ANALISIS SENTIMEN DEBAT CALON PRESIDEN DAN WAKIL PRESIDEN INDONESIA 2019 MENGGUNAKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

DOI: https://doi.org/10.22236/semnas.v1i1.103

Rizky Zein Adam*, Atiqah Meutia Hilda, Rachel Yukabit Rosyidah Ilahi
[1,2,3]*Universitas Muhammadiyah Prof. DR. HAMKA*
**rizkyzeinadam@gmail.com*

Abstract*. News is an interesting timely report for a large number of people. The information in the news can be a opinions or facts, both positive and negative. Sentiment analysis is a part of text mining research to classify an entity in a text document. This research was conducted by classifying sentiments using a headline dataset and news content about the debates candidates of the presidential and vice presidential for the Republic of Indonesia in 2019. Sentiment analysis in this research is divided into two classes, positive and negative. The data used was taken from the news media Detik, Kompas, Sindonews, Viva, Republika, and CNNIndonesia related to research. Weighting is done using TF-IDF and the algorithm used in this research is the Naïve Bayes Classifier algorithm with the Naïve Bayes Multinomial model. Based on the calculation that has been done, the comparison value of each debating activity is obtained where 42.3% is positive and 57.7% is negative from 130 datasets for the first debate, 48.0% is positive and 52.0% is negative from 122 datasets for the second debate, 44.0% positive and 56.0% negative from 124 datasets for the third debate, and 44.0% positive and 56.0% negative from 100 datasets for the fourth debate. While the accuracy value obtained for each debate is 76.923% for the first debate, 80% for the second debate, 84% for the third debate, and 95% for the fourth debate.*

Key words*: Sentiment Analysis, Naïve Bayes Classifier, Debates Candidates of The Presidential and Vice Presidential For The Republic of Indonesia In 2019*

Abstrak. Berita merupakan laporan tepat waktu yang menarik untuk sejumlah orang banyak. Informasi yang terdapat pada berita dapat berupa opini atau fakta, baik yang bersifat positif maupun negatif. Analisis sentimen merupakan cabang penelitian *text mining* untuk mengklasifikasikan suatu entitas pada dokumen teks. Penelitian ini dilakukan dengan mengklasifikasikan sentimen menggunakan *dataset headline* dan isi berita mengenai debat calon presiden dan wakil presiden Republik Indonesia tahun 2019. Analisis sentimen dalam penelitian ini terbagi menjadi dua kelas, yaitu positif dan negatif. Data yang digunakan diambil dari media berita Detik, Kompas, Sindonews, Viva, Republika, dan CNNIndonesia yang terkait dengan penelitian. Pembobotan dilakukan menggunakan TF-IDF dan algoritma yang digunakan dalam penelitian ini adalah algoritma *Naïve Bayes Classifier* dengan model *Multinomial Naïve Bayes*. Berdasarkan perhitungan yang telah dilakukan, maka diperoleh nilai perbandingan dari setiap kegiatan

debat dimana 42,3% positif dan 57,7% negatif dari 130 *dataset* untuk debat pertama, 48,0% positif dan 52,0% negatif dari 122 *dataset* untuk debat kedua, 44,0% positif dan 56,0% negatif dari 124 *dataset* untuk debat ketiga, dan 44,0% positif dan 56,0% negatif dari 100 *dataset* untuk debat keempat. Sementara nilai akurasi yang didapat pada setiap debat sebesar 76,923% untuk debat pertama, 80% untuk debat kedua, 84% untuk debat ketiga, dan 95% untuk debat keempat.

Kata kunci*: Analisis Sentimen, Naïve Bayes Classifier, Debat Calon Presiden dan Wakil Presiden Republik Indonesia Tahun 2019*

