KLASIFIKASI SENTIMEN PADA TWITTER DENGAN NAIVE BAYES CLASSIFIER

Sigit Suryono, Ema Utami, Emha Taufiq Luthfi
Program Magister Teknik Informatika
Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Condong Catur, Depok, Sleman, Yogyakarta
sigitharsy25@gmail.com, emma@nrar.net, emhataufiqluthfi@amikom.ac.id

Abstract

Sentiment classification is the one branch of the field of Text Mining. Sentiment classification can be an important in the process of evaluations about something problem. The main of sentiment classification are to find out the polarity of positive, negative and neutral sentiments. The sentiment classification obtained from Twitter. In this paper, the tweets related to predefined keyword are collected using the tools provided by Twitter. The data that has been collected is processed by using Natural Language Toolkit that run on Python programming language. After that, the data will be classified by using Naive Bayes Classifier to find out about the sentiment. The result of classification will be measured accurate level. Based on the experimental result for three times trial, the result obtained accuracy level in the first is 64.95%, the second is 66.36%, and the third is 66.79%. Another result obtained is percentage of sentiment are positive sentiment is 28%, negative is 20% and neutral is 52%. Based on percentage result of sentiment classes, neutral sentiment is the most sentiment that related to Joko Widodo and his government topic.

Keyword: Sentiment classification, Opinion Mining, Naive Bayes Classifier, NLTK, Twitter Sentiment.

Abstrak

Klasifikasi sentimen merupakan salah satu cabang dari Text mining. Klasifikasi sentimen dapat menjadi sesuatu yang penting dalam proses evaluasi terhadap sebuah topik permasalahan. Tujuan utama dari klasifikasi sentimen adalah untuk mencari tahu polaritas dari sentimen positif, negatif dan netral. Klasifikasi sentimen salah satunya dapat diperoleh melalui tweet yang ada pada Twitter. Dalam tulisan ini, tweet yang berhubungan dengan kata kunci yang dicari dihimpun dengan menggunakan tools yaitu API Twitter. Data yang didapat dari proses penghimpunan akan diolah dengan menggunakan Natural Language Toolkit yang berjalan diatas bahasa pemrograman Python. Data selanjutnya akan dilakukan klasifikasi sentimen dengan menggunakan Naive Bayes untuk melihat sentimen yang dihasilkan. Dari proses klasifikasi yang telah dilakukan akan diukur tingkat akurasi. Dari hasil uji coba sebanyak 3 kali, didapatkan tingkat akurasi pada percobaan pertama 64.95%, kedua 66.36% dan ketiga 66.79% Hasil lain yang didapatkan dari proses klasifikasi yaitu Sentimen positif 28% sentimen negatif 20% dan sentimen netral 52%. Berdasarkan hasil persentase kelas sentimen, sentimen neutral merupakan sentimen yang paling banyak apabila dikaitkan dengan topik Presiden Joko Widodo dan Pemerintahannya.

Kata Kunci: Klasifikasi Sentimen, Opinion Mining, Naive Bayes Classifier, Natural Language Toolkit, Sentimen Twitter.

1. Latar Belakang Masalah

Penggunaan internet di Indonesia sudah dapat dikatakan berkembanga secara merata. Adanya koneksi internet dapat dimanfaat sebagai sarana untuk menyalurkan hobi baik itu jual beli secara online, menulis atau menyampaikan pendapat terhadap sesuatu melalui social media. Social media yang cukup popular di Indonesia salah satunya yaitu Twitter. Melalui Twitter pengguna menyampaikan pendapatnya secara bebas. Twitter juga menyediakan fitur tranding topik untuk kawasan atau wilayah tertentu sesuai dengan preferensi pengguna masing-masing.

