Meme Opinion Categorization by Using Optical Character Recognition (OCR) and Naïve Bayes Algorithm

Amalia Amalia[1], Amer Sharif[1], Fikri Haisar[1], Dani Gunawan[2], Benny B Nasution[3]
[1]Department of Computer Science, [2]Department of Information Technology, [3]Politeknik Negeri Medan
Universitas Sumatera Utara
Medan, Indonesia
amalia@usu.ac.id

Abstract— **Generally, a meme is an image which is produced by the society which is used to comment a certain event, followed with a particular template from the decent online images. The distribution of meme becomes the phenomenon and very popular in the last few years. The problem is, some of these memes contain negative contents to harm others. One type of meme image that trends in social media is aimed at government, either to support the performance of the government or to insinuate and dislike of a government. Therefore, Political view by the citizen can be identified by viral memes on the internet. The aim of this research is classifying the types of a meme by applying image processing and OCR Tesseract which are combined with Naïve Bayes Algorithm. OCR Tesseract is required to recognize text in an image, meanwhile Naïve Bayes algorithm which is used to find the highest probability to classify the testing dataset into the correct category. This research uses a meme as the dataset. The result is meme which is successfully classified. The accuracy depends on the OCR result which utilizes tesseract engine.**

Keywords-sentiment analysis; meme classification; ocr; naïve Bayes

I. INTRODUCTION

The Web 2.0 is a new set of communication technologies that enable democratic sphere. The web 2.0 rises the citizen journalism trends (Citizen Journalism) that allows everyone to express the opinion in the form of text, photos, and videos on the internet. One of the most prevalent, distinct and understudy phenomenon that represent the potential mediated cultural participation is internet meme [1]. Internet meme or term “meme” was first popularized by Richard Dawkins in his book The Selfish Gene (1976). But the term of ‘memes’ has evolved and is shifting in public discourse like in social media. In this emerging sense, ‘memes’ are amateur media artifacts, remixed extensively and recirculated by different participants on social media networks [1]. Memes are images with a caption that reflex the author feels, and the author wants the viewer who sees this image will get the same feeling.

In Indonesia, the memes distribution is varying such as humor, politics, quotes, education, and recent news. One of the most viral memes in Indonesia is about politics. Citizen spread and share memes that contain the emergent political situation in Indonesia. These memes are aimed at government, either to support the performance of the government or to insinuate and dislike of a government. Opinion mining or sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc., is positive or negative [3]. Usually, the sentiment analysis is extracted from a collection of texts in document level or sentence levels [2] such as webpages, social media or product reviews. Sentiment analysis can be defined by semantic orientation extraction that refers to the subjectivity, polarity, and strength of words, phrases or texts [3].

The growth of Web 2.0 or social media makes the development of sentiment analysis is become famous for industry and academia in this recent years [2] because the advent of social media platforms has lowered the cost of information generation, boosting the potential reach among online users [4].

The viral Memes can be described as a tool for bringing together the small contributions of millions of people and making them matter. The viral memes are memes that shares from one person to others, this phenomenon can be identified as a participatory culture. As the potential for mediated cultural participation in public imagination [1] identifying internet memes can locate public opinion mining in a country. Based on this phenomenon, the viral memes collection from the internet can also be an indicator of public assessment to the government.

The problem is, opinion extraction from meme is different from opinion extraction from social media status or product reviews because memes are images with caption thus we need a process to set apart the caption from the image.

To solve this problem, in this study we implemented Optical Character Recognition (OCR) to extract the text caption of the memes. OCR is a process for converting the image of a text document or scanned handwriting into ASCII characters (machine-readable characters). In a typical OCR system, input characters are digitized by an optical scanner. Each character is then located and segmented, and the resulting character is fed into a pre-processor for noise reduction and normalization [5]. Based on this caption, we can identify and classify public opinion mining towards the government.

Classification of opinion mining into positive and negative can be solved by machine learning or classifier methods like naïve Bayes classifier.

Naive Bayes is an implemented algorithm that can find the highest probability value to classify test data in the most appropriate category. In this research, the test data are meme image documents.

In this paper, we propose a method to categorize opinion mining not from user's status but memes. Our contribution, we combined image processing and Natural Language Processing (NLP) methods to extract opinion mining from internet memes. This research is conducting sentiment analysis to the collection of Indonesian memes that aimed to government originated from the internet. This research is consisting of several stages, such as image and caption segmentation of memes, text recognition, text extraction and opinion mining classification.

This paper is organized as follows. In section 2, we describe related works from previous researchers that related to this study. In part 3, we describe research methodology of this study. In section 4, we describe result and discussion, finally, we draw some conclusions and future works in section 5.

II. Related Works

Some previous research has analyzed internet memes like the study by [1] and [4]. A study by [1] conducted internet meme as a public discourse that transforms established cultural texts to negotiate the worth of diverse identities, and to engage in unconventional arguments about public policy and current events. This study explained memes are an essential parameter of social influence. Meanwhile, research by [4] characterized the nonlinear interactions of online users with meme ranking in social media, which mark the process of meme diffusion and meme popularity through the lens of social influence.

This study declared ranking memes based on their popularity could promote content distribution. Previous research about opinion mining had done by [2] [3] [6]. A study by [2] presented a review of NLP technique to extract opinion mining. This study also discusses challenges and problems in opinion mining. A study by [3] showed a lexicon-based approach to extracting sentiment from text with semantic orientation calculation. Survey by [3] extracted sentiment-bearing words like adjectives, verbs, nouns, and adverbs and use them to calculate semantic orientation. A study by [3] declared that this lexicon-based method performs well, and also robust across domains and texts. Although sentiment focused on adjectives or adjective phrases as the primary source of subjective content in text [7][8][9] But some of the previous studies like [10] also consider that verbs and adverbs also influence to obtain the polarity. In our study, we consider all the words to be calculated to achieve the meme sentiment. We implemented a supervised classification task for text classification approach that involves building classifiers from a labeled instance of texts like the previous study by [11]. A study by [6] implemented information extraction on Twitter status in Bahasa Indonesia by using sentiment analysis method to assess the quality of a TV program. In our research, we implemented NLP techniques to extract opinion mining, particularly for sentences, level opinion mining as described by [2] and we also used a lexicon-based method like the study by [3]. However, review by [2] and [3] is techniques for English opinion mining extraction, therefore, we also implemented methods by [6] to Bahasa Indonesia opinion mining extraction.

This study implements Tesseract OCR engine to extract text or caption from a meme. Tesseract is an open source OCR engine with a good result, many previous studies implement and discuss Tesseract like [12] and [13].

III. Research Methodology

This research retrieves the desired object from an image. In this case, the sentences in the meme are extracted and analyzed to obtain the sentiment. This research architecture is shown in Fig. 1.

A. Memes Collection

Memes collection is dataset for this study. These memes were harvested from the internet especially social media like Twitter, Facebook, and Instagram. Because of the aim of this study to find opinion mining from viral memes towards Indonesian Government, therefore we only collected memes in Bahasa Indonesia towards to Indonesian governments. The number of memes is 100, divided into 70 memes for training and 30 memes for testing. The example of a meme is shown in Fig. 2.

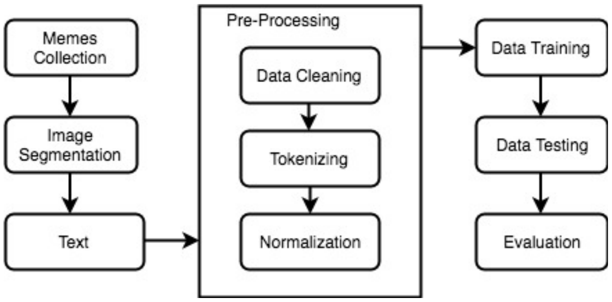

Fig. 1. Research architecture

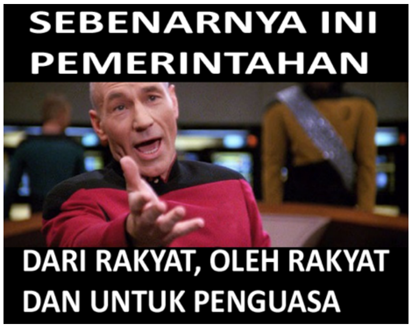
SEBENARNYA INI PEMERINTAHAN
DARI RAKYAT, OLEH RAKYAT DAN UNTUK PENGUASA

Fig. 2. The example of a meme

B. Image Segmentation

Image segmentation in this study is a stage to extract meme's caption. The image in a meme will be eliminated. Therefore, we apply the process of separating the area of observation (region) on each character detected.

