IDENTIFIKASI KARAKTER PRESIDEN MELALUI ANALISIS SENTIMEN PADA TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER DAN POS TAGGING

Bastomy[1], Anisa Herdiani[2], Indra Lukmana Sardi[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]bastomyy@student.telkomuniversity.ac.id, [2]anisaherdiani@telkomuniversity.ac.id,
[3]indraluk@telkomuniversity.ac.id

Abstrak

Media sosial merupakan salah satu media untuk menyampaikan opini tentang politik, salah satunya tentang presiden. Opini atau pandangan masyarakat dapat berupa opini positif dan negatif. Menurut KBBI karakter merupakan kata sifat, dengan demikian opini yang menjadi fokus utama adalah opini yang memiliki jenis kata sifat. Kita dapat meneliti sentimen yang terdapat pada Twitter berupa *tweet* yang menandai presiden untuk mendapatkan karakter. Inti dari penelitian ini menggunakan metode Naïve Bayes Classifier (NBC) untuk mengklasifikasikan *tweet* dan POS *tagging* untuk mengetahui jenis kata dari setiap *tweet* positif. Dari pengujian yang telah dilakukan menghasilkan *pre-processing* seperti *casefolding*, *tokenizing*, penghapusan kata yang tidak memiliki makna, simbol atau tanda baca. Untuk meningkatkan akurasi NBC digunakan metode N-gram yang bertujuan menggabungkan kata negasi dan kata selanjutnya untuk menghindari perubahan makna dari kata tersebut. Hasil pengujian klasifikasi dengan menggunakan metode cross-validation menghasilkan rata-rata akurasi sebesar 80,29% dan POS *tagging* menghasilkan akurasi sebesar 73,3% dalam menentukan karakter presiden. pengujian di atas menunjukkan bahwa identifikasi karakter melalui analisis sentimen menggunakan NBC dan POS *tagging* dapat digunakan untuk mendapatkan karakter presiden. Hasil akhir penelitian ini berupa daftar kata berjenis kata sifat yang telah diurutkan berdasarkan polaritas kemunculannya yang telah divalidasi oleh ahli Bahasa.

Kata kunci : Analisis sentimen, Naïve Bayes Classifier, N-gram, POS tagging, *Preprocessing*, Twitter

Abstract

Social media is one of the media to express opinions about politics, one of which is about the president. Public opinion or opinion can take the form of positive and negative opinions. According to KBBI characters are adjectives, thus opinions that are the main focus are opinions that have the type of adjectives. We can consider the sentiments on Twitter in the form of tweets needed by the president to get the character. The core of this research uses the Naïve Bayes Classifier (NBC) method to classify tweets and POS markings to understand the type of words of each positive tweet. From the testing that has been done, it produces pre-processing such as casefolding, tokenizing, deletion of words that do not have meaning, symbols or punctuation. To improve the accuracy of NBC the N-gram method is used which replaces the negation words and subsequent words to avoid changing the meaning of the word. Test results using the cross-validation method produce an average accuracy of 80.29% and POS marking produces an accuracy of 73.3% in determining the character of the president. Learn above How to examine characters through sentiment analysis using NBC and POS marking can be used to get the president's character. The final results of this study contain a list of adjective type words that have been sorted based on the polarity of their appearance which has been validated by language experts.

Keywords: Naïve Bayes Classifier, N-gram, POS tagging, *Preprocessing*, *Sentiment analysis*, Twitter

1. Pendahuluan

Latar Belakang

Presiden adalah seorang pemimpin negara yang memiliki dua peran utama yaitu sebagai kepala negara dan kepala pemerintahan. Presiden memiliki banyak tugas yaitu menetapkan peraturan pemerintahan, mengangkat menteri, pengesahan rancangan undang-undang dan menetapkan kebijakan lainnya yang berhubungan dengan kepentingan negara. Kebijakan yang diambil tak lepas dari pro dan kontra terutama kebijakan yang menyangkut kepentingan masyarakat secara luas, hal ini akan menimbulkan penilaian dari masyarakat terkait kinerja seorang presiden. Penyampaian penilaian masyarakat yang berhubungan dengan presiden bisa melalui media cetak, media masa dan media internet.

Media internet merupakan media komunikasi yang dapat menyebar luaskan informasi secara luas dan cepat, salah satu komunikasi melalui media internet adalah media sosial. Media sosial merupakan media yang mudah dan murah yang digunakan masyarakat untuk menyampaikan suatu opini dan juga penilaian terhadap suatu kebijakan yang diambil. Hal inilah yang membuat setiap kebijakan yang diambil akan mendapatkan respons yang cepat dari masyarakat baik itu respons positif atau negatif. Kata sifat yang berasal dari *tweet* positif hasil klasifikasi akan menjadi tolak ukur penilaian masyarakat terkait karakter presiden.

Pada penelitian ini media sosial yang dimaksud adalah twitter. Twitter adalah layanan *microblogging* yang memiliki fitur membaca dan menulis status, status yang dapat ditulis maksimal memiliki 140 karakter, status bisanya juga disebut *tweet* atau kicauan [1]. Pada *twitter* juga memungkinkan pengguna Menggunakan tanda "@" untuk membalas status orang lain dan tanda "#" untuk *hashtag* atau topik yang dibahas.

Penelitian ini akan berfokus pada analisis sentimen terkait presiden Indonesia dengan menganalisis tweet-tweet yang me-*mention* presiden sekarang atau sebelumnya, tetapi dikarenakan keterbatasan data yang dapat di ambil maka penelitian ini hanya akan mengambil data presiden Joko Widodo dan Susilo Bambang Yudhoyono.

Analisis sentimen adalah bagian dari Natural Language Processing (NLP) yang berkembang dengan penelitian mulai dari tingkat klasifikasi dokumen [2], analisis sentimen sendiri merupakan proses mengidentifikasi dan mengekstrak data sentimen yang akan dikategorikan berdasarkan polaritasnya, apakah itu positif atau negatif [3]. Sentimen terhadap presiden akan memberikan informasi tentang pandangan masyarakat terhadap presiden dan dari hasil itu sentimen positif akan dilakukan proses POS *tagging* untuk mengetahui jenis kata sifat untuk mendapatkan kriteria presiden.

Untuk melakukan klasifikasi sentimen digunakan metode Naïve Bayes Classifier metode ini memiliki akurasi sebesar 80,29%% [4] dan untuk POS *tagging* menggunakan CRF *tagger* menggunakan korpus yang telah di tag dan memiliki akurasi sebesar 73,3%. Penelitian terkait sebelumnya pernah dilakukan oleh Fam Rashel [5]. POS *tagging* yang memiliki jenis kata sifat akan di hitung polaritasnya yang kemudian akan divalidasi oleh ahli Bahasa untuk mendapatkan karakter presiden.

Topik dan Batasannya

Berdasarkan penjelasan latar belakang di atas maka didapat beberapa rumusan masalah, yaitu bagaimana mendapatkan model klasifikasi sentimen menggunakan Naïve Bayes Classifier pada twitter dan bagaimana mendapatkan karakter presiden berdasarkan klasifikasi sentimen.

Batasan masalah pada penelitian ini adalah data yang digunakan hanya dari twitter. Data yang digunakan berupa tweet yang di ambil dari 12 Januari 2019 sampai dengan 17 Februari 2019. Data yang digunakan hanya yang *memention* akun Joko Widodo dan Susilo Bambang Yudhoyono yaitu @jokowi dan @SBYudhoyono.

Tujuan

Dengan adanya masalah yang disebutkan, maka tujuan penelitian ini adalah mendapatkan model klasifikasi sentimen menggunakan Naïve Bayes Classifier pada twitter. Mendapatkan karakter presiden berdasarkan sentimen masyarakat terhadap presiden Joko Widodo dan Susilo Bambang Yudhoyono.

2. Studi Terkait

2.1 Naive Bayes Classifier

Naïve Bayes Classifier (NBC) merupakan salah satu metode klasifikasi dengan tolak ukur probabilitas. Teorema *naïve bayes* ditemukan oleh ilmuan inggris Thomas Bayes pada abad ke 18,yaitu untuk memprediksi masa depan berdasarkan pengalaman masa lalunya. Naïve bayes dapat mengklasifikasikan sebuah data dengan menggunakan probalitas, metode ini terkenal dengan sebutan Naïve Bayes Classifier. Metode ini memiliki akurasi yang tinggi yaitu sekitar 81% [4].

Naïve bayes classifier memiliki ciri utama yaitu setiap atribut tidak memiliki pengaruh terhadap atribut lainnya atau dengan kata lain setiap atributnya independen. Metode NBC memiliki dua tahap yaitu *training* dan klasifikasi. Teorema NBC bisa dinyatakan sebagai berikut [1] [6].

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (1)$$

Keterangan:

- A = kelas-kelas hasil klasifikasi
- B = sampel data yang tidak diketahui labelnya
- P(B|A) = probabilitas terjadinya B jika A diketahui, disebut *likelihood function*, karena peluang B tergantung dengan peluang data *sample* A.
- P(A|B) = probabilitas terjadinya A jika B diketahui. Disebut probabilitas posterior, karena peluang A bergantung dari nilai B tertentu.
- P(A) = probabilitas prior A, dan bertindak sebagai *normalizing constant*. Secara intuitif , teorema Bayes menggambarkan bahwa perubahan pada "B" dapat diamati apabila "A" terlebih dahulu diamati.
- P(B) = probabilitas B merupakan probabilitas dari *sample* yang mempunya kelas B.

2.2 POS Tagging

Part of speech (POS) atau kategori tata bahasa adalah pemberian jenis kata seperti kata benda, kata sifat, kata kerja dan lainnya ke sebuah kata dalam kalimat [5]. Terdapat dua pendekatan yaitu *rule-based* dan *stochastic*, pada penelitian ini digunakan pendekatan *rule-based*, korpus berasal dari POS Tag Indonesia yang memiliki sekitar 250.000 *token* [7]. Daftar jenis kata yang dihasilkan dapat dilihat pada tabel 1.