Twitter berkembang dengan cepat dalam segi pengguna. Pada tahun 2013, terdapat lebih dari 500 juta pengguna terdaftar dan 200 juta diantaranya merupakan penggunakan aktif (Twitter, 2017). Perkembangan pengguna pada Twitter terjadi ketika terdapat sebuah kejadian populer didunia tanpa mengabaikan perkembangan pengguna media sosial lain seperti Facebook, Instagram dan lain sebagainya. Pengguna lainnya berasal dari pihal developer. Twitter cukup popouler dikalangan developer karena kemudahanya dalam mengambil data-data yang diperlukan oleh developer.

Klasifikasi sentimen merupakan salah satu cara untuk mengetahui pendapat seseorang atau sekelompok orang terhadap isu, produk, layanan atau golongan tertentu. Klasifikasi sentimen dapat dilakukan dengan mengumpulkan data melalui Twitter denga topik tertentu. Pada penelitian ini kasus topik yang diangkat adalah tentang Presiden Republik Indonesia Joko Widodo beserta pemerintahan yang sedang berjalan. Berdasarkan topik yang telah ditentukan akan dilakukan klasifikasi dengan menggunakan metode klasifikasi data yaitu Naive Bayes. Data hasil klasifikasi akan diukur tingkat akurasinya dengan menggunakan Split Validation. Data hasil klasifikasi juga akan digunakan sebagai gambaran bagaimana sentimen sebagian pengguna Twitter terhadap presiden Joko Widodo dan pemerintahan yang sedang berjalan.

2. Metodologi Penelitian

Metode penelitian yang digunakan adalah metode eksperimen. Pada penelitian data yang telah didapatkan akan dibagi menjadi 2 data set yaitu data set *training* dan data set *testing*. Beberapa metode yang digunakan dalam penelitian ini antara lain sebagai berikut.

a. Natural Language Toolkit (NLTK)

NLTK merupakan salah satu tools pengolahan bahasa natural yang berjalan pada bahasa pemrograman Python (NLTK Project, 2017). NLTK menawarkan tampilan antar muka yang mudah dipahami dan menyediakan lebih dari 50 data serta kamus data yang dapat digunakan. Beberapa kamus data yang dapat digunakan antara lain *WordNet*, *TextProcessing* serta untuk proses klasifikasi yaitu *tokenization*, *stemming, tagging, parsing* dan juga *semantic reasoning*.

b. Naive Bayes

Naive Bayes merupakan sebuah model yang dapat bekerja dengan baik pada proses pembagian kategori untuk teks (Manning & Schuetze, 1999). Naive bayes melakukan klasifikasi dengan menggunakan dua proses yaitu proses *training* dan proses *testing*.

c. Analisis Sentimen

Analisis sentimen merupakan salah satu riset yang kompleks. Adapun karakteristik dari Analisis Sentimen adalah sebagai berikut (Pozzi et al, 2017).

1. Kategorisasi Sentimen yaitu membedakan antara kalimat subjektif dan kalimat objektif.

2. Tingkat analisis. Tingkat analisis dibagi menjadi 3 tingkatan yaitu Message level, Sentence Level dan Entity and Aspect Level.

3. Pendapat yang membandingkan sesuatu serta pendapat yang hanya sekedar pendapat. Artinya setiap orang dapat memberikan pendapat dengan membandingkan satu hal dengan hal lain atau hanya memberikan pendapatnya saja.

4. Pendapat eksplisit dan pendapat implisit. Pendapat yang diungkapkan secara jujur, tegas dan jelas atau pendapat yang diungkapkan secara tidak jelas.

