Jurnal JAMIKA Manajemen Informatika

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Pembobotan TF-IDF Menggunakan Naïve Bayes Pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH

TF-IDF Weighting Using Naïve Bayes on Public Sentiment on The Issue of Rising BIPIH

Risa Wati[1*], **Siti Ernawati**[2], **Hilda Rachmi**[3]
[1,3]Program Studi Sistem Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia
[2]Program Studi Sistem Informasi, Universitas Nusa Mandiri, Jakarta, Indonesia
*E-mail: risawati.rwx@bsi.ac.id

Abstrak

Kementerian agama mengusulkan untuk menaikkan Biaya Perjalanan Ibadah Haji (Bipih) tahun 1444 H/2023 M menjadi Rp.69,19 juta. Terdapat kenaikan biaya yang cukup tinggi dibandingkan tahun 2022. Hal ini menimbulkan sentimen pada masyarakat, terdapat opini masyarakat yang pro dan kontra terhadap isu kenaikan Bipih di media sosial twitter. Tujuan dari penelitian ini adalah untuk menganalisa sentimen terhadap isu kenaikan Biaya Perjalanan Ibadah Haji dan untuk membuktikan apakah Naive Bayes merupakan pengklasifikasi text yang baik pada sentimen isu kenaikan Bipih. Naive Baye merupakan salah satu algoritma pengklasifikasi teks yang baik. Data diambil dari media sosial twitter. Data dikelompokkan menjadi opini pro dan opini kontra kemudian diolah menggunakan bahasa pemrograman python dan jupyter sebagai teks editor. Data yang digunakan sebanyak 850 data. Data dibagi menjadi data training dan data testing dengan perbandingan 80:20. Dengan jumlah data training sebesar 679 data dan jumlah data testing 170 data. Selanjutnya mengimplementasikan algoritma Multinominal Naive Bayes (MNB) sebagai pengklasifikasi teks serta dilakukan pembobotan kata menggunakan TF-IDF. Hasil uji coba diperoleh nilai akurasi sebesar 89% dan nilai ROC sebesar 0,91. Terbukti bahwa algoritma Multinominal Naive Bayes (MNB) merupakan pengklasifikasi teks yang baik untuk sentiment analisis opini isu kenaikan Biaya Perjalanan Ibadah Haji karena masuk kedalam Excellent Classification.

Kata kunci: BIPIH; Naïve Bayes; Analisis Sentimen; TF-IDF; Twitter.

Abstract

The Ministry of Religious Affairs proposes to increase the cost of Hajj Travel (Bipih) in 1444 H/2023 M to Rp.69.19 million. There is a fairly high increase in costs compared to 2022. This raises sentiment in the community, there are public opinions for and against the issue of rising Bipih on social media twitter. The purpose of this study was to analyze the sentiment on the issue of increasing the cost of Hajj Travel and to prove whether Naive Bayes is a good classifier of text on the issue of incremental sentiment. Naive Baye is one of the best text classifier algorithms. Data taken from social media twitter. The Data are grouped into pro and Contra opinions and then processed using python programming language and jupyter as text editor. Data used as much as 850 data. The Data is divided into training data and testing data with a ratio of 80:20. With the number of training data of 679 data and the number of testing data of 170 data. Then implement Multinominal Naive Bayes algorithm (MNB) as text classifier and word weighting using TF-IDF. The test results obtained accuracy value of 89% and ROC value of 0.91. It is proven that Multinominal Naive Bayes algorithm (MNB) is a good classifier of text for sentiment analysis of opinion on the issue of increasing the cost of Hajj travel because it is included in the Excellent Classification.

Keywords: BIPIH; Naïve Bayes; Sentiment Analysis; TF-IDF; Twitter.

Naskah diterima 9 Mar. 2023; direvisi 9 Apr. 2023; dipublikasikan 12 Apr. 2023.

I. PENDAHULUAN

Kementerian Agama mengusulkan Biaya Perjalanan Ibadah Haji (Bipih) tahun 1444 H/2023 M yang harus dibayar calon jamaah yang akan berangkat, yang awalnya hanya sebesar Rp. 39,89 juta menjadi Rp. 69,19 juta. Terdapat kenaikan biaya yang cukup tinggi dibandingkan tahun 2022. Usulan kenaikan biaya haji dipicu oleh Arab Saudi yang menetapkan adanya kenaikan Biaya Masyair dengan jumlah yang sangat signifikan. Pemerintah mengusulkan komposisi antara Bipih dan penggunaan Nilai Manfaat adalah 70%:30% karena nilai manfaat dana Jemaah haji bukan hanya untuk Jemaah haji yang akan berangkat saja tetapi untuk Jemaah haji yang belum diberangkatkan. Isu mengenai kenaikan biaya haji menjadi polemik dan perbincangan

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

publik yang menimbulkan pro kontra diberbagai media salah satunya adalah media sosial [1], [2]. Salah satu media sosial yang banyak digunakan adalah Twitter karena twitter dapat digunakan untuk mengungkapkan opini maupun memberikan saran dan kritik pada kebijakan pemerintah [3]. Media sosial twitter memungkinkan pengguna untuk berkomunikasi dengan seluruh dunia [4]. Tweets yang berisi opini maupun komentar, merupakan resource yang dapat digunakan untuk menganalisis sentimen terhadap suatu instansi maupun perorangan [5]. Analisis sentimen dikenal sebagai Opinion Mining atau emosi Artificial Intelligence dan pemanfaatan NLP, Text Mining, linguistik komputasi dan pengukuran bio untuk mengenali, mengevaluasi dan memeriksa keadaan emosional dan informasi subjektif [6].