We use OCR methods at this stage. Optical Character Recognition (OCR) is a process for converting printed documents or handwriting of scan/image into ASCII characters (machine-readable characters). The automatic text recognition using OCR is the process of converting the text document image to digital text. Therefore, the text can be used for the editing process. The first stage is converting the image to grayscale. This technology allows the machine to recognize characters automatically through the same optical mechanism as humans who use the eyes to see objects in the world [5].

OCR has several sub-processes that are normalization, feature extraction, and recognition. Normalization is the process of changing the dimensions of the region of each character and character thickness. Characteristic extraction is the process of extracting specific characteristics of the observed character. Recognition is a process for recognizing the observed character. In this study, we implemented OCR process with Tesseract engine. Tesseract is an open source OCR engine. Tesseract recognizes character with training process from a dictionary. One of the available dictionaries in Tesseract engine is English alphabet. The alphabet for Bahasa Indonesia is the same as the alphabet in English so that this tesseract machine can be implemented for Bahasa Indonesia.

C. Text Pre-processing

Text Pre-processing is a process to prepare text into next process. There are some pre-processing tasks such as data cleaning, tokenizing and normalization.

- Data cleaning can be interpreted as removing characters, punctuation, and symbols that have no meaning or information required. Some of the symbols appear when meme segmentation process with tesseract engine. This unnecessary symbol needs to be removed. This stage also removed whitespace that appeared when meme segmentation process. The resulting example from Tesseract can be seen in Table 1.

Table 1. Tesseract and Data Cleaning Result

Meme Example	OCR Result	After Data Cleaning
KEMBALIKAN MEDIA ISLAMA DASAR PEMERINTAH TIDAK TANGGUNG JAWAB	KEMBALIKAN MEDIA ISLAMA ' x L/ DASAR'PEMERINTAH TIDAK TANGGUNG JAWAB	KEMBALIKAN MEDIA ISLAM DASAR PEMERINTAH TIDAK TANGGUNG JAWAB

- Tokenizing is the separation of the sequence of characters in a set of documents into pieces of words or characters that match the system requirement. These pieces are known as tokens.
- Normalization

Normalization is a process to normalize the tokens. This process is necessary because many memes contain unstructured sentences. For examples, memes that contain unstructured abbreviations and memes with slang terms. We utilize a dictionary. This dictionary was created manually. The example of this process for terms like "kira2" will be normalized to "kira-kira". In this stage, we also process negation terms in Bahasa Indonesia. For words that start with negation words "tdk", "tidak", "ga" "engga" will be grouped into the next word. These grouped words will be count as one token. Examples of this process for tokens "tidak" and "suka" will be grouped as a token "tidak_suka".

D. Data Training

One of data training stage is labeling. Labeling is a process to determine meme's caption has negative or positive opinion towards the government. We decided these classified manually with human judgments. We use 60 memes for training, contains 30 of negatives memes and 30 of positives memes. The excerpt of this process can be described in Table 2.

Based on the positive and negative classification results in the labeling process, the next step is to calculate the number of vocabulary frequencies that appear in the positive class and negative class. Some of tokens or words appear in both class, for example, a word "government" appears in the negative and also on the positive class. Based on the dataset from this study, we found the total number of words for negative is 167, and the total number of words for positive is 142. The Excerpt of this calculation can be seen in the following Table 3.

Table 2. The excerpt of Meme labeling

No. of Meme	Meme's Caption	Classification Opinion by Human Judgements
1	Bbm akhir nya naik kita di bohong lagi	Negatives
2	Jangankan hujan air hujan peluru pun kami hadang demi keutuhan indonesia	Positives
3	Terimakasih pak jokowi mudik lancar keluarga tenang	Positives
4	Pemerintah suka bohong ya	Negatives
...	.....	.....
...	.....	.....
59	Terimakasih pak jokowi mudik lancar keluarga tenang	Positives
60	Selamat datang di indonesia surga bagi para koruptor	Negatives