d. Python

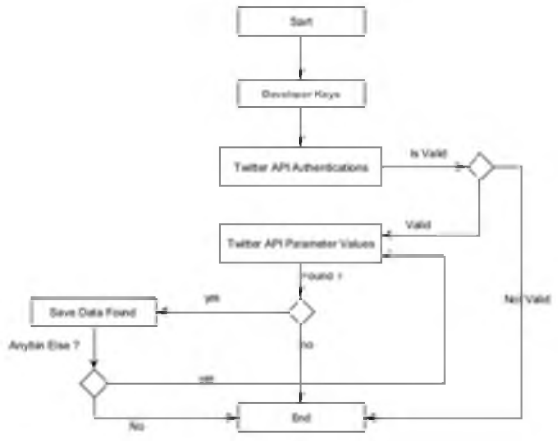
Python merupakan bahasa pemrograman tingkat tinggi. Hal ini disebabkan karena kode yang dituliskan akan di compile menjadi byte code dan dieksekusi sehingga Python cocok digunakan untuk scripting language, aplikasi web dan lain sebagainya. Hal lain yang menjadikan bahasa ini menjadi bahasa pemgrograman tingkat tinggi adalah Python dapat di extend kedalam bahasa C dan C++ serta bahasa pemgrograman ini memiliki struktur konstruksi yang kuat (blok kode, fungsi, class, modules, dan packages) dan serta konsisten menggunakan konsep Object Oriented Programming (OOP) (Kuhlman, 2015).

e. Split Validation

Split validation merupakan sebuah operator yang menerakan simple validation yakni memisahkan (split) data set secara acak kedalam sebuah data set training dan dataset testing dan evaluasi model yang terbentuk (Rapid Miner, 2014). Split validation juga melakukan validasi untuk mengukur performa dari algoritma atau metode yang digunakan. Pengukuran performa yang pada umumnya dilakukan mengukur seberapa tingkat akurasi sebuah model dalam menjalankan fungsinya.

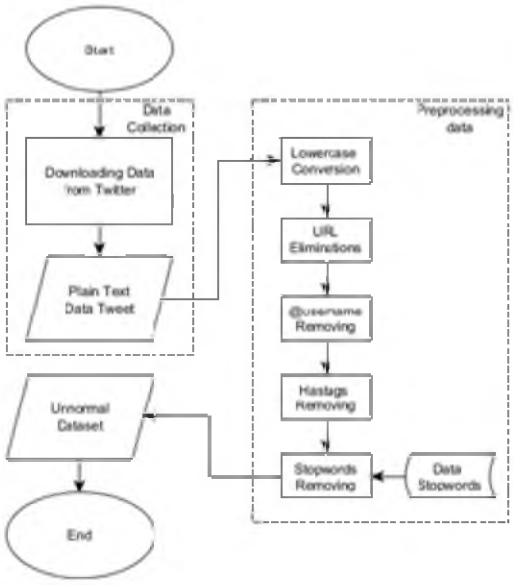
3. Hasil dan Pembahasan

Proses pertama yang dilakukan dalam melakukan klasifikasi sentimen adalah proses pengumpulan data. Pengumpulan data dilakukan dengan menggunakan Twitter API yang diimplementasikan pada bahasa pemrograman Pyhton. Topik yang dicari pada proses pengumpulan data ini adalah Presiden Joko Widodo dan Pemerintahannya. Proses pengumpulan data akan ditunjukkan pada gambar 1 berikut ini.



Gambar 1 Proses Pengumpulan Data

Setelah data terkumpul maka akan dilakukan proses preprocessing data. Pada proses preprocessing data terdapat 5 sub tahapan yang harus dilalui yaitu Lowercase Conversion, URL Eliminations, @username Removing, Hashtags Removing dan Stopwords Removing. Proses preprocessing data dan implementasi dalam bahasa pemrograman Python masing-masing akan ditunjukkan pada gambar 2 dan 3 berikut ini.



