Analysis of Interest Pattern of Youtube Browsing in Indonesia Using Web Crawling and Supervised Learning

Analisis Pola Minat Tayangan Youtube DI Indonesia dengan Web Crawling dan Supervised Learning

Nofriani
BPS-Statistics of Bengkulu Province
Jl. Adam Malik, KM. 8 Bengkulu City, Bengkulu Province

nofriani@bps.go.id

Naskah diterima: 18 Maret 2018, direvisi: 7 September 2018, disetujui: 19 November 2018

Abstract

YouTube is a popular video sharing website, specifically in Indonesia. Every day, in every country, the list of trending videos is updated on YouTube's Trending page. The data of trending videos can be used for information exploration, such as analysis of user interest pattern of YouTube browsing. This research aims to grab and analyse the metadata of trending videos to generate a classifier model and statistics of trending YouTube videos in Indonesia. Every day, the data are collected from YouTube's Trending page using Scraper and Screaming Frog SEO Spider tools for 10 consecutive days. The data are later classified into video categories. The approach used for this purpose is rule-based classification using J48 tree algorithm and TF-IDF filter. The result of this research shows that videos about people, blogs, sports, news, politics, comedy, entertainment and music are what interest the Indonesian people the most.

Keywords: classification, video category, trending, crawling, supervised learning

Abstrak

YouTube adalah website yang populer untuk berbagi video, khususnya di Indonesia. Setiap hari, di setiap negara, daftar video-video yang trending di-update di halaman Trending YouTube. Data video trending dapat digunakan untuk analisis informasi, salah satunya analisis pola minat tayangan YouTube. Penelitian ini bertujuan untuk mengambil dan menganalisis metadata video untuk membangun model dan data statistik minat video tayangan YouTube di Indonesia. Data diambil dari halaman Trending YouTube dengan menggunakan alat Scraper dan Screaming Frog SEO Spider, setiap hari selama 10 hari berturut-turut. Data-data tersebut diklasifikasikan ke dalam berbagai kategori video. Pendekatan yang digunakan adalah klasifikasi berbasis aturan menggunakan algoritma J48 tree dan filter TF-IDF. Hasil penelitian ini menunjukkan bahwa video mengenai masyarakat, blog, olahraga, berita, politik, komedi, hiburan dan musik adalah yang paling menarik penonton di Indonesia.

Kata kunci: *klasifikasi, kategori video, trending, crawling, pembelajaran terarah*

INTRODUCTION

YouTube is a free video sharing website that allows users to watch online videos or upload their own. It was first established in 2005, and is now one of the most popular sites on the Web, with visitors watching around 6 billion hours of video every month (GCF, 2018).

Indonesia is one of the countries of which the people watch YouTube every day. According to Hootsuite, in Indonesia, YouTube is ranked 6th among top websites visited every month, with monthly traffic at 558,900,000 and 23 minutes spent per visit.

Every day, for every country, the list of trending videos is updated on YouTube's featured page: Trending. According to YouTube Help, the list shows the videos that meet the criteria in YouTube algorithm to be in Trending; 1) view count, 2) the growth rate in views, 3) where the views are coming from, and 4) the age of the video. The list chooses videos that are most relevant to viewers' interests.

Viewers' interests also help youtubers across the world to upload better contents that are more relevant to their viewers. The educators can also use the information to better teach their students based on their features of interest. What's happening on YouTube relatively reflects what's happening in the real world, so even a regular civilian can use the information to get a brief description of what interests Indonesians to watch.

According to Backlinko, there are several contributing factors that affect viewers' interests in watching a YouTube video, which in turn affect the ranking of the video: 1) The length of video, 2) Video view count, 3) The number of video shares, 4) Video likes, 5) Quality of videos.

One of the approaches to understand the viewers' interests on YouTube is supervised machine learning. It is an approach in which an algorithm is trained on documents with known categories (Hulth, 2006). The basic idea of supervised machine learning is that a computer can automatically learn from experience (Pojon, 2017). Supervised machine learning has been a great success in real-world applications. It is used in almost every domain, including text and web domains (Liu, 2011).

Previous works have used supervised learning to study patterns on a series of data. One research in 2016 by Lya Hulliyyatus Suadaa used supervised learning to find the pattern of tweets on Twitter about commuter line. The research concluded that supervised learning can predict commuter line intrusion in the future by accuracy up to 90%. Another research in 2018 by Bening Herwijayanti et. al. also used machine learning to classify online news based on TF-IDF weighting and cosine similarity. The research obtained 80% accuracy in predicting the online news classification using machine learning.

This research aims to analyse the pattern of interests in YouTube browsing using the properties of the videos in YouTube's trending page. The information can be used for youtubers, educators, researches, civilians and the government to determine the types of contents that interest Indonesians the most. This research's question is to find out the way to extract the information from YouTube's trending page to analyse and conclude the pattern of interests in YouTube browsing. The text analytics helps to extract meanings, patterns and structure hidden in unstructured textual data (Chakraborty, 2013), which is derived from YouTube's trending page.

