Available online at JDR Website: http://journal.unublitar.ac.id/jdr

Journal Of Development Research, 5 (2), November 2021, Pages 94-105

DOI: https://doi.org/10.28926/jdr.v5i2.149

Implementation of EM Algorithm in Opinion Mining Movies Review Case Studies

[(1)]Muhammad Danial Romadloni, [(2)]Indra Gita Anugrah

Universitas Muhammadiyah Gresik, Indonesia

E-mail: [(1)]danial_170602@umg.ac.id, [(2)]indragitaanugrah@umg.ac.id

Received: 14 May 2021; Revised: 17 October 2021; Accepted: 15 November 2021

Abstract

One of the websites that often used to review movies we had watched is IMDb. That data review can be used for opinion mining, whether the movie being reviewed was good or not. One of the algorithms that are often used for opinion mining is Naïve Bayes, apart from being easy to implement and used, Naïve Bayes is also known to be very fast to predict classes on a test dataset. The purpose of this study is to see how much influence the Expectation-Maximization to increase accuracy on opinion mining movies review case studies. From the results of this study uses 878 training datasets that consist of the half with the positive label and another half with the negative label from IMDb's movie reviews and 577 as testing datasets. Accuracy that Naïve Bayes got is 67% but after the uses of EM as the alternative algorithm of Naïve Bayes, the accuracy that EM got is 71%. Using the EM method, it was found that the accuracy increased by 4% compared to using Naïve Bayes.

Keywords: Opinion mining; Sentiment analysis; Expectation-Maximization; Naïve Bayes;

Introduction

Opinion mining becomes one of the most widely adopted topics by researchers. This is mainly supported by the easier it is to get datasets from the internet and if we talk about the internet, we could imagine how big the data is. Especially the data reviews for movie domain which one of the common websites that holds this kind of datasets is IMDb. In IMDb, one could see that all the reviews written by the users, not only movie title that released recently but also from years ago. Thus, confirms that how easy it is to get the datasets for this particular domain which is movie domain for opinion mining. The general algorithm that used for opinion mining and sentiment analysis is Naïve Bayes. There are three main components of opinion mining and analysis methods such as pre-processing, feature extraction and classification (Kulkarni & Rodd, 2018. Many research that uses Naïve Bayes as classifier algorithm for instances (Vijay et al., 2020 for product review data from amazon, (Harahap et al., 2019 for predicting purchase, (Ilham Esa Tiffani, 2020 for hotel review, (Pugsee & Chatchaithanawat, 2020 for laptop reviews and (Pugsee et al., 2019 skin care products on twitter, also (Poovaraghan et al., 2019 and (Granik & Mesyura, 2017 for fake news detection applications. Another research that similar (with this study in movie domain) although not identical (Novendri et al., 2020 uses data from YouTube comments also using Naïve Bayes for the classification. All the previous studies that the author mentions share the same though that Naïve bayes do the job well. Unfortunately, most of the data from the internet are unstructured. In (Kulkarni & Rodd, 2018 the author states that the accurate identification of content features gathered from the unstructured textual data is major research challenge. Therefore, it is a must to design an efficient technique which will be able to identify such features from the unstructured datasets (Kulkarni & Rodd, 2018. Therefore, an alternative or at least an optimization of the current algorithm is needed in some applications. As addition, in (Zhou et al., 2020 also states that enhancing the accuracy of sentiment analysis is still an essential issue in product evaluation. Several researches had used Expectation Optimization (EM) Algorithm to make some improvements in some applications, such as in (Zhou et al., 2020 uses Bayesian-based model to analyze the comments and get the inference of user comments. Then use EM algorithm to infer the evaluation of the product (Zhou et al., 2020. Also in pattern recognition world, in (Asheri et al., 2021 a coordinate-descent EM Algorithm was used for estimating the parameters of flexibly tied GMMs that outperform basic GMM and kmeans algorithm. From several previous research it can be interpreted that EM algorithm could be used to improve the performance of Bayesian in some aspect including the accuracy. Therefore, in this study, an EM algorithm is used to analyze movie reviews and whether it could perform better than Bayesian in term of accuracy in predicting positive or negative reviews.

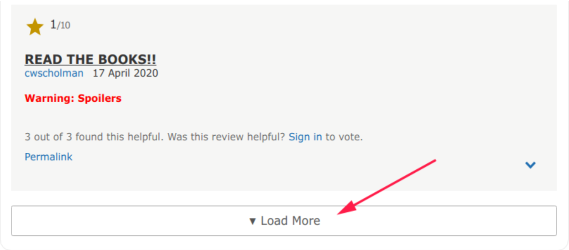
Material and Method

The methodology that the author uses is EM by utilizing the results of the Naïve Bayes calculation as a reference or basis for the E-step. The steps that the author takes in this study have several steps, starting from collecting material or data review from the IMDb website to testing and evaluating the accuracy results of the Naïve Bayes algorithm and EM. The following are the materials and tools that the author uses in this study, followed by a detailed description of the steps the author has taken.

Materials and Tools

The programming language that I use is python by using several libraries such as PySastrawi, beautifulsoup4, and pandas for processing datasets. Selenium webdriver to help the author get 100 reviews for each movie title because by default when opening the IMDb review page, it only loads 25 reviews, with the help of Selenium the author can mimic the mouse click on the 'Load More' button at the bottom of the review page.

Figure 1 shows the 'Load More' button which loads 25 additional reviews so that when the author is scraping, the author has to click the button three times which will make the page load 100 reviews.


★ 1/10

READ THE BOOKS!!

cwscholman 17 April 2020

Warning: Spoilers

3 out of 3 found this helpful. Was this review helpful? Sign in to vote.

