ANALISIS SENTIMEN PADA MEDIA DARING TENTANG PEMILIHAN PRESIDEN INDONESIA TAHUN 2019 MENGGUNAKAN METODE NAÏVE BAYES

Chandra Jaya R[1], Kemas Muslim L[2]

Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[1]chunjr@students.telkomuniversity.ac.id, [2]kemasmuslim@telkomuniversity.ac.id

Abstrak

Kompas dan Detik merupakan beberapa contoh media daring yang menjadi wadah informasi satu arah masyarakat untuk dapat memperoleh informasi dan membahas tentang berbagai macam berita terkini. Analisis sentimen dilakukan untuk mengukur kecenderungan opini masyarakat terhadap suatu kejadian yang sedang atau telah terjadi. Salah satu kasus yang diangkat untuk dianalisis adalah Pemilihan Presiden tahun 2019 di Indonesia. Sebelum proses analisis sentimen, dilakukan terlebih dahulu pengambilan data berupa teks dengan metode *web scraping*, kemudian data tersebut kemudian diolah dengan melakukan *text pre-processing* pada data teks tersebut. Keluaran dari analisis sentimen ini berupa *confusion matrix*. Tugas akhir ini dibangun dengan tujuan dapat mendeteksi sebuah topik berita dari satu atau lebih portal berita yang memiliki kecenderungan konflik sentimen positif dan negatif pada tiap *headline* berita di masing-masing portal berita yang membahas tentang pemilihan presiden 2019, dengan akurasi sebesar 70% untuk Jokowi-Ma'ruf dan 65% untuk Prabowo-Sandiaga pada portal berita *kompas.com*, dan akurasi sebesar 70% untuk Jokowi-Ma'ruf dan 80% untuk Prabowo-Sandiaga pada *detik.com*. Penelitian ini memberikan informasi data yang diperoleh dari hasil klasifikasi menggunakan metode *Naïve Bayes.*

Kata kunci: analisis sentimen, *Naïve Bayes*, media daring, Kompas, Detik, *web scraping*, *text pre-processing*, *confusion matrix.*

Abstract

Kompas and *Detik* are some examples of online media that become a one-way information container for the public to obtain information and discuss various kinds of news. Sentiment analysis is carried out to measure tendencies of public opinion towards an event that has been or has occurred. One of the cases raised for analysis was the 2019 Presidential Election in Indonesia. Before the sentiment analysis process, data is taken first in the form of text using the web scraping method, the data is then processed by doing text pre-processing in the text data. The output of this sentiment analysis is in the form of a confusion matrix. This final project was built with the aim of being able to detect a news topic from one or more news portals that have a positive and negative sentiment conflict tendency on each news headline in each news portal that discusses the 2019 presidential election, with 70% accuracy for news about Jokowi-Ma'ruf and 65% accuracy about Prabowo-Sandiaga on the news portal *kompas.com* and 70% accuracy for news about Jokowi-Ma'ruf and 80% accuracy about Prabowo-Sandiaga on *detik.com*. This study provides information on data obtained from the classification using the Naïve Bayes method.

Keywords: *sentiment analysis*, *Naïve Bayes*, *online media*, Kompas, Detik, *web scraping*, *text pre-processing*, *confusion matrix*.

1. Pendahuluan

Media daring merupakan suatu wadah informasi yang memanfaatkan koneksi Internet dimana informasi mengenai berbagai hal dan isu sedang diberitakan. Beberapa informasi seperti berita Pemilihan Presiden di Indonesia pada tahun 2019 merupakan salah satu contoh informasi di media daring yang banyak diperbincangkan oleh masyarakat Indonesia [1].

Pada penelitian sebelumnya, analisis sentimen berfokus pada media sosial *Twitter* dengan metode yang sama yaitu *Naïve Bayes Classifier*. Untuk mengetahui polaritas satu atau beberapa media daring, dibutuhkan analisis sentimen terhadap berita yang disajikan oleh media daring tersebut. Tugas akhir ini dibuat dengan tujuan mengamati dan menganalisis data dari hasil klasifikasi menggunakan metode *Naïve Bayes*, dimana data yang akan diolah diperoleh dari media daring. *Naïve Bayes Classifier* merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidaktergantungan (independen) yang tinggi.

