Sentiment analysis towards Jokowi`s government using twitter data with convolutional neural network method

H Tsaniya, R Rosadi and A S Abdullah*

Department of Computer Science, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

* atje.setiawan@unpad.ac.id

Abstract. In this current digital era, internet has become commonly used in society, it become an important aspect in life, one of its function are as a communication platform and channel for public opinion. In which utilize social media platforms. Indonesia really active in communication and convey expression using social media. In such a way that Indonesia`s government takes advantage of it as a way to communicate and get feedback from Indonesian people. In this research, as jokowi lead his second period as a president of Indonesian`s government, we try to see how Indonesian sentiment toward his leadership. Using Convolutional Neural Network method from 7228 row data crawled from twitter obtained 38% positive sentiment, 51% negative sentiment, and 18% neutral sentiment. Overall we have 57% prediction accuracy using CNN model.

1. Introduction

Communication and dissemination of information is an important aspect of leadership, especially in the field of government [1]. Some of the important roles of the government are in the dissemination of information, achievements, the existing cabinet work design, and performance progress to the public, in various ways, both through traditional media, conventional media and digital media. In the technological era where almost all people are familiar with the internet, digital media is one of the most effective delivery media for the community [2].

The rapid dissemination of information through the internet has resulted in the community experiencing a change in mindset that is increasingly critical in responding to existing situations and conditions, the existence of such conditions is a demand for the government to be able to meet the various needs of society in all aspects. With these demands, the government through the Office of Communication and Information Technology (Diskominfo) collects feedback from the public as a source of information to improve performance. The various feedbacks received from the community are not only positive, but also negative.

One way to use technology to get feedback from the public is by using social media. Based on statistical data, Indonesian internet users grew by 12.6% from 2018 with the most access to social media [3]. Especially in the use of Twitter. Indonesia is ranked 5th as the most Twitter social media users [4]. Advantage of this situation, the integration of data, information and knowledge from various sources and organizations can be a source of study on government social media in the coming years. So that in this study, sentiment analysis will be carried out by analyzing twitter about public opinion on the government.

Data in the form of complaints or criticisms, information or suggestions on Twitter social media can be used as a source of data to monitor the results of feedback from the public. To get feedback
results that can be seen clearly, a classification process is used for public opinion. We can process this public opinion data with data mining techniques, namely classification. The classification that will be carried out in this study is the classification of public sentiment using one of the methods in Deep Learning, namely, Convolutional Neural Network as a test of the accuracy of the application of the method to the sentiment analysis of the Jokowi government. The data taken are tweets related to Jokowi's verification account.

Sentiment is textual information that resides on the web and contains facts and opinions. Sentiment is a subjective statement that reflects a person's perception of an event [5]. Data mining is the process of looking for patterns or interesting information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The choice of the right method or algorithm is very much dependent on the objectives and the overall Knowledge Discovery in Database (KDD) process [6].

Convolutional Neural Network is a machine learning method from the development of Multi Layer Perceptron (MLP) which is designed to process two-dimensional data. CNN is included in the type of Deep Neural Network because of its deep network level and is widely implemented in image data. CNN has two methods; namely the classification using feedforward and the learning stage using backpropagation. The way CNN works is similar to MLP, but in CNN, each neuron is presented in two dimensions, unlike MLP, where each neuron is only one dimension [7].

2. Method

2.1. Object of Research
The object of this research is the Indonesian government which is currently led by President Jokowi. Jokowi is 7th President of Indonesia who elected since October 20, 2014. One of Jokowi’s strengths in politics is the use of social media. During his campaign and running of government, he used social media a lot as a dissemination of information and a means of interaction with the public. Data collection used in this study utilizes Jokowi’s social media.

The data collection used comes from Jokowi’s verified account. To obtain data, researchers used the Twitter API application with the python programming language and the Tweepy library. It takes an API key, API secret key, access token and access token secret to be able to access the Twitter crawler provided by API search. Access key is obtained by registering via the Twitter application. To get this access right, you must be registered as a Twitter user and then submit a request for access by filling in the application form.