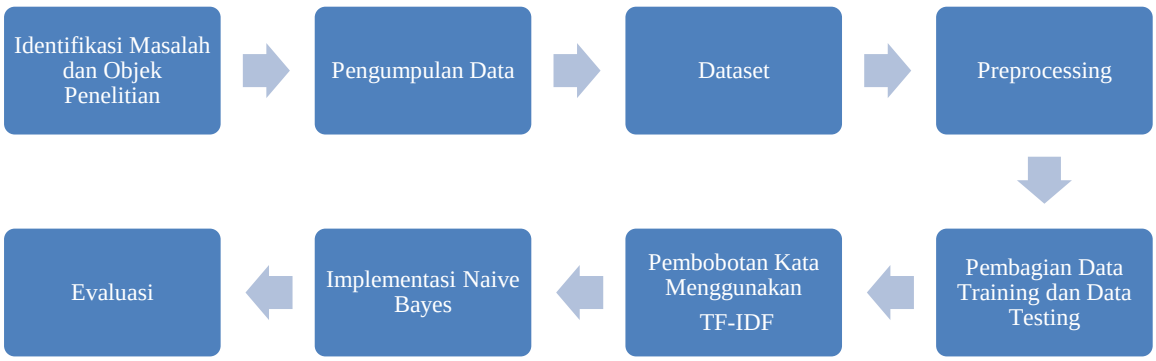
Tujuan dari penelitian ini adalah untuk menganalisa sentimen terhadap isu kenaikan Biaya Perjalanan Ibadah Haji dan untuk membuktikan apakah Naive Bayes merupakan pengklasifikasi text yang baik pada sentimen isu kenaikan Biaya Perjalanan Ibadah Haji. Algoritma Naïve Bayes dipilih karena salah satu metode pengklasifikasi teks yang telah banyak digunakan serta mudah dan cepat dalam penerapannya[7]. Pengklasifikasi Naïve Bayes digunakan untuk klasifikasi tujuan, Naïve Bayes juga mengklasifikasikan ulasan berdasarkan probability[8].

Beberapa penelitian pengklasifikasi text menggunakan algoritma Naive Bayes seperti penelitian dengan judul Perbandingan Metode Naïve Bayes dan *Support Vector Machine* pada Analisis Sentimen Twitter terbukti bahwa metode Naive Bayes memiliki hasil akurasi, presisi, recall dan F1-Score yang lebih baik dibandingkan metode Support Vector Machine [5]. Naïve bayes memiliki kinerja yang baik dalam memprediksi sentimen tetang kebijakan pemerintah dalam penerapan new normal [9]. Pada penelitian sentimen analisis menggunakan algoritma Naive Bayes data crawler twitter dilakukan percobaan dengan membandingkan tiga metode pengklasifikasi, yaitu metode Naive Bayes, SVM dan KNN, metode Naive Bayes memiliki tingkat akurasi yang lebih baik dibandingkan menggunakan metode lain, yaitu memperoleh nilai akurasi sebesar 80.90% [12].

Pada penelitian ini menggunakan model dengan fitur pembobotan TF-IDF yang diterapkan dengan algoritma Naïve Bayes terhadap review berbahasa Indonesia mengenai polemik masyarakat terhadap isu kenaikan Biaya Perjalanan Ibadah Haji. Penelitian ini juga memanfaatkan library sastrawi untuk pemrosesan teks dalam bahasa Indonesia. Evaluasi yang akan dilakukan terhadap hasil eksperimen yang dilakukan pada penelitian ini menggunakan confusion marix dan nilai AUC yang ada pada kurva ROC yang dihasilkan.

II. METODE PENELITIAN

Dalam penelitian ini terdapat beberapa tahapan penelitian yang digambarkan pada gambar 1. Tahapan penelitian dimulai dari identifikasi masalah dan objek penelitian, pengumpulan data, dataset, preprocessing, pembagian data training dan data testing, pembobotan kata menggunakan TF-IDF, implementasi Naïve Bayes, dan evaluasi.