Permalink

▾ Load More

Figure 1 Load More Button

Crawling or Scraping

At this step, the author collected data by scraping data reviews from the IMDb website. Starting by choosing ten movie titles randomly, where the ten titles have their respective titleid and each title is taken as many as 100 data reviews so that the total data review obtained is 1000 data. This data review retrieval also includes labeling for each data review. The criteria that the author applies is if the rating that the reviewer gives is 1 to 5 it will be labeled as negative while if it is 6 to 10 it will be labeled as positive. Scraping the 1000 data reviewed above, the author saves the scraped data using Comma-separated value (CSV) format. The reason the author chose the CSV format is that the Pandas library that the author use has very good support (built-in function) in processing data in the CSV format. The following is the template that the author uses to store the data review set and its labels.

Table 1 shows that there are three labels on the y-axis or column, namely the review contains data review while the label contains a label whether the data review is labeled positive or negative, if it's negative then will be labeled as N and labeled P if it is positive. The username is useful for the author when translating in the next step.

Translate

The author uses google translate to translate all the reviews. Of course, translating 1000 data reviews one by one will be quite stressful, therefore by utilizing the programming language The

Table 1 Template Dataset

username	Review	label
username1	Data review ke-1	P
username2	Data review ke-2	N
username-n	Data review ke-n	…

Almighty Python coupled with a toolkit called Selenium Webdriver, the author can translate 1000 data reviews with ease. There is one more problem, namely, in google translate itself, the maximum number of characters that can be translated at one time is 5000 characters. As expected, 5000 characters are not enough for some reviews. Before translating, the author calculates each review whether there is a review that has some characters that exceed 5000. The following is a snippet of source code that the author uses.

Table 2 Source Code

```
src_path: str = "/path/to/dataset/reviews.csv"
dataframe = pd.read_csv(src_path)
longest_char: int = 0
more_than_5000: int = 0
for _, df in dataframe.iterrows():
    review: str = df.get("review")
    if len(review) > 5000:
        more_than_5000 += 1
    if len(review) > longest_char:
        longest_char = len(review)
print(longest_char)
print(more_than_5000)
```

Table 2 shows source code above provides two pieces of information, namely the longest number of characters in the review as many as 9456 characters, and the number of reviews with more than 5000 characters as many as 8 reviews. This information shows that the previous author's assumption that the maximum number of characters allowed by google translate is 5000 characters is insufficient for some data reviews, thus making the author make a ploy so that the data review with a total of more than 5000 characters still can be translated. The strategy that the author does is to break the review string into batches consisting of words that are no more than 2300 characters, then translate them by batch per batch so that one batch of translation will not exceed the google translate limit of 5000 characters

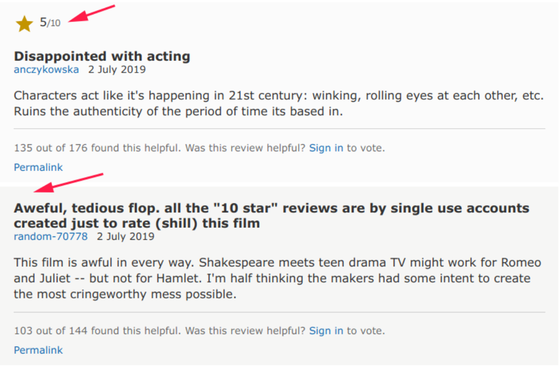
Join All Reviews

All translated data review is ready to be combined after the previous step the author did with a separate dataset based on each titleid, so it needs to be joined into one dataset with a total of 1000 reviews.

Cleaning NULL Label

Starting with randomly choosing ten titleid. Figure 2 shows that not all reviews have a rating which makes the label column in the dataset labeled NULL as shown in figure 3. Since the data review does not have a P or N label, the author decided not to use the data and delete the data labeled NULL from the dataset. There are also some of the data reviews that have a format that is not suitable as shown in figure 4 after the translation step.

Figure 2. Review without Rating



★ 5/10

Disappointed with acting
anczykowska 2 July 2019

Characters act like it's happening in 21st century: winking, rolling eyes at each other, etc. Ruins the authenticity of the period of time its based in.

135 out of 176 found this helpful. Was this review helpful? Sign in to vote.

Permalink

Aweful, tedious flop. all the "10 star" reviews are by single use accounts created just to rate (shill) this film
random-70778 2 July 2019

This film is awful in every way. Shakespeare meets teen drama TV might work for Romeo and Juliet -- but not for Hamlet. I'm half thinking the makers had some intent to create the most cringeworthy mess possible.

103 out of 144 found this helpful. Was this review helpful? Sign in to vote.

Permalink

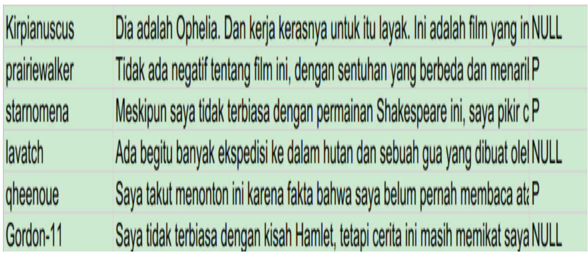
Case Folding

This step is quite simple, which is to transform all the data review's strings to lowercase.