Penelitian ini juga bertujuan untuk memberikan informasi yang akurat dari data yang diolah dengan metode *Naïve Bayes* ini, bukan membandingkan pengolahan data dengan metode lain [2]. Keuntungan penggunaan metode *Naïve Bayes Classifier* adalah metode ini membutuhkan jumlah *data training* yang lebih kecil, dengan performa yang sama dengan metode pendekatan lain serta dengan akurasi yang baik.

Dengan ditentukannya batasan masalah, akan mempermudah jalannya penelitian ini agar tidak terlampau luas dan berjalan dengan baik. Batasan masalah yang diangkat adalah menganalisa opini masyarakat tentang Pemilihan Presiden di Indonesia pada 2019, di media daring Kompas dan Detik dengan *web scraping* menggunakan aplikasi ekstensi dari Google Chrome yaitu *Data Miner*, setelah itu dilakukan *pre-processing*, lalu pengklasifikasian dengan metode *Naïve Bayes.*

2. Studi Terkait

2.1 *Text Mining*

Text mining adalah proses ekstraksi pola informasi dari sejumlah besar sumber data tak terstruktur [3]. Pertama-tama yang dilakukan adalah mengumpulkan data terkait Pemilihan Presiden 2019, kemudian pengumpulan data dilakukan dengan *web scraping*, sebuah proses penambangan data dari *website* menggunakan aplikasi ekstensi Google Chrome, *Data Miner*. Data yang diambil merupakan data dari media daring Kompas dan Detik.

2.2 *Text Pre-Processing*

Text Pre-Processing adalah suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan untuk proses *mining* yang lebih akurat [4]. Berikut adalah tahapan *Text Pre-Processing*, antara lain:

a. *Stop Words Removal*

Stop Words Removal merupakan penghapusan kata sambung, karena kata-kata tersebut dianggap tidak memiliki arti dan membuat teks menjadi sulit untuk di analisa [5].

b. *Case Folding*

Case Folding berfungsi untuk mengubah semua huruf pada teks menjadi huruf kecil.

c. *Tokenizing*

Tokenizing berfungsi sebagai pemecah teks menjadi bentuk kata dengan melakukan penghilangan tanda baca [6].

Dibawah ini merupakan **Tabel 1 Contoh *Text Pre-processing*** yang dilakukan pada data set

Data hasil *web scraping*	*Stop Words Removal*	*Case Folding*	*Tokenizing*
Selesai coblos, Pemilu berujung ricuh.	Selesai coblos, Pemilu berujung ricuh.	selesai coblos, pemilu berujung ricuh.	selesai coblos pemilu berujung ricuh

2.3 Analisis Sentimen

Analisis sentimen adalah teknik klasifikasi teks yang digunakan untuk mengelompokkan teks berdasarkan pada opini yang terkandung dalam teks tersebut. Teknik ini umumnya digunakan untuk membantu memilih keputusan pada perdagangan, investasi, saham dan pemilu [7]. Dalam penelitian ini, analisis sentimen dibagi menjadi dua kategori yaitu negatif dan positif, yang banyak datanya diperoleh dari media daring Kompas dan Detik.

2.4 *Naïve Bayes Classifier*

Naïve Bayes adalah sebuah klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. *Naïve Bayes Classifier*, berdasarkan teorema Bayes, memiliki kemampuan klasifikasi serupa seperti pohon keputusan dan jaringan saraf [8], memiliki bentuk umum sebagai berikut:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (1)$$

P(x|y) = peluang hipotesis y bersyarat x
P(y) = peluang hipotesis kejadian x
P(y|x) = peluang hipotesis x bersyarat y
P(x) = peluang kejadian x

$$\hat{C} = argmax\ P(c|d) = argmax\ \frac{P(d|c)P(c)}{P(d)} \quad (2)$$

P(c|d) = peluang hipotesis d bersyarat c
P(d) = peluang hipotesis kejadian d
P(d|c) = peluang hipotesis c bersyarat d
P(c) = peluang kejadian c
d = dokumen
c = kelas
$\hat{C}$ = kelas yang telah diberi nilai positif/negatif [9].

2.4.1 Contoh pengerjaan Naïve Bayes

Tabel 2. Contoh Data.