PENDAHULUAN

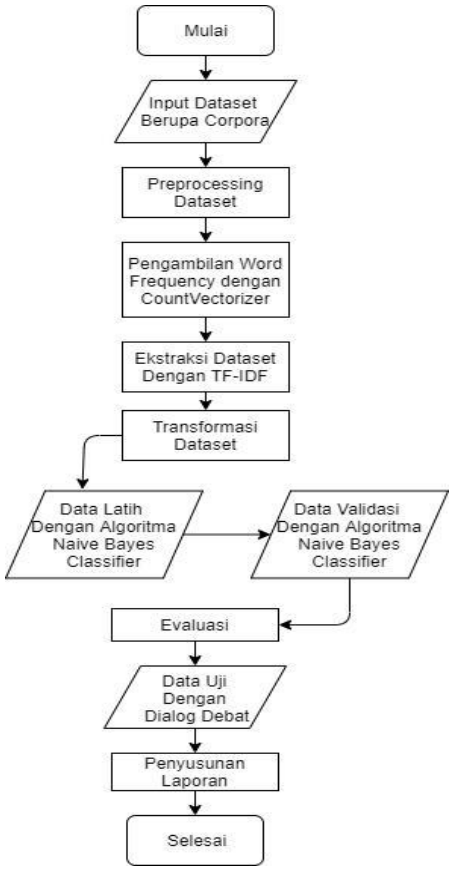
Debat capres dan cawapres pada tahun 2019 meraih lebih banyak penonton dibandingkan debat capres dan cawapres pada tahun 2014. Hal ini dibuktikan berdasarkan hasil pantauan Nielsen TV *Audience Measurement* (TAM) yang mengungkapkan bahwa kegiatan debat capres dan cawapres pada tahun 2019 mendapatkan jangkauan sebesar 67,9%, lebih besar dibandingkan periode 2014 (naik 62,9%). Dalam hal *rating*, debat capres dan cawapres pada tahun 2019 mencapai angka *rating* gabungan yang jauh lebih tinggi karena jumlah stasiun yang menyiarkan program debat di tahun ini jumlahnya lebih banyak dibandingkan dengan 2014. *Rating* tertinggi di 2019 adalah di debat kedua yaitu Debat Capres Jokowi versus Prabowo sebesar 18,8% (Malia, 2019).

Tingginya antusiasme masyarakat selama masa debat capres dan cawapres 2019, mengakibatkan banyaknya berita yang dirilis selama masa debat. Berita menjadi data latih yang digunakan karena setiap berita yang disebarluaskan harus berdasarkan fakta, adil, dan tidak memihak (Fachruddin, 2017). Berita dirilis dalam bentuk *text* sehingga dapat dilakukan analisis sentimen dengan mengelompokkan apakah reaksi yang diberikan bersifat postif atau negatif selama masa debat capres dan cawapres 2019. Percobaan dari data yang dilatih menggunakan dialog debat capres dan cawapres 2019 selama 4 kegiatan yang telah disediakan "bahasakita.co.id".

Untuk mendukung analisis sentimen diperlukan algoritma untuk menguji data hasil debat dan algoritma yang akan digunakan pada penelitian ini adalah *Naïve Bayes Classifier*. Algoritma *Naïve Bayes Classifier* dianggap sebagai metode yang berpotensi baik untuk melakukan klasifikasi data dari pada metode klasifikasi lainnya dalam hal akurasi dan komputasi (B, 2018). Setelah algoritma diimplementasikan maka akan dihitung nilai akurasi yang dihasilkan dari algoritma tersebut.

METODE PENELITIAN

Analisis sentimen debat calon presiden dan wakil presiden Republik Indonesia 2019 menggunakan algoritma *Naïve Bayes Classifier* dilakukan berdasarkan diagram alir yang ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alir Diagram alir pada

Gambar 1 menunjukkan :

1. *Input Dataset* Berupa *Corpora*, *dataset* dalam penelitian ini diambil dari media Detik, Kompas, Sindonews, Viva, Republika, dan CNNIndonesia. *Dataset* tersebut terbagi menjadi 4 dataset (sesuai dengan jumlah kegiatan debat). *Dataset* yang digunakan dapat dilihat pada https://bit.ly/dataset_skripsi . Pengumpulan *dataset* dilakukan dengan durasi seperti yang digambarkan pada Tabel 1.

Tabel 1. Durasi Pengumpulan Dataset

Nama *Dataset*	Durasi Pengumpulan *Dataset*
Dataset 1	14 Januari 2019 – 19 Januari 2019
Dataset 2	14 Februari 2019 – 19 Februari 2019
Dataset 3	14 Maret 2019 – 19 Maret 2019
Dataset 4	28 Maret 2019 – 04 April 2019

