Jurnal CoreIT, Vol.7, No.2, Desember 2021
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Pemanfaatan Metode Klasifikasi *Naïve Bayes* Untuk Pendeteksi Berita *Hoax* Pada Artikel Berbahasa Indonesia

Soleman[1*]
[1]Program Studi Manajemen Informatika, Universitas Borobudur, Indonesia
solemediagrafik@gmail.com (**corresponding author*)

Abstrak – Berita hoax telah banyak tersebar di internet. Kemudahan dalam membuat dan berbagi informasi merupakan salah satu faktornya. Berita hoax menjadi ancaman dan konsentrasi banyak pihak, permasalahan muncul dalam mengidentifikasi atau mengklasifikasikannya karena tidak ada pola yang dapat diidentifikasi, dan adanya gaya penulisan yang bebas dan tidak kaku. Akurasi yang rendah dari sistem pendeteksi hoax yang saat ini ada ditemukan metode dan atribut yang dapat digunakan untuk mengklasifikasikan berita hoax dengan akurasi yang tinggi. Atas dasar itulah penelitian ini dilakukan, seperti pada kebanyakan klasifikasi berita hoax yang dijadikan acuan dalam penelitian ini, dilakukan tahapan *preprocessing (case folding, tokenization, stemming, dan stopword removal)*, ekstraksi fitur dan penambahan atribut serta artikel tentang preprocessing. Hal ini diambil dari situs web tempat artikel diposting, publikasi dan status situs. Hasil dari penelitian ini diperoleh akurasi sebesar 72% yang ternyata mengalami penurunan sebesar 6,6% dibandingkan penelitian sebelumnya yaitu 78,6% karena satu situs hanya menerbitkan satu artikel hoax dan memungkinkan domain dari situs tersebut sudah kadaluarsa sehingga mengurangi bobot dari nilai klasifikasi.
Kata Kunci: Artikel Hoax, Klasifikasi, *Naïve Bayes, Preprocessing.*

Abstract *– Hoax news has been widely spread on the internet. Ease of creating and sharing is one factor. Hoax news is a threat and concentration of many parties, problems arise in identifying or classifying them because there is no identifiable pattern, and the writing style is free and not rigid. The lack of accuracy of the existing hoax detection system found the methods and attributes used to classify hoax news with high accuracy. On this basis, this research was carried out, as in most classifications of hoax news used as reference in this study, preprocessing was carried out (case folding, tokenization, stemming, and stopword removal), feature extraction and adding attributes other than preprocessing articles. done like the website where the article is posted. publication and site status. The results of this study obtained an accuracy of 72% which turned out to be a decrease of 6.6% compared to previous research, namely 78.6% because one site only publishes one hoax article and allows the site's domain to expire thereby reducing the weight of the classