# TICATE2019

*by* Viny M

**Submission date:** 08-Nov-2020 03:22PM (UTC+0700)
**Submission ID:** 1439414997
**File name:** TICATEVinyScrap.docx (967.65K)
**Word count:** 2120
**Character count:** 11015

# COMMENTS SCRAPING APPLICATION FOR REVIEW YOUTUBE CONTENT

**Viny Christanti M.***, Walda, Tri Sutrisno

Computer Science Department, Faculty of Information Technology, Universitas Tarumanagara

* viny@untar.ac.id

**Abstract**. YouTube is one of social media that allows people to upload, view, and share videos through website. YouTube's mission statement is to give everyone a voice and show them the world. Through YouTube, people can express their voices or show their creativity effortlessly. YouTube provides service for users to give opinions through comment section. The comments of users can be used to decide either the video has good ratings or not. Web Scraping can be used to take the comments of users from comment section. Web Scraping is a method to take some information from website by extracting data through tags in html. By inspecting code the YouTube website, the information of comments is contained in 'ytd-comment-thread-renderer'. Information of comments is divided into reviewer and content of comments. Reviewer's information is contained in id 'author-text' and the content is in 'content-text'.

## 1. Introduction

In the new era of globalization, business's competitions are likely increased, both in national and international. It affects the company to think about how to promote products or services in creative and innovative way. One of the ways to promote products or services is promoting some contents through social media.

YouTube is one of the social media that is popular nowadays. Lembaga riset pasar statista predicts that the users of YouTube will reach around 1.8 billion in 2021 [1]. YouTube's mission statement is to give everyone a voice and show them the world. Through YouTube, people can express their voices or show their creativity effortlessly. People can involve indirectly in YouTube by viewing the videos, giving comments, and sharing the videos to another people. By the advantages of YouTube, company decide to do promotion through YouTube. The promotions are shared by video that has a good content inside.

Company can get the feedback for users by seeing the comments of users in the comment section. The comments can be used to decide either the content of video has good ratings of not. Checking the comments one by one in the comment section will take a long time, so there is a need to scrap all of the comments and use it to do some predictions. Scraping the comments in YouTube can be done by web scraping method. Web Scraping is a method to take some information from website by extracting data through tags in HTML. Web Scraping can be done by inspecting the code of YouTube and see position of the information that will be taken from it. The reviewer and the comments from You Tube will be extracted from web.

Anand, Kedar, and Shweta use web scraping method to mine information from different and unstructured websites and transform it into a comprehensible structure like spreadsheets, database, or

comma-separated values (CSV) file [2]. The data such as item pricing, stock pricing, reports, market pricing, and product details can be used to take effective decisions in business process.

Ahmat, Leon, and Suryayusra use web scraping to make an application for searching scientific article [3]. The application is run by inputting the keywords and the machine will search the information through some portals, such as Garuda, ISJD, and Google Scholar. The machine will show the information if the keywords are matching with the data in database. If the machine did not find the keywords then it will show "the information are not found"

Valdivia et. al. uses web scraping to get hotel reviews in Trip Advisor [4]. But Gunawan et. al. collects data from Tripadvisor and Traveloka manually [5]. The process of retrieving data manually is certainly going to take time. For example how much time does it take for Gunawan et. al. to collect 1400 data from Traveloka and Tripadvisor manually, maybe need one weeks [5]. Where should the core of the research is not to collect documents but to conduct an analysis of hotel reviews. The results of the review analysis can decide the voice of customer materials that comes from reviews or survey responses.

Mahek, Ricky, and Nishil uses web scraping to get reviews of different mobile company such as Apple, Samsung, and Oneplus [6]. The reviews are likely consist of "not Satisfied and satisfied" reviews. The reviews are classified based on keywords, ratings, and emotions that decide the final review score of the products.

In this research we build a system that is used dynamically to scrap and clean text data from the YouTube automatically. The system is equipped with a validation function that can be used as a form to carry out the comment validation process. The result of this application is to prepare comment data for analysis in the opinion mining process.

## 2. Methodology

The method in this research is the waterfall method in the System Development Life Cycle (SDLC) which is one of the methods of making application programs. SDLC consists of planning, system analysis, system design, system development, system testing by researchers and maintenance. The stages used in this research are the stages in text processing from web that pay attention to the web structure in crawling and scraping.