2. *Preprocessing Dataset*, terdapat beberapa proses pada tahap ini, diantaranya adalah sebagai berikut :
   a. *Case folding*, tahap mengubah semua huruf kapital menjadi huruf kecil.
   b. *Cleansing*, tahap menghilangkan kata yang tidak berpengaruh.
   c. *Stemming*, tahap menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda.
3. Pengambilan *Word Frequency* dengan *CountVectorizer*, tahap ini dilakukan dengan menghitung banyaknya kata yang muncul dalam sebuah *corpora* dari *dataset* yang telah melewati tahap *preprocessing*. Selanjutnya dengan *CountVectorizer* akan diubah fitur teks menjadi representasi matriks.
4. Ekstraksi *Dataset* dengan TF-IDF, tahap ini dilakukan untuk memberikan pembobotan kata dalam dokumen dengan menghitung frekuensi kemunculan *term* pada dokumen. Nilai TF-IDF meningkat secara proporsional sesuai dengan berapa kali suatu kata muncul pada dokumen dan diimbangi oleh frekuensi kata dalam *corpora*.
5. Transformasi *Dataset*, pada tahap ini algoritma yang digunakan adalah *Naïve Bayes Classifier* dengan model *Multinomial Naïve Bayes*. *Dataset* yang digunakan adalah data latih dan data validasi yang diambil dari berita selama masa kegiatan debat calon presiden dan wakil presiden Republik Indonesia 2019. Perbandingan antara data latih dan data validasi yang

digunakan adalah sebesar 80:20. Data latih digunakan untuk mempelajari lebih banyak data sehingga apabila dilakukan validasi maka akan menampilkan akurasi yang baik.

6. Evaluasi, Tahap ini dilakukan dengan menghitung nilai *f-measure* (*f1-score*) yang merupakan kombinasi dari *recall* dan *precision* (Nadia, Nhita, Informatika, Telkom, & Online, 2018).

   Persamaan *f-measure*

   $$= 2 \times \frac{\times}{}$$

   Persamaan *Recall* (1)

$$( ) = \frac{}{(+)} \quad (2)$$

dengan TP : *true positive* yaitu jumlah data positif yang terklasifikasi benar oleh sistem, FN : *false negative* yaitu jumlah data negatif yang terklasifikasi salah oleh sistem

Persamaan *Precision*

$$( ) = \frac{}{(+)} \quad (3)$$

dengan FP : *false positive* yaitu jumlah data positif yang terklasifikasi salah oleh sistem

7. Data Uji Dialog Debat, tahap ini dilakukan dengan menguji data menggunakan dialog debat capres dan cawapres 2019 selama empat kegiatan yang disediakan oleh bahasakita.co.id.
8. Penyusunan Laporan, tahap ini dilakukan dengan menulis apa saja hasil yang telah didapatkan dari proses awal pengumpulan data sampai dengan penarikan kesimpulan.

HASIL DAN PEMBAHASAN

Dataset yang telah dikumpulkan akan dibagi menjadi dua kelas yaitu, *corpora*_pos dan *corpora*_neg untuk setiap kegiatan debat seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Corpora Seluruh Kegiatan Debat

Kegiatan	***Corpora***	**Jumlah**
Debat Pertama	Positif	65
	Negatif	65
	Total	130
Debat Kedua	Positif	61
	Negatif	61
	Total	122
Debat Ketiga	Positif	62
	Negatif	62
	Total	124
Debat Keempat	Positif	50
	Negatif	50
	Total	100

Selanjutnya akan dilakukan *preprocessing dataset* secara manual. Hasil proses tersebut akan disimpan dalam folder *clean*_neg dan *clean*_pos seperti yang ditunjukkan pada Gambar 2.

```
import string
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()
save_to_file = open('corpora_clean/clean_neg/neg50.txt', 'w')
with open('corpora/neg/neg50.txt', 'r') as fileinput:
    for line in fileinput:
        line = line.lower()
        table = str.maketrans({key: None for key in string.punctuation})
        line = line.translate(table)
        line = line.strip()
        line = stemmer.stem(line)
        save_to_file.write(line)
```

Gambar 2. Preprocessing Dataset

Dataset yang telah selesai melalui tahap *preprocessing* selanjutnya memasuki tahap dimana *dataset* dipanggil dalam program agar bisa mengetahui banyaknya *dataset* yang akan digunakan. Proses ini ditunjukkan pada Gambar 3.

```
[2]: dataset = 'corpora_clean'

[3]: berita_train = load_files(dataset, shuffle=True)

[4]: # Berapa total data corpora
     len(berita_train.data)
```