The objectives of this research are as follows:

1. To analyse the statistics of YouTube daily trending videos.
2. To provide the data classifier model to classify the videos on YouTube trending page.
3. To develop a prototype method to grab and analyse the data on YouTube trending topic.

METHOD

The methodology used to achieve the objectives of this research is a set of information extraction processes shown in Figure 1.

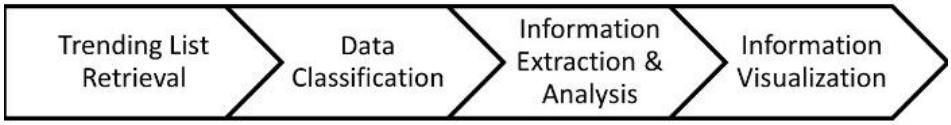


Figure 1. Information Extraction Processes

Process 1: Trending List Retrieval grabs the video properties such as titles, keywords, the character length of the video, etc. *Process 2: Data Classification* classify the videos based on words in the titles and keywords. *Process 3: Information Extraction and Analysis* perform statistical analysis on the classification results and other properties of the video. *Process 4: Information Visualization* present tables and figures that explain the results of the previous processes.

Trending List Retrieval

Information Retrieval is obtaining information resources relevant to an information need from huge amount of information available in digital form (George, 2014). In this process, the information retrieval is performed on the Trending List, the daily-updated list consisting of at most fifty trending videos is collected (scraped) every day at random hours from 12 pm to 6 pm, for ten consecutive days, from 2nd to 11th of March 2018. Since YouTube shows different list of trending videos for each country, the location is set to "Indonesia".

YouTube do not currently provide the API *(Application Programming Interface)* to facilitate users in crawling the trending video list. Therefore, this research uses Scraper tool to scrape the video URLs from the Trending page and Screaming Frog SEO Spider to crawl the video metadata corresponding to each scraped URL. Metadata is a kind of highly structured document summary (Witten, 2004). This research uses three types of metadata: 1) Words in the video title, 2) Video keywords, 3) The character length of video title

Data Classification

Data mining includes many techniques that can unveil inherent structure in the underlying data (Hammouda, 2004). One of these techniques is data classification. It is the process of classifying the text documents based on words, phrases and word combinations with respect to a set of predefined categories (Menaka, 2013). For the data classification process, this research uses two types of data; words in the video title and words in the video keywords (if there are any). Both terms will be combined and referred to as "video contents".

The video contents are classified into several categories, using 8 out of 18 categories on YouTube webpage's official list, "YouTube Lesson: Video Categories" (YouTube, 2018):

0: Comedy and Entertainment
1: Films
2: Music
3: News and Politics
4: People and Blogs (including Celebrities)
5: Sports
6: Technology
7: Others

For this purpose, this research uses Weka application version 3.8.2. It is a free open-source software containing a collection of machine learning algorithms for data mining tasks developed by Waikato University of New Zealand (Waikato University, 2018). It uses ARFF file format as the

source file to do the data processing (Noviyanto, 2015). Using the features in Weka, the data go through cleaning, pre-processing, tokenization, association, clustering, classification, and visualization.

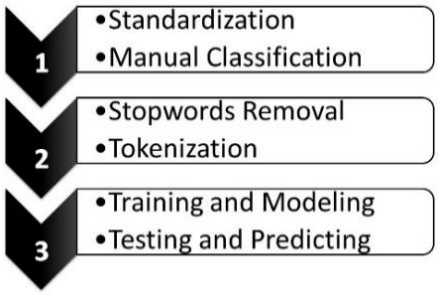


Figure 2. Process Flow of Data Classification

The details of data classification steps shown in Figure 2 are explained below:

1. Step 1: Pre-Processing

   Most of YouTube video contents written in Bahasa Indonesia contain texts with non-standard spelling, equipped with numbers (0-9) and non-alphanumeric characters. This step's objective is to ease the data processing and improve the accuracy in data classification by standardizing the texts, removing all numbers and non-alphanumeric characters then manually classifying each field in training data set.

   a. Standardization

      In this phase, the informal language style in video contents is manually formalized with the help of VBA (Visual Basic for Applications). Figure 3 shows the VBA syntax used to remove all numbers, fullstops, commas and other non-alphanumeric characters.

```
Sub RemoveNotAlphasNotNum()
Dim Rng As Range
Dim WorkRng As Range
On Error Resume Next
xTitleId = "KutoolsforExcel"
Set WorkRng = Application.Selection
Set WorkRng = Application.InputBox("Range", xTitleId, WorkRng.Address, Type:=8)
For Each Rng In WorkRng
    xOut = ""
    For i = 1 To Len(Rng.Value)
        xTemp = Mid(Rng.Value, i, 1)
        If xTemp Like "[a-z]" Or xTemp Like "[A-Z]" Then
            xStr = xTemp
        Else
            xStr = " "
        End If
        xOut = xOut & xStr
    Next i
    Rng.Value = xOut
Next
End Sub
```

Figure 3. Visual Basic Syntax to Remove Numbers and Non-alphanumeric Characters

```
Sub ReplaceCharacters()
    Sheet1.Select

    Sheet1.Range("E2:E101").Select
    Selection.Replace What:=" dan ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" ya ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" ini ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" ga ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" tidak ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" and ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" di ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
    Selection.Replace What:=" on ", Replacement:=" ", LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
```

Figure 4. VBA Syntax to Remove Stopwords

b. Manual Classification

Text classification classifies documents on the basis of words, phrases, combination of words, etc. into predefined class labels (Kawade, 2016).