Stemming

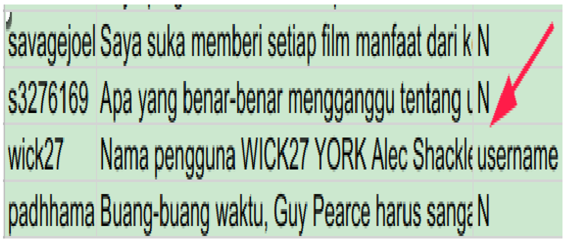
The last step is stemming. Stemming is where the process of mapping words in affixed sentences be the original word (without prefix, suffix, insertion, combination) which is executed by specific algorithm (Rofiqi et al., 2019 which is Naïve Bayes and EM algorithm in this case. Using the Pysastrawi library which is a python implementation of the Sastrawi library using PHP. Stemming makes all existing words reviewed into basic words as shown in Table 3.

Figure 3 Dataset with NULL Label



Kirpianuscus	Dia adalah Ophelia. Dan kerja kerasnya untuk itu layak. Ini adalah film yang in	NULL
prairiewalker	Tidak ada negatif tentang film ini, dengan sentuhan yang berbeda dan menaril	P
starnomena	Meskipun saya tidak terbiasa dengan permainan Shakespeare ini, saya pikir c	P
lavatch	Ada begitu banyak ekspedisi ke dalam hutan dan sebuah gua yang dibuat olel	NULL
qheenoue	Saya takut menonton ini karena fakta bahwa saya belum pernah membaca at	P
Gordon-11	Saya tidak terbiasa dengan kisah Hamlet, tetapi cerita ini masih memikat saya	NULL

Figure 4 Dataset with Unsuitable Format



savagejoel	Saya suka memberi setiap film manfaat dari k	N
s3276169	Apa yang benar-benar mengganggu tentang ı	N
wick27	Nama pengguna WICK27 YORK Alec Shackl	username
padhhama	Buang-buang waktu, Guy Pearce harus sang	N

Table 3 Stemming Data Review

before	after
kirim ini ke tempat sampah! movie penghancur diri ini.	*kirim sampah movie hancur*

This stemming step not only makes all words into basic words but also eliminates stopwords such as symbols, hyphens, and numbers.

Testing and Evaluation

This step where the author performs data training and data testing with both algorithms Naïve Bayes and Expectation-Maximization. Evaluation comes later, and compares the resulting accuracy. Additionally, the author also counts the time that is needed by both algorithms starting from training to testing the dataset.

The author uses the same template as the training dataset for the testing dataset which then iterates over every row in the dataset and compares if the resulting of the performed testing by algorithm does not the same as the label in the same row as the review, then it will be regarded as missed and otherwise as correct. Besides missed and correct, the author also found the third condition which the author calls unknown. Unknown is the condition when the results of the testing performed by the algorithm reach zero and couldn't be calculated any further which either positive or negative get zero and that's why it is unknown. The author decided that the unknown result will be counted as missed.

Naïve Bayes

The Naïve Bayes algorithm is a classification method based on odds by calculating the total odds based on the number of trials with a combination of values that appear. Another definition of the Naïve Bayes algorithm is statistical classification that can be used to predict the probability of membership of a class. Bayesian classification is based on Bayes' theorem which has similar classification capabilities with decision trees and neural networks (Annur, 2018. The Naïve Bayes algorithm assumes that all attributes have no dependence on the value of each class variable (Aggarwal & Zhai, 2012, or it can be said that the final probability value is the result of the initial probability which is a collection of individual probabilities (Anugrah & Rosyid, 2019. The author uses equation from (Anugrah & Rosyid, 2019 for the Naïve Bayes algorithm in this research. Equation (1) shows the calculation of this algorithm.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

W $$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$ here:

- X is samp X le data from class (label) which is not yet known,
- H is the H predicted data from the class that already known,
- $P(H)$ is the $P(H)$ probability value of, H
- $P(X)$ is t H he $P(X)$ probability value of the data sample observed,
- $P(X|H)$ is the $P(X|H)$ sample probability of X data, a X ssuming that the guess is correct.

Expectation-Maximization

EM mainly used by Probabilistic Latent Semantic Analysis (PLSA) to determined their parameters. In topic modeling using PLSA is able to create document contexts that are differentiated based on words that have many meanings by grouping words based on topics (Revindasari et al., 2017). The EM algorithm is designed to maximize the likelihood ln p ($\mathbf{X}$, $\mathbf{Z}|\boldsymbol{\theta}$) by two steps, i.e., the E step and the M step. In the E step, we use the current parameters $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of $\mathbf{Z}$ given by p ($\mathbf{X}$, $\mathbf{Z}|\boldsymbol{\theta}$). Then we use the posterior distribution to find the expectation of the complete data likelihood Q ($\boldsymbol{\theta}\cdot\boldsymbol{\theta}^{\text{old}}$) (Li et al., 2019).

The EM algorithm is an algorithm for estimating a parameter, the initial stage is to search for the expectation value. After the expectation stage, it is continued with the stage of updating the parameter values using the Maximization algorithm. In the process of finding the expected value using the Naive Bayes algorithm, this algorithm is a class prediction method from data proposed by Thomas Bayes. In this research the author uses equation from (Anugrah & Sarno, 2017) for the EM calculation. EM algorithm consists of two phases :

1. E Step, this phase is used to find the ap-

proximate value of the probability of the topics in the document originating from the initial probability value (Anugrah & Sarno, 2017). Equation (2) shows this step's calculation.

$$P(c_i|d_j) = \frac{p(c_i)\prod_{k_j}^{|d_j|} p(c_i)}{\sum_{r=1}^{|c|} p(c_i)\prod_{k_j}^{|d_j|} p(c_i)} \quad (2)$$

Where:

$p(c_i)$ is probability of category c_i,

$P(c_i|d_j)$ is probability of category c_i in document d_j,

$p(c_i)\prod_{k_j}^{|d_j|} p(c_i)$ is probability of category c_i in term k_j to document d_j.