Tabel 1 Jenis kata

Tag	Deskripsi	Contoh
CC	Coordinating conjunction	dan, tetapi, atau
CD	Cardinal number	dua, juta, enam, sepertiga, banyak, kedua, ribuan
OD	Ordinal number	ketiga, ke-4, pertama
DT	Determiner / article	Para, Sang, Si
FW	Foreign word	climate change, terms and conditions
IN	Preposition	dalam, dengan, di, ke, oleh, pada, untuk
JJ	Adjective	bersih, panjang, hitam, lama, jauh, marah, suram, nasional, bulat
MD	Modal and auxiliary verb	boleh, harus, sudah, mesti, perlu
NEG	Negation	tidak, belum, jangan
NN	Noun	monyet, bawah, sekarang, rupiah
NNP	Proper noun	Boediono, Laut Jawa, Indonesia, India, Malaysia, Bank Mandiri, BBKP, Januari, Senin, Idul Fitri, Piala Dunia, Liga Primer, Lord of the Rings: The Return of the King
NND	Classifier, partitive, and measurement noun	orang, ton, helai, lembar
PR	Demonstrative pronoun	ini, itu, sini, situ
PRP	Personal pronoun	saya, kami, kita, kamu, kalian, dia, mereka
RB	Adverb	sangat, hanya, justru, niscaya, segera
RP	Particle	pun, -lah, -kah
SC	Subordinating conjunction	sejak, jika, seandainya, supaya, meski, seolah-olah, sebab, maka, tanpa, dengan, bahwa, yang, lebih ... daripada ..., semoga
SYM	Symbol	IDR, +, %, @
UH	Interjection	brengsek, oh, ooh, aduh, ayo, mari, hai
VB	Verb	merancang, mengatur, pergi, bekerja, tertidur
WH	Question	siapa, apa, mana, kenapa, kapan, di mana, bagaimana, berapa
X	Unknown	statemen
Z	Punctuation	"...", ?, .

Berdasarkan KBBI karakter merupakan jenis kata sifat untuk itu jenis kata yang digunakan adalah jenis kata JJ (kata sifat). Kata sifat ini berasal dari *tweet* positif yang sebelumnya telah di klasifikasi, hasil dari klasifikasi selanjutnya dilakukan POS *tagging* untuk mendapatkan jenis kata sifat. Kata yang memiliki jenis kata sifat akan dihitung polaritas kemunculannya. Karakter akan dilihat dari nilai polaritasnya, semakin tinggi kemunculannya maka semakin besar kemungkinan kata tersebut menjadi karakter presiden dan begitu sebaliknya.

2.3 Measuring Performance

Measuring performance merupakan sebuah metode statistika yang mengevaluasi dan membandingkan algoritma *learning* dengan membagi data menjadi dua bagian yaitu data latih dan data validasi. Data latih akan digunakan untuk pembelajaran sebuah model sedangkan data validasi digunakan untuk memvalidasi model tersebut. *Cross validation* digunakan untuk mendapatkan performansi dari model tersebut dan untuk

mendapatkan fleksibilitas (model *selection*) dari model tersebut. Untuk menghitung nilai akurasi dapat menggunakan *confusion matrix* [8]. Bentuk umum pada *confusion matrix* dapat dilihat pada tabel 2**.**

Tabel 2 Confusion Matrix

	Relevant	Not-relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not-Retrieved	False Negative (FN)	True Negative (TN)

Precission adalah jumlah *tweet* yang dengan benar telah diklasifikasikan ke dalam sebuah kelas dibagi dengan jumlah total *user* yang diklasifikasikan dalam kelas tersebut [8]. *Precision* juga sering disebut dengan ketepatan antara informasi yang diminta oleh pengguna dengan hasil yang diberikan oleh sistem. Berikut adalah rumus *precision*:

$$Precision = \frac{TP}{(TP+FP)} * 100\% \quad (2)$$

Recall adalah jumlah *tweet* yang dengan benar diklasifikasikan dalam sebuah kelas dibagi dengan jumlah total *tweet* dalam kelas tersebut [8]. *Recall* juga sering disebut dengan tingkat keberhasilan sistem dalam menemukan kembali informasi. Berikut rumus *recall*:

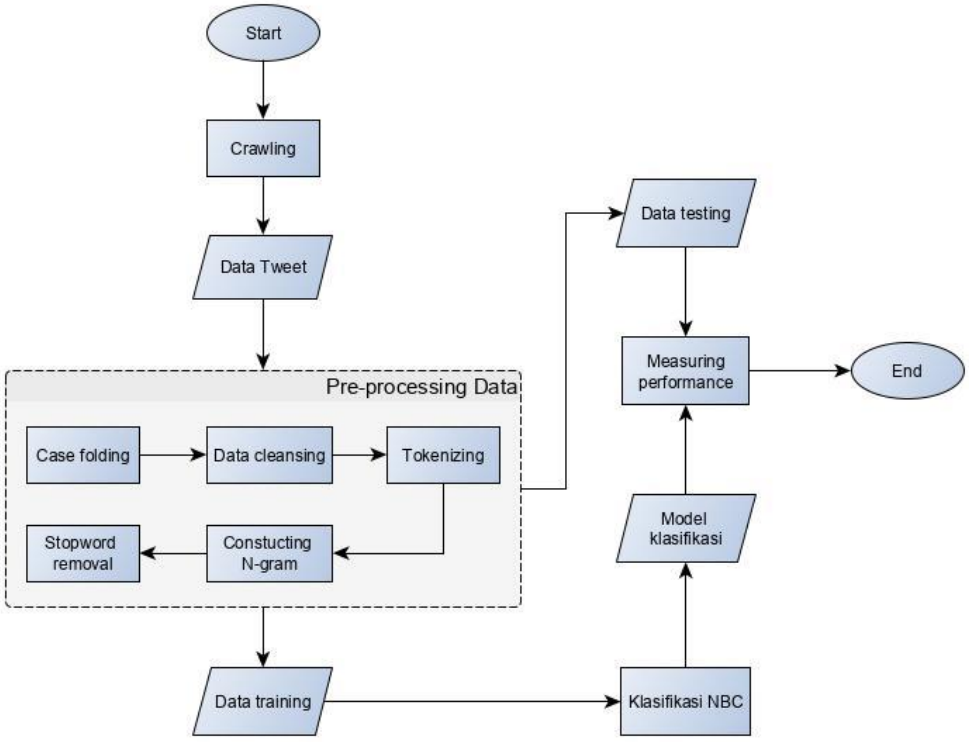
$$Recall = \frac{TP}{(TP+FN)} * 100\% \quad (3)$$

