, comprehensible, and easy to interpret [16]. However, the number of extracted rules sometimes can be excessive [16]. Therefore, the minimum support and confidence values will be introduced to limit the number of outputs. Although each category needs to be provided with a pair of the aforementioned values, the number of required runs is acceptable as many categories are absent or have insufficient videos.

3.1.2 Naïve Bayes classifiers

We choose the Naïve Bayes classifier as a method to classify videos based on descriptions because it is one of the most popular text classification methods [2]. The technique has shown excellent performances in many other textual data classification problems [3-7]. The method's implementation is easier compared to other classification techniques [8]. Furthermore, the Naïve Bayes classifier has a quick processing speed [9], and its effectiveness is also surprisingly high [10]. However, the method's accuracy is heavily affected by the sample size [10] and the conditional dependency of the dataset. Despite its disadvantages, the Naïve Bayes classifier is still a suitable classification method to prove our research's goal.

3.2 RESULTS AND ANALYSIS

3.2.1 Data Pre-processing

Dimension Reduction

To reduce the dimension of the dataset, only columns of used were retrieved. These useful attributes include "category_id" as the target and "description" as the predictor in the classification task. Duplicate videos were removed, resulting in a dataset of 1571 rows (videos).

```
# A tibble: 16 x 2
   category_id total_videos
         <dbl>        <int>
 1           1           79
 2           2            5
 3          10          207
 4          15           22
 5          17          124
 6          19            9
 7          20           46
 8          22          123
 9          23          149
10          24          459
11          25           65
12          26          148
13          27           62
14          28           70
15          29            1
16          43            2
```

Figure 3.1: Video counts per category in the dataset

Figure 3.1 shows the video counts in each category in the dataset. Several categories absent from the dataset (compared to the total of 32 YouTube categories) will not be classified. Also, there was an apparent skewness in the dataset where some categories had notably fewer videos than others. This might affect the accuracy of the model later in the classification task.

Text Cleaning

Video classification based on their description involves the issue of text classification. To perform classification, the text must be pre-processed regardless of the model used. Before tokenising the text description of each video, several steps have been taken and some of them include: -

- Converting to lowercase: Capitalisation is not of importance in the sematic of a word.
- Removal of numbers and punctuations.
- Removal of non-English words and "stop words": These refer to words that are important in terms of the sentence construction but not for text classification such as "is" and "the".
- Lemmatisation: Converting words to their base meaning.
- Removal of extra white spaces.
- Words regarding other social media names such as "Facebook", "follow", "Instagram": These words do not contribute to the classification task.

Feature Engineering

The text is then represented as a numeric matrix called a Document-Term Matrix (DTM), in which the frequency of the terms in the set of documents is described. For each document, a "1" represents the existence of a term in the document while a "0" represents its absence. A fraction of the DTM for the dataset is as follows:

```
<<DocumentTermMatrix (documents: 1571, terms: 15501)>>
Non-/sparse entries: 76621/24275450
Sparsity           : 100%
Maximal term length: 37
Weighting          : term frequency (tf)
Sample             :
      Terms
Docs   episod jimmi kimmel late live love music product star watch
  1064      0     0      0    0    1    1     1       0    0     2
  1111      0     0      0    0    0    0     0       0    0     0
  1140      1     0      0    0    0    1     0       0    0     0
  1222      2     0      0    0    1    0     2       2    1     0
  1355      2     0      0    0    1    0     0       2    2     0
  191       2     0      0    0    1    0     0       2    1     1
  247       2     0      0    0    1    0     2       2    2     1
  46        2     0      0    0    1    0     0       2    2     0
  481       0     0      0    0    0    0     0       1    2     3
  98        0     0      0    0    0    1     0       1    4     3
```

Figure 3.2: Document-Term matrix for the first 10 documents and terms

3.2.2 Association Rules for Text Mining

In this step, Apriori is a common algorithm in text analysing to extract the frequent patterns and create association rules from the dataset. Videos are then analysed separately to generate a list of rules appearing the most in each category. There are two important values that dramatically affect the outcomes of any association rule mining process, namely the support and confidence value. The support value indicates the ratio of the occurrence of a rule to the number of total videos, while the confidence value shows the ratio of the occurrence of rule head to the occurrence of rule body.

