Sentimen Analisis Pada Media Online Mengenai Pemilihan Presiden 2019 Dengan Menggunakan Metode Naive Bayes

Mehdi Mursalat Ismail[1], Kemas Muslim Lhaksamana[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]mehdimi@students.telkomuniversity.ac.id, [2]kemasmuslim@telkomuniversity.ac.id,

Abstrak

Baru-baru ini sedang ramai pemberitaan mengenai pemilihan presiden di media *online*. Dengan maraknya pemberitaan mengenai pemilihan presiden tersebut, media *online* digunakan sebagai penggiring opini publik yang efektif. Maka dari itu penelitian ini mengimplementasikan metode Naive Bayes Classifier pada Sentiment Analysis yang memungkinan kita untuk mengetahui kecondongan politik suatu media *online*. Penelitian ini akan ditujukan kepada teks yang berhubungan dengan pemilu 2019. Sebelum proses *sentiment analysis*, dilakukan terlebih dahulu pengambilan data berupa teks dengan metode *web scraping*, lalu dilakukan *text preprocessing* pada data teks tersebut. *Output* dari *sentiment analysis* ini berupa *confusion matrix*. Penelitian ini membangun sistem yang dapat mendeteksi sebuah berita memiliki sentimen positif atau negatif pada salah satu pasangan calon presiden tahun 2019 dengan akurasi sebesar 79,5% untuk berita mengenai Jokowi-Ma'ruf dan 64% untuk berita mengenai Prabowo-Sandi.

Kata Kunci : ***sentiment analysis, naive bayes classifier, web scraping, text preprocessing, confusion matrix***

Abstract

Recently there has been a lot of news about the presidential election in online media. With the proliferation of news about the presidential election, online media is used as an effective guide to public opinion. Therefore, this research was carried out by implementing the Naive Bayes Classifier method on Sentiment Analysis that allows us to know the political bias of an online media. This research will be directed to texts relating to the 2019 election. Before the sentiment analysis process, data collection is done in the form of text using the web scraping method, then a text preprocessing is performed on the text data. The output of this sentiment analysis is in the form of a confusion matrix.This study build a system that can detect positive or negative sentiment of the news with 79,5% accuracy for news about Jokowi-Ma'ruf and 64% for news about Prabowo-Sandi .

Keywords: sentiment analysis, naive bayes classifier, web scraping, text preprocessing, confusion matrix

1. Pendahuluan

1.1. Latar Belakang

Media *online* adalah sebuah sarana komunikasi satu arah yang disajikan secara *online* yang berupa *website*. Semakin dekatnya pemilihan presiden dan wakil presiden 2019 pemberitaan mengenai pemilihan presiden terbagi kedalam dua kubu, yaitu kubu Jokowi atau Prabowo, contohnya pada *media online* Viva ada berita berjudul "Fahri Hamzah Anggap Kritik Prabowo pada Jokowi Mencerdaskan" dan pada *media online* Okezone terdapat berita dengan judul "Sempat Diprotes, Ma'ruf Amin Sebut Kalangan Tunanetra Sekarang Berbalik Mendukung" mungkin dari judul sudah terlihat polaritas dari berita tersebut. Dengan jumlah berita mengenai pemilihan presiden yang besar sulit untuk mengetahui polaritas suatu media online secara manual. *Media online* yang seharusnya bersifat netral, beberapa malah menjadi alat untuk menggiring opini publik ke suatu kubu, baik berupa opini positif maupun negatif.

Pada penelitian sebelumnya *sentiment analysis* berfokus pada sosial media *tweeter* dengan metode yang sama yaitu *Naïve Bayes Classifier*. Untuk mengetahui polaritas sebuah media *online*, maka dibutuhkan *sentiment analysis* terhadap berita di terbitkan di media online tersebut. Penenelitan ini menggunakan metode klasifikasi *Naïve Bayes Classifier* untuk melakukan klasifikasi teks berbahasa Indonesia. *Naïve Bayes Classifier* merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidaktergantungan *(independent)* yang tinggi [1]. Keuntungan penggunan *Naïve Bayes Classifier* dibandingkan dengan metode lain yaitu *SVM, ANN, dan Decision Tree* metode ini hanya membutuhkan jumlah *data training* yang lebih kecil, dengan performa yang sama dengan metode pendekatan yang lain disertai akurasi yang baik [2] .