Akurasi merupakan parameter evaluasi terhadap sistem yang dibangun dalam penelitian ini. Akurasi sering disebut dengan tingkat kedekatan antara prediksi dengan nilai yang sebenarnya. Berikut rumus akurasi.

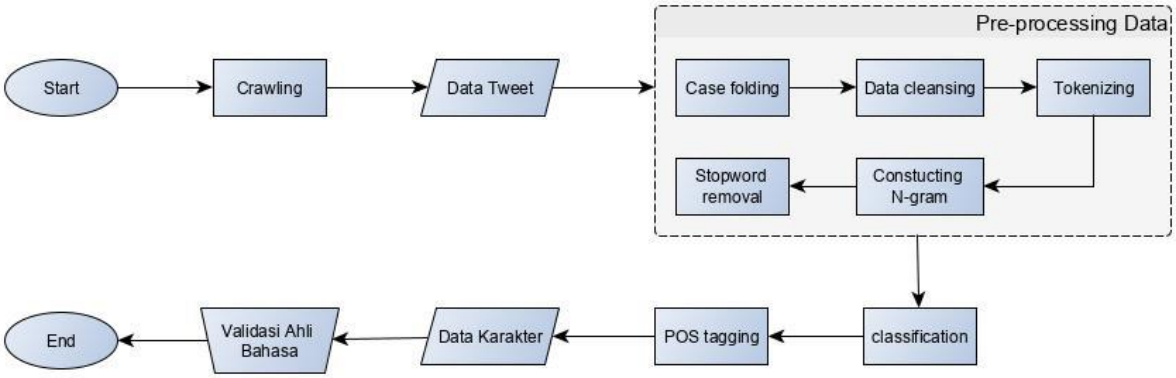
$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (4)$$

3. Sistem yang Dibangun

Sistem yang akan di bangun pada tugas akhir ini adalah sebuah sistem dengan tujuan untuk mengetahui sentimen masyarakat terhadap Joko Widodo dan Susilo Bambang Yudhoyono berdasarkan *tweet* yang di miliki pengguna. Setiap tweet akan di klasifikasikan menjadi dua bagian yaitu positif dan negatif yang kemudian hasil dari sentimen positif tersebut akan di lakukan POS tagging untuk mendapatkan jenis kata sifat dimana jenis kata tersebut akan menjadi karakter yang diharapkan oleh masyarakat. Untuk mendapatkan data *tweet* akan diperoleh dari API yang di sediakan *twitter*, selanjutnya data *twitter* akan melalui tahap *pre-proccesing data* sebelum di klasifikasi menggunakan metode NBC. Data keluaran yang dihasilkan adalah kata sifat karakter presiden yang diharapkan oleh masyarakat. Berikut merupakan alur rancangan sistem klasifikasi dan POS tagging yang dilakukan pada penelitian ini dapat dilihat pada gambar 1 dan gambar 2.



Gambar 1 Flowchart Klasifikasi Sentimen



Gambar 2 Blok Diagram POS tagging

3.1 Crawling

Crawling data adalah program yang bekerja dengan metode tertentu untuk mengumpulkan informasi secara otomatis yang ada dalam sebuah penyimpanan data. Pada penelitian ini data yang akan di ambil pada tahap *crawling data* adalah *twitter*. Data yang diperoleh dari Twitter berupa data *user* dan *tweet*, untuk memperkecil pencarian data *tweet* difokuskan pada *tweet* yang me-*mention* akun presiden seperti yang telah ditentukan pada batasan masalah. Dengan menggunakan API *twitter* yang telah di terapkan pada pemrograman *python* data *twitter* akan diperoleh dengan cara mengunduh secara otomatis dari parameter yang telah ditentukan [9]. Keluaran pada tahap ini adalah data *user* dan *tweet*. Contoh data *tweet* dapat dilihat pata tabel 3.