2.2. Dataset Preparation
Before classification process, there is a steps to make the dataset before it qualify enough to get into classification model.

2.2.1. Collecting Data
The data used were sourced from Jokowi’s verified account from November 22, 2019 to December 22, 2019. To obtain data, researchers used the Twitter API application with the python programming language and the Tweepy library. It takes an API key, API secret key, access token and access token secret to be able to access the Twitter crawler provided by API search.

2.2.2. Labelling Data
After the data got collected, begin to labelling each sentiment in the data to sentiment classification, in this research, the data got labelled into 3 sentiments, which is positive, negative, and neutral. The labelling process done manually by the researcher.

2.2.3. Preprocessing Data
Preprocessing is done so that the raw data obtained from the crawling process can be trained and tested with good accuracy. There are three steps taken during preprocessing, which is case folding, stemming, and remove unimportant word from the data. Case folding is cleaning the initial tweet from numbers, URLs, emoticons, special characters, and changing the tweet to lowercase or lowercase.
In the stemming process, words in sentences that have affixes will be converted into their root words. This process is carried out using the literary library in python which can change the Indonesian affixes to the root words. Compiled unimportant word that have no influence on the sentimental polarity of an existing tweet are called stopwords. Stopwords can be made by yourself, or you can use existing modules such as literature.

2.3. Embedding Model
In learning about an object, it is necessary to have the characteristics of the object or what is commonly known as a feature. If the object received is in the form of text, it is necessary to extract the text into a numeric vector that represents each word which is known as the word embedding process. The process of converting text to numbers is needed because deep learning algorithms are not able to perform analysis on string or text input. One of word embedding model is Word2vec.

Word2Vec consists of two techniques, namely the Continuous Bag of Words (CBoW) and the Skip Gram Model. Both of these methods use the concept of neural networks that map words to target variables which are also words. Both of these techniques use weight as a word vector representation [8]. The Continuous Bag of Word is an architecture in word2vec to predict words that are associated with multiple context inputs [8].

![Figure 1. CBoW architecture](image)

2.4. Convolutional Neural Network
On CNN, the convolutional matrix used to filter text input is generated based on the number of words that are used as the input matrix. The input matrix will be filtered with a convolutional matrix to produce an activation function.

Mathematically, the convolution process on a matrix with dimensions $N \times N$ using a matrix filter with dimensions $m \times m$ with weighting $\omega$ will produce a convolutional output matrix of size $(N - m + 1) \times (N - m + 1)$ and the addition of non-linear activation to the results a convolutional matrix, can be formulated as follows:

$$x_{ij}^f = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \omega_{ab} y_{ij}(i+a)(j+b)^t$$

(1)

where

- $x$ : result of convolution process in matrix shape,
- $m$ : dimension of convolution matrix
- $a, b$ : are index starts of matrix
- $\omega$ : weight of the data that implemented in hidden layer randomly
- $y$ : linear function to adjust matrix numerical value

In the convolution process, the convolutional matrix will divide the input matrix into several size parts to produce a feature mapping to the input matrix. From the existing feature mapping, it will be
linked to form a feature vector on the neuron layers. From the resulting feature vectors, the softmax function is used to classify text based on the final output matrix.

One of the important aspects of CNN is the convolution process, where the convolutional matrix will shift to map the features of the input matrix. In the convolution process, the pooling layer plays an important role in producing a feature map. The pooling layer works by taking the highest value (max) or average value (avg) of the input matrix to produce a feature vector. The pooling layer will produce an output matrix with a clear size, clear output that will simplify the text classification process. For example, in the N × N dimension input matrix, pooling will only take the k × k part of the input matrix which will produce an output by simplifying the value on the input matrix which is the most important summary value of the input [9].