Gambar 3. Proses Pengambilan Dataset

Dilanjutkan dengan perhitungan *Word Frequency* untuk mengetahui banyaknya kata yang mucul setelah dilakukannya *Word Frequency* kata diubah menjadi matriks dengan menggunakan *CountVectorizer*. Proses ini dapat dilihat pada Gambar 4.

```
[13]: berita_vec = CountVectorizer(min_df=1, tokenizer=nltk.word_tokenize)
      berita_counts = berita_vec.fit_transform(berita_train.data)

[14]: # Jumlah kata 'Prabowo' dalam dataset
      berita_vec.vocabulary_.get('prabowo')

[14]: 3620

[15]: # Jumlah kata 'Jokowi' dalam dataset
      berita_vec.vocabulary_.get('jokowi')

[15]: 2101

[16]: berita_counts.shape

[16]: (130, 5030)
```

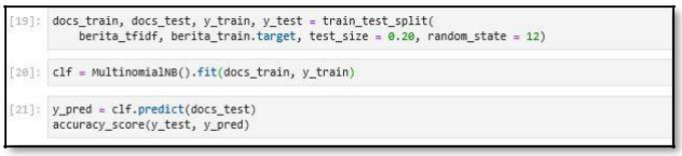
Gambar 4. Proses Word Frequency dan CountVectorizer

Setelah dilakukannya *Word Frequency* untuk pencarian kata selanjutnya dilakukan pembobotan dengan TF-IDF. Fungsi TF-IDF dapat dilihat pada Gambar 5.

```
[17]: tfidf_transformer = TfidfTransformer()
      berita_tfidf = tfidf_transformer.fit_transform(berita_counts)
      berita_tfidf.toarray()
```

Gambar 5. Fungsi TF-IDF

Setelah proses pembobotan nilai TF-IDF selanjutnya data terlebih dahulu di *split* 80:20 dimana data latih 80% sementara data validasi 20% kemudian dilakukan *fitting* untuk algoritma *Naïve Bayes Classifier*. Untuk mendapatkan hasil akurasi maka dilakukan prediksi pada data validasi. Proses ini dapat dilihat pada Gambar 6.



Gambar 6. Implementasi Naïve Bayes Classifier

Lalu akan dilakukan proses evaluasi terhadap model yang telah dibentuk. Model tersebut perlu dievaluasi untuk memastikan apakah model tersebut sudah sesuai dan performansinya dalam menjalankan tugas. Proses evaluasi dilakukan dengan mendapatkan nilai *precision*, *recall*, dan *f1-score*. Proses ini ditunjukkan pada Gambar 7.

Gambar 7. Proses Evaluasi

Selanjutnya akan didapatkan hasil percobaan berupa nilai akurasi (A), *precision* (P), *recall* (R), dan *f1-score* (F) untuk empat kegiatan debat dengan kelas *corpora* (C) yaitu positif dan negatif seperti yang ditunjukkan Tabel 3.

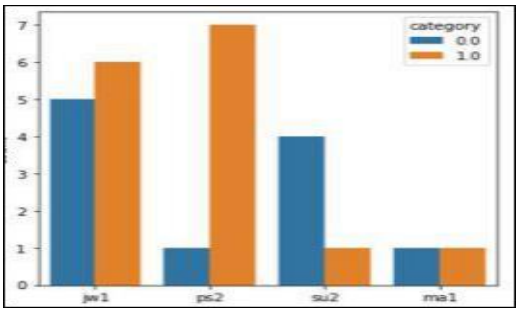
Tabel 3. Hasil Percobaan

Kegiatan	A	C	P	R	F
Debat Pertama	76.923%	Positif	67%	91%	77%
		Negatif	91%	67%	77%
Debat Kedua	80%	Positif	85%	79%	81%
		Negatif	75%	82%	78%
Debat Ketiga	84%	Positif	91%	77%	83%
		Negatif	79%	92%	85%
Debat	95%	Positif	91%	100%	95%

Keempat		Negatif	100%	90%	95%

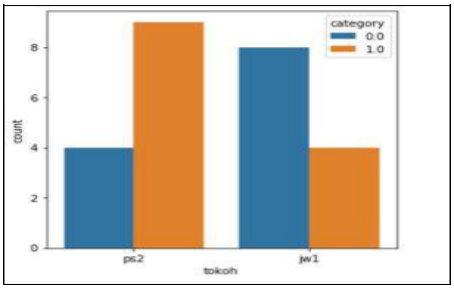
Berdasarkan proses yang telah dilakukan maka akan didapatkan hasil pengujian yang terbagi menjadi empat hasil diantaranya adalah sebagai berikut :

1. Debat Pertama, hasil prediksi keseluruhan 42,3% positif dan 57,7% negatif. Dari debat pertama, dilakukan perbandingan kedua pasangan calon yaitu Jokowidodo-Ma'aruf Amin dan Prabowo Subianto-Sandiaga Uno yang disingkat menjadi jw1-ma1 dan ps2-su2 dimana untuk warna biru menunjukan hasil negatif sedangkan yang berwarna oranye menunjukan hasil positif. seperti yang ditunjukkan Gambar 8.