Gambar 1. Tahapan Penelitian

Identifikasi Masalah dan Objek Penelitian

Kementerian Agama mengusulkan untuk menaikan Biaya Perjalanan Ibadah Haji tahun 2023 menjadi Rp. 69,19 Juta, hal ini menimbulkan pro dan kontra pada masyarakat. Permasalahan dalam penelitian ini adalah untuk menganalisa sentimen masyarakat terhadap pro kontra isu kenaikan Biaya Perjalanan Ibadah Haji. Objek dalam penelitian ini adalah opini masyarakat mengenai isu kenaikan Biaya Perjalanan Ibadah Haji pada media sosial twitter.

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan opini masyarakat mengenai isu kenaikan biaya perjalanan ibadah haji pada media sosial twitter yang diperoleh dari link https://netlytic.org/index.php dengan kata kunci "dana haji" dan "dana haji naik" kemudian file disimpan dalam format .CSV (Comma-Separated Values). Data dikumpulkan selama kurang lebih satu bulan mulai dari 06 Januari 2023 sampai dengan 01 Februari 2023. Gambar 2. menunjukan salah satu contoh review yang diberikan masyarakat terhadap isu kenaikan biaya haji pada twitter.

Gambar 2. Sample Review Masyarakat terhadap isu kenaikan biaya Haji Pada Twitter

Dataset

Dalam penelitian ini data yang digunakan adalah sebanyak 850 data. Data diambil secara acak pada media sosial twitter kemudian data diklasifikasikan secara manual ke dalam dua kriteria, yaitu opini masyarakat yang pro terhadap isu kenaikan biaya perjalanan ibadah haji dan opini masyarakat yang kontra terhadap isu kenaikan biaya perjalanan ibadah haji. Berdasarkan data yang diambil secara acak diperoleh opini pro berjumlah 189 data dan opini kontra berjumlah 661 data. Selanjutnya data diolah menggunakan bahasa pemrograman python dan jupyter sebagai text editor. Python merupakan salah satu bahasa pemrograman yang banyak digunakan, hal ini membuat python menjadi bahasa pemrograman yang mulai banyak dipelajari[11]. Jupyter merupakan perangkat lunak yang bersifat *open source* untuk mendukung ilmu *data science* dan *scientific computing* [12].

Preprocessing

Dalam bidang *text mining* maupun data mining tahapan *preprocessing* sangat berperan penting. Tahapan *preprocessing* ini dilakukan untuk melakukan proses seleksi data pada setiap dokumen yang akan diolah. Hasil *preprocessing* sangat menentukan hasil data yang diolah dan dapat mempengaruhi keakuratan klasifikasi dokumen[13], [14]. Dalam tahapan ini menggunakan library sastrawi. Library ini berisi kumpulan algoritma dan aturan untuk melakukan pemrosesan teks dalam bahasa Indonesia [15]Library sastrawi ditulis dalam bahasa pemrograman PHP dan dapat digunakan untuk melakukan *stemming* atau pengembalian kata dasar, *stopword removal* atau penghapusan kata-kata umum yang tidak memiliki makna khusus, tokenisasi atau

Jurnal JAMIKA Manajemen Informatika
Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

pemisahan teks menjadi kata-kata. Dalam implementasinya, menginstall library sastrawi pada python dapat menggunakan pip (package installer for python) dengan mengetikan kode pip install Sastrawi. Berikut proses *preprocesing* yang dilakukan:

1) *Cleaning Text*
Dalam proses Cleaning Text, berisi beberapa tahapan diantaranya adalah tahap *Case Folding* adalah proses mengubah dataset menjadi huruf kecil [13]. Contoh dari Case Folding, yaitu "DANA" menjadi "dana", "HAJI" menjadi "haji", "MODUS" menjadi "modus" dan lain-lain. Menghapus tanda baca (remove punct) seperti koma, titik, tanda tanya, tanda seru dan tanda baca yang lainnya. Menghapus hashtag, situs website, angka, karakter kosong (remove whitespace).

2) *Tokenization*
Tokenization adalah proses pemecahan teks pada kalimat menjadi potongan kata [5]. Proses *Tokenization* selain memisahkan teks, juga dapat menafsirkan dan mengelompokan token yang terisolasi untuk membuat token dengan tingkat yang lebih tinggi [15].

3) *Stopword Removal*
Stopword Removal adalah proses untuk menghapus kata yang tidak memiliki makna seperti pada, dan, hingga, yang dan lain-lain [13]. *Stopword* dapat diartikan juga untuk menghilangkan kata yg kurang efektif [17]. Pada proses *Stopword Removal* dilakukan penghapusan kata yang tidak relevan dengan kamus bahasa indonesia, seperti seperti wiiih, emmm, kalo, ngga, oleh, masy , yaaa, pdhal, sngj, malapraktik, bgmn, bhkn, wakk, adlh, dan lain-lain.

4) *Stemming*
Stemming adalah proses mengubah kata berimbuhan menjadi kata dasar [13]. Stemming menyaring kata yang terdapat kata sambung, kata ganti, kata depan menjadi kata dasar, yaitu dengan menghilanmgkan awalan dan akhiran kata [17].