Label	Data
-	Selesai coblos pemilu berujung ricuh
+	Prabowo klaim menang
-	Pemilu banyak pengalihan isu
+	Pemilu bersih aman damai
+	Prabowo sandi menang tps wiranto
?	Pemilu jokowi prabowo siapa menang

Keterangan label :

- "+" = data bernilai positif
- "-" = data bernilai negatif
- "?" = data yang belum mempunyai nilai/label
- Data set diatas berjumlah 6 data, terdiri dari 5 data training dan 1 data testing.

- Data training yang berjumlah 5 tersebut dikelompokkan berdasarkan label positif dan negatif.
- Setelah dikelompokkan, data tersebut dihitung menggunakan rumus metode *Naïve Bayes*.
- Contoh perhitungan data negatif dan data positif seperti pada **Tabel 3** dan **Tabel 4**.
- Untuk keterangan mengenai **Tabel 3** dan **Tabel 4**, dapat dilihat pada **Tabel 5**.

Tabel 3. Contoh Perhitungan Data Negatif

P(-)	(2/5)
pemilu (-)	(2+1)/(9+21)=0.1
jokowi (-)	(0+1)/(9+21)=0.03
prabowo (-)	(0+1)/(9+21)=0.03
siapa (-)	(0+1)/(9+21)=0.03
menang (-)	(0+1)/(9+21)=0.03
P(-).P(S|-)	(2/5) x 0.1 x 0.03 x 0.03 x 0.03 x 0.03
	= 3.24 x 10^{-8}

Tabel 4. Contoh Perhitungan Data Positif

P(+)	(3/5)
pemilu (+)	(1+1)/(12+21)=0.06
jokowi (+)	(0+1)/(12+21)=0.03
prabowo (+)	(2+1)/(12+21)=0.09
siapa (+)	(0+1)/(12+21)=0.03
menang (+)	(2+1)/(12+21)=0.09
P(+).P(S|+)	(3/5) x 0.06 x 0.03 x 0.09 x 0.03 x 0.09
	= 2.62 x 10^{-7}

Tabel 5. Keterangan mengenai Tabel 3.

Keterangan	
2	Jumlah kata "pemilu" dalam label negatif pada tabel
1	Default dari rumus
9	Jumlah seluruh kata dalam label negatif
21	Jumlah seluruh kata dalam data set
0.1	Nilai hasil rumus *Naïve Bayes*

- Setelah didapatkan hasil P x P(S|-) dan P x P(S|+)
- P x P(S|-) **= 3.24 x 10^{-8}**
- P x P(S|+) **= 2.62 x 10^{-7}**
- Bandingkan mana yang lebih besar dari kedua hasil tersebut,
- Nilai yang lebih besar adalah yang berlabel positif, maka
- Maka, kalimat data testing "pemilu jokowi prabowo siapa yang menang" berlabel **positif.**

2.5 Evaluasi Akurasi

Untuk mengukur akurasi dari setiap metode sebelumnya, diukur menggunakan nilai *accuracy, precision,* dan *recall.* Terdapat empat istilah sebagai representasi hasil proses klasifikasi. Berikut tabel dari empat istilah tersebut [10]:

Tabel 6. *Confusion Matrix*

Prediksi Aktual	**Klasifikasi**	
Kelas Positif	*True* Positif (TP)	*False* Negatif (FN)
Kelas Negatif	*False* Positif (FP)	*True* Negatif (TN)

Dimana:

True Positif (TP) : Kelas yang diprediksi positif, dan faktanya adalah positif.
True Negatif (TN) : Kelas yang diprediksi negatif, dan faktanya adalah negatif.
False Negatif (FN) : Kelas yang diprediksi positif, dan faktanya adalah negatif.
False Positif (FP) : Kelas yang diprediksi negatif, dan faktanya adalah positif.

a. *Accuracy*

Accuracy adalah perhitungan ketepatan kelas prediksi dengan kelas aktual. Jika hasil *accuracy* semakin besar maka klasifikasi semakin baik. Perhitungan *accuracy* sebagai berikut [11]:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}\ x\ 100\% \quad (3)$$

b. *Precision*

Precision adalah tingkat ketepatan antara informasi yang diinginkan pengguna dengan jawaban dari sistem. Rumus dari *precision* adalah sebagai berikut [11]:

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (4)$$

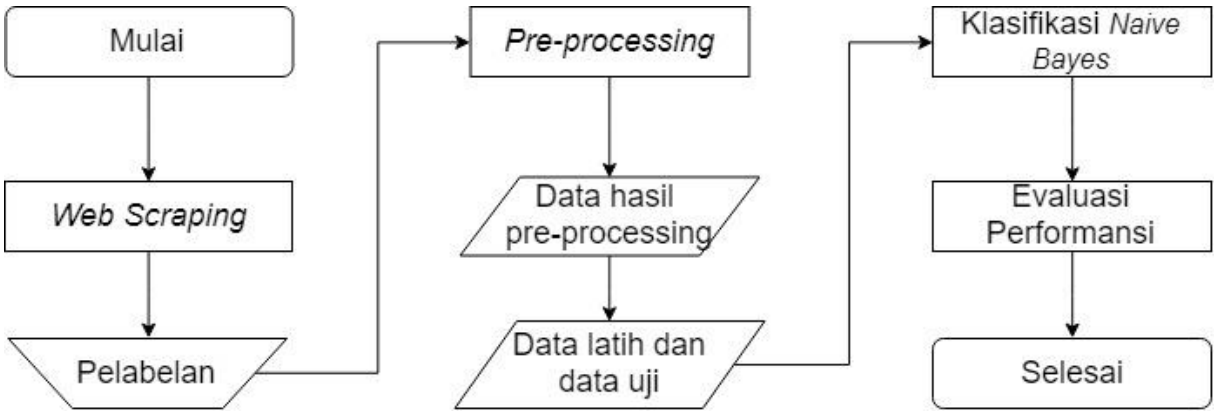
c. *Recall*

Recall adalah pengukuran presentase seberapa banyak data aktual yang di prediksi benar oleh sistem. Perhitungan *recall* dapat dilakukan dengan rumus sebagai berikut [11]:

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (5)$$

3. Sistem yang dibangun

Pada penelitian ini, proses dimulai dengan penambangan data dengan cara *web scraping*, kemudian data yang telah diperoleh tersebut melalui tahap pelabelan. Pelabelan dilakukan dengan memberi label positif dan negatif pada setiap data yang diperoleh dari *web scraping*. Setelah itu, tahap *pre-processing* dilakukan dan dihasilkan data latih dan data uji. Lalu data latih dan data uji dianalisa menggunakan algoritma *Naïve Bayes*, dan dihasilkan keluaran berupa kelas positif dan negatif. Berikut merupakan diagram yang menjelaskan keseluruhan kerja sistem:



Gambar 1. Flowchart Sistem Yang Dibangun.

4. Evaluasi

Bagian ini berisi dua sub-bab bagian, yaitu hasil pengujian dan analisis hasil pengujian. Pelabelan, pengujian dan analisis yang dilakukan selaras dengan tujuan tugas akhir sebagaimana dinyatakan dalam Pendahuluan.

4.1 Hasil dan Analisis

Bagian ini berisi hasil akhir berupa *confusion matrix*, akurasi, presisi dan *recall*. Hasil klasifikasi *Naïve Bayes* dapat dilihat di bagian lampiran.

4.2 Analisis Hasil Pengujian

Analisis dilakukan berdasarkan pada keluaran yang diperoleh pada tahap klasifikasi, diperoleh tabel sebagai berikut:

Tabel 1. Data set dan akurasi media daring Detik.

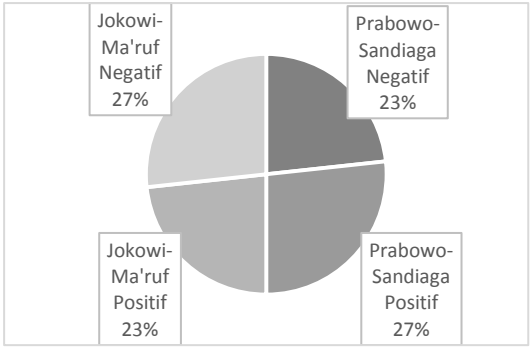
Jokowi-Ma'ruf		Prabowo-Sandiaga	
Data Set	Akurasi (%)	Data Set	Akurasi (%)
90% : 10%	70	90% : 10%	80
80% : 20%	66.25	80% : 20%	73.75
70% : 30%	64.16	70% : 30%	70
60% : 40%	61.25	60% : 40%	68.125
50% : 50%	59.5	50% : 50%	64.5