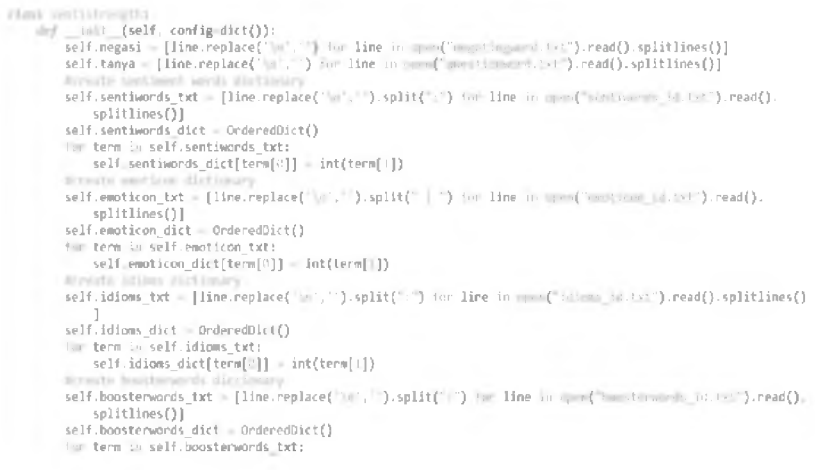
Gambar 2 Proses Preprocessing Data

```
def preprocessingTweet(tweet):
    tweet = tweet.lower()

    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', 'URL', tweet)
    tweet = re.sub('@[^\s]+' , 'AT_USER' , tweet)
    tweet = re.sub('[\s]+' , ' ' , tweet)
    tweet = re.sub(r'#([^\s]+)' , r'\1' , tweet)
    tweet = tweet.strip('\'"')
    return tweet
```

Gambar 3 Implementasi Preprocessing Data dalam bahasa pemrograman Python

Data yang didapatkan dan telah dilakukan preprocessing data yaitu sebesar 3485 baris. Data yang telah didapatkan ini akan selanjutnya akan dilakukan proses klasifikasi dengan menggunakan Naive Bayes. Sebelum melakukan klasifikasi data yang ada akan dibagi menjadi 50% untuk dilabeli terlebih dahulu. Data yang dilabeli akan digunakan sebagai acuan untuk melakukan proses klasifikasi dengan menggunakan Naive Bayes. Proses implementasi pelabelan data akan ditunjukkan pada gambar 4 berikut ini.



Gambar 4 Implementasi Proses Pelabelan Data

Proses klasifikasi dengan menggunakan Naive Bayes dilakukan dengan mempersiapkan data yang telah dilabeli sebagai referensi atau acuan untuk melakukan proses klasifikasi. Data yang telah dilabeli akan dilakukan proses *feature extraction* untuk mendapatkan kata-kata unik dari dari setiap kalimat yang ada. Data pada proses *feature extraction* akan digunakan untuk menentukan sentimen setiap tweet yang belum dilabeli. Implementasi dari proses klasifikasi dengan Naive Bayes akan ditunjukkan pada gambar 5 berikut ini.

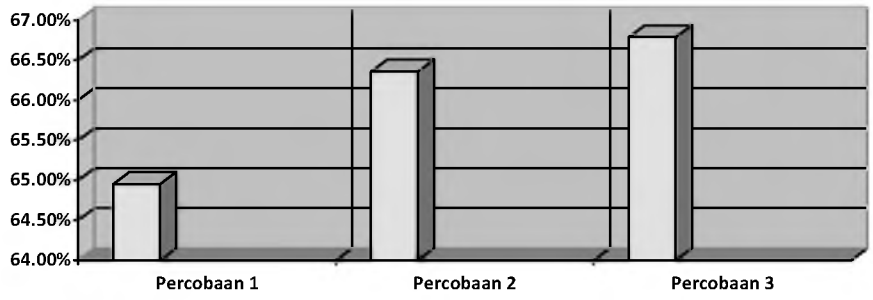
Gambar 5 Implementasi klasifikasi Naive Bayes pada pemrograman Python

Setelah semua data memiliki label, selanjutnya adalah proses untuk mengukur tingkat akurasi dari proses klasifikasi yang telah dilakukan. Proses ini akan dilakukan oleh RapidMiner. Adapun skenario pengujian untuk mengukur tingkat akurasi akan ditunjukkan pada tabel 1 berikut ini.