Dalam penelitian ini dilakukan proses preprocessing data, yaitu *cleaning text*, *tokenization*, *stopword* dan *stemming*. Tabel 1 menunjukan hasil *preprocessing* data yang telah dilakukan. Diambil beberapa sample opini masyarakat yang pro dan opini masyarakat yang kontra terhadap kenaikan biaya perjalanan ibadah haji. Terlihat bahwa hasil dari *preprocessing* berjalan sesuai dengan fungsi dari masing-masing tahapan.

TABEL 1
HASIL PREPROCESSING DATA

Review	Cleaning Text	tokenization	Stopword	Stemming
Tapi kalo ngembat dana Haji dosa kan pak ???https://t.co/KBMx0eyHRK	tapi kalo ngembat dana haji dosa	['tapi', 'kalo', 'ngembat', 'dana', 'haji', 'dosa']	ngembat dana haji dosa	kalo ngembat dana haji dosa
@Tan_Mar3M Yg benalu tuh ,yg bawa kabur uang negara,koruptor ,naikan dana haji ,bemsin listrik, para buzzer yg di gaji oleh negara,..salafy wahabi salah nya apa ..	benalu bawa kabur uang negara koruptor naikan dana haji bemsin listrik para buzzer gaji oleh negara salafy wahabi salah	['benalu', 'bawa', 'kabur', 'uang', 'negara', 'koruptor', 'naikan', 'dana', 'haji', 'bemsin', 'listrik', 'para', 'buzzer', 'gaji', 'oleh', 'negara', 'salafy', 'wahabi', 'salah']	benalu bawa kabur uang negara koruptor naik dana haji listrik buzer gaji negara salafy wahabi salah	benalu bawa kabur uang negara koruptor naik dana haji bemsin listrik buzzer gaji negara salafy wahabi salah
@OposisiCerdas @msaid_didu Usut penyelewengan dana haji !	usut penyelewengan dana haji	['usut', 'penyelewengan', 'dana', 'haji']	usut seleweng dana haji	usut seleweng dana haji
@democrazymedia Yang boleh mengelola dana haji hanya yang sanggup dan amanah	yang boleh mengelola dana haji hanya yang sanggup amanah	['yang', 'boleh', 'mengelola', 'dana', 'haji', 'hanya', 'yang', 'sanggup', 'amanah']	kelola dana haji sangup amanah	kelola dana haji sanggup amanah
@Hoshrii BPKH sudah benar dalam pengelolaan dana haji.. Mantab	bpkh sudah benar dalam pengelolaan dana haji mantab	['bpkh', 'sudah', 'benar', 'dalam', 'pengelolaan', 'dana', 'haji', 'mantab']	bpkh kelola dana haji mantab	bpkh kelola dana haji mantab

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Review	Cleaning Text	tokenization	Stopword	Stemming
BPKH Mendukung Adanya Dana Haji yang Berkeadilan dan Berkelanjutan https://t.co/wr1sKmEgXp	bpkh mendukung adanya dana haji yang berkeadilan berkelanjutan	['bpkh', 'mendukung', 'adanya', 'dana', 'haji', 'yang', 'berkeadilan', 'berkelanjutan']	bpkh dukung dana haji adil lanjut	bpkh dukung dana haji adil lanjut

a. Pembagian Data Training dan Data Testing

Setelah dilakukan preprocessing data langkah selanjutnya, yaitu membagi data menjadi data traning (data latih) dan data testing (data uji). jumlah data training sebesar 679 data dan data testing sebesar 170 data. Data training digunakan untuk melatih algoritma untuk mencari model yang sesuai dan data testing digunakan untuk menguji model yang didapat setelah tahapan testing.

Proses pembagian data dalam python dapat mengimport library Sklearn dan sublibrary train_test_spilt. Sublibrary ini berasal dari modul model_selection.

```
from sklearn.model_selection import train_test_split
train_X, test_X, train_Y, test_Y = model_selection.train_test_split(df_finish['stopword'], df_finish['Label'], test_size = 0.2, random_state = 42)
```

Proses definisi dilakukan, yaitu train_X sebagai data X yang akan dilatih dan test_X sebagai dataX yang adakan dites. Variabel train_Y dependen yang akan dilatih dan test_Y merupakan variabel dependen yang akan diuji. Dalam penelitian ini dilakukan pembagian data training (data latih) sebesar 80% dan data uji (data testing) sebesar 20%. Selain pembagian data, parameter lain, yaitu random_state. Parameter ini merupakan parameter RNG (random number generator) yang diisi dengan nilai 42.

b. Pembobotan Kata Menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF)