If {X => Y} is an extracted rule, then X is the rule body and Y is the rule head. (n is the total number of transactions)

$$support = \frac{(X \cup Y).count}{n.count}$$
$$confidence = \frac{(X \cup Y).count}{X.count}$$

Category ID	Minimum Support	Minimum Confidence
1	0.13	0.80
10	0.10	0.80
15	0.30	0.98
17	0.40	0.87
19	0.23	0.83
20	0.13	0.80
22	0.25	0.83
23	0.40	0.87
24	0.10	0.80
25	0.19	0.83
26	0.19	0.83
27	0.25	0.83
28	0.14	0.90

Table 3.1: Minimum Support and Confidence used to extract rules for each category.

Apriori algorithm requires users to specify the minimum the support and confidence values to limit the number of outcomes and only show rules that are truly important. The minimum support and confidence for each category are altered to optimise the outcomes.

Results and Explanation

The outcomes of this process are shown in the table below. Each row shows a relationship between terms used in a category and other related attributes, namely support, confidence, lift and count. **Table 3.1** is in the decreasing order of lift values. However, 19 out of 32 categories which have at most five videos will not be shown since they do not consist of any video or any outstanding pattern after analysing.

Category 1 – Film & Animation

	lhs		rhs	support	confidence	lift	count
[1]	{entertain}	=>	{trailer}	0.1392405	1.0000000	2.548387	11
[2]	{product}	=>	{trailer}	0.1518987	0.9230769	2.352357	12
[3]	{offici}	=>	{trailer}	0.1898734	0.8823529	2.248577	15
[4]	{produc}	=>	{trailer}	0.1645570	0.8125000	2.070565	13

Category 10 – Music

	lhs		rhs	support	confidence	lift	count
[1]	{itun}	=>	{music}	0.1256039	1.0000000	2.009709	26
[2]	{play,spotifi}	=>	{music}	0.1062802	1.0000000	2.009709	22
[3]	{play}	=>	{music}	0.1835749	0.9268293	1.862657	38
[4]	{exclus}	=>	{music}	0.1159420	0.8571429	1.722607	24
[5]	{spotifi}	=>	{music}	0.1449275	0.8571429	1.722607	30

Category 15 – Pets & Animals

	lhs		rhs	support	confidence	lift	count
[1]	{anim,websit}	=>	{connect}	0.3181818	1	3.142857	7
[2]	{anim,visit}	=>	{connect}	0.3181818	1	3.142857	7
[3]	{anim,visit,websit}	=>	{connect}	0.3181818	1	3.142857	7
[4]	{connect}	=>	{websit}	0.3181818	1	2.750000	7
[5]	{connect,visit}	=>	{websit}	0.3181818	1	2.750000	7
[6]	{anim,connect}	=>	{websit}	0.3181818	1	2.750000	7
[7]	{anim,visit}	=>	{websit}	0.3181818	1	2.750000	7
[8]	{anim,connect,visit}	=>	{websit}	0.3181818	1	2.750000	7
[9]	{connect}	=>	{visit}	0.3181818	1	2.444444	7
[10]	{websit}	=>	{visit}	0.3636364	1	2.444444	8
[11]	{connect,websit}	=>	{visit}	0.3181818	1	2.444444	7
[12]	{anim,connect}	=>	{visit}	0.3181818	1	2.444444	7
[13]	{anim,websit}	=>	{visit}	0.3181818	1	2.444444	7
[14]	{anim,connect,websit}	=>	{visit}	0.3181818	1	2.444444	7
[15]	{discov}	=>	{anim}	0.3181818	1	1.833333	7
[16]	{connect}	=>	{anim}	0.3181818	1	1.833333	7
[17]	{connect,websit}	=>	{anim}	0.3181818	1	1.833333	7
[18]	{connect,visit}	=>	{anim}	0.3181818	1	1.833333	7
[19]	{connect,visit,websit}	=>	{anim}	0.3181818	1	1.833333	7