The cleaned training data set is then prepared for the data training process by manually classifying every field in the set.

2. Step 2: Features Extraction

Feature extraction is a process of extracting new features from the original features using functional mapping (Langgeni, 2010). Such extraction is widely applied in text summarization, information retrieval and other aspects (Li-gong, 2013).

a. Stopwords Removal

Stopwords are the words in a document that are frequently occurring but meaningless in terms of Information Retrieval (Lo, 2005). Since they are meaningless, stopwords must be removed from the document to optimize the classification accuracy. Figure 4 shows the VBA syntax to remove stopwords from the training data set.

b. Tokenization

Tokenization is a process of splitting a set of words into tokens separated by whitespaces. The result of this step is a bag of important words for model development. To choose the words, a text-mining approach is used, i.e. Term Frequency/ Inverse Document Frequency (TF-IDF). As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in (Ramos, 2003). It scores the importance of words in the data set based on how frequently they appear across multiple data sets (Loria, 2013). Figure 5 shows the use of this method in Weka application.

In calculating the importance of words, the TF of every word is weighed 1. While the IDF value is formulated as follows:

$$IDF(word) = log\frac{td}{df}$$

$IDF_{(word)}$ is the value of IDF of each word, *td* is the total of documents, and *df* is the total appearance of words in all documents (Herwijayanti, 2018). The more frequently a word appear in the document, the higher the $IDF_{(word)}$ value of the word (Fitri, 2013).

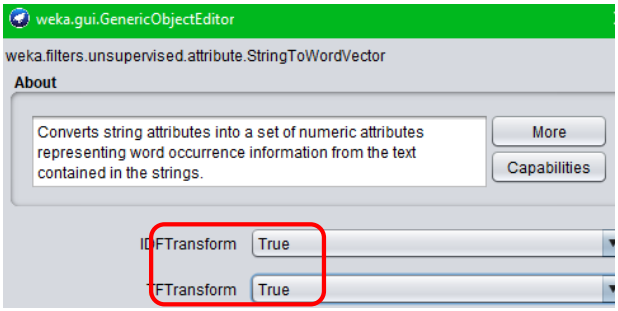


Figure 5. TF-IDF

3. Step 3: Data Processing
   a. Training and Modeling

      The training process is held using a filtered classifier feature provided in Weka application. The classifier uses J48 tree algorithm to classifiy each set of words and is filtered using TF-IDF method.

   b. Predicting and Testing

      The resulted classifier model from the training process is then applied to label the text classifications on test data set. The purpose of text classification automation is to perform effectively as humans when classifying text in knowledge fields (Afonso, 2014). Automated text classification has been considered a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing (Ikonomakis, 2005). Such analysis can help provide us with a better understanding of the data at large (Han, 2012).

DISCUSSION

Video Statistics

Figure 6 shows the word cloud of video contents. Figure 7 shows fifteen most frequently reoccurring words in video contents.