Proses selanjutnya, dilakukan pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), pembobotan kata berfungsi untuk mengubah data berupa teks menjadi numerik. Bidang pembelajaran mesin atau *Machine learning* dan *deep learning* akan bekerja dengan baik dan maksimal jika data yang diolah berupa data numerik. Maka setiap penelitian apda bidang ini khususnya analisis sentimen harus mengubah data berupa kumpulan kata-kata atau kalimat menjadi numerik. Proses tersebut bisasnya disebut dengan proses pembobotan kata. Proses ini memiliki banyak metode yang dapat digunakan, diantaranya BoF (*Bag of Words*, N-gram, Word2Vec dan TF-IDF [16]. TF-IDF merupakan salah satu teknik yang digunakan dalam pengolahan teks untuk memberikan bobot pada kata-kata dalam sebuah dokumen. Tujuan dari TF-IDF adalah untuk mengidentifikasi kata-kata yang paling penting dalam suatu dokumen atau kumpulan dokumen. *Term Frequency* (TF) adalah nilai frekuensi kemunculan suatu kata dalam sebuah dokumen. Nilai TF dapat dihitung dengan menggunakan rumus:

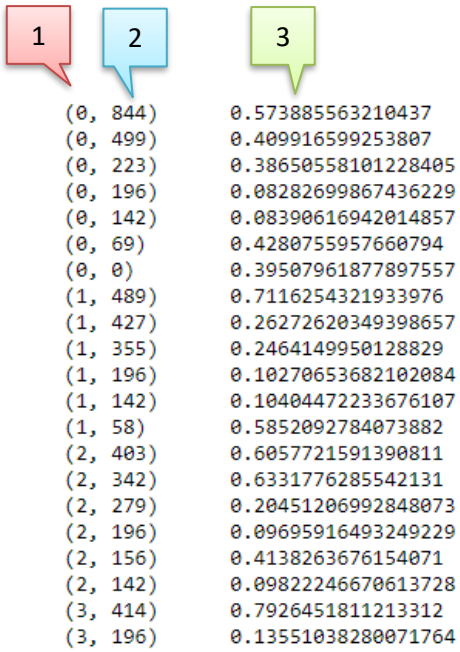
$$\text{TF} = \frac{(\text{jumlah kemunculan kata dalam dokumen})}{(\text{jumlah kata dalam dokumen})} \quad (1)$$

Namun, nilai TF ini tidak memberikan informasi tentang pentingnya kata tersebut dalam dokumen. Kata-kata yang sering muncul dalam sebuah dokumen seperti kata hubung atau kata umum mungkin memiliki nilai TF yang tinggi, tetapi sebenarnya tidak memiliki makna yang penting dalam dokumen tersebut. Oleh karena itu, dibutuhkan teknik lain, yaitu *Inverse Document Frequency* (IDF) yang memberikan bobot pada kata-kata yang muncul jarang di seluruh dokumen. Nilai IDF dapat dihitung dengan menggunakan rumus:

$$\text{IDF} = log\,\frac{N}{n} \quad (2)$$

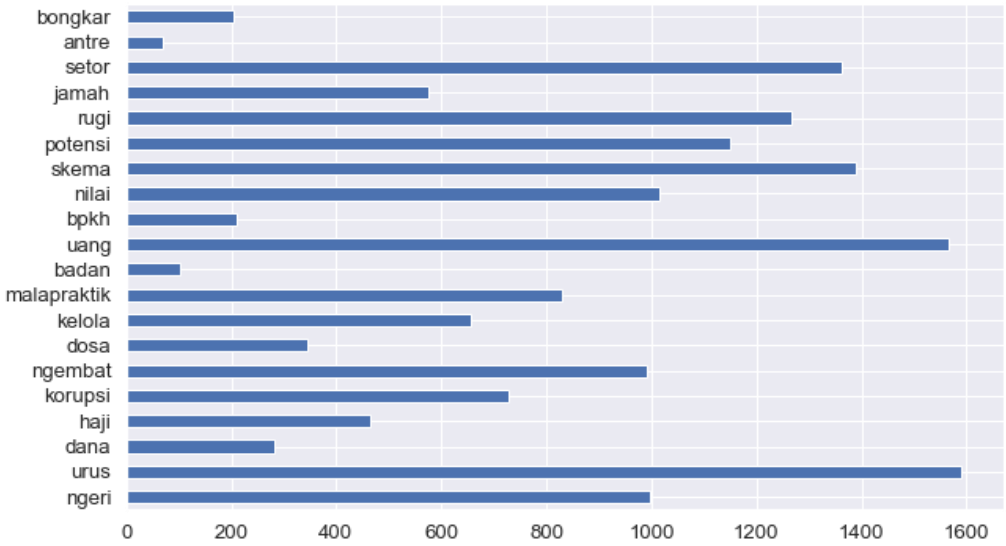
Dimana N adalah jumlah dokumen dalam kumpulan dokumen dan n adalah jumlah dokumen yang mengandung kata tersebut. Setelah nilai TF dan IDF diperoleh, nilai TF-IDF dapat dihitung dengan mengalikan nilai TF dengan nilai IDF. Kata-kata yang memiliki nilai TF-IDF yang tinggi dianggap penting dan memberikan kontribusi yang lebih besar dalam menentukan topik dokumen atau kumpulan dokumen tersebut. Gambar 3 adalah hasil pembobotan kata menggunakan TF-IDF pada python.