Tabel 3 Data Tweet

ID tweet	Id User	Username	Tweet	created at
1097157291147649025	57328078	echaluceria	Sebar hoax tanda tak mampu ..\n Dukung pakdhe @jokowi karena dia mampu membangun Indonesia jadi lebih baik lagi	17-02-2019 15:34:17

			\xf0\x9f\x98\x8d\xf0\x9f\x91\x8d\n\n @iAmLogan95\n@KarinaArdiana\n @permadiaktivis\n@RizmaWidiono\n\n #DebatPintarJokowi https://t.co/OzjzPWsGxl	

3.2 Pre-processing data

Tahap pre-processing data merupakan tahap yang bertujuan untuk mempersiapkan data sebelum di olah ke tahap klasifikasi. Pada tahap ini dilakukan *case folding, data cleansing, tokenizing, construc n-grams,* dan *stopword removal*. Data akan dilakukan *cleansing* yang bertujuan untuk menghapus data *tweet* yang tidak digunakan pada penelitian seperti *link, emoticon, number, special character* kecuali `#` karena sebagai simbol *hastag*. Hastag tidak di hapus dikarenakan meningkatkan nilai akurasi dari klasifikasi. Pada tahap data cleansing dilakukan penghapusan ekpresi tertawa seperti `haha`,`hehe`,`hoho`,`wkwk` dan kombinasi dari keempat ekpresi tertawa tersebut, seperti `he` dan `hehehe`. Penerapan *pre-processing* dapat dilihat pada tabel 4.

Tabel 4 Pre-processing

Tahap	Hasil
Casfolding	sebar hoax tanda tak mampu haha..dukung pakdhe @jokowi karena dia mampu membangun indonesia jadi lebih baik lagi \xf0\x9f\x98\x8d\xf0\x9f\x91\x8d @iamlogan95@karinaardiana@permadiaktivis@rizmawidiono#debatpintarjokowi https://t.co/ozjzpwsgxl
Data Cleansing	sebar hoax tanda tak mampu dukung pakdhe karena dia mampu membangun indonesia jadi lebih baik lagi #debatpintarjokowi
Tokenizing	['sebar', 'hoax', 'tanda', 'tak', 'mampu', 'dukung', 'pakdhe', 'mampu', 'membangun', 'indonesia', '#debatpintarjokowi']
N-gram	['sebar', 'hoax', 'tanda', 'tak mampu', 'dukung', 'pakdhe', 'mampu', 'membangun', 'indonesia', '#debatpintarjokowi']
Stopword removal	['sebar', 'hoax', 'tanda', 'tak mampu', 'dukung', 'mampu', 'membangun', 'indonesia', '#debatpintarjokowi']

3.3 Klasifikasi NBC

Naïve Bayes Classifier pada penelitian ini berfungsi untuk menghitung probabilitas bersyarat pada setiap atribut(kata) dari setiap kelas. Tahapan yang akan dilakukan pada proses klasifikasi adalah dengan cara *bag of word* yang bertujuan untuk mengumpulkan kata yang ada pada setiap *tweet* berdasarkan frekuensi kemunculannya di *tweet* tersebut. Setelah tahap *bag of word* tahap selanjutnya adalah melakukan *training* untuk mendapatkan model dalam bentuk probabilitas yang akan digunakan. Untuk menghindari peluang 0 yang ada di data *training* dan merusak hasil *testing* maka dibutuhkan tahap *laplace smoothing*. *Laplace smoothing* adalah menambahkan angka 1 dibagi dengan semua fitur yang ditambahkan ke semua fitur sehingga tidak ada yang nilainya 0 [10] . setelah mendapatkan model probabilitas dari data *training* selanjutnya lakukan pengujian model dengan menggunakan data yang berbeda untuk menguji model yang telah dihasilkan. Berikut contoh penerapan NBC dapat dilihat pada tabel 5.

Tabel 5 Dataset

No	Tweet	Label
1	Jokowi kerja ikhlas demi Indonesia maju karena jujur dalam bekerja	Positif
2	Pembangunan saja terus rakyat kecil tidak diperhatikan hanya pro pengusaha	Negatif
3	Jokowi kerja ikhlas jujur demi rakyat	-

Conditional probability

- $P(jokowi\ |positif) = \frac{1+1}{10+29}$ = 0.068
- $P(kerja\ |positif) = \frac{1+1}{10+19}$ = 0.068
- $P(ikhlas\ |positif) = \frac{1+1}{10+19}$ = 0.068
- $P(jujur\ |positif) = \frac{1+1}{10+19}$ = 0.068
- $P(jokowi\ |negatif) = \frac{1}{9+19}$ = 0.035
- $P(kerja\ |negatif) = \frac{1}{9+19}$ = 0.035
- $P(ikhlas\ |negatif) = \frac{1}{9+19}$ = 0.035
- $P(jujur\ |negatif) = \frac{1}{9+19}$ = 0.035

- $P(demi\ |positif) = \frac{1+1}{10+19} = 0.068$
- $P(rakyat\ |positif) = \frac{1}{10+10} = 0.034$
- $P(demi\ |negatif) = \frac{1}{9+19} = 0.035$
- $P(rakyat\ |negatif) = \frac{1+1}{9+10} = 0.071$

Probability

- $P(Positif|d3) = 0.5 * (0.068)^5 * 0.034 = 0.00024$
- $P(negatif|d3) = 0.5 * (0.035)^5 * 0.071 = 0.00018$

Prior

- $P(positif) = 1/2 = 0.5$
- $P(negatif) = 1/2 = 0.5$

Dari data di atas terdapat dua contoh data yang telah diberi label dan satu data yang belum memiliki label. Klasifikasi dilakukan dengan cara menghitung setiap probabilitas bersyarat untuk setiap kelas, setiap probabilitas akan dikalikan dengan *prior* sesuai dengan labelnya. Label ditentukan oleh besarnya nilai probabilitas setiap kelas, pada perhitungan di atas menghasilkan nilai probabilitas positif lebih besar maka data ke tiga akan diklasifikasikan ke label positif.