Figure 6. Word Cloud of Video Contents

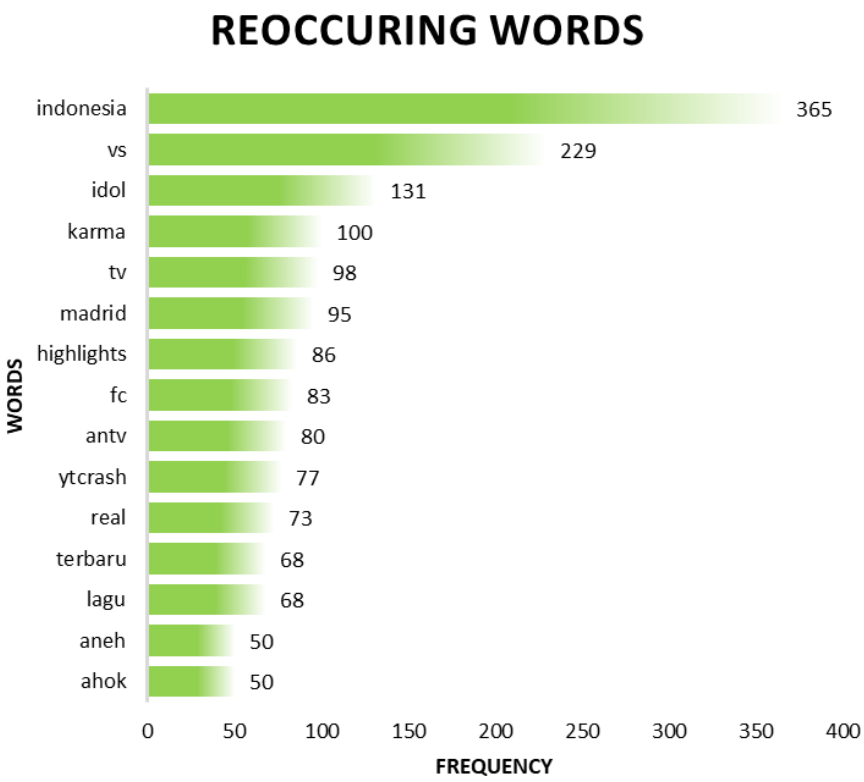


Figure 7. Fifteen Most Frequently Reoccurring Words in Video Keywords

The figures show that out of 11,979 words in the video contents (titles and keywords), the most frequently reoccuring words are *indonesia, vs (versus), idol, karma, tv, madrid, highlights, fc (football club), antv,* and *ytcrash (youtube crash).* This means that Indonesians mostly watch videos about:

Sports, because video titles about sport matches contain the words *vs*, *highlights, real,* and *madrid*.

FC, which stands for Football Club or Fans Club. It also represents **sports**.

Indonesian Idol, the trending singing talent show in Indonesia at the time.

Karma TV, the trending television show in Indonesia at the time. It is aired in a private TV station **ANTV.**

YouTube Crash, the trending section of YouTube that shows feeds about unique facts and trending events across the world.

Updated News, because the frequently emerged word *terbaru* means *newest/updated* in English.

Songs, because the frequently arised word *lagu* means *songs.* It is possible that this is included in Trending because of the the singing talent show **Indonesian Idol.**

Strange/weird things, because the frequently reoccured word *aneh* means *strange* or *weird* in English.

Politics, because the frequently appeared word *ahok* is a name of former governor of DKI Jakarta, which frequently shows up on the news.

The average of character length of video titles is 75.28. This means that the most intriguing video titles contain about 75 characters (letters and numbers) including the whitespaces. That is equal to about 8-15 words (Nirmaldasan, 2008). The number suggests that the relevant video titles that will likely interest the viewers contain no more than 15 words.

Classification Training Results

The classification is performed on video contents (titles and keywords) using decision tree algorithm in data mining. A decision tree is a tree whose internal nodes are tests (on input patterns) and whose leaf nodes are categories (of patterns) (Nilsson, 1998). J48 pruned tree is used to build the decision tree of the data set. Decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data (Sharma, 2015). The dependent variable used in this research is video classification consisting of eight classes (video categories enlisted in Methodology Section). Figure 8 shows the classifier model resulted from the training process.

```
Classifier Model
J48 pruned tree
------------------

galaxy <= 0
|   vs <= 0
|   |   animation <= 0
|   |   |   black <= 0
|   |   |   |   idol <= 0
|   |   |   |   |   vallen <= 0
|   |   |   |   |   |   amazing <= 0
|   |   |   |   |   |   |   coaster <= 0
|   |   |   |   |   |   |   |   cinta <= 1.750702
|   |   |   |   |   |   |   |   |   pk <= 0
|   |   |   |   |   |   |   |   |   |   top <= 0
|   |   |   |   |   |   |   |   |   |   |   berita <= 0
|   |   |   |   |   |   |   |   |   |   |   |   konser <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   banjarnegara <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   net <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   antv <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   kaget <= 0: 4 (28.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   kaget > 0: 0 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   antv > 0: 0 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   net > 0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   banjarnegara > 0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   konser > 0: 0 (2.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   berita > 0: 3 (8.0/2.0)
|   |   |   |   |   |   |   |   |   |   top > 0: 3 (2.0)
|   |   |   |   |   |   |   |   |   pk > 0: 3 (6.0)
|   |   |   |   |   |   |   |   cinta > 1.750702: 2 (2.0/1.0)
|   |   |   |   |   |   |   coaster > 0: 7 (2.0)
|   |   |   |   |   |   amazing > 0: 7 (2.0)
|   |   |   |   |   vallen > 0: 2 (3.0)
|   |   |   |   idol > 0: 2 (9.0)
|   |   |   black > 0: 1 (2.0/1.0)
|   |   animation > 0: 1 (2.0)
|   vs > 0
|   |   indonesia <= 0: 5 (18.0/1.0)
|   |   indonesia > 0: 3 (2.0)
galaxy > 0: 6 (2.0)

Number of Leaves  :     19

Size of the tree :      37
```