Table 3. The excerpt of Words Occuration Frequency Calculation

Classificasion	Words Occurance Frequency (n_k)	Total (n)
Negatif	Bbm(2), akhir(1), naik(2), kita(2), bohong(2), lagi(4), akibat(2), pelajaran(1), tidak(7), masuk(1), yang(2), ada(2), malah(3), kecapekan(1), bangsa(1), ini(5), kekurangan(3), orang(2), pintar(1), tidak_jujur(1), Sby(1), tidak_suka(2), ya(1), sholat(1), sering(1), telat(1), minta(2), jodoh(1), cepat(1), selamat(1), datang(1), indonesia(3), surga(1), bagi(1), para(1), koruptor(1), pergi(1), ke(3), dukun(1), ... kayak(1), buronan(1), habisi(1), duit(1), hasil(1), anggota(1), dpr(1), banyak(1), bicara(1), sedikit(1), pendidikan(1), mahal(1), dengan (1), malam(1), hari(2), susah(2), pagi(1), bangun(1), kadang(1), merasa(1), sedih	167
Positif	Sungguh(6), luarbiasa(1), polisi(5), ini(8), mengajarkan(2), siswa(2), paduan(1), mantap(1),selama(1), jadi(7),wagub(1), sandiaga(1), akan(2), yatim(1), piatu(1), dan(2), dhuafa(1), mulia(4), sekali(3), membiarkan(1), motor(2), secara(1), Cuma(1), kepada(3), seorang(1), pemuda(2), ... permainan(1), tradisional(1), ketapel(1), kalian(2), sendiri(1), menyalahkan(1), soal(1), banjir(1), tapi(2),buang(1), sampah(1), sembarangan(1), jangan(2), salahkan(1), anggota(1), dpr(1), mereka(1), ingin(2), cerdas(1), rapat(1), selanjutnya(1), daripada(1), sahabat(1), super(1), lihat(4), ganteng(1), bergaulah(1), dengan(1), orang(5),	142

The next step the result of this training data will be processed by probability model of Naive Bayes. The first step is to calculate the probability of each text document against a set of documents with the equation 1.

$$P(vj) = \frac{|Docs\ j|}{|TotDoc|} \qquad (1)$$

Where :

$P(vj)$: the probability of each document against a set of documents

$|Docs\ j|$: number of training data documents from each category

$|TotDoc|$: a total number of training data

The result of this calculation can be seen in Table 4.

Table 4. $P(vj)$ for Positive and Negative Classification

Klasifikasi	$P(vj)$
Negatif	$\frac{30}{60}$= 0.5
Positif	$\frac{30}{60}$= 0.5

The next process is to calculate the probability value of the appearance of the word on the document if the document classification is already known. This calculation can be described with equation 2.

$$P(wk|vj) = \frac{n_k+1}{n+|Voc|} \qquad (2)$$

Where :

$P(wk|vj)$: the probability value of the appearance of the word on the document

n_k : frequency of word appearing in each document of a particular category.

n : the number of all words in a particular category

$|Voc|$: the number of words in data training

The excerpt of this calculation can be seen in Table 5.

E. *Testing Dataset*

Testing stage is a process to determine the memes into positive or negative with Naïve Bayes classifier. Naive Bayes algorithm is an algorithm to find the highest probability value to classify test data in the most appropriate category. We then calculated the probability of the Vmap tendency in each of the following classification categories This classification based on pattern and calculation from training stage. We use 40 memes for testing.

The example of this calculation can be seen in Table 6 and Table 7.

Table 5. The probability value of the Appearance of a Words

Tokens / Words	Documents $P(wk|vj) = \frac{n_k+1}{n+|kosakata|}$	
	Positive (V_1)	**Negative (V_2)**
Bbm	3/451	1/476
Akhir	2/451	1/476
Naik	3/451	1/476
Kita	3/451	2/476
Bohong	3/451	1/476
Lagi	5/451	1/476
Akibat	3/451	1/476
Pelajaran	2/451	1/476
.....	....	....

Table 6. Case Example to Determine Meme's Opinion Class

Dokumen	Teks hasil OCR dan Preprocessing	Classificasion
Testing Data	Harga naik lagi dasar pemerintah tidak peduli rakyat	???

Table 7. The probability of the Vmap tendency in each of the following classification categories

	Tokens	Negative	Positive
P(V_j)		1/2	1/2
$P((wk\|vj)) = \frac{n_k + 1}{n + \|kosakata\|}$	Harga	1/451	1/476
	BBM	4/451	2/476
	Naik	4/451	2/476
	Lagi	6/451	2/476
	Dasar	1/451	1/476
	Pemerintah	4/451	4/476
	Tidak_Peduli	1/451	1/451
	Rakyat	6/451	2/476
	Vmap	$1{,}4 \times 10^{-21}$	$2{,}5 \times 10^{-23}$

Based on the calculation in Table 7, we can see Vmap value for the negative class is more significant than Vmap value from the positive class. Therefore, this meme is classified as a negative sentiment.