3.4 Measuring Performance

Data hasil klasifikasi positif dan negatif menggunakan metode NBC akan diuji akurasinya berserta nilai *precision* dan *recall*-nya dengan metode Cross Validation dengan nilai 10 fold [8]. Pengujian 10 fold bertujuan untuk mendapatkan nilai akurasi yang optimal akan di ambil rata-ratanya dari setiap *fold.* Hal ini berarti membagi data *training* ke 10 bagian, setiap bagian akan di *cross* ke dalam data *training* dan testing secara bergantian. Missal pada *fold* 1, bagian 1 digunakan sebagai testing dan 9 bagian sisanya digunakan sebagai *training* dan begitu seterusnya hingga semua bagian teruji.

3.5 POS Tagging

Data yang telah di klasifikasi selanjutnya akan dilakukan POS *tagging* yaitu memberi *tagging* jenis kata pada setiap *tweet* hasil klasifikasi. Tweet yang akan di proses yaitu yang memiliki nilai positif. Pemberian *tag* jenis kata bertujuan untuk mendapatkan karakter presiden. Karakter merupakan jenis kata sifat. Setiap kata yang memiliki jenis kata sifat akan dikalkulasikan untuk menghitung polaritas kemunculannya. Semakin besar nilai kemunculannya maka semakin besar peluang kata tersebut merupakan karakter presiden yang di harapkan. Pada tahap ini menghasilkan daftar karakter yang telah terurut berdasarkan polaritasnya yang kemudian akan di validasi oleh ahli Bahasa untuk mendapatkan hasil akhir karakter presiden.

Untuk membangun POS *tagging* digunakan dataset yang diambil dari penelitian sebelumnya yaitu sebanyak 250.000 *token* yang telah diberi label [7]. Data tersebut digunakan sebagai data *training* untuk membangun *tagger* dengan menggunakan *library* CRF (Conditional Random Field model) yang disediakan oleh NLTK (Natural Language Toolkit). Berikut contoh hasil dari POS tagging dapat dilihat pada tabel 6.

Tabel 6 POS tagging

Tahap	Hasil
Input	kerja ikhlas demi Indonesia maju, karena jujur dalam bekerja
Tokenizing	[‘kerja’, ’ikhlas’, ‘Indonesia’ ,’maju’, ’jujur’ ,’bekerja’]
Tagging	('kerja', 'NN'), ('ikhlas', JJ), ('Indonesia', 'NNP'), ('maju', 'VB'), (`jujur`, 'JJ') , (`bekerja`, 'VB')

Hasil dari proses POS tagging adalah kata yang telah diberi jenis kata, jenis kata yang digunakan hanya yang memiliki jenis kata sifat. Setiap kata akan dikalkulasikan polaritas kemunculannya untuk diurutkan berdasarkan kemunculannya. Tahap ini bertujuan untuk mendapatkan karakter presiden yang berasal dari polaritas paling tinggi, setelah mendapatkan daftar kata dilakukan validasi oleh ahli Bahasa untuk menentukan jenis kata sifat mana yang termasuk kedalam karakter.

4. Evaluasi

4.1 Data Acuan

Data acuan yang digunakan sebagai data model merupakan data *tweet* yang diperoleh dari tanggal 12 Januari 2019 sampai dengan 17 Februari 2019 yaitu sebanyak 3.995 data *tweet.* Data telah di

beri kelas dengan teknik *labeling* secara manual dengan total 2253 positif dan 1742 negatif. Selain data untuk membangun model terdapat juga data *tweet* untuk mendapatkan karakter presiden yaitu sebanyak 584.851 *tweet*. Adapun jumlah *tweet* yang digunakan terdiri dari 534.937 *tweet* mengenai Joko Widodo dan 49.914 *tweet* mengenai Susilo Bambang Yudhoyono.

Data *tweet* berupa teks yang diperoleh merupakan hasil *crawling* yang diambil dari pengguna Twitter. Data tersebut merupakan *tweet* yang me-*mention* salah satu akun presiden seperti pada batasan masalah yang telah dijabarkan di atas. Data *tweet* diambil dengan menggunakan API yang disediakan Twitter dengan di implementasikan ke pemrograman *python* menggunakan *library tweepy*. Data hasil *tweet* disimpan dalam bentuk file CSV.