In CNN we use convolutions over the input layer to compute the output. This results in local connections, where each region of the input is connected to a neuron in the output. Each layer applies different filters typically hundreds or thousands, and combines their results. During the training phase, a CNN automatically learns the values of its filters.

The input to most natural language processing tasks are sentences or documents represented as a matrix. Each row of the matrix corresponds to one token, typically a word, but it could be a character. That is, each row is a vector that represents a word. Typically, these vectors are weighted using word embedding method that index the word into a vocabulary. The illustration can be seen in figure 2, whereas for natural language processing we convert the based on text input into vector matrix and apply linear activation function in the neuron layer.

![CNN in natural language processing illustration](http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/)

### 3. Result and Discussion

#### 3.1. Data Collection Results

The data collection process from November 22, 2019 to December 22, 2019 with the keyword @jokowi, with the limit that tweets only came from Indonesian and Indonesian-language accounts, resulted in 7,228 lines of tweet data related to Jokowi’s verification account. In table 1 is the head of the data, can be seen top 5 data from the dataset which contain date of crawled data, username, and raw tweet.
Table 1. Data collection results

| Date          | Username          | Tweet                                                                 |
|---------------|-------------------|----------------------------------------------------------------------|
| 11/22/2019    | b'agung7709'      | @hukumdan @jokowi @sumbarpemprov @irwanprayitno kata erick jokowi gila kerja https://t.co/gqs2pfya0w' |
|               |                   | b'@scsuroso @sardiridwan78 @ustadtengkuzul @jokowi congorna penipu' |
|               | b'rezfatti'       | 11/22/2019 4:41:16 AM                                               |
|               |                   | b'tonton dan simak kata\xc2\xb2nya ya, karna dengan secara sadar mereka telah mengatakan tidak lg menjadi bagian dari nkri.ln @jokowi @gerindra @pksejahtera @kemenag_ri https://t.co/dlx5gigumf' |

3.2. Data Labeling
The data that has been collected is carried out by a manual labeling process based on the implied sentiment of the tweet. In table 2, you can see the results of numerical labeling of data and the number of sentiments. From the labeled data, a dataset will be pre-processed for the weighting process.

Table 2. Numerical sentiment labeling of the dataset

| Sentiment | Label | Total |
|-----------|-------|-------|
| Positive  | 1     | 2,253 |
| Negative  | 0     | 3,655 |
| Neutral   | 2     | 1,321 |

3.3. Weighting with Word2Vec
After the preprocessing process, the feature weighting of the dataset will be carried out using the Word2Vec method using the CBoW architecture. Before modeling, in Table 3 we can see 5 words with the highest frequency of appearance in the data.

Table 3. The word with the highest frequency of occurrence

| Word     |
|----------|
| Presiden |
| Tidak    |
| Jokowi   |
| Terima kasih |
From the data, all 45,770 vocabularies are found in the data and will be used for model building. The model is trained as many as 30 iterations, and the resulting weighting of 1,488,120 words from all data.

The results of building the CBoW model can be seen from the representation of words in the vector dimension of the dataset. In figure 3 you can see the word representation of the word ‘Jokowi’ as the target word for the dataset. In the image the word representation is seen in the form of vector dimensions with red color is the target word whose weighting representation can be seen according to the word neighbor which is often used in the data, blue is the word neighbor with the closest weight representation to the target word, and green is the random input word which is used to visualize the position of the word according to the weight on the graph, the position of the word on the vector graph shows the weight value numerically.

**Figure 3.** Vector image of word representation in dimensions

### 3.4. Classification with CNN

The establishment of a fully connected layer using 5 neuron layers for the convolution and pooling processes. Before the classification process, the feature weighting of the dataset is implemented using CBoW modeling, so that the data that will go through the convolution process on neurons already has a weight on the features to be extracted according to the modeling in the embedding.