2. Maximization Step (M Step), this phase is used to update the probability value so get the maximum probability value (Anugrah & Sarno, 2017). Equation (3) shows this step's calculation.

$$P(w_{kj}|c_i) = \frac{\sum_{r=1}^{|D|} N(w_{kj}.d_j)p(c_i|d_j)}{\sum_{s=1}^{|w|}\sum_{r=1}^{|D|} N(w_{kj}.d_j)p(c_i|d_j)} \quad (3)$$

Where:

$N(w_{kj}.d_j)$ is number of words W_k in the document d_j,

$|w|$ is total number of words / features used,

$N(w_{kj}.d_j)$ is total number of training documents.

Results and Discussion

Scraping

Starting with randomly choosing ten titleids as data training and five titleids as data test results in 1000 data review and 500 data review which for data training and data testing respectively. By default, IMDb only shows 25 data reviews on every page, so the author needs to mimic a mouse click that makes the review page load more 25 data reviews for every click. The author needs to click the same load more button three times so the final result will be 100 data reviews.

Figure 5 shows that chrome noticed the browser being controlled by some 'test software' which is a selenium webdriver. After getting 100 data reviews, the author saves that data to CSV format according to the template in Table 1 which makes a total 15 datasets with CSV format, 10 datasets for data training and 5 datasets for testing. Because from initial number dataset to ready-to-use dataset might differ greatly so the author keeps tracking the number for every step, Table 4 shows currently the dataset's number. The result of scraping with titleid tt5689068 is shown in Table 5 that is filled based on template in Table 1 except for the No column is only to show how many data review per titleid that got scraped.

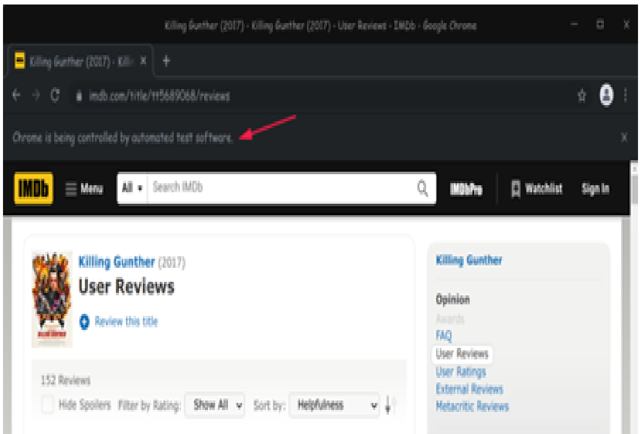
Translate

This is quite a tricky one because of the limited number of characters allowed by google translate. Figure 6 shows how the author translates using google translate via Selenium Web-

Table 4 The Current Amount of Dataset

titleid	1# scraping	type
tt5689068	1000	Data training
...		
tt5717194		
tt6148156	500	Data testing
...		
tt5813916		

Figure 5 Scraping with Selenium Webdriver



Chrome is being controlled by automated test software.

Table 5 Scraped Review With titleid tt5689068

No	username	review	label
1	bekariiii	I probably never laughed more in my life except Kung Fury Of course it's trash, but it's very funny trash, once in a while you need this kind of movie, not serious, funny, which is made just to make you lough. Can't understand why are you people rating it so low.	P
2	matze-23967	It's what you can call a marmite movie, you either love it or hate it. Best to watch with some friends, having some nibbles and beers. I felt entertained and had a good time, never expected it to be an Oscar contestant	P
…	…	…	…
…	…	…	…
99	Draadityabajaj	A very different action comedy, also has elements of 'found footage'. I was surprised by the poor rating. I thoroughly enjoyed the movie, although Arnold arrives a bit late in the picture. Please watch with an open mind and don't pay heed to the reviews.	P
100	Brunovanael	In a comedy I expect to laugh, or have a smile on my face at least once. May be a matter of taste but for me this movie wasn't funny or fun at all.	N

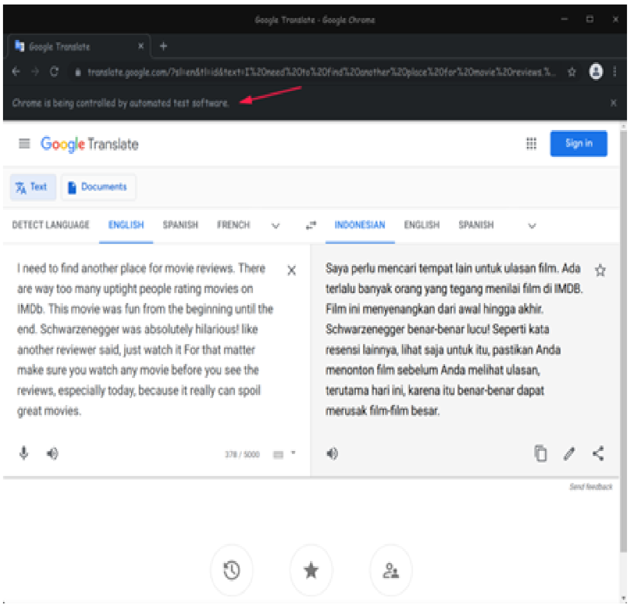
driver. Just like the previous step, the author also uses selenium to translate every single data review in every dataset. Because this step only translates data in column review which contain the actual review. The template and file format to save the data will be the same as before as shown in Table 1.

For some reason that the author is not well aware of, the dataset's number decreased as shown in Table 6. Table 7 shows the result of the translated dataset with titleid tt5689068.

