Sentimen Analisis

By Alif Almuqsit

I. Problem Scoping

In selling or making a product, users will gain experience with the product. Of course, customers have an assessment of their experience using the product. these assessments need to be obtained and analyzed to develop the product. One way that can be done is to use the sentiment analysis method. Sentiment analisis adalah suatu cara untuk mengetahui penilaian positif atau negatif dari pengguna produk tersebut.

II. Data Acquisition

Data is provided and obtained from PT Prosa Solusi Cerdas. Data is provided and obtained from PT Prosa Solusi Cerdas. There are 2 types of files or data, namely for training and testing, each of which is a tsv file type. The data can be access here (https://drive.google.com/file/d/1rfjmie2tTNwOkBmCaZ7phZiEWqLOxriB/view?usp=sharing)

III. Data Exploration

In the training data there are 1780 sentences which are divided into 2 classes, namely positive and negative, of which 1200 are positive and 580 are negative. In the test data there are 185 sentences which are divided into 2 classes, namely positive and negative, of which 120 are positive and 65 are negative.

On the training and test data, I preprocessed the data by clearing a few words that didn't have much effect on user sentence or ratings sentiment analysis. I am using a library that has been developed by someone else, Sastrawi. From the Sastrawi library, I took the available Indonesian stopwords and added some words. I added this because the sentences in the data did not follow the standard language (KBBI). I also removed the symbols in the sentences.

After cleaning the data, the data will be converted into vectors using the help of the SciKit library (TFIDFVectorizer) (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html). The parameter used in converting sentences into vectors is it can consist of one to 3 words (ngram_range=(1,3)) and a minimum of 10 occurrences of all sentences (min_df=10). Of course the class or label is assigned to a different variable.

IV. Modelling

I am using Logistic Regression architecture (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html). I chose architecture because I had studied it at PT. Orbit Ventura Indonesia in Magang dan Studi Independen Kampus Merdeka 2021 Program entitled Artificial

Intelligence for Gen Y. This architecture was also chosen because it does not use the concept of deep learning so it is lighter and suitable for sentiment analysis cases.

In this model, I use the following parameters. I chose the inverse of regulation strength of 3 (C=3) because the data or sentences can be trusted or match the real world. I'm using liblinear as an optimization function. I did this because of the recommendation from the SciKit documentation that provides this architecture. Finally, the maximum iteration I chose was 150.

V. Evaluation

At this stage, I test the model using test data that has been converted into vectors and has been separated by labels or classes. To find out the performance and of the model, I used the F1-Score matrix. I chose this matrix because in the dataset there was an unbalanced data so that the accuracy matrix was not suitable for use. The F1-Score on the model scores around 0.83. That's a pretty good result.

I also tested manually using sentences that might not be in the dataset. Finally, I provide a function to use the model.