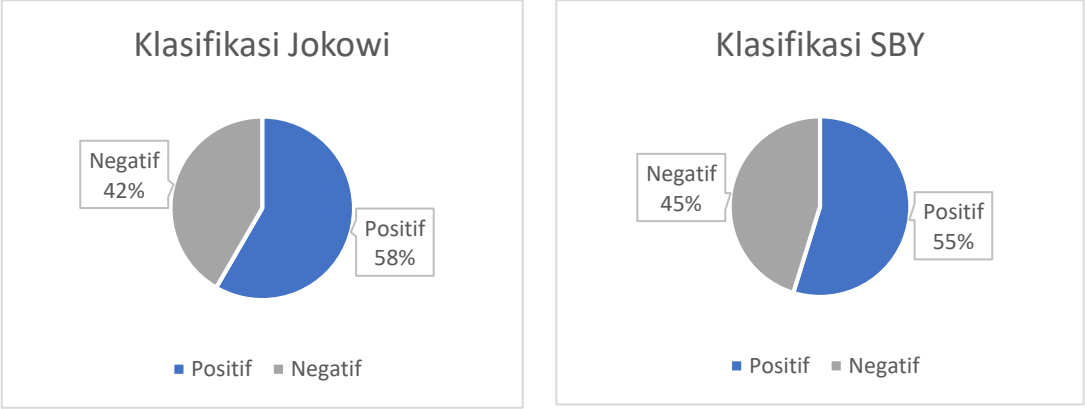
4.2 Pengujian Sistem

Untuk mengukur akurasi pada penelitian ini digunakan persamaan (4). Akurasi digunakan untuk mengukur ketepatan klasifikasi yang dilakukan dengan *classifier* yang telah dibangun. *Classifier* dibangun menggunakan data yang telah diberi label yaitu sebanyak 3.995 tweet yang akan menjadi model. Metode yang digunakan untuk menghitung akurasi menggunakan *cross-validation* dengan nilai 10-fold, metode ini akan membagi data menjadi 10 bagian, setiap bagian akan dilakukan *cross* sebagai data *training* dan testing. Misal data bagian 1 merupakan data testing maka 9 bagian data lainnya akan digunakan sebagai *training*, begitu seterusnya hingga seluruh data dilakukan pengujian. Hasil dari 10 perulangan pengujian tersebut akan diambil rata-ratanya. Selain mencari nilai akurasi pada metode ini juga dilakukan pencarian nilai *precision* dan *recall*. Hasil lengkap dari pengujian sistem dapat dilihat pada tabel 7.

Tabel 7 Measuring Performance

	Precision	Recall	Akurasi
Rata-rata	83,93 %	81,03 %	80,29 %

Data yang akan digunakan sebagai identifikasi karakter berasal dari data tweet yang mention akun Jokowi dan SBY yang telah diklasifikasi menggunakan model yang telah dibangun. Hasil dari klasifikasi tersebut menghasilkan 337.963 *tweet* positif dan 247.076 negatif. Adapun detail dari data *tweet* terdiri dari 312.253 sentimen positif untuk Joko Widodo dan 27.345 sentimen positif untuk Susilo Bambang Yudhoyono. Data yang akan digunakan hanya yang memiliki nilai positif untuk lanjut ke tahap *tagging*. Berikut adalah diagram hasil klasifikasi sentimen karakter presiden Joko Widodo dan Susilo Bambang Yudhoyono yang dapat dilihat pada gambar 3 atau pada lampiran 1 untuk lebih detailnya.



Gambar 3 Persentase hasil klasifikasi sentimen

Data sentimen positif selanjutnya akan diberi *tagging* dengan metode POS *tagging*. Data yang akan menjadi karakter presiden adalah kata yang memiliki jenis kata sifat (JJ) yang telah divalidasi oleh ahli Bahasa. Berdasarkan hasil validasi, POS *tagging* menghasilkan akurasi sebesar 73,3% dalam menentukan karakter dengan 22 data sebagai karakter dari 30 data teratas yang telah diurutkan berdasarkan kemunculannya. Berikut hasil validasi karakter dapat dilihat pada tabel 8 atau pada lampiran 2 untuk lebih detailnya.

Tabel 8 Validasi Karakter

No	Kata	Kemunculan	Karakter (ya/tidak)	No	Kata	Kemunculan	Karakter (ya/tidak)
1	Maju	8443	Ya	16	Damai	1078	Ya
2	Sehat	7304	Tidak	17	Bersih	1032	Ya
3	Terbaik	3296	Ya	18	Rendah	1000	Ya
4	Cepat	3089	Ya	19	Tulus	962	Ya
5	Sukses	3049	Tidak	20	Positif	893	Ya
6	Bagus	2343	Ya	21	Tenang	634	Ya
7	Keras	2235	Ya	22	Setia	558	Ya
8	Jujur	1836	Ya	23	Patut	556	Ya
9	Layak	1592	Tidak	24	Ikhlas	496	Ya
10	Kuat	1561	Ya	25	Kreatif	404	Tidak
11	Akal	1418	Ya	26	Konsisten	328	Ya
12	Sosial	1379	Ya	27	Produktif	297	Tidak
13	Adil	1358	Ya	28	Netral	282	Tidak
14	Berani	1229	Ya	29	Sepuh	282	Tidak
15	Bebas	1084	Ya	30	Maksimal	266	Tidak

4.3 Analisis

Pre-processing merupakan tahap paling penting dalam penelitian ini. Tweet yang diperoleh memiliki data yang tidak beraturan seperti banyaknya tanda baca, *link*, *emoticon* dan kata yang disingkat. Data *training* juga memiliki peran yang penting, semakin besar data *training* yang digunakan maka semakin besar akurasi yang dihasilkan. Hal ini disebabkan semakin besar data *training* maka semakin banyak kata yang terbentuk dan akan mempengaruhi hasil klasifikasi. Hasil pengujian dapat dilihat pada tabel 9.