Category 17 – Sports

	lhs		rhs	support	confidence	lift	count
[1]	{nba}	=>	{highlight}	0.4354839	0.9000000	1.488000	54
[2]	{nba}	=>	{news}	0.4354839	0.9000000	1.665672	54
[3]	{media}	=>	{watch}	0.4032258	0.9803922	1.814457	50
[4]	{news}	=>	{highlight}	0.5000000	0.9253731	1.529950	62
[5]	{sport}	=>	{watch}	0.4516129	0.9032258	1.671642	56
[6]	{highlight,nba}	=>	{news}	0.4274194	0.9814815	1.816473	53
[7]	{nba,news}	=>	{highlight}	0.4274194	0.9814815	1.622716	53

Category 19 – Travel & Events

	lhs		rhs	support	confidence	lift	count
[1]	{visit}	=>	{life}	0.3333333	1	2.25	3
[2]	{cook}	=>	{food}	0.3333333	1	1.50	3
[3]	{travel}	=>	{food}	0.3333333	1	1.50	3
[4]	{len}	=>	{food}	0.3333333	1	1.50	3

Category 20 - Gaming

	lhs		rhs	support	confidence	lift	count
[1]	{switch}	=>	{nintendo}	0.1304348	1.0000000	7.666667	6
[2]	{nintendo}	=>	{switch}	0.1304348	1.0000000	7.666667	6
[3]	{awesom}	=>	{game}	0.1304348	1.0000000	2.705882	6
[4]	{content}	=>	{game}	0.1304348	0.8571429	2.319328	6
[5]	{epic}	=>	{game}	0.1304348	0.8571429	2.319328	6

Category 22 – People & Blogs

	lhs		rhs	support	confidence	lift	count
[1]	{awesom}	=>	{buzzfe}	0.2520325	0.9687500	3.404464	31
[2]	{buzzfe}	=>	{awesom}	0.2520325	0.8857143	3.404464	31
[3]	{audio}	=>	{audioblock}	0.2926829	0.9729730	3.068607	36
[4]	{audioblock}	=>	{audio}	0.2926829	0.9230769	3.068607	36

Category 23 - Comedy

	lhs		rhs	support	confidence	lift	count
[1]	{highlight,nba}	=>	{news}	0.4274194	0.9814815	1.816473	53
[2]	{media}	=>	{watch}	0.4032258	0.9803922	1.814457	50
[3]	{sport}	=>	{watch}	0.4516129	0.9032258	1.671642	56
[4]	{nba}	=>	{news}	0.4354839	0.9000000	1.665672	54
[5]	{nba,news}	=>	{highlight}	0.4274194	0.9814815	1.622716	53
[6]	{news}	=>	{highlight}	0.5000000	0.9253731	1.529950	62
[7]	{nba}	=>	{highlight}	0.4354839	0.9000000	1.488000	54

Category 24 – Entertainment

	lhs		rhs	support	confidence	lift	count
[1]	{late,live}	=>	{night}	0.1023965	0.9791667	6.242188	47
[2]	{late}	=>	{night}	0.1045752	0.9411765	6.000000	48
[3]	{execut}	=>	{produc}	0.1067538	0.9423077	4.914991	49
[4]	{david}	=>	{produc}	0.1045752	0.9230769	4.814685	48
[5]	{late,night}	=>	{live}	0.1023965	0.9791667	4.539773	47
[6]	{late}	=>	{live}	0.1045752	0.9411765	4.363636	48
[7]	{night}	=>	{live}	0.1307190	0.8333333	3.863636	60
[8]	{pm}	=>	{episod}	0.1002179	0.8214286	2.992347	46
[9]	{pm}	=>	{watch}	0.1002179	0.8214286	2.618304	46