In table 4, you can see the CNN modeling, input with a vector dimension limit of 3,043, which will change with the adjustment of parameters in the fully connected layer of convolutional neurons. The feature extraction process on neurons uses max pooling which takes the highest numerical value from the data matrix as feature information on the data to be processed on neurons. Each layer has different limitations and parameters according to the training process in the model. Adding dense and dropout layers to the model to reduce neurons that are considered useless or disturbing in the classification process, the process is carried out backpropagated in training to avoid overfitting the model. Meanwhile, the dense layer in the model is used as a matrix value adjustment for classification of outputs with the activation function.

| Tipe Layer | Output Shape | Parameter |
|------------|--------------|-----------|
| Input layer| 3,043        | 0         |

**Table 4.** Convolutional neural network modeling
From the modeling results, it can be seen from the training performance data in figure 3, using cross-entropy, the value of acc and loss from training is 8 iterations with the distribution of 10% of validation data and 90% of training data. starting to follow the motion of the training data chart in the 4th iteration, which changes back to the 7th iteration where there is a difference in the direction of the chart movement. In the loss graph between validation data and training data, the movement of validation data values and training shows a unidirectional movement, but there is a change after the 6th iteration, where the difference in value between validation and training data, this difference can cause overfitting. Based on [10], dropout and dense layers are added to avoid these conditions, so that the model is not too static in the training data and can produce fairly good accuracy on the test data.

![Figure 4](image.png)

**Figure 4.** The cross entropy graph of the CNN model

The results of data training with the CNN model are predicted against the test data and the accuracy is obtained with the following calculations:

\[
\text{accuracy} = \frac{\text{correct predicted label}}{\text{sum of all label data}}
\]  

(2)

From the results of the test data prediction with the model, the accuracy value of the model is 57% from 743 test data, with a comparison between model prediction and manual labeling as shown in figure 4. In figure 4 it can be seen that the model has a fairly good positive sentiment prediction ability, while on negative sentiment and neutral, there is a fairly far gap with the number of neutral sentiments predicted to be negative by the CNN model used. Indicating that the model predicts the
classification of neutral and negative labels has almost similar features, so that the number of sentiments labeled neutral is predicted to be negative by the model.

![Figure 5. Predictive label graph with data labels](image)

4. Conclusion
Based on the sentiment classification research that has been carried out on 7,228 rows of data taken from November 22, 2019 to December 22, 2019 using the Convolutional Neural Network method with word weighting using the CBoW method, it can be seen that the accuracy of using this method on the dataset with the topics used reaches 57%, with the performance value of the accuracy and loss graph for the model is good enough to avoid overfitting. The percentage of sentiment obtained was 51% negative sentiment, 31% positive sentiment, and 18% neutral sentiment.

Acknowledgments
The authors thank Rector Universitas Padjadjaran who provided financial assistance for the dissemination through Academic Leadership Grant 2020 with contract number: 1427/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.

References
[1] Judson J 2018 Importance Aspect Leadership Judson Group
[2] Oyza I and Edwin A 2015 Effectiveness of Social Media Networks as a Strategic Tool for Organizational Marketing Management Journal of Internet Banking and Commerce 3 201
[3] Jayani H 2019 Pengguna Internet di Indonesia databoks katadata
[4] Jayani H 2020 Media Sosial di Indonesia databoks katadata hootsuit
[5] Lawrence S, Pennock D, and Dave P D M 2003 Mining the peanut gallery: Opinion extraction and semantic classification of product reviews Int. World Wide Web Conf.(WWW’03)
[6] Mardi Y 2019 Data Mining: Algoritma Klasifikasi Jurnal Edik Informatika 3 213-219
[7] Sofia N 2018 Convolutional Neural Network Deep Learning Method 2 203-206
[8] Abdullah and Husain A 2018 Word Embedding RStudio 20
[9] Gibiansky A 2014 Fullcy Connected Convolutional Neural Network Math Code
[10] Labach A and Salehinejad H 2019 Dropout Methods for Deep Neural Networks (USA: University of Toronto)