Figure 8. Classifier Model

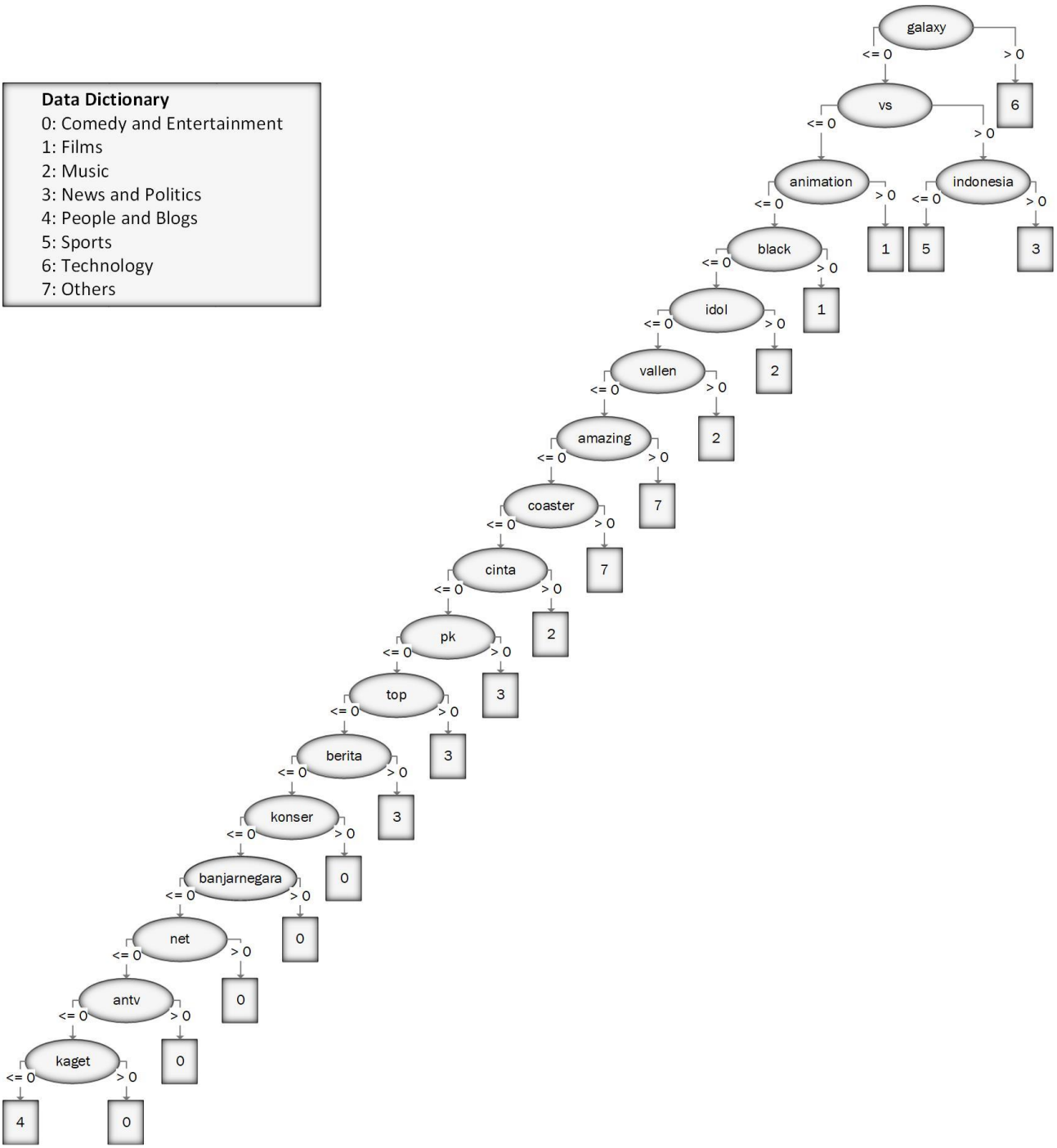


Figure 9. Decision Tree of Video Classification

Figure 9 shows the visualized decision tree based on the result shown in Figure 8. The tree is interpreted as follows:

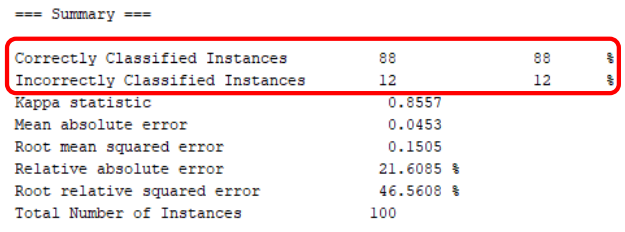
1. If the video has the word *galaxy*, then it is classified as Category 6: Technology.
2. If the video has no word *galaxy*, then:
    a. If it has the word *vs* and *indonesia*, it is classified as Category 3: News/Politics.
    b. If it has the word *vs* but not *indonesia*, it is classified as Category 5: Sports
3. If the video has no word *galaxy* nor *vs* then:
    a. If it has the word *animation*, then it is classified as Category 1: Films
    b. If it has no word *animation,* but has the word *black,* it is also classified as Films.
4. If it has no word *galaxy, vs, animation,* nor *black,* then:
    a. If it has the word *idol,* then it is classified as Category 2: Music
    b. If it has no word *idol,* but has the word *vallen,* it is also classified as Music.
5. If it has no word *galaxy, vs, animation, black, idol,* nor *vallen,* then:
    a. If it has the word *amazing,* then it is classified as Category 7: Others.
    b. If it has no word *amazing,* but has the word *coaster,* it is also classified as Others.
6. If it has no word *galaxy, vs, animation, black, idol, vallen, amazing,* nor *coaster,* then:
    a. If it has the word *cinta,* then it is classified as Category 2: Music.
    b. If it has no word *cinta,* but has the word *pk* (pakar hukum), it is classified as Category 3: News and Politics
7. If it has no word *galaxy, vs, animation, black, idol, vallen, amazing, coaster, cinta,* nor *pk,* then:
    a. If it has the word *top,* then it is classified as Category 3: News and Politcs
    b. If it has no word *top* but has the word *berita,* it is also classified as News and Politics.
8. If it has no word *galaxy, vs, animation, black, idol, vallen, amazing, coaster, cinta, pk, top,* nor *berita* then:
    a. If it has the word *konser,* then it is classified as Category 0: Comedy and Entertainment
    b. If it has no word *konser,* but has *banjarnegara,* it is also classified as Comedy and Entertainment.
9. If it has no word *galaxy, vs, animation, black, idol, vallen, amazing, coaster, cinta, pk, top, berita, konser,* nor *banjarnegara* then:
    a. If it has the word *net,* then it is classified as Category 0: Comedy and Entertainment
    b. If it has no word *net,* but has *antv,* it is also classified as Comedy and Entertainment.
10. If it has no word *galaxy, vs, animation, black, idol, vallen, amazing, coaster, cinta, pk, top, berita, konser, banjarnegara, net,* nor *antv,* then:
    a. If it has the word *kaget,* then it is classified as Category 0: Comedy and Entertainment
    b. If it has no word *kaget,* it is classified as People and Blogs.

Figure 10 shows the summary of Classifier Model. It shows that the model has 88% prediction accuracy. The number is adequate considering that the pattern of YouTube video titles is grammatically unstructured and contains irrelevant clickbaits. Clickbaits refer to a certain kind

of web content advertisement that is designed to entice viewers into clicking an accompanying link (Potthast et al, 2016).

Even though the classifier model has 88% prediction accuracy, this classification technique might not be accurate enough if the additional test data set contains words that are not in the training data set. This is because the classification algorithm only classifies the data based on the previously identified categories in the training data set, without taking into account the context or meaning of the words. Interpretation of human written language generally requires high levels of cognitive functionalities, and computers are not originally designed for this task (Bezerra, 2006). Therefore, this classifier may not work well on a higher level of test data set.

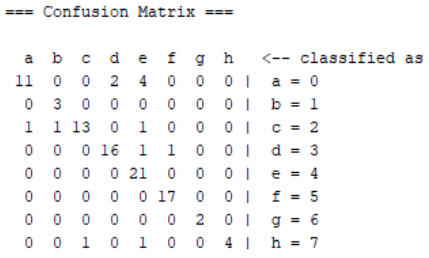
If the pattern of the test data set is entirely different, then the classifier will have difficulties in recognizing the pattern and the entire process will have to stop. This is the major problem faced by most of the machine learning process and algorithms (Talwar, 2013). Figure 11 shows the details on misclassified data on the training set. There are 12 YouTube video categories that are misclassified by the classifier model. Figure 11 shows the details on misclassified data on the training set.



```
=== Summary ===

Correctly Classified Instances          88               88      %
Incorrectly Classified Instances        12               12      %
Kappa statistic                          0.8557
Mean absolute error                      0.0453
Root mean squared error                  0.1505
Relative absolute error                 21.6085 %
Root relative squared error             46.5608 %
Total Number of Instances              100
```

Figure 10. Summary of Classifier Model



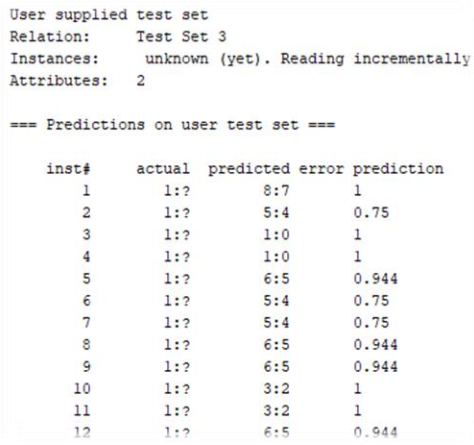
```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 11  0  0  2  4  0  0  0 |  a = 0
  0  3  0  0  0  0  0  0 |  b = 1
  1  1 13  0  1  0  0  0 |  c = 2
  0  0  0 16  1  1  0  0 |  d = 3
  0  0  0  0 21  0  0  0 |  e = 4
  0  0  0  0  0 17  0  0 |  f = 5
  0  0  0  0  0  0  2  0 |  g = 6
  0  0  1  0  1  0  0  4 |  h = 7
```