Category 25 – News & Politics

	lhs		rhs	support	confidence	lift	count
[1]	{websit}	=>	{headlin}	0.2153846	1	4.642857	14
[2]	{headlin}	=>	{websit}	0.2153846	1	4.642857	14
[3]	{check}	=>	{news}	0.2000000	1	2.031250	13

Category 26 – Howto & Style

	lhs		rhs	support	confidence	lift	count
[1]	{affili,sponsor}	=>	{link}	0.2229730	1.0000000	2.846154	33
[2]	{link,sponsor}	=>	{affili}	0.2229730	0.8684211	2.794050	33
[3]	{affili}	=>	{link}	0.2972973	0.9565217	2.722408	44
[4]	{link}	=>	{affili}	0.2972973	0.8461538	2.722408	44
[5]	{sponsor}	=>	{link}	0.2567568	0.9047619	2.575092	38

Category 27 – Education

	lhs		rhs	support	confidence	lift	count
[1]	{alexand}	=>	{support}	0.2580645	1.0000000	2.296296	16
[2]	{patron}	=>	{support}	0.3387097	1.0000000	2.296296	21
[3]	{patreon,patron}	=>	{support}	0.2741935	1.0000000	2.296296	17
[4]	{kevin}	=>	{support}	0.2580645	0.9411765	2.161220	16

Category 28 – Science & Technology

	lhs		rhs	support	confidence	lift	count
[1]	{affili,click}	=>	{mail}	0.1428571	1.0000000	7.000000	10
[2]	{click,link}	=>	{mail}	0.1428571	1.0000000	7.000000	10
[3]	{affili,click,link}	=>	{mail}	0.1428571	1.0000000	7.000000	10
[4]	{mail}	=>	{click}	0.1428571	1.0000000	6.363636	10
[5]	{click}	=>	{mail}	0.1428571	0.9090909	6.363636	10
[6]	{affili,mail}	=>	{click}	0.1428571	1.0000000	6.363636	10
[7]	{link,mail}	=>	{click}	0.1428571	1.0000000	6.363636	10
[8]	{affili,link}	=>	{mail}	0.1428571	0.9090909	6.363636	10
[9]	{affili,link,mail}	=>	{click}	0.1428571	1.0000000	6.363636	10
[10]	{mail}	=>	{affili}	0.1428571	1.0000000	5.833333	10
[11]	{click,mail}	=>	{affili}	0.1428571	1.0000000	5.833333	10
[12]	{link,mail}	=>	{affili}	0.1428571	1.0000000	5.833333	10
[13]	{click,link}	=>	{affili}	0.1428571	1.0000000	5.833333	10
[14]	{link,product}	=>	{affili}	0.1428571	1.0000000	5.833333	10
[15]	{click,link,mail}	=>	{affili}	0.1428571	1.0000000	5.833333	10
[16]	{affili,link}	=>	{click}	0.1428571	0.9090909	5.785124	10
[17]	{click}	=>	{affili}	0.1428571	0.9090909	5.303030	10
[18]	{mail}	=>	{link}	0.1428571	1.0000000	4.666667	10
[19]	{descript}	=>	{link}	0.1571429	1.0000000	4.666667	11
[20]	{click,mail}	=>	{link}	0.1428571	1.0000000	4.666667	10
[21]	{affili,mail}	=>	{link}	0.1428571	1.0000000	4.666667	10
[22]	{affili,click}	=>	{link}	0.1428571	1.0000000	4.666667	10
[23]	{affili,product}	=>	{link}	0.1428571	1.0000000	4.666667	10
[24]	{affili,click,mail}	=>	{link}	0.1428571	1.0000000	4.666667	10
[25]	{affili}	=>	{link}	0.1571429	0.9166667	4.277778	11
[26]	{click}	=>	{link}	0.1428571	0.9090909	4.242424	10
[27]	{click}	=>	{product}	0.1428571	0.9090909	3.977273	10
[28]	{affili,link}	=>	{product}	0.1428571	0.9090909	3.977273	10