Join All Reviews

The author decides to combine all reviews or datasets into two datasets in this step, obviously for training and testing to make the further step a lot easier to perform. How the author handles the combined datasets filename is shown in Table

Figure 6 Translating to Indonesian



Google Translate - Google Chrome
Google Translate
translate.google.com/?sl=en&tl=id&text=I%20need%20to%20find%20another%20place%20for%20movie%20reviews.%...
Chrome is being controlled by automated test software.
Google Translate
Sign in
Text
Documents
DETECT LANGUAGE
ENGLISH
SPANISH
FRENCH
INDONESIAN
ENGLISH
SPANISH
378 / 5000
Send feedback

I need to find another place for movie reviews. There are way too many uptight people rating movies on IMDb. This movie was fun from the beginning until the end. Schwarzenegger was absolutely hilarious! like another reviewer said, just watch it For that matter make sure you watch any movie before you see the reviews, especially today, because it really can spoil great movies.

Saya perlu mencari tempat lain untuk ulasan film. Ada terlalu banyak orang yang tegang menilai film di IMDB. Film ini menyenangkan dari awal hingga akhir. Schwarzenegger benar-benar lucu! Seperti kata resensi lainnya, lihat saja untuk itu, pastikan Anda menonton film sebelum Anda melihat ulasan, terutama hari ini, karena itu benar-benar dapat merusak film-film besar.

Table 6 The Current Amount of Dataset

titleid	2# translate	type
tt5689068	997	Data training
…		
tt5717194		
tt6148156	449	Data testing
…		
tt5813916		

Table 7 Translated Dataset With titleid tt5689068

No	username	review	label
1	bekariiii	*Saya mungkin tidak pernah tertawa lebih dalam hidup saya kecuali Kung Fury tentu saja sampah, tapi itu sampah yang sangat lucu, sesekali Anda membutuhkan movie seperti ini, tidak serius, lucu, yang dibuat hanya untuk membuat Anda lough. can'tPahami mengapa kalian memberi peringkat begitu rendah.*	P
2	matze-23967	*Itulah yang dapat Anda sebut movie Marmite, Anda menyukainya atau membencinya. Terbaik untuk ditonton dengan beberapa teman, memiliki beberapa camilan dan bir. Saya merasa terhibur dan bersenang-senang, tidak pernah mengharapkannya menjadi kontestan Oscar*	P
...	...	...	…
...	...	...	...
99	Draadityabajaj	*Komedi aksi yang sangat berbeda, juga memiliki elemen 'found facage'. Saya terkejut dengan peringkat yang buruk. Saya benar-benar menikmati movie itu, meskipun Arnold tiba agak terlambat dalam gambar. Harap tonton dengan pikiran terbuka dan jangan bayar Head to tinjauan.*	P
100	brunovanael	*Dalam sebuah komedi saya berharap untuk tertawa, atau tersenyum di wajah saya setidaknya sekali. Mungkin masalah selain bagi saya movie ini tidak lucu atau menyenangkan sama sekali.*	N

8. As the author already mentioned earlier that the username column is only needed when translating so it's no longer needed and the author removes the column in this step.

Table 8 Combine Dataset

filename	combined filename	type
tt5689068.csv	reviews.csv	Data training
...		
tt5717194.csv		
tt6148156.csv	datatest.csv	Data testing
...		
tt5813916.csv		

Since this step only combines datasets and does not do any modifications to the datasets itself that affect the dataset's number, the dataset's number will be the same as shown in Table 6.

Cleaning NULL Label

This will be the step that greatly reduces the dataset's number. Figure 2 and 4 shows the reasons why cleaning is needed and the cleaning, in this case, means removing every single data review that has either the NULL label or unsuitable format with the template in Table 1. Manually labeling tens of data reviews is quite troublesome so the author decides to remove it from the dataset and resulting in the dataset's number decreased as shown in Table 9.

From Table 10 that contains the original training dataset before cleaned and Table 11 training dataset after getting cleaned we could see that the last two data reviews shifted up because 24 data reviews either had a NULL label or unsuited format removed from the dataset. This also applied to the testing dataset.

Case Folding

This step will not change the dataset's number since it only modifies the reviews in

Table 9 The Current Amount of Dataset

titleid	4# cleaning	type
tt5689068	973	Data training
...		
tt5717194		
tt6148156	482	Data testing
...		
tt5813916		

Table 10 Original Training Dataset

No	review	label
1	*Saya perlu mencari tempat lain untuk ulasan movie. Ada terlalu banyak orang yang tegang menilai movie di IMDB. movie ini menyenangkan dari awal hingga akhir. Schwarzenegger benar-benar lucu!Seperti kata resensi lainnya, lihat saja untuk itu, pastikan Anda menonton movie sebelum Anda melihat ulasan, terutama hari ini, karena itu benar-benar dapat merusak movie-movie besar.*	P
2	*Ketika saya dan pacar saya melihat peringkat 4. 7 di sini, harapan kami di mana rendah tetapi kami mencobanya karena arnold. Dan betapa senangnya kami, kami melakukannya!Ini menunjukkan lagi bahwa Anda tidak dapat memberikan skor IMDB terlalu berat. Batu movie ini!Ini menghibur, lucu sekali dan langkahnya bagus. Satu-satunya kekecewaan adalah waktu layar pendek Arnold. Abaikan skornya!Tonton saja!*	P
205	*movie ini mengerikan dalam segala hal. Shakespeare Meets Remaja Drama TV mungkin bekerja untuk Romeo dan Juliet - tetapi tidak untuk Hamlet. Saya setengah berpikir bahwa pembuat memiliki niat untuk menciptakan kekacauan yang paling berani.*	NULL
...	...	…
995	*Aaron Stanford bukan bagian dari para pemeran movie ini, nama aktor adalah Aaron Swafford, bukan hal yang sama. Lihatlah.*	NULL
996	*Adegan yang buruk berasal dari naskah yang buruk. Skenario bodoh. Saya terkejut, produksi movie ini adalah 2020. Dumb Marshal bertemu perampok bisu. movie terburuk yang pernah saya lihat !!!Terburuk dari setiap aspek sinematografi :(*	P
997	*Aktingnya mengerikan dan banyak ketidakkonsistenan. Saya tidak tahu apa-apa tentang membuat movie tetapi bahkan saya bisa melakukan pekerjaan yang lebih baik.*	N