Figure 11. Confusion Matrix of Classification Output

Classification Testing Results

The 392 fields of test data set are labeled and classified in Weka application using the classifier model developed out of the training data set. Figure 12 shows the result of data testing.



```
User supplied test set
Relation:     Test Set 3
Instances:     unknown (yet). Reading incrementally
Attributes:   2

=== Predictions on user test set ===

    inst#     actual  predicted error prediction
        1        1:?        8:7       1
        2        1:?        5:4       0.75
        3        1:?        1:0       1
        4        1:?        1:0       1
        5        1:?        6:5       0.944
        6        1:?        5:4       0.75
        7        1:?        5:4       0.75
        8        1:?        6:5       0.944
        9        1:?        6:5       0.944
       10        1:?        3:2       1
       11        1:?        3:2       1
       12        1:?        6:5       0.944
```

...

```
384        1:?        6:5        0.944
385        1:?        5:4        0.75
386        1:?        5:4        0.75
387        1:?        5:4        0.75
388        1:?        5:4        0.75
389        1:?        1:0        1
390        1:?        5:4        0.75
391        1:?        5:4        0.75
392        1:?        4:3        1

=== Summary ===

Total Number of Instances                0
Ignored Class Unknown Instances                392
```

Figure 12. Predicted Classifications

The result shows that the predicted confidence interval is around 75-100%, which might be caused by issues on OOV (Out of Vocabularies) or OOR (Out of Rules). Out of Vocabulary is the condition where the information extraction system cannot find the match for a token in the list of vocabulary in the database (Endarnoto, 2011). Nevertheless, the number still prove that classifier model can accurately predict the training data set with 88% prediction accuracy.

Supervised machine learning techniques involve a process that creates a predictive model from training samples that are representatives of its domain (Russell, 2014) Thus, the training data set is customizable should one desires to improve it for increased accuracy. Figure 13 shows the statistics of video classifications

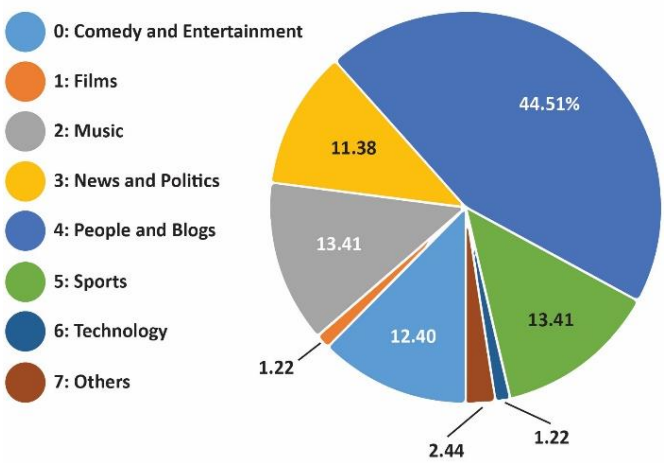


Figure 13. Frequent Video Classifications

The pie chart shows that Indonesians love to watch videos about people or blogs, sports, music, news or politics, and comedy or entertainment. Meanwhile, videos about technology, films and other topics are not quite interesting to the viewers. This might be caused by the worlwide trending topics. For example, if sport matches and mobile phones are trending worldwide, they will also be a trending topic in Indonesia. The people and blogs are in Trending because of the frequent cases happening in Indonesia that are shown on the news. Latest trending TV shows also contribute to the large share in comedy, music and entertainment categories. Meanwhile, news and politics are always interesting to most viewers because they are trending topic across the year.

CLOSING

In this research, information extraction via rule-based approach is performed to understand the pattern of interests of YouTube browsing in Indonesia from YouTube's Trending Page. The rule is developed in classification training using 100 video metadata containing video titles and keywords. The decision tree of J48 algorithm resulted from the classification training is used as classifier model to classify the remaining video metadata. The classifier model has 88% prediction accuracy and 75-100% confidence in predicting the video classifications. Based on the result of the classification and video statistics, it is concluded that there are three video categories that interest people in Indonesia the most, i.e. videos about people and blogs, sports, and music. However, considering that this task of research scope is far from being solved, our future work will be on building a system that can show real-time video classifications on YouTube trending page.

REFERENCES

Afonso, Alexandre Ribeiro and Claudio Gottschalg Duque. "Automated Text Clustering of Newspaper and Scientific Texts in Brazilian Protuguese: Analysis and Comparison of Methods." *JISTEM – Journal of Information Systems and Technology Management.* Brazil: University of Sao Paulo, 2014.

Backlinko. "We Analyzed 1.3 Million YouTube Videos. Here's Wat We Learned About YouTube SEO." Accessed on March 3rd, 2018. https://backlinko.com/youtube-ranking-factors/

Bezerra, G. B., Barra, T. V., Ferreira, H. M., & Von Zuben, F. J. A hierarchical immune-inspired approach for text clustering. *Selected papers based on the presentations at the 2006 conference on Information Processing and Management of Uncertainty.* France, 2006.

Chakraborty, Goutam et al. "Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS." USA, 2013.

Endarnoto, S K., et al, "*Traffic Condition Information Extraction and Visualization from Social Media Twitter for Android Mobile Application*", International Conference on Electrical Engineering and Informatics, 2011.