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika



Gambar 3. Hasil Pembobotan Kata Menggunakan TF-IDF

Berdarkan gambar 3 dapat disimpulkan bahwa kode 1 menunjukan nomor baris dari setiap data yang diolah. 2 merupakan nomor integer unik setiap kata yang ada pada baris. 3 merupakan hasil pembobotan (skor) yang dihitung menggunakan TF-IDF. Setelah dilakukan analisa terhadap proses *preprocessing* dan pembobotan TF-IDF maka dapat dilihat kata-kata yang sering muncul atau top word yang disajikan dalam bentuk diagram plot dan dapat dilihat pada gambar 4.



Gambar 4. Diagram Plot Menunjukan Kata-Kata Yang Sering Muncul Atau Top Word

c. Implementasi Naive Bayes

Naive Bayes merupakan algoritma klasifikasi yang didasarkan oleh teorema Bayes [5]. Naïve Bayes juga merupakan salah satu metode machine learning yang memanfaatkan perhitungan probabilitas dan statistika [19]Pada penelitian yang dilakukan menggunakan algoritma multinominal naïve bayes (MNB). Algoritma ini mengimplementasikan algoritma naïve bayes untuk data yang didistribusikan secara multinasional dan merupakan salah satu dari dua varian naive bayes klasik yang digunakan dalam klasifikasi teks [20]. MNB cocok untuk mengklasifikasikan teks yang memiliki banyak fitur (misalnya

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

kata-kata dalam dokumen) dengan menggunakan metode probabilitas. Algoritma ini bekerja dengan menghitung probabilitas kemunculan setiap kata di setiap kelas dan kemudian menggunakan probabilitas ini untuk memprediksi kelas baru dari suatu teks. Implementasi Multinominal Naïve Bayes biasayanya digunakan pada klasifikasi email sebagai spam atau bukan spam, klasifikasi dokumen sebagai topik tertentu, atau klasifikasi sentimen pada teks (positif, negatif, atau netral). Kode yang digunakan saat proses eksperimen dengan mengimport kelas MultinomialNB dari modul naïve bayes pada library scikit-learn. Dengan menggunakan MultinomialNB dapat melakukan pelatihan dan prediksi dengan mudah pada data teks.

Rumus Multinomial Naive Bayes digunakan untuk menghitung probabilitas dokumen terhadap setiap kategori yang ada. Probabilitas ini digunakan untuk memprediksi kategori mana yang paling mungkin untuk dokumen tersebut.

Rumus dasar Multinomial Naive Bayes adalah sebagai berikut:

$$\mathrm{P}(\frac{c}{d}) = \mathrm{P}(c) * \mathrm{P}(\frac{d}{c})/\mathrm{P}(d) \quad (3)$$

Keterangan:

$\mathrm{P} = \left(\frac{c}{d}\right)$ adalah probabilitas dokumen d masuk ke dalam kategori c.

$\mathrm{P} = (\mathrm{c})$ adalah probabilitas prior untuk kategori c.

$\mathrm{P} = \left(\frac{d}{c}\right)$ adalah probabilitas kemunculan fitur (kata atau istilah) dalam dokumen d, jika dokumen d masuk dalam kategori c.

$\mathrm{P} = (d)$ adalah probabilitas dari dokumen d.

Untuk menghitung probabilitas P(d|c), digunakan rumus sebagai berikut:

$$\mathrm{P}\big((d|c)\big) = \prod(P(t|c)\text{^}nt) \quad (4)$$

Keterangan:

$P = \left(\frac{t}{c}\right)$ adalah probabilitas kemunculan fitur t (kata atau istilah) dalam kategori c.

nt adalah jumlah kemunculan fitur t dalam dokumen d.

Untuk menghitung probabilitas P(c) dan P(d), digunakan rumus sebagai berikut:

$$P(c) = \frac{Nc}{N} \quad (5)$$

Keterangan:

Nc adalah jumlah dokumen dalam kategori c.

N adalah jumlah total dokumen.

$P(d)$dapat dihitung dengan cara yang sama seperti menghitung $P(d|c)$, dengan asumsi bahwa dokumen tersebut termasuk dalam semua kategori yang ada. Setelah semua probabilitas diperoleh, dokumen dapat diklasifikasikan ke dalam kategori dengan probabilitas tertinggi. Kategori dengan probabilitas tertinggi dianggap sebagai kategori dokumen yang paling mungkin.

d. Evaluasi

Langkah terakhir dalam penelitian ini, yaitu evaluasi, apakah algoritma Naive Bayes merupakan pengklasifikasi teks yang baik sehingga dapat memperoleh nilai akurasi yang tinggi pada analisis sentimen masyarakat terhadap isu kenaikan Biaya Perjalanan Ibadah Haji. Proses evaluasi dilakukan menggunakan confusion matrix dengan melihat nilai akurasi dan ROC dari hasil eksperimen yang dilakukan.

III. HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan pengujian sentimen analisis terhadap opini masyarakat terkait isu kenaikan Biaya Perjalanan Ibadah Haji (Bipih). Data akan diolah menggunakan Bahasa pemrograman Python dan jupyter sebagai teks editor. Data diperoleh dari media social twitter dengan jumlah 850 data. Data dikelompokkan secara manual kedalam opini pro dan opini kontra. Selanjutnya dilakukan preprocessing data, yaitu proses Cleaning Text, Tokenization, Stopword Removal dan Stemming untuk menghilangkan noise pada data. Data dibagi menjadi 2, yaitu data training dan data testing dengan perbandingan data 80:20. Data training

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

berjumlah 679 data dan data testing berjumlah 170 data. Selanjutnya, dilakukan pembobotan kata menggunakan TF-IDF.

Tahapan selanjutnya, yaitu penerapan algoritma Naive Bayes terhadap data yang telah selesai melalui proses preprocesing dan pembobotan. Proses evaluasi dalam penelitian ini menggunakan Confusion Matrix yang berfungsi untuk memvisualisasikan hasil kinerja dari suatu algoritma. Dari hasil uji coba diperoleh nilai akurasi sebesar 0.89 dengan nilai True Negatif (data negatif yang diprediksi benar) sebesar 74.12% dengan jumlah data 126 data, False Positif (data negatif namun diprediksi sebagai data positif) sebesar 0.59% dengan jumlah data 1 data, False Negatif (data positif namun diprediksi sebagai data negatif) sebesar 10.59% dengan jumlah data 18 data dan True Positif (data positif yang diprediksi benar) sebesar 14.71% dengan jumlah data 25 data. Dapat dilihat pada gambar 5.

Gambar 5. Confusion Matrix

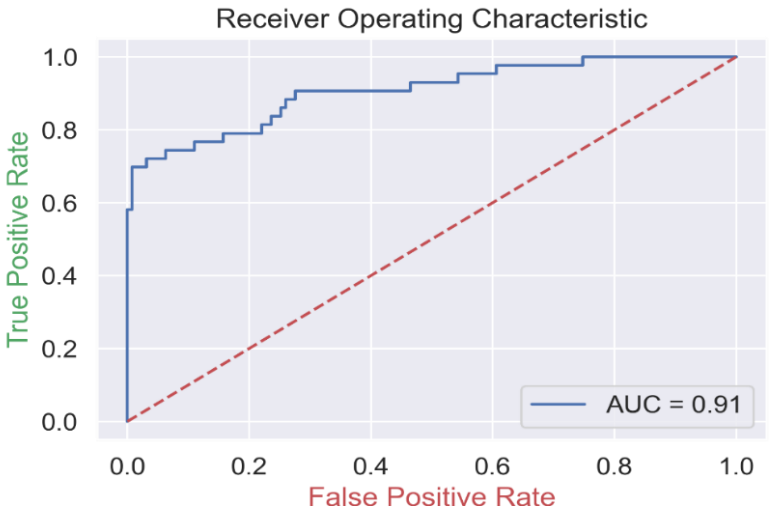
Berdasarkan confusion matriks yang dihasilkan maka dapat dilihat hasil dari nilai akurasi yang didapatkan. Hasil akurasi yang didapatkan, yaitu 0.89, diperoleh hasil akurasi yang tinggi karena False Positif dan False Negatif memiliki nilai yang lebih kecil dari nilai True Negatif dan True Positif. Hasil akurasi dapat dijabarkan dalam bentuk perhitungan menggunakan persamaan di bawah ini.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\% \quad (6)$$

$$Accuracy = \frac{14.71 + 74.12}{14.71 + 74.12 + 0.59 + 10.59} x\ 100\% = \frac{88.83}{100} = 0.89$$

Grafik *Receiver Operating Characteristic* (ROC) banyak digunakan untuk menilai hasil prediksi. ROC akan menunjukkan hasil prediksi dari algoritma pengklasifikasi yang digunakan. Kurva ROC ini dijadikan metrik evaluasi untuk model klasifikasi yang digunakan dan mengukur seberapa baik model dapat membedakan antara kelas positif dan negatif. Kurva ROC didapatkan dari nilai antara False Positive Rate dan True Positive Rate pada confusion matrix. Kinerjanya dari model yang digunakan dapat dilihat dari luas area dibawah kurva atau disebut *Area Under Curve* (AUC).

Jurnal Manajemen Informatika (JAMIKA)
Volume 13 Nomor 1 Edisi April 2023
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika



Gambar 6. Grafik ROC

Nilai AUC akan terlihat bagus kinerja jika luasnya lebih besar. Pada penelitian ini diperoleh nilai AUC sebesar 0,91 dan terbukti bahwa luas area di bawah kurva lebih besar dibandingkan dengan luas area di atas kurva. Nilai AUC yang didapatkan masuk kedalam kelompok *Excellent Classification* berdasarkan tingkat nilai diagnosa pada kurva ROC. Nilai AUC dengan range 0.90-1.00 masuk ke dalam *Excellent Classification* Penyajian kurva ROC dapat dilihat pada gambar 6.