Tabel 2. Data set dan akurasi media daring Kompas

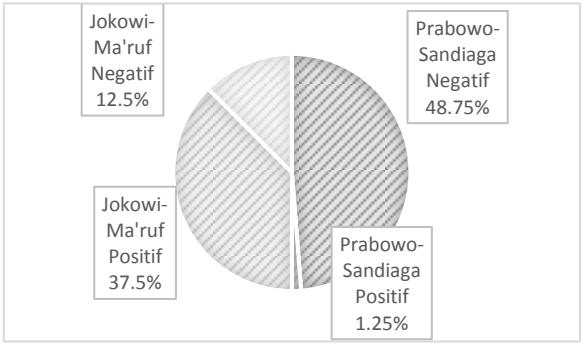
Jokowi-Ma'ruf		Prabowo-Sandiaga	
Data Set	Akurasi (%)	Data Set	Akurasi (%)
90% : 10%	70	90% : 10%	65
80% : 20%	67.5	80% : 20%	57.5
70% : 30%	66.66	70% : 30%	55.83
60% : 40%	62.5	60% : 40%	54.37
50% : 50%	60.5	50% : 50%	51.5

Dari hasil percobaan dapat dilihat pada **Tabel 1** dan **Tabel 2**, diperoleh akurasi terbaik untuk model *Naïve Bayes Classifier*, yaitu untuk model yang disajikan oleh portal berita Detik pada Jokowi-Ma'ruf sebesar 70%, pada Prabowo-Sandiaga memperoleh sebesar 80%. Sedangkan dari portal berita Kompas pada Jokowi-Ma'ruf sebesar 70% dan pada Prabowo-Sandiaga sebesar 65%, dilakukan pada data set 90% data latih dan 10% data uji.

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



Gambar 2. Persentase sentimen berita media daring Detik.



Gambar 3. Persentase sentimen berita media daring Kompas.

5. Kesimpulan

Pada percobaan ini disimpulkan bahwa model *Naïve Bayes Classifier* dapat digunakan untuk menentukan positif negatif konflik suatu sentimen pada berita. Selain menyajikan informasi aktual, pada pengujian ini dapat dilihat media daring Kompas dan Detik memiliki sentimen berbeda pada kedua capres dan cawapres. Dalam pengklasifikasian model *Naïve Bayes Classifier*, dapat dikatakan cukup baik dikarenakan hasil akurasi yang cukup besar yaitu 80% pada portal berita Detik dan 70% pada Kompas dengan data set 90% data latih dan 10% data uji. Untuk kedepannya, riset ini dapat dikembangkan lagi dengan data yang lebih seimbang dan sistem yang lebih optimal.

Daftar Pustaka

[1] Ratna, Lidwina Galih Puspa (2012). "*MEDIA ONLINE SEBAGAI PEMENUH KEPUASAN INFORMASI (Studi Analisis Deskriptif Kualitatif Mengenai Kepuasan Informasi bagi Kaum Wanita pada Media Online wolipop.com)".*

[2] Akbar, D.A. (2019). "*Analisis Sentimen Pada Tweet Masyarakat Tentang Pemilihan Gubernur Jawa Barat Menggunakan Metode Decision Tree*".

[3] Mardius, Bukhari (Universitas Widyatama, 2016). "*Sentiment Analiys Dalam Menentukan Emosi Karyawan Menggunakan Text Mining Naïve Bayes*".

[4] Universitas Brawijaya. (2016). Text Pre-Processing. Malang, Jawa Timur: M Ali Fauzi. Retrieved from lecture.ub.ac.id: http://malifauzi.lecture.ub.ac.id/files/2016/02/Text-Pre-Processing.pdf

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

[6] G.S., Aditya. (2019). "*Sentimen Analisis Politik Berita Media Online Dalam Pemilihan Presiden 2019 Menggunakan Metode Support Vector Machine*".

[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] Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. (2009). *"Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages".*

[9] J, Daniel, H. Martin, James. (2016). "*Speech and Language Processing.*"

[10] M. Ismail, Mehdi. (2019). "*Sentimen Analisis pada Media Online Mengenai Pemilihan Presiden 2019 Menggunakan Metode Naïve Bayes*".

[11] J. Han, M. Kamber and J. Pei. (2012). "*Data Mining Concepts and Techniques*".
USA: Elsevier.