1.2. Tujuan

Tujuan dari penelitian ini adalah:

a. Menghasilkan sistem yang dapat menentukan sebuah teks berita memiliki sentimen positif atau negatif kepada salah satu calon presiden dan wakil presiden.
b. Mengimplementasikan metode *Naïve Bayes Classifier* untuk melakukan klasifikasi sebuah teks berita memiliki sentimen positif atau negatif pada salah satu pasangan calon presiden.
c. Mengetahui kecondongan politik dari media online Republika.

1.3. Batasan Masalah

Batasan masalah untuk penelitian ini adalah:

a. Teks berita yang akan diolah hanya teks berbahasa Indonesia.
b. Teks yang di analisis adalah teks yang diterbitkan di *website www.*republika.co.id dan hanya teks berita yang mengandung tag berita dengan nama calon presiden dan calon wakil presiden.
c. Tugas Akhir ini akan memfokuskan pada hasil klasifikasi yang di hasilkan dengan menggunakan metode Naïve Bayes Classifier.
d. Klasifikasi teks berita hanya kedalam positif dan negatif.
e. Berita yang diambil dimulai dari 1 Desember 2018 – 28 February 2019.

2. Studi Terkait

2.1. *Data Collecting*

Pertama-tama yang dilakukan adalah mengumpulkan data terkait pemilihan presiden tahun 2019. Pengumpulan data dilakukan dengan *web scraping* menggunakan bahasa pemrograman python. *Web scraping* merupakan data *scraping* yang digunakan untuk mengambil data pada *website* [3]. Data yang diambil merupakan data dari media *online* Republika.

2.2. *Text Preprocessing*

Text Preprocessing merupakan pengolahan data berupa teks agar teks lebih mudah untuk diproses saat *Sentiment Analysis* dilakukan. Berikut *text preprocessing* yang dilakukan:

a. Stop Words Removal

Stop Words Removal merupakan penghapusan *stop words* (kata sambung) seperti ke, yang, di, karena kata-kata tersebut dianggap tidak memiliki arti dan membuat teks menjadi sulit untuk di analisis [4].

b. *Case Folding*

Mengubah huruf kapital menjadi huruf biasa.

c. Stemming

Mengubah term yang berimbuhan menjadi term berbentuk kata dasar [12].

d. Tokenizing

Menghilangkan tanda baca dan menguraikan kalimat menjadi perkata.

e. Filtering

Setelah proses *Tokenizing* dilanjutkan dengan proses *filtering. Filtering* yaitu proses pemilihan kata-kata penting [12].

Tabel 1. Tabel Alur Preprocessing

Awal	Stop word removal	Case folding	Stemming	Tokenizing	Filtering
Adi akan bepergian ke luar kota	Adi akan bepergian luar kota.	adi akan bepergian luar kota.	adi akan pergi luar kota	“adi”, “akan”, “pergi”, “luar”, “kota”	“adi”, “pergi”, “luar”, “kota”

2.3. *Term Frequency - Inverse Document Frequency*

Menghitung bobot berdasarkan frekuensi kata kunci yang muncul di dalam sebuah teks [5,6].
Berikut persamaan yang digunakan:

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{maksimum\ kemunculan\ kata} \quad (1)$$

$$idf(t,d) = log\ \frac{|D|}{banyaknya\ dokumen\ term\ t\ muncul} \quad (2)$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,d) \quad (3)$$

$tf(t,d) = jumlah\ kemunculan\ term\ t\ pada\ dokumen\ d$
$idf(t,d) = imverse\ document\ frequency$

2.4. *Sentiment Analysis*

Sentiment Analysis adalah teknik klasifikasi teks yang digunakan mengelompokkan teks berdasarkan pada opini yang dikandungnya. Sentiment Analysis memainkan peranan penting dari *Natural Language Processing*. *NLP* adalah bidang ilmu komputer dan kecerdasan buatan yang berhubungan dengan interaksi bahasa manusia-komputer. Teknik ini umumnya digunakan dalam membantu memilih keputusan pada perdagangan, investasi saham, dan pemilu [7].