Tabel 9 Hasil Pengujian

No	Penggunaan Data	Selisih	Akurasi
1	Hastag	+ 6,1 %	78.07 %
2	Construct N-gram	+ 2 %	73.97 %
3	Stemming	- 1.4 %	70.57 %
4	Data Cleansing	+ 0.22 %	72.19 %
5	Hashtag, Construct N-Gram, Data Cleansing	+ 8.32 %	80.29 %

Pemilihan metode POS tagging menjadi penting karena dibutuhkan kecepatan dan ketepatan dalam tagging mengingat jumlah data yang diproses sangat banyak. Beberapa pengujian dilakukan pada tahap ini. Berikut pengujian POS tagging dapat dilihat pada tabel 9.

Tabel 9 Pengujian POS Tagging

No	Metode	Akurasi	Keterangan
1	Baseline	43%	Membutuhkan proses training berulang
2	Unigram	95%	Membutuhkan proses training berulang
3	Rule Based	97%	Proses training hanya sekali dilakukan saat pembuatan model

Berdasarkan hasil pengujian menunjukkan penggunaan *rule-based* memiliki kelebihan yaitu metode ini memiliki akurasi yang lebih bagus dan hanya memerlukan sekali *training* untuk membuat model yang nantinya langsung bisa digunakan untuk *tagging*.

5. Kesimpulan

Berdasarkan dari hasil pengujian, kesimpulan yang didapatkan sebagai berikut.

1. Naïve Bayes Classifier dapat digunakan untuk model klasifikasi sentimen pendapat masyarakat terhadap presiden Joko Widodo dan Susilo Bambang Yudhoyono dengan rata-rata akurasi yang didapat sebesar 80,29 %.
2. POS *tagging* dapat digunakan untuk mengenali karakter presiden berdasarkan sentimen positif dengan akurasi sebesar 73,3%. Sehingga dapat disimpulkan bahwa karakter presiden dapat diidentifikasi berdasarkan jenis kata dan diurutkan berdasarkan polaritas kemunculannya untuk mendapatkan karakter yang paling diharapkan masyarakat.

Saran

Tahap POS *tagging* terbatas pada pengenalan kata perkata karena menggunakan tokenizing untuk menghitung kemunculan kata, untuk meningkatkan nilai akurasi dalam menentukan karakter presiden akan lebih baik jika *tweet* melalui tahap pengenalan frase sebelum masuk ke tahap POS tagging. Pengenalan frase bertujuan untuk menggabungkan kata yang seharusnya tidak dipisahkan saat *tokenizing* untuk mendapatkan makna dari kata tersebut.

Daftar Pustaka

[1] H. Kwak, C. Lee, H. Park dan S. Moon, "What is Twitter, a Social Network or a News Media?," p. 591, 2010.

[2] E. Kouloumpis, T. Wilson dan J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," dalam *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Edinburgh, 2011.

[3] H. S. Ginting, . K. M. Lhaksmana dan . D. T. Murdiansyah, "Klasifikasi Sentimen Terhadap Bakal Calon Gubernur Jawa Barat 2018 di Twitter," dalam *e-Proceeding of Engineering*, Bandung, 2018.

[4] A. Pak dan P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," pp. 1320-1326.

[5] F. Rashel, A. luthfi, A. Dinakaramani dan R. Manurung, "Building an Indonesian Rule-Based Part-of-Speech Tagger," dalam *Intenational Conference on Asian Language Processing (IALP)*, Kuching, 2014.

[6] F. dan R. Manurung, "Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews," dalam *Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications (ICACIA)*, Depok, Indonesia, 2008.

[7] A. Dinakaramani, F. Rashel, A. Luthfi dan R. Manurung , "Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus," dalam *International Conference on Asian Language Processing (IALP)*, Kuching, 2014.

[8] P.-N. Tan, M. Steinbach dan V. Kumar, Introduction to Data Mining, Boston: Pearso Education, 2006.

[9] J. E. Sembodo, E. B. Setiawan dan Z. A. Baizal, "Data Crawling Otomatis pada Twitter," dalam *Indonesian Symposium on Computing (Indo-SC)*, 2016.

[10] A. Juan dan H. Ney, "Reversing and Smoothing the Multinomial Naive Bayes Text Classifier," *In Proc. 4th International Workshop on Pattern Recognition in Information Systems, PRIS,* pp. 108-117, 2004.

[11] A. Ziani, N. Azizi, D. Schwab, M. Aldwairi dan N. Chekkai, "Recommender System Through Sentiment Analysis," 2018.

[12] A. Kumar dan . T. M. Sebastian, "Sentiment Analysis on Twitter," *IJCSI International Journal of Computer Science Issues,* vol. 9, no. 4, pp. 372-378, 2012.

[13] T. Pang-Ning, M. Steinbach dan K. Vipin, Introduction to Data Mining, Boston: Pearson Education, 2006.

[14] H. Jiawei, K. Micheline dan P. Jian, Data Mining Concepts and Techniques Third Edition, Waltham: Morgan Kaufmann, 2012.

[15] M. K. Andreas dan M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business Horizons, Paris, 2010.