Each category appears to have a different set of extracted rules which are frequently present in the video descriptions. The set of rules shows that the relationships between videos within each category have different strengths which are shown by the fluctuation in the lift value. We believe a further analysis on the topic should be conducted to prove the predetermined goal, which leads to the next section.

3.2.3 Naïve Bayes Classifier for Text Classification

Naïve Bayes classifier has been applied in numerous multi-class classification problems where instances are classified into at least three categories such as this YouTube video classification problem. It assumes that there is a conditional independence among the attributes and has been found to perform very well on dataset with this property.

Let Y be the response label of a video and X be the set of terms as predictors for classification (i.e. features of a document). Based on the Bayes classifier, our predicted class $\hat{y}$ for a document is the value of $y \in Y$ so that the conditional property of y that is $P(y|x)$ is maximised. The formal decision rule is as follows:

$$\hat{y} = \underset{y \in Y}{\operatorname{argmax}} P(y|\boldsymbol{x})$$

$$= \underset{y \in Y}{\operatorname{argmax}} \frac{P(\boldsymbol{x}|y)P(\boldsymbol{x})}{P(y)} \quad \textit{(Bayes' Rule)}$$

$$= \underset{y \in Y}{\operatorname{argmax}} P(\boldsymbol{x}|y)P(\boldsymbol{x}) \quad \textit{(as all probabilities have P(y) as their denominator)}$$

The videos will be classified using their description according to the decision rule above. The class with the highest probability will be our predicted class. The dataset is divided into two smaller sets. The first set containing 70% of the initial dataset is the training set while the test set consists of the other 30%. It is ensured that the training set contains videos from all categories, including those with only a few samples. By ensuring each category has sufficient samples, it helps the model understand the variance among these instances.

Results and Explanation

Confusion matrices can be useful tools for evaluating a classification model's performance against the actual targets in the dataset. Two confusion matrices on the training and test sets are generated to evaluate the model.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2  10  15  17  19  20  22  23  24  25  26  27  28  29  43
        1   50   0   0   0   0   0   0   0   0   3   0   0   0   0   0   0
        2    0   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        10   2   1 144   4   3   0   0  12  18  51   1   7   2   3   0   0
        15   1   0   0  14   0   0   0   0   0   0   0   0   0   0   0   0
        17   0   0   0   0  71   0   0   2   0   0   0   0   0   0   0   0
        19   0   0   0   0   0   9   0   0   0   0   0   1   0   0   0   0
        20   0   0   0   0   0   0  30   0   1   8   1   1   0   0   0   0
        22   0   0   0   0   0   0   0  69   0   6   0   8   0   0   0   0
        23   0   0   0   0   0   0   0   0  83  14   0   0   0   0   0   0
        24   0   0   0   0   0   0   0   0   0 217   0   1   0   0   0   0
        25   0   0   0   0   0   0   0   2   0   4  41   0   0   0   0   0
        26   0   0   0   0   0   0   0   2   0   9   0  79   0   0   0   0
        27   0   0   0   0   0   0   0   0   0   0   0   0  47   0   0   0
        28   0   0   0   0   0   0   0   1   2  11   0   1   0  56   0   0
        29   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
        43   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2

Overall Statistics

               Accuracy : 0.8336
                 95% CI : (0.8103, 0.8552)
    No Information Rate : 0.2936
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.81