Fitri, Meisya. Perancangan Sistem Temu Balik Informasi dengan Metode Pembobotan Kombiasi TF-IDF untuk Pencarian Dokumen Berbahasa Indonesia. Semarang: Universitas Tanjungpura, 2013.

GCF Global. "What is YouTube?". Accessed on March, 2nd 2018. https://www.gcflearnfree.org/youtube/what-is-youtube/1/. (Official website)

George, Soumya and Shibily Joseph. "Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature. *IOSR Journal of Computer Enngineering.* India: IOSR Journals, 2014.

Hammouda, K. M., & Kamel. M. S. Efficient Phrase-Based Document Indexing for Web Document Clustering. *IEEE Transactions on Knowledge and Data Engineering.* USA, 2004.

Han, Jiawei et al. Data Mining Concepts and Techniques, 3rd Edition. USA: Elsevier Inc., 2012.

Herwijayanti, Bening et al. Klasifikasi Berita Online dengan Menggunakan Pembobotan TF-IDF dan Cosine Similarity. *Jurnal Pengembangan Teknologi Informasi dan Komputer.* Malang: Universitas Brawijaya, 2018.

Hootsuite Media. *Indonesia Digital Landscape 2018,* 2018. Accessed on February 8th, 2018. https://hootsuite.com/resources/digital-in-2018-apac/

Hulth Anette, Beata B Megyesi. A Study on Automatically Extracted Keywords in Text Categorization. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*. Sydney, 2006.

Ikonomakis, M et al. "Text Classification Using Machine Learning Techniques." *WSEAS Transactions on Computers Issue 8, Volume 4, August 2005.*

Kawade, Dipak Ramchandra and Kavita S.Oza. "News Classification: A Data Mining Approach." *Indian Journal of Science and Technology.* India: Indian Society of Education and Environment, 2016.

Langgeni, Diah Pudi, et al. "Clustering Artikel Berita Bahasa Indonesia Menggunakan Unsupervised Feature Selection." *Seminar Nasional Informatika.* Yogyakarta: UPN, 2010.

Li-gong Y, et al. "Keywords extraction based on text classification." *Proceedings of the 2nd International Conference On Systems Engineering and Modeling*. 2013.

Liu, Bing. "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data." 2nd Edition. USA: Springer-Verlag Berlin Heidelberg, 2011.

Lo, Tsz-wai Rachel, et al. "Automatically Building A Stopword List for An Information Retrieval System." Glasgow: University of Glasgow, 2005.

Loria, Steven. "Tutorial: Finding Important Words in Text Using TF-IDF". 2013. Accessed on March 15th, 2018. https://stevenloria.com/tf-idf/

Menaka S, Radha N. Text Classification Using Keyword Extraction Technique. *International Journal of Advanced Research in Computer Science and Software Engineering.* 2013.

Nilsson, Nils J. "Introduction to Machine Learning." Standford University, 1998.

Nirmaldasan. "The Average Sentence Length." 2008. Accessed on March 15th, 2018. https://strainindex.wordpress.com/2008/07/28/the-average-sentence-length/

Noviyanto, Hendri, et al. "Pengklasifikasian Laman Web Berdasarkan Genre Menggunakan URL Feature." *Seminar Nasional Teknologi Informasi dan Komunikasi.* Yogyakarta: UGM, 2015.

Pojon, Murat. "Using Machine Learning to Predict Student Performance." *M.Sc. Thesis.* Finland: University of Tampere, 2017.

Potthast, Martin, et al. "Clickbait Detection." *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 16).* The Netherlands: Bauhaus Universität Weimar, 2016.

Ramos, Juan. "Using TF-IDF to Determine Word Relevance in Document Queries." Rutgers University, 2003.

Russell, Matthew A., Mining the Social Web, Second Edition. O'Reilly Media, Inc, USA, 2014.

Sharma, Priyanka. "Comparative Analysis of Various Decision Tree Classification Algorithms using Weka." *International Journal on Recentand Innovation Trends in Computing and Communication Volume: 3 Issue: 2.* India: Auricle Technologies Pvt. Ltd, 2015.

Suadaa, Lya Hulliyyatus. "Tracking Commuter Train Intrusion Through Twitter Crawling." *Jurnal Aplikasi Statistika dan Komputasi Statistik.* Jakarta: Politeknik Statistika STIS, 2016.

Talwar, Anish and Yogesh Kumar. "Machine Learning: An Artificial Intelligence Methodology." *International Journal of Engineering and Computer Science Volume 2 Issue 12, Dec. 2013.* India: 2013.

Waikato University. "Weka 3: Data Mining Software." Accessed on March 3rd, 2018. https://www.cs.waikato.ac.nz/ml/weka/

Witten, Ian H. Text Mining. New Zealand: Waikato University, 2004.

YouTube. *YouTube Lesson: Video Categories.* Accessed on March 5th, 2018. https://creatoracademy.youtube.com/page/lesson/overview-categories