IV. KESIMPULAN

Hasil penelitian mengenai analisis sentimen masyarakat terhadap isu kenaikan Biaya Perjalanan Ibadah Haji, terbukti bahwa algoritma Naive Bayes merupakan pengklasifikasi teks yang baik dengan memperoleh nilai akurasi sebesar 89% dan nilai ROC sebesar 0,91 masuk kedalam Excellent Classification. Penelitian ini diharapkan dapat menjadi bahan pertimbangan pemerintah dalam menentukan kenaikan biaya perjalanan ibadah haji, karena dengan adanya penelitian ini menunjukan bahwa banyak masyarakat yang tidak setuju (kontra) atas isu kenaikan biaya perjalanan ibadah haji ini. Diharapkan pada penelitian selanjutnya dapat dilakukan peningkatan nilai akurasi dengan menambahkan seleksi fitur seperti menggunakan algoritma Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Information Gain dan lain-lain.

UCAPAN TERIMA KASIH

Puji dan Syukur peneliti panjatkan kepada Tuhan Yang Maha Esa yang telah melimpahkan Rahmat dan Karunia-Nya sehingga peneliti dapat menyelesaikan penelitian dengan judul Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan Biaya Perjalanan Ibadah Haji. Peneliti juga mengucapkan terima kasih kepada institusi, yaitu Universitas Bina Sarana Informatika dan Universitas Nusa Mandiri yang telah memberikan dukungan kepada para peneliti sehingga penelitian ini dapat terselesaikan.

DAFTAR PUSTAKA

[1] A. Yusuf, "Kontroversi Biaya Haji," *Badan Pengelola Keuangan Haji*, Feb. 09, 2023. https://bpkh.go.id/kontroversi-biaya-haji/ (accessed Feb. 27, 2023).

[2] M. Khoeron, "BPIH, antara Kalkulasi Biaya dan Kebijakan Politik," *Kementerian Agama*, Feb. 21, 2023. https://kemenag.go.id/read/bpih-antara-kalkulasi-biaya-dan-kebijakan-politik-v5bm1 (accessed Mar. 06, 2023).

[3] E. Febriyani and H. Februariyanti, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Algoritma Naive Bayes Classifier Di Twitter," *Tekno Kompak*, vol. 17, no. 1, pp. 25–38, 2023.

[4] S. A. El Rahman, F. A. AlOtaibi, and W. A. AlShehri, "Sentiment Analysis of Twitter Data," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019.

[5] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *SMATIKA*, vol. 10, no. 2, pp. 71–76, 2020.

[6] A. Alsaeedi and M. Z. Khan, "A study on sentiment analysis techniques of Twitter data," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.

[7] M. Abbas, K. Ali, A. Jamali, K. Ali Memon, and A. Aleem Jamali, "Multinomial Naive Bayes Classification Model for Sentiment Analysis Wireless Sensor Networks View project Analyzing Distributed Denial of Service Attacks in Cloud Computing Towards the Pakistan Information Technology Industry View project Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS International Journal of Computer Science and Network Security*, vol. 19, no. 3, p. 62, 2019, doi: 10.13140/RG.2.2.30021.40169.

[8] S. P. PM and S. B, "Sentimental Analysis using Naive Bayes Classifier," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 2019, pp. 1–5.

[9] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 1, p. 55, Jan. 2021, doi: 10.22146/ijccs.60718.

[10] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler : Twitter," in *International Conference on Informatics and Computing (ICIC)*, 2019.

[11] M. Romzi and B. Kurniawan, "Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma," no. 2, pp. 37–44, 2020.

[12] "Project Jupyter's origins and governance," Mar. 13, 2023. https://jupyter.org/about (accessed Mar. 13, 2023).

[13] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," in *IOP Conference Series: Materials Science and Engineering*, Jul. 2020, vol. 874, no. 1. doi: 10.1088/1757-899X/874/1/012017.

[14] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy," in *International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019, pp. 112–117.

[15] "Sastrawi 1.0.1." https://pypi.org/project/Sastrawi/ (accessed Mar. 13, 2023).

[16] A. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification Pattern Recognition View project Improvement text classification using log(TF-IDF) with K-NN Algorithm View project," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 16, no. 6, pp. 22–36, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/

[17] A. Rahman Isnain, A. Indra Sakti, D. Alita, and N. Satya Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma SVM," *JDMSI*, vol. 2, no. 1, pp. 31–37, 2021, [Online]. Available: https://t.co/NfhnfMjtXw

[18] M. Chiny, M. Chihab, Y. Chihab, and O. Bencharef, "LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, pp. 265–275, 2021, [Online]. Available: www.ijacsa.thesai.org

[19] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional," vol. 15, no. 1, 2021.

[20] "1.9. Naive Bayes." https://scikit-learn.org/stable/modules/naive_bayes.html (accessed Mar. 13, 2023).