Table 11 Cleaned Training Dataset

No	review	label
1	*Saya perlu mencari tempat lain untuk ulasan movie. Ada terlalu banyak orang yang tegang menilai movie di IMDB. movie ini menyenangkan dari awal hingga akhir. Schwarzenegger benar-benar lucu! Seperti kata resensi lainnya, lihat saja untuk itu, pastikan Anda menonton movie sebelum Anda melihat ulasan, terutama hari ini, karena itu benar-benar dapat merusak movie-movie besar.*	P
2	*Ketika saya dan pacar saya melihat peringkat 4. 7 di sini, harapan kami di mana rendah tetapi kami mencobanya karena arnold. Dan betapa senangnya kami, kami melakukannya!Ini menunjukkan lagi bahwa Anda tidak dapat memberikan skor IMDB terlalu berat. Batu movie ini!Ini menghibur, lucu sekali dan langkahnya bagus. Satu-satunya kekecewaan adalah waktu layar pendek Arnold. Abaikan skornya!Tonton saja!*	P
...	...	...
...	...	…
972	*Adegan yang buruk berasal dari naskah yang buruk. Skenario bodoh. Saya terkejut, produksi movie ini adalah 2020. Dumb Marshal bertemu perampok bisu. movie terburuk yang pernah saya lihat !!! Terburuk dari setiap aspek sinematografi :(*	P
973	*Aktingnya mengerikan dan banyak ketidakkonsistenan. Saya tidak tahu apa-apa tentang membuat movie tetapi bahkan saya bisa melakukan pekerjaan yang lebih baik.*	N

every single row to lowercase.

The snippet of source code above shows how convenient it is to use the Pandas library since it has very good support in handling CSV format. The author only needs to prepare a function to do the convert to lowercase, read CSV format, inject the prepared function to the converters parameter and Pandas will do the rest for the author to convert every single data review in the dataset, then save the already converted Dataframe to new CSV file. Table 12 shows case foldsing step. Table 13 shows the result of this step.

Stemming

After this step is done, the datasets will be ready for training and testing then evaluate the accuracy. This step consists of some steps, besides the stemming itself and stop word removal, it also has a step that distinct data review that has an either positive or negative label. The author needs equal numbers of a positive and negative

Table 12 Case Folding Step

```
src_path: str = "/path/to/src/reviews.csv"
dst_path: str = "/path/to/dst/reviews.csv"

def to_lower_case(string: str) -> str:
   return string.lower()

dataset    =    pd.read_csv(src_path,    converters=
      {"review": to_lower_case}
)
dataset.to_csv(dst_path, index=False)
```

Table 13 Case Folding Training Dataset

No	before case folding	after case folding	label
1	*Saya perlu mencari tempat lain untuk ulasan movie. Ada terlalu banyak orang yang tegang menilai movie di IMDB. movie ini menyenangkan dari awal hingga akhir. Schwarzenegger benar-benar lucu!Seperti kata resensi lainnya, lihat saja untuk itu, pastikan Anda menonton movie sebelum Anda melihat ulasan, terutama hari ini, karena itu benar-benar dapat merusak movie-movie besar.*	*saya perlu mencari tempat lain untuk ulasan movie. ada terlalu banyak orang yang tegang menilai movie di imdb. movie ini menyenangkan dari awal hingga akhir. schwarzenegger benar-benar lucu!seperti kata resensi lainnya, lihat saja untuk itu, pastikan anda menonton movie sebelum anda melihat ulasan, terutama hari ini, karena itu benar-benar dapat merusak movie-movie besar.*	P
2	*Ketika saya dan pacar saya melihat peringkat 4. 7 di sini, harapan kami di mana rendah tetapi kami mencobanya karena arnold. Dan betapa senangnya kami, kami melakukannya!Ini menunjukkan lagi bahwa Anda tidak dapat memberikan skor IMDB terlalu berat. Batu movie ini!Ini menghibur, lucu sekali dan langkahnya bagus. Satu-satunya kekecewaan adalah waktu layar pendek Arnold. Abaikan skornya!Tonton saja!*	*ketika saya dan pacar saya melihat peringkat 4. 7 di sini, harapan kami di mana rendah tetapi kami mencobanya karena arnold. dan betapa senangnya kami, kami melakukannya!ini menunjukkan lagi bahwa anda tidak dapat memberikan skor imdb terlalu berat. batu movie ini!ini menghibur, lucu sekali dan langkahnya bagus. satu-satunya kekecewaan adalah waktu layar pendek arnold. abaikan skornya!tonton saja!*	P
...	...		...
...	...		…
972	*Adegan yang buruk berasal dari naskah yang buruk. Skenario bodoh. Saya terkejut, produksi movie ini adalah 2020. Dumb Marshal bertemu perampok bisu. movie terburuk yang pernah saya lihat !!!Terburuk dari setiap aspek sinematografi : (*	*adegan yang buruk berasal dari naskah yang buruk. skenario bodoh. saya terkejut, produksi movie ini adalah 2020. dumb marshal bertemu perampok bisu. movie terburuk yang pernah saya lihat !!!terburuk dari setiap aspek sinematografi : (*	P
973	*Aktingnya mengerikan dan banyak ketidakkonsistenan. Saya tidak tahu apa-apa tentang membuat movie tetapi bahkan saya bisa melakukan pekerjaan yang lebih baik.*	*aktingnya mengerikan dan banyak ketidakkonsistenan. saya tidak tahu apa-apa tentang membuat movie tetapi bahkan saya bisa melakukan pekerjaan yang lebih baik.*	N