F. *Evaluation*

Evaluation is a step to calculate the accuracy of this method. Based on the experiments in this study, from 40 testing memes, 30 memes are classified in right classification or the accuracy is 75%.

IV. RESULT AND DISCUSSION

After the testing process, the results from OCR Tesseract Engine can recognize the image without additional pre-processing. However, OCR has not yet been able to acknowledge posts in italics. The accuracy of this methods is 75%, we found the drawbacks of this method because some of the tokens in the testing stage have not recognized yet in training process. Therefore, to increase the accuracy, we have to increase the amount of training data. Many of the features of tokens were quite predictable for particular classification, for example, tokens or words like “jelek”, “bohong”, “koruptor”, “BBM” are among in negative classification, whereas “cinta”, “pembangunan”, “hebat" are indicated as positive sentiment.

V. CONCLUSION

According to the result and discussion of this research, we summarize that meme sentiment analysis system is successfully created by utilizing Optical Character Recognition method and Naive Bayes algorithm. Naive Bayes algorithm has fairly good accuracy in classifying sentiment analysis on the meme. We obtain 75% accuracy of 100 datasets consisting of 60 data training and 40 data testing. OCR Tesseract Engine is not able to recognize italic sentences in the meme.

There is still much room for improvement. There are many methods such as Neural Network, K-Nearest Neighbor, and others. For the future research, it is required to utilize additional text pre-processing like n-gram tokenization to improve the accuracy.

ACKNOWLEDGMENT

The authors gratefully acknowledge that the present research is supported by Ministry of Research and Technology and Higher Education Republic of Indonesia. The support is under the research grant PTUPT of Year 2018 Contract Number 254/UN5.2.3.1/PPM/KP-DRPM/2018.

REFERENCES

[1] R. M. Milner, J. Childers, and B. Chappell, “The World Made Meme : Discourse and Identity In Participatory Media,” University of Kansas, 2012.

[2] S. Sun, C. Luo, and J. Chen, *A review of natural language processing techniques for opinion mining systems*, vol. 36. Elsevier B.V., 2017.

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-Based Methods for Sentiment Analysis,” *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, 2011.

[4] S. He, X. Zheng, and D. Zeng, “A model-free scheme for meme ranking in social media,” *Decis. Support Syst.*, vol. 81, pp. 1–11, 2016.

[5] P. K. Y. Sandeep Tiwari, Shivangi Mishra, Priyank Bhatia, “Optical Character Recognition using MATLAB,” *Int. J. Adv. Res. Electron. Commun. Eng.*, vol. 2, no. 5, pp. 579–582, 2013.

[6] A. Amalia, W. Oktinas, I. Aulia, and R. F. Rahmat, “Determination of quality television programmes based on sentiment analysis on Twitter,” in *2nd International Conference on Computing and Applied Informatics 2017*, 2018.

[7] V. Hatzivassiloglou, K. R. McKeown, V. Hatzivassiloglou, and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the 35th annual meeting on Association for Computational Linguistics* -, 1997, pp. 174–181.

[8] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’04*, 2004, p. 168.

[9] M. Taboada, C. Anthony, and K. Voll, “Methods for Creating Semantic Orientation Dictionaries *,” pp. 427–432.

[10] V. S. Subrahmanian and D. Reforgiato, “AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis,” *IEEE Intell. Syst.*, vol. 23, no. 4, pp. 43–50, Jul. 2008.

[11] B. Pang, L. Lee, … S. V. the A.-02 conference on, and undefined 2002, “Thumbs up?: sentiment classification using machine learning techniques,” *dl.acm.org*.

[12] R. Smith, D. Antonova, and D.-S. Lee, “Adapting the Tesseract open source OCR engine for multilingual OCR,” in *Proceedings of the International Workshop on Multilingual OCR - MOCR ’09*, 2009, p. 1.

[13] R. Smith, “An Overview of the Tesseract OCR Engine,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, 2007, pp. 629–633.