 Mcnemar's Test P-Value : NA
```

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2 10 15 17 19 20 22 23 24 25 26 27 28 29 43
        1  17  0  0  0  0  0  1  0  0  2  0  0  0  0  0  0
        2   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
        10  4  0 60  1  5  0  1  7 14 21  3  2  1  1  0  0
        15  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0
        17  0  0  0  0 36  0  0  3  0  0  0  0  0  0  0  0
        19  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0
        20  0  0  0  0  1  0  7  0  1  9  0  1  0  0  0  0
        22  0  0  0  0  2  0  0 12  2  3  0  4  0  0  0  0
        23  0  0  0  0  0  0  0  0 23  6  0  0  0  0  0  0
        24  4  0  2  0  4  0  4  6  2 80  0  2  1  1  0  0
        25  0  0  1  0  1  0  0  0  0  2 18  0  0  0  0  0
        26  0  0  0  0  0  0  0  4  0  6  0 37  0  0  0  0
        27  0  0  0  0  0  0  0  1  1  2  0  0 11  0  0  0
        28  1  0  0  0  1  0  3  2  2  5  1  3  0  9  0  0
        29  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
        43  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.6645
                 95% CI : (0.6199, 0.7071)
    No Information Rate : 0.2887
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6117

 Mcnemar's Test P-Value : NA
```

Figure 3.4: Confusion matrices for the model on the training set (left) and test set (right)

Figure 3.4 illustrates the confusion matrices which can be used to evaluate the performance of the model on the training and test sets. The model performs relatively well on the training set with an accuracy of around 83.36%. However, its performance on the test set is only mediocre at around 66.45%. This indicates that an overfitting problem may have occurred.

4.0 Discussion and Conclusion

In this paper, two techniques for categorising YouTube videos based on their description were discovered. These are Association Rules and Naïve Bayes classifier. The application of Association Rules into finding the most frequent terms used for each category was studied. As a result, we were able to determine a set of different frequent words appearing together for each category. Also, there was an uncertainty in whether these rules consistently appear in each category as proven by the fluctuation in the lift value. This could be due to the small size of the samples. However, it is possible that distinct frequent words can be extracted for each category. Similar researches such as **[24]** and **[25]** also find that Association Rule mining could be adapted to textual descriptions of the videos to identify their topic. Furthermore, as in the case of **[19]** and **[23],** the user comments on the videos could also be taken into consideration for video categorisation.

Also, we have adapted the Naïve Bayesian model for this classification task. As a result, the model performs very well on the training set while its overall accuracy on the test set is mediocre. This indicates a possible overfitting problem. The problem could be due to the skewness of the initial dataset where one category has significantly more samples than another. For instance, category 24 has significantly more samples than category 29. **[18]** pointed out that the skewed initial dataset, resulting in a skewed training set, creates a bias problem where over-represented classes have a higher bias. A larger number of False Negatives (FNs) for classes that have higher bias is observed. For example, in the matrix on the test set, category 24 has notably more FNs (56 FNs) compared to the other classes. Thus, to improve the model's accuracy, there needs to be a larger and balanced dataset for training and testing. Also, the low performance of the model could be because many of the descriptions could contain words of other categories. This means that these samples have the features of other classes besides their true classes, which makes the prediction task more difficult. While a simple Naïve Bayes model is unbiased, **[17]** introduces a new variable for text classification that creates a biased model but with low variance. An application of the model in this dataset can improve the prediction accuracy.

In conclusion, it is possible to extract the frequent terms used in the video descriptions for each category with Association Rules. This makes classifying the videos based on their textual descriptions possible by applying different techniques such as Naïve Bayes classifier. Future in-depth research should investigate into improving the model's performance or applying a hybrid model to improve the prediction accuracy.

5.0 References

[1] Simonsen, T.M., 2011. Categorising YouTube. MedieKultur: Journal of media and communication research, 27(51), pp.23-p.

[2] Dilrukshi, I. and De Zoysa, K., 2013, December. Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithms. In *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 278-278). IEEE.

[3] Xiao, L., Wang, G. and Liu, Y., 2018, December. Patent Text Classification Based on Naive Bayesian Method. In *2018 11th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 1, pp. 57-60). IEEE.