Sentiment Analysis melibatkan pengklasifikasian opini dalam teks ke dalam kategori seperti "positif" atau "negatif" atau "netral"[8].

2.5. *Naïve Bayes Classifier*

Naïve Bayes Classifier adalah algoritma dalam teknik *data mining* yang menerapkan teori Bayes dalam klasifikasinya [10]. Berikut ini model naïve bayes classifier [11]:

$$p(C|x) = \frac{p(c)p(x|C)}{p(x)} \quad (4)$$

$p(C|x) = peluan\ kejadian\ C\ bersyarat\ x$
$p(C) = peluang\ kejadian\ C$
$p(x|C) = peluang\ kejadian\ x\ bersyarat\ C$
$p(x) = peluang\ kejadian\ x$

2.6. Evaluasi Akurasi

Untuk mengukur akurasi dari setiap metode sebelumnya maka akan diukur menggunakan nilai accuracy, precision, dan recall. Terdapat empat istilah sebagai representasi hasil proses klasifikasi. Berikut tabel dari empat istilah tersebut[9]:

Tabel 2. ***Confusion Matrix***

Actual Class	Predicted Class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

- TP (True Positive) : Merupakan data positif yang terdeteksi positif.
- TN (True Negative) : Merupakan data negatif yang terdeteksi negatif.
- FP (False Positive) : Data negatif yang terdeteksi positif.
- FN (False Negative) : Data positif yang terdeteksi negatif.

a. Accuracy

Tingkat kesamaan antara nilai prediksi dengan nilai aktual, jika semakin besar maka klasifikasi semakin baik. Berikut persamaannya.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

b. Precision

Precision adalah tingkat ketepatan antara informasi yang diinginkan pengguna dengan jawaban dari sistem.

$$precision = \frac{TP}{TP + FP} \quad (5)$$

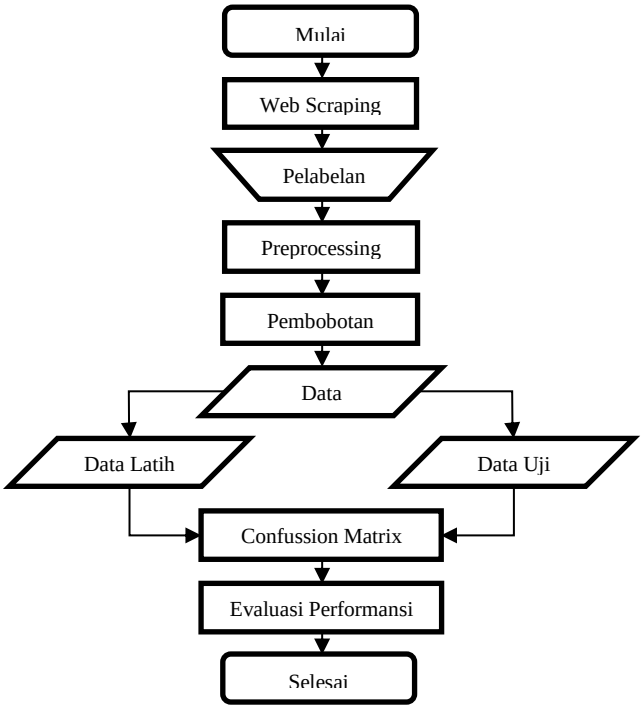
c. Recall

Recall merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

$$recall = \frac{TP}{TP + FN} \quad (6)$$

3. Sistem yang Dibangun

Pada penelitian ini, proses dimulai dengan memasukan data yang didapatkan dari hasil *web scraping*, lalu data akan dianalisis menggunakan algoritma Naïve Bayes, dan dihasilkan output berupa kelas (positif, negatif, netral). Berikut diagram yang menjelaskan keseluruhan sistem secara singkat:



Gambar 1. Sistem yang dibangun.