Tabel 1 Skenario Uji Coba

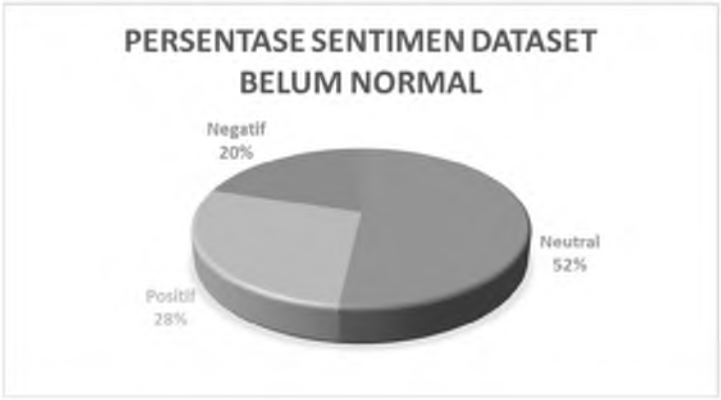
No	Number of Trials	Persentase pembagian Data	
		Training	Testing
1	1	40%	60%
2	2	50%	50%
3	3	60%	40%

Adapun hasil ujicoba berdasarkan tabel 1 akan ditunjukan pada gambar 6 berikut ini.



Gambar 6 Tingkat Akurasi Proses Klasifikasi

Pada gambar 6 dapat dijelaskan bahwa pada percobaan pertama tingkat akurasi yang dihasilan yaitu sebesar 64.95%. Pada percobaan kedua tingkat akurasi yang dihasil adalah sebesar 66.36% dan pada percobaan ketiga tingkat akurasi yang dihasilkan adalah *sebesar* 66.79%. Hasil lain yang akan ditunjukkan pada penelitian ini adalah besar persentase untuk masing-masing kelas sentiment berdasarkan hasil klasifikasi. Kelas yang dihasilkan pada penelitian ini adalah 3 kelas yaitu kelas positif, negatif dan neutral. Adapun besar persentase untuk masing-masing kelas sentimen akan ditunjukkan pada gambar 7 berikut ini.



Gambar 7 Besar persentase Kelas Sentimen

Pada gambar 7 dapat dijelaskan bahwa kelas sentimen neutral mendapat nilai terbesar dengan 52%. Selanjutnya yaitu kelas sentimen positif dengan 28% dan 20% untuk kelas sentimen negatif. Berdasarkan gambar 7 dapat disimpulkan bahwa sentiment pengguna Twitter terhadap presiden Joko Widodo dan Pemerintahannya yaitu bersentimen neutral.

4. Kesimpulan

Berdasarkan hasil pembahasan dapat disimpulkan bahwa tingkat akurasi terbesar apabila didasarkan pada skenario uji coba yaitu pada percobaan ketiga dengan tingkat *akurasi* sebesar 66.79% dilanjutkan oleh ujicoba kedua dengan 66.36% dan uji coba pertama denga 64.95%. Berdasarkan hasil klasifikasi besar persentase untuk ketiga kelas, ditempat pertama yaitu dengan kelas sentimen neutral dengan 52%, kedua sentimen positif dengan 28% dan ketiga sentimen negatif dengan 20%. Berdasarkan hasil persentase kelas sentimen, sentimen neutral merupakan sentimen yang paling banyak apabila dikaitkan dengan topik Presiden Joko Widodo dan Pemerintahannya.

Daftar Pustaka

Twitter. 2017. About Twitter. Twitter. [Online] September 28, 2017. https://about.twitter.com/.

NLTK Project. 2017. www.nltk.org. www.nltk.org. [Online] November 02, 2017. http://www.nltk.org/.

Manning, Christopher and Schuetze, Hinrich. 1999. Foundations of Statistical Natural Language Processing. London : MIT Press, 1999.

Pozzi, Federico Alberto, et al. 2017. Sentiment Analysis in Social Networks. Cambridge : Todd Green, 2017

Kuhlman, Dave. 2015. A Python Book: Beginning Python, Advanced Python and Python Exercises. s.l.: MIT, 2015.

Rapid Miner. 2014. Operator Refence Manual. Boston: Rapid Miner Inc., 2014.