label, which means a positive label must has the same number as a negative label. Although the training dataset has 973 data as shown in Table 6, it does not necessarily have an equal number of positive and negative. The author checks the dataset, it consists of 439 negative labels so the author needs to balance the number that makes the remaining 95 needed to be separated from the main training dataset. The author decides to join the remaining data from the training dataset to the testing dataset which makes the testing dataset's number increasing as shown in Table 14. As for the result of stemming training dataset is shown in Table 15.

Testing and Evaluation

It's time for the main course, datasets are ready and this is where the performance of both Naïve Bayes and EM algorithm is compared. The author uses 878 training datasets with a half

Table 14 Stemming Training Dataset

type	before split	splitting	after split	type
Data training	973	878	878	Data training
		95	577	Data testing
Data testing	482	482		

positive and half negative label, 577 as testing dataset. Besides 577 as the maximum number of testing datasets, the author also uses another range of a number of the testing dataset to see whether EM consistently has better accuracy than Naïve Bayes.

Instead of confusion matrix, because of the immense number of the data, the author uses the simple way, if the predicted result has the same label as the actual result, then the author count this as correct, maybe it's called True Positive and True Negative in the confusion matrix, and if the predicted result has a different label as an actual result, then it's counted as missed. There is another result besides correct and missed, the author calls this result unknown. The unknown is the condition when the predicted result reaches zero and couldn't be calculated any further which either positive or negative get zero and that's why it is unknown. The author counts this unknown as missed.

Table 16, 17, 18, 19 and 20 shows that the author uses different numbers of dataset for testing. How the author calculates the final accuracy is shown in (4).

$$Accuracy = \frac{Correct}{Correct+Missed+Unkown} \ x\ 100 \qquad (4)$$

Table 16, 17, 18, 19, and 20 are clearly shows that EM consistently has accuracy better than Naïve Bayes, although the larger the testing dataset is, the lower the EM's accuracy, but still, it consistently has accuracy more than 70% meanwhile Naïve Bayes accuracy never reach 70%

Table 15 Stemming Training Dataset

No	before stemming	after stemming	label
1	*Saya perlu mencari tempat lain untuk ulasan movie. Ada terlalu banyak orang yang tegang menilai movie di IMDB. movie ini menyenangkan dari awal hingga akhir. Schwarzenegger benar-benar lucu!Seperti kata resensi lainnya, lihat saja untuk itu, pastikan Anda menonton movie sebelum Anda melihat ulasan, terutama hari ini, karena itu benar-benar dapat merusak movie-movie besar.*	*cari ulas movie tegang nilai movie imdb movie senang schwarzenegger benarbenar lucuseperti resensi pasti tonton movie ulas benarbenar rusak moviemovie*	P
2	*Ketika saya dan pacar saya melihat peringkat 4. 7 di sini, harapan kami di mana rendah tetapi kami mencobanya karena arnold. Dan betapa senangnya kami, kami melakukannya!Ini menunjukkan lagi bahwa Anda tidak dapat memberikan skor IMDB terlalu berat. Batu movie ini! Ini menghibur, lucu sekali dan langkahnya bagus. Satu-satunya kekecewaan adalah waktu layar pendek Arnold. Abaikan skornya!Tonton saja!*	*pacar peringkat 4 7 harap rendah coba arnold betapa senang melakukannyaini skor imdb berat batu movie iniini hibur lucu langkah bagus satusatunya kecewa layar pendek arnold abai skornyatonton*	P
...	...		...
...	...		…
972	*Adegan yang buruk berasal dari naskah yang buruk. Skenario bodoh. Saya terkejut, produksi movie ini adalah 2020. Dumb Marshal bertemu perampok bisu. movie terburuk yang pernah saya lihat !!!Terburuk dari setiap aspek sinematografi :(*	*adegan buruk asal naskah buruk skenario bodoh kejut produksi movie 2020 dumb marshal temu rampok bisu movie buruk buruk aspek sinematografi*	P
973	*Aktingnya mengerikan dan banyak ketidakkonsistenan. Saya tidak tahu apa-apa tentang membuat movie tetapi bahkan saya bisa melakukan pekerjaan yang lebih baik.*	*akting keri ketidakkonsistenan apaapa movie kerja*	N

Table 16 Testing Result With 150 Datasets

		Algorithm	
		Naïve Bayes	EM
Result	Unknown	10	12
	Missed	40	27
	Correct	100	111
	Accuracy	67%	74%

Table 17 Testing Result With 250 Datasets

		Algorithm	
		Naïve Bayes	EM
Result	Unknown	26	29
	Missed	64	41
	Correct	160	180
	Accuracy	64%	72%

Table 18 Testing Result With 350 Datasets

		Algorithm	
		Naïve Bayes	EM
Result	Unknown	35	38
	Missed	85	63
	Correct	230	249
	Accuracy	66%	71%

Table 19 Testing Result With 450 Datasets

		Algorithm	
		Naïve Bayes	EM
Result	Unknown	41	44
	Missed	112	91
	Correct	297	315
	Accuracy	66%	70%

Table 20 Testing Result With 577 Datasets

		Algorithm	
		Naïve Bayes	EM
Result	Unknown	43	47
	Missed	149	118
	Correct	385	412
	Accuracy	67%	71%

Table 21 Time Consumed by Naïve Bayes and EM

	Dataset	Naïve Bayes	EM
Time	150	20.9s	2215.4s
	250	20.5s	3351.3s
	350	23s	3515.6s
	450	18s	4473.3s
	557	20.5s	5858.7s

and thus proved that EM algorithm has a better result than Naïve Bayes in this particular case which is movie review that has so many tokens that eventually produce a huge bag of words.