4. Evaluasi

Bagian ini berisi dua sub-bagian, yaitu Hasil Pengujian dan Analisis Hasil Pengujian. Pengujian dan analisis yang dilakukan selaras dengan tujuan TA sebagaimana dinyatakan dalam Pendahuluan.

4.1. Hasil Pengujian

Bagian ini berisi hasil akhir yang berupa *confusion matrix* dan akurasi.

Hasil Klasifikasi Naïve Bayes:

Tabel 5. Confusion Matrix Jokowi Ma'ruf dengan Data Set 0.9 : 0.1

Data Set 0.9 : 0.1		Prediction Class	
		Positive	Negative
Actual Class	Postitive	0	9
	Negative	3	28

Accuracy (%)	Precision	Recall
70	0	0

Tabel 6. Confusion Matrix Jokowi-Ma'ruf dengan Data Set 0.8 : 0.2

Data Set 0.8 : 0.2		Prediction Class	
		Positive	Negative
Actual Class	Postitive	0	14
	Negative	5	61

Accuracy (%)	Precision	Recall
76.25	0	0

Tabel 7. Confusion Matrix Jokowi-Ma'ruf dengan Data Set 0.7 : 0.3

Data Set 0.7 : 0.3		Prediction Class	
		Positive	Negative
Actual Class	Postitive	1	19
	Negative	7	94

Accuracy (%)	Precision	Recall
78.5124	0.125	0.05

Tabel 8. Confusion Matrix Jokowi-Ma'ruf dengan Data Set 0.6 : 0.4

Data Set 0.6 : 0.4		Prediction Class	
		Positive	Negative
Actual Class	Postitive	1	28
	Negative	7	124

Accuracy (%)	Precision	Recall
78.125	0.125	0.034483

Tabel 9. Confusion Matrix Jokowi-Ma'ruf dengan Data Set 0.5 : 0.5

Data Set 0.5 : 0.5		Prediction Class	
		Positive	Negative
Actual Class	Postitive	4	36
	Negative	5	155

Accuracy (%)	Precision	Recall
79.5	0.444444	0.1

Tabel 10. Confusion Matrix Prabowo-Sandi dengan Data Set 0.9 : 0.1

Data Set 0.9 : 0.1		Prediction Class	
		Positive	Negative
Actual Class	Postitive	3	14
	Negative	4	19

Accuracy (%)
55

Precision	Recall
0.428571	0.176471

Tabel 11. Confusion Matrix Prabowo-Sandi dengan Data Set 0.8 : 0.2

Data Set 0.8 : 0.2		Prediction Class	
		Positive	Negative
Actual Class	Postitive	7	23
	Negative	8	42

Accuracy (%)
61.25

Precision	Recall
0.466667	0.233333

Tabel 12. Confusion Matrix Prabowo-Sandi dengan Data Set 0.7 : 0.3

Data Set 0.7 : 0.3		Prediction Class	
		Positive	Negative
Actual Class	Postitive	11	37
	Negative	14	59

Accuracy (%)
57.8512

Precision	Recall
0.44	0.229167

Tabel 13. Confusion Matrix Prabowo-Sandi dengan Data Set 0.6 : 0.4

Data Set 0.6 : 0.4		Prediction Class	
		Positive	Negative
Actual Class	Postitive	12	48
	Negative	16	84

Accuracy (%)
60.0000

Precision	Recall
0.428571	0.2

Tabel 14. Confusion Matrix Prabowo-Sandi dengan Data Set 0.5 : 0.5

Data Set 0.5 : 0.5		Prediction Class	
		Positive	Negative
Actual Class	Postitive	14	58
	Negative	14	114