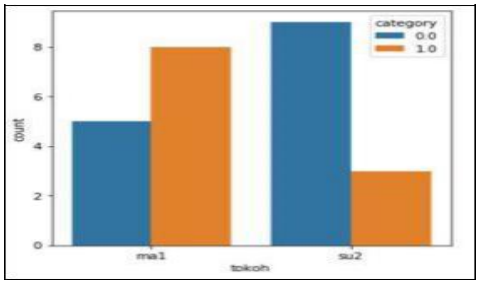
Gambar 8. Hasil Perbandingan Debat Pertama

2. Debat Kedua, hasil prediksi keseluruhan 48,0% positif dan 52,0% negatif. Dialog kedua antara Jokowidodo dan Prabowo Subianto memperoleh hasil perbandingan seperti yang ditunjukkan Gambar 9.



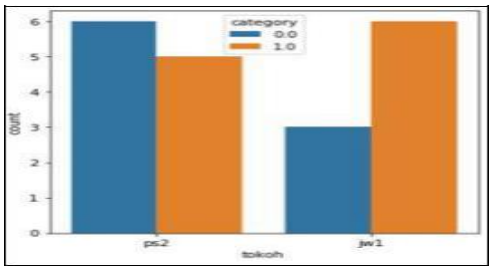
Gambar 9. Hasil Perbandingan Debat Kedua

3. Debat Ketiga, hasil prediksi keseluruhan 44,0% positif dan 56,0% negatif. Dialog ketiga antara Ma'aruf Amin dan Sandiaga Uno memperoleh hasil perbandingan seperti yang ditunjukkan Gambar 10.



Gambar 10. Hasil Perbandingan Debat Ketiga

4. Debat Keempat, hasil prediksi keseluruhan 44,0% positif dan 56,0% negatif. Dialog ketiga antara Jokowidodo dan Prabowo Subianto memperoleh hasil perbandingan seperti yang ditunjukkan Gambar 11.



Gambar 11. Hasil Perbandingan Debat Keempat

KESIMPULAN

Berdasarkan hasil dari percobaan yang dilakukan maka dapat disimpulkan :

1. Algoritma *Naïve Bayes Classifier* (NBC) mampu melakukan prediksi serta perbandingan analisis sentimen pada 4 kegiatan debat capres dan cawapres 2019 dengan merepresentasikan hasil yang cenderung lebih banyak hasil negatif dalam 4 kegiatan debat dimana perbandingannya adalah 42,3% positif dan 57,7% negatif dari 130 *dataset* untuk debat pertama, 48,0% positif dan 52,0% negatif dari 122 *dataset* untuk debat kedua, 44,0% positif dan 56,0% negatif dari 124 *dataset* untuk debat ketiga, dan 44,0% positif dan 56,0% negatif dari 100 *dataset* untuk debat keempat.
2. Nilai akurasi yang didapat pada setiap debat sebesar 76,923% untuk debat pertama, 80% untuk debat kedua, 84% untuk debat ketiga, dan 95% untuk debat keempat jadi hasil performa dari model untuk analisis sentemen tergantung pada ketepatan *dataset* pada proses kategori *corpora* positif atau *corpora* negatif.

DAFTAR PUSTAKA

B, F. S. (2018). *2018_Prediction of Song Popularity Based on BILLBOARD Chart Using The NAÏVE BAYES Algorithm*. *4*(1), 120–122.

Fachruddin, A. (2017). *Dasar-dasar Produksi Televisi: Produksi Berita, Feature, Laporan Investigasi, Dokumenter, dan Teknik Editing*.

Malia, I. (2019). Selama Pemilu 2019, Total Belanja Iklan Capres-Cawapres Rp206,6 Miliar. Retrieved December 6, 2019, from https://www.idntimes.com/news/indonesia/indianamalia/selama-pemilu-2019-total-belanja-iklan-capres-cawapres-rp2066-miliar

Nadia, R., Nhita, F., Informatika, F., Telkom, U., & Online, M. (2018). *Analisis Dan Implementasi Algoritma Naïve Bayes Classifier Terhadapa Pemilihan Gubernur Jawa Barat 2018 Pada Media Online*. *5*(1), 1678–1700.