### *2.1. YouTube*

YouTube is a media for uploading videos and sharing information in the form of videos [7]. YouTube provides a commentary column for comments on shared videos. YouTube is often used as a place to market products and get opinions through the number of likes or comments written by readers. In the Figure 1 we can see video about Samsung note 10 and at figure 2 we can see review from people.


**Figure 1.** Video about product

**Figure 2.** Review product from reader

YouTube service provides a column for writing comments that can contain opinions. But sometimes, the word is not an opinion but only meaningless term. In the comments column, other

readers can reply the comment to indicate whether they agree or not, or add other comments. The sentence in this comment can also not be related and not mutually supportive. Language inconsistency, inaccuracy in the form of the comment sentence requires validation from humans so that the data entered for analysis is the right data in the form of comments rather than others

In principle, opinions on YouTube are almost similar to opinions on blogs. Where on the blog there is also a comment column that can be used by users to write opinions about the writings listed by the author. On YouTube the comment service is limited in character length. An opinion can be considered as one large document that contains many contents or as a small document that contain one content [8].

Unlike collections from news sources where each document is written briefly and contains topics related to the writing, a comment post is usually slightly improved and tends to be irregular. In the use of comment words for example the word "sedap" is written as "sedaaap", so that the detection of opinion words becomes more difficult. In addition the comments writing page often contains comments that can extend the writing without adding text related to the topic being discussed. Comments collection is also vulnerable to spam content in the form of spam blogs (splogs) and irrelevant comments [8].

### *2.2. Web HTML Structure*

HTML is an abbreviation of Hypertext Markup Language is a script for compiling Web documents [16]. HTML documents are stored in regular text format and contain tags that instruct the web browser to execute the specified commands. The basic structure of an HTML document can be seen in Figure 3. The structure HTML consists of Tag, Element and Attribute.

```
<html>
<head>
<title>Di sini Judul Dokumen HTML</title>
</head>
<body>
   Disini penulisan  informasi Web
</body>
</html>
```

**Figure 3.** HTML Structure

On YouTube it also consists of existing structures in HTML. In figure 4, we can see comments that are visible on the web. But when the source is seen the form consists of html structure. So we must create application that only takes comments without taking other data that is not needed. Some of comments not only consist of letters but can contain images. The sentence “ya ampun senyumnya nic itu sumpah ya bikin meleleh....😍😍😍😍😍” consist of emoticon so the comment must be cleaned up before it can be processed in opinion mining.

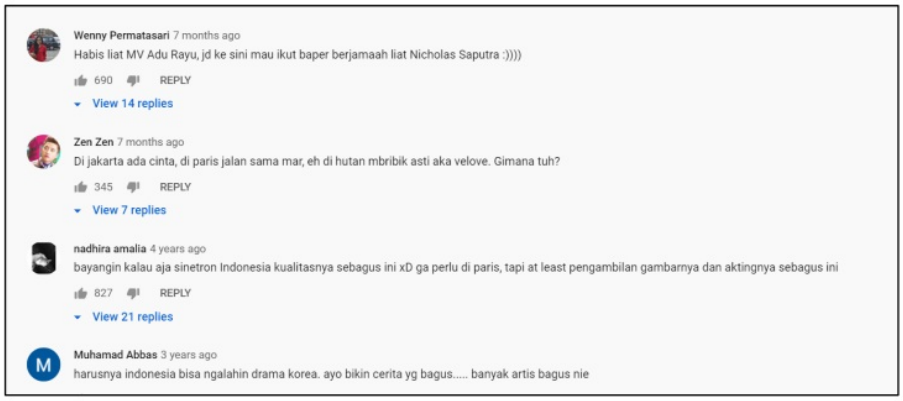



**Figure 4.** Example of YouTube comment

## 3. Result

This research makes a comment scraping program on YouTube by using the PHP and HTML programming languages. This name of this application is "Rukomindo" (Pengeruk Komentar Indonesia) equipped with a validation function. The goal of this application is scrap the comment, clean the comment, validation the comment and prepare comment data for analysis. The program consists of 5 main modules namely scraping module, comment results, validation process, and summary of comments validation.

At figure 5, we can see interface of scraping module. At scraping module, we only copy paste the link of YouTube video that we will scrap. We can see the video and the application will automatically take out the comment. User can arrange what they want to scrap, comment only or comment with all replies.