[4] BaygIn, M., 2018, September. Classification of text documents based on Naive Bayes using N-Gram features. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-5). Ieee.

[5] Bužić, D. and Dobša, J., 2018, May. Lyrics classification using naive bayes. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1011-1015). IEEE.

[6] Lv, L. and Liu, Y.S., 2005, December. Research and realization of naive Bayes English text classification method based on base noun phrase identification. In *2005 International Conference on Information and Communication Technology* (pp. 805-812). IEEE.

[7] Kulkarni, A.R., Tokekar, V. and Kulkarni, P., 2012, September. Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining. In *2012 CSI sixth international conference on software engineering (CONSEG)* (pp. 1-4). IEEE.

[8] Kim, S.B., Han, K.S., Rim, H.C. and Myaeng, S.H., 2006. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, *18*(11), pp.1457-1466.

[9] Bhargavi, P. and Jyothi, S., 2009. Applying naive bayes data mining technique for classification of agricultural land soils. *International journal of computer science and network security*, *9*(8), pp.117-122.

[10] Huang, Y. and Li, L., 2011, September. Naive Bayes classification algorithm based on small sample set. In *2011 IEEE International Conference on Cloud Computing and Intelligence Systems* (pp. 34-39). IEEE.

[11] Siswanto, B & Thariqa, P 2018, 'Association rules mining for identifying popular ingredients on YouTube cooking recipes videos', in *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, pp. 95-8.

[12] Zhang, C, Wu, X, Shyu, M-L & Peng, Q 2013, 'Adaptive association rule mining for web video event classification', in *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, pp. 618-25.

[13] Reddy, GS, Rajinikanth, TV & Rao, AA 2014, *A frequent term based text clustering approach using novel similarity measure*, IEEE, 978-1-4799-2571-1

[14] Zhou, Y 2009, 'Searching and Clustering on Social Tagging Sites', in *2009 Fifth International Conference on Semantics, Knowledge and Grid*, pp. 99-105.

[15] Beil, F, Ester, M & Xu, X 2002, 'Frequent term-based text clustering', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 436-42.]

[16] Cherfi, H., Napoli, A. and Toussaint, Y., 2006. Towards a text mining methodology using association rule extraction. *Soft Computing*, *10*(5), pp.431-441.]

[17] Chen, J, Dai, Z, Duan, J, Matzinger, H & Popescu, I 2019, 'Naive Bayes with Correlation Factor for Text Classification Problem', in 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 1051-6.

[18] Jeni, LA, Cohn, JF & De La Torre, F 2013, 'Facing Imbalanced Data Recommendations for the Use of Performance Metrics', International Conference on Affective Computing and Intelligent Interaction and workshops : [proceedings]. ACII (Conference), vol. 2013, pp. 245-51.

[19] Madden, A., Ruthven, I. and McMenemy, D., 2013. A classification scheme for content analyses of YouTube video comments. *Journal of Documentation*, 69(5), pp.693-714.

[20] Xu, S., 2016. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1), pp.48-59.

[21] Rekik, R., Kallel, I., Casillas, J. and Alimi, A., 2018. Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management*, 38(1), pp.201-216.

[22] Pérez, J., Iturbide, E., Olivares, V., Hidalgo, M., Martínez, A. and Almanza, N., 2015. A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases. *Journal of Medical Systems*, 39(11).

[23] Padmanabhan, D. and Kummamuru, K., 2007. Mining conversational text for procedures with applications in contact centers. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(3-4), pp.227-238.

[24] Güder, M & Çiçekli, NK 2018, 'Multi-modal video event recognition based on association rules and decision fusion', Multimedia Systems, vol. 24, no. 1, pp. 55-72.

[25] Zhang, C, Wu, X, Shyu, M & Peng, Q 2013, 'Adaptive association rule mining for web video event classification', in 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), pp. 618-25.