Accuracy (%)
64.0000

Precision	Recall
0.5	0.194444

4.2. Analisis Hasil Pengujian

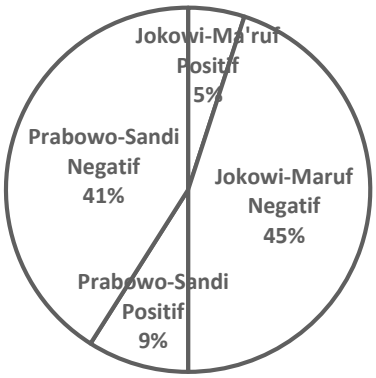
Analisis dilakukan berdasarkan pada output yang didapatkan pada tahap klasifikasi, didapatkan tabel sebagai berikut:

Tabel 15. Perbandingan Data Set dan Akurasi

Jokowi-Ma'ruf		Prabowo-Sandi	
Data Set	Akurasi (%)	Data Set	Akurasi (%)
0.9 : 0.1	70	0.9 : 0.1	55
0.8 : 0.2	76.25	0.8 : 0.2	61.25
0.7 : 0.3	78.5124	0.7 : 0.3	57.8512
0.6 : 0.4	78.125	0.6 : 0.4	60
0.5 : 0.5	**79.5**	**0.5 : 0.5**	**64**

Dari hasil percobaan,dapat dilihat pada tabel 15 bahwa didapatkan akurasi terbaik untuk model naïve bayes classifier, yaitu untuk model Jokowi-Ma'ruf sebesar 79,5% dan Prabowo-Sandi sebesar 64% terjadi pada data set 50% data latih dan 50% data uji.

Jika diambil informasi dari confusion matrix masing-masing kubu dengan akurasi tertinggi, maka didapatkan tabel sebagai berikut:



Gambar 2. Persentase sentiment berita.

Dilihat dari gambar 2 dapat dilihat bahwa sentiment positif untuk kubu Prabowo-Sandi (9%) lebih besar dibandingkan kubu yang hanya Jokowi-Ma'ruf (5%).

5. Kesimpulan

Pada percobaan ini dapat disimpulkan bahwa model *Naïve Bayes Classifier* dapat digunakan untuk menentukan positif negatif suatu *sentiment* pada berita. Dalam pengklasifikasian model *Naïve Bayes Classifier* dapat dikatakan cukup baik dikarenakan menghasilkan akurasi yang cukup besar yaitu untuk model Jokowi-Ma'ruf sebesar 79,5% dan Prabowo-Sandi sebesar 64% terjadi pada data set 50% data latih dan 50% data uji. Untuk kedepannya riset ini dapat dikembangkan lagi dengan menambahkan data yang seimbang.

Daftar Pustaka

[1] R. Mujib, S. Suyono, M. Sarosa, 2013, “Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier”.

[2] G. Pablo, G. Marcos, 2014, “A Naive-Bayes Strategy for Sentiment Analysis on English Tweets”.

[3] Boeing, G.; Waddell, P. (2016). "New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings". Journal of Planning Education and Research.

[4] M.F. Porter, An Algorithm for Suffix Stripping, Program, vol. 14, no. 3, pp. 130-137, 1980.

[5] Menaka S and Radha N, Text Classification using Keyword Extraction Technique, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013, ISSN: 2277 128X.

[6] S.Charanyaa, K.Sangeetha, Term Frequency Based Sequence Generation Algorithm for Graph Based Data Anonymization, International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 2, February 2014, ISSN(Online): 2320-9801.

[7] M. D. Devika, C. Sunitha, G. Amal, 2016, Sentiment Analysis:A Comparative Study On Different Approaches, Procedia Computer Science 87 (2016) 44 – 49

[8] Vishal A. Kharde, S.S. Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications (0975 – 8887), Volume 139 – No.11, April 2016.

[9] Fawcett, Tom, 2006. "An Introduction to ROC Analysis" (PDF). Pattern Recognition Letters. 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010.

[10] Santosa, B. 2007. Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu. Yogyakarta.

[11] Stuart, A.; Ord, K. (1994), Kendall's Advanced Theory of Statistics: Volume I—Distribution Theory, Edward Arnold, §8.7.

[12] Wahyudi, Dwi.,Susyanto, Teguh., nugroho, didik., “Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter Pada Dokumen Berbahasa Indonesia”.