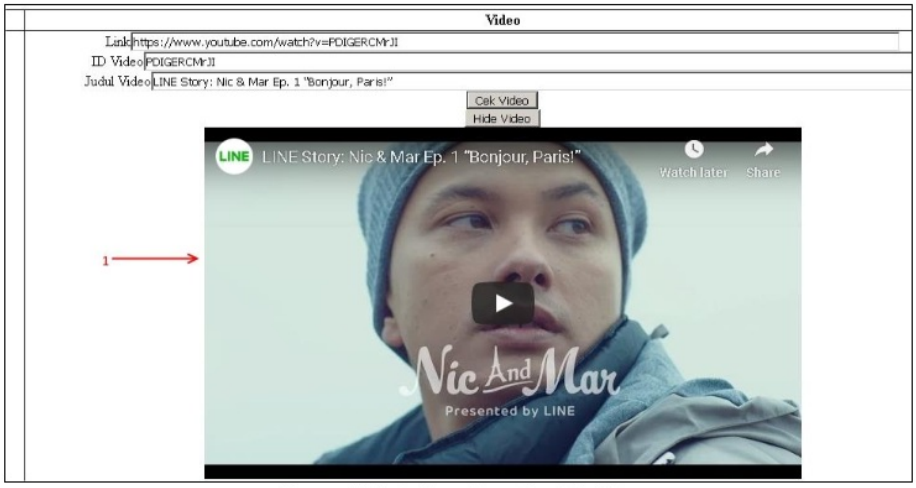



Figure 5. Interface Scraping Module

After put the YouTube link, next step is see the comment result. We can see the comment result at figure 6. There are lists comment that already represent in table format. We can validation the comment and prepare for analysis step. We only save the comment without the name ID who wrote of comment.

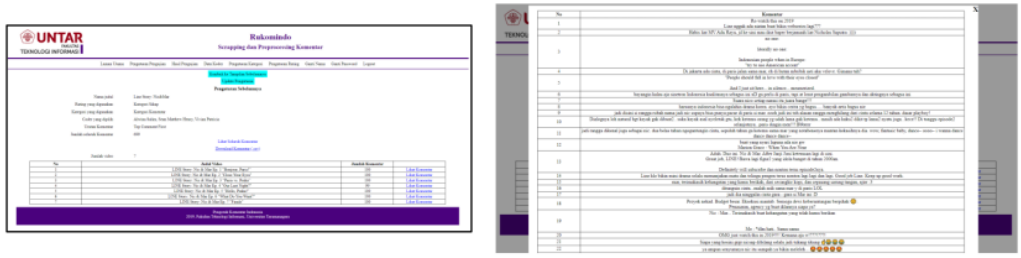
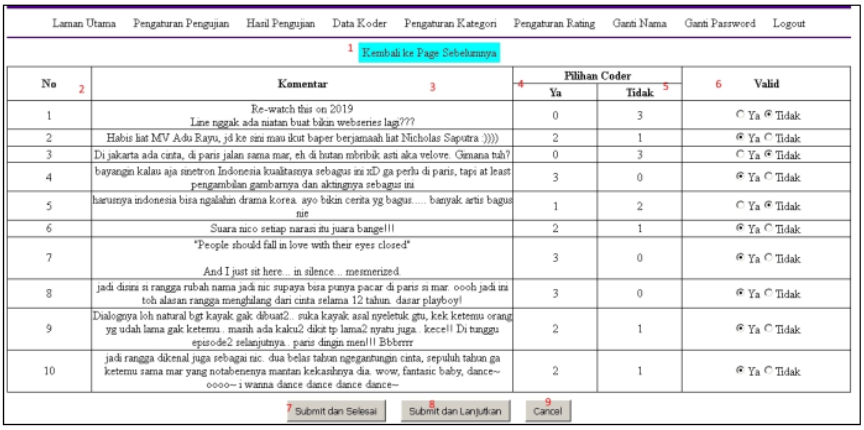


Figure 6. Interface Comment Result

At figure 7 we can see interface of validation process. We can arrange how many validators will give voting for each comment. In the end user will make decision is the comment is opinion or not. User will prepare the data separate the sentence that not consists of statement or opinion. Only opinion will be used for sentiment analysis.