Additionally, the author also notes the time that both algorithms took to perform the training and testing, unsurprisingly Naïve Bayes has the upper hand in this field. As the author already mentioned at the very beginning of this paper, Table 21 shows that how fast Naïve Bayes compared to EM, no matter how much the dataset is, has no real effect on how fast Naïve Bayes is. Although EM really did a good job to have better accuracy but it also comes with a tradeoff which is the time that it consumes.

Conclusion

In this research, the author uses EM as the alternative to the most used algorithm for opinion mining Naïve Bayes. This research uses 878 training datasets that consist of the half with the positive label and another half with the negative label from IMDb's movie review and 577 as testing datasets. Accuracy that Naïve Bayes got is 67% but after the uses of EM as the alternative algorithm of Naïve Bayes, the accuracy that EM got is 71%, there is an increase of accuracy by 4% and thus also proves that EM has better accuracy than Naïve Bayes on opinion mining in movie reviews case studies. The only tradeoff that EM comes with is that it takes a much longer time and consumes a lot of resource which is the processor than it took to perform the same dataset as Naïve Bayes.

References

Kulkarni, D. S., & Rodd, S. F. (2018). Extensive study of text based methods for opinion mining. *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, *Icisc*, 523–527. https://doi.org/10.1109/ICISC.2018.8399127

Vijay, R., Vangara, B., Thirupathur, K., & Vangara, S. P. (2020). Opinion Mining

Classification using Naive Bayes Algorithm. *International Journal of Innovative Technology and Exploring Engineering*, *9* (5), 495–498. https://doi.org/10.35940/ijitee.e2402.039520

Harahap, F., Harahap, A. Y. N., Ekadiansyah, E., Sari, R. N., Adawiyah, R., & Harahap, C. B. (2019). Implementation of Naïve Bayes Classification Method for Predicting Purchase. *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, *Citsm*, 1–5. https://doi.org/10.1109/CITSM.2018.8674324

Ilham Esa Tiffani. (2020). Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review. In *Joscex* (Vol. 1, pp. 1–7).

Pugsee, P., & Chatchaithanawat, T. (2020). Opinion mining for laptop reviews using naïve bayes. *Current Applied Science and Technology*, *20*(2), 278–294. https://doi.org/10.14456/cast.2020.16

Pugsee, P., Nussiri, V., & Kittirungruang, W. (2019). Opinion mining for skin care products on twitter. *Communications in Computer and Information Science*, *937*, 261–271. https://doi.org/10.1007/978-981-13-3441-2_20

Poovaraghan, R. J., Keerti Priya, M. V., Sai Surya Vamsi, P. V., Mansi Mewara, M. M., & Loganathan, S. (2019). Fake news accuracy using naive bayes classifier. *International Journal of Recent Technology and Engineering*, *8*(1C2), 962–964.

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings*, 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

Novendri, R., Callista, A. S., Pratama, D. N., & Puspita, C. E. (2020). Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes. *Bulletin of Computer Science and Electrical Engineering*, *1*(1), 26–32. https://doi.org/10.25008/bcsee.v1i1.5

Zhou, T., Wang, X., & Fang, Y. (2020). HARK: Harshness-Aware Sentiment Analysis Framework for Product Review. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 13993–13994.

Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., & Liu, H. (2019). Expectation-maximization attention networks for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 9166–9175. https://doi.org/10.1109/ICCV.2019.00926

Asheri, H., Hosseini, R., & Araabi, B. N. (2021). A new EM algorithm for flexibly tied GMMs with large number of components. *Pattern Recognition*, *114*. https://doi.org/10.1016/j.patcog.2021.107836

Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(2), 58–64. https://doi.org/10.28926/ilkomnika.v1i2.18

Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes. *ILKOM Jurnal Ilmiah*, *10*(2), 160–165. https://doi.org/10.33096/ilkom.v10i2.303.160-165

Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text classification algorithms. *Mining Text Data*, *9781461432*, 163–222. https://doi.org/10.1007/978-1-4614-3223-4_6

Anugrah, I. G., & Rosyid, H. (2019). Penerapan Information Retrieval Menggunakan Pemodelan Topik Pada Deskripsi Portal Multimedia. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *2*(1), 48. https://doi.org/10.32672/jnkti.v2i1.1057

Revindasari, F., Sarno, R., & Solichah, A. (2017). Traceability between business process and software component using Probabilistic Latent Semantic Analysis. *2016 International Conference on Informatics and Computing, ICIC 2016*, *Icic*, 245–250. https://doi.org/10.1109/IAC.2016.7905723

Anugrah, I. G., & Sarno, R. (2017). Business Process model similarity analysis using hybrid PLSA and WDAG methods. *Proceedings of 2016 International Conference on Information and Communication Technology and Systems, ICTS 2016*, 231–236. https://doi.org/10.1109/ICTS.2016.7910304