Lamun Utama Pengaturan Pengujian Hasil Pengujian Data Koder Pengaturan Kategori Pengaturan Rating Ganti Nama Ganti Password Logout

1 Kembali ke Page Sebelumnya

| No 2 | Komentar 3 | Pilihan Coder 4 Ya | 5 Tidak | 6 Valid |
|---|---|---|---|---|
| 1 | Re-watch this on 2019<br>Line nggak ada niatan buat bikin webseries lagi??? | 0 | 3 | ○ Ya ◉ Tidak |
| 2 | Habis liat MV Adu Rayu, jd ke sini mau ikut baper berjamaah liat Nicholas Saputra :))))) | 2 | 1 | ◉ Ya ○ Tidak |
| 3 | Di jakarta ada cinta, di paris jalan sama mar, eh di hutan mbribik asti aka velove. Gimana tuh? | 0 | 3 | ○ Ya ◉ Tidak |
| 4 | bayangin kalau aja sinetron Indonesia kualitasnya sebagus ini xD ga perlu di paris, taps at least pengambilan gambarnya dan aktingnya sebagus ini | 3 | 0 | ◉ Ya ○ Tidak |
| 5 | harusnya indonesia bisa ngalahin drama korea. ayo bikin cerita yg bagus..... banyak artis bagus nie | 1 | 2 | ○ Ya ◉ Tidak |
| 6 | Suara nico setiap narasi itu juara bangeⅠⅠⅠ | 2 | 1 | ◉ Ya ○ Tidak |
| 7 | "People should fall in love with their eyes closed"<br><br>And I just sit here... in silence... mesmerized. | 3 | 0 | ◉ Ya ○ Tidak |
| 8 | jadi disini si rangga rubah nama jadi nic supaya bisa punya pacar di paris si mar. oooh jadi ini toh alasan rangga menghalang dari cinta selama 12 tahun. dasar playboy! | 3 | 0 | ◉ Ya ○ Tidak |
| 9 | Dialognya loh natural bgt kayak gak dibuat2.. suka kayak asal nyeletuk gtu, kek ketemu orang yg udah lama gak ketemu.. masih ada kaku2 dikit tp lama2 nyatu juga.. kecee!! Di tunggu episode2 selanjutnya.. paris dingin men!!! Bbberrrr | 2 | 1 | ◉ Ya ○ Tidak |
| 10 | jadi rangga dikenal juga sebagai nic. dua belas tahun nggantungin cinta, sepuluh tahun ga ketemu sama mar yang notabenenya mantan kekasihnya dia. wow, fantastic baby, dance~ oooo~ i wanna dance dance dance dance~ | 2 | 1 | ◉ Ya ○ Tidak |

7 Submit dan Selesai 8 Submit dan Lanjutkan 9 Cancel

Figure 7. Interface validation process

After validation process, we can see at figure 8 the summary of validation process. We can download recapitulation from validation process. So we can prepare the data for another process like sentiment analysis or mining process.

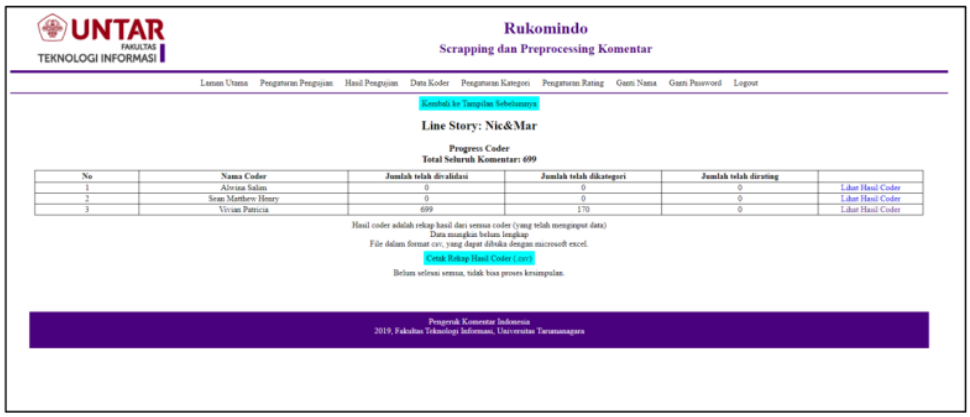



Figure 8. Interface Summary of Comment Validation

## 4. Conclusion

In this research a scraping application has been built to obtain comment data on YouTube. This application can be used to pull comments from a video and provide validation for these comments. So the results of comments that have been collected can be used for mining or further analysis. This application helps users in taking comments from YouTube movies just by filling in the links from the desired YouTube.

This application has been used by the management team who wants to process the film data of Nic and Mar 7 episodes and obtained 699 comments from the video. The research team from the management faculty has also tried to provide validation for the comments in the film.

For further research we will try to create scraping for another social media such as Twitter, Facebook, Instagram and etc. After create scraping, this application must develop for pre-processing step, like separate reply and real comment, separate which negative and positive comment or save the account id to evaluate hoax or non-hoax comment.

## Acknowledgments

We would like to gratitude to DPPM Untar, for funding this research. Thank you to research assistant that carries out this research. And we would like to thank to Dr. Cokki from Economics Faculty Untar who used our application for their research.

## References

[1] Translation from Praditya, Diaz. (2018). 3 Fakta menarik dari Riset Google tentang Perkembangan Youtube di Indonesia, https://id.techinasia.com/fakta-perkembangan-youtube-di-indonesia, Accessed on November 2019, 16th .

[2] Saurkar, Anand V; Pathare, Kedar G; Gode Shweta A. (2018). An Overview On Web Scraping Techniques and Tools. International Journal on Future Revolution in Computer Science & Communication Engineering Volume 4.

[3] Translation from Josi, Ahmat; Abdillah, Leon A.; Suryayusra. (2014). Penerapan Web Scraping pada Mesin Pencari Ilmiah. Bina Darma University.

[4] Valdivia, Ana, M. Victoria Luzón, and Francisco Herrera. "Sentiment analysis on tripadvisor: Are there inconsistencies in user reviews?." In *International Conference on Hybrid Artificial Intelligence Systems*, pp. 15-25. Springer, Cham, 2017.

[5] Gunawan, Eric, Viny Christanti Mawardi, and Naga, S. Dali. " Analisis Sentimen Pada Ulasan Hotel Online Bahasa Indonesia Dengan Support Vector Machine." In *Proceedings of SNTI*

*XV*, ISSN 1829-9156, 2018.

[6] Merchant, Mahek; Shah, Nishil; P.Boominathan. (2016). Sentiment Analysis of Web Scraped Product Reviews using Hadoop. IJRASET Volume 4 Issue VIII

[7] Hopkin Jim. "Surprise! There's a third YouTube co-founder." *USA Today* 11, no. 10 (2006).

[8] Elsas, Jonathan L., Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. "Retrieval and feedback models for blog feed search." In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 347-354. ACM, 2008.

# TICATE2019

ORIGINALITY REPORT

| 13% | 10% | 5% | 11% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to CSU, Chico<br>Student Paper | 2% |
| 2 | www.cs.cmu.edu<br>Internet Source | 2% |
| 3 | www.ijfrcsce.org<br>Internet Source | 2% |
| 4 | ajonksb.blogspot.com<br>Internet Source | 2% |
| 5 | Submitted to Universiti Teknologi MARA<br>Student Paper | 1% |
| 6 | Submitted to London Metropolitan University<br>Student Paper | 1% |
| 7 | digilib.isi.ac.id<br>Internet Source | 1% |
| 8 | M. Viny Christanti, Walda, Tri Sutrisno. "Comments Scraping Application For Review Youtube Content", IOP Conference Series: Materials Science and Engineering, 2020<br>Publication | 1% |

| | | |
|---|---|---|
| 9 | www.encyclopedia.com<br>Internet Source | 1% |
| 10 | hrmars.com<br>Internet Source | 1% |
| 11 | Chairisni Lubis, Felicia Gondawijaya. "Heart Sound Diagnose System with BFCC, MFCC, and Backpropagation Neural Network", IOP Conference Series: Materials Science and Engineering, 2019<br>Publication | 1% |
| 12 | "Hybrid Artificial Intelligent Systems", Springer Science and Business Media LLC, 2017<br>Publication | <1% |
| 13 | "Systems development life cycle (SDLC)", Salem Press Encyclopedia of Science, 2016<br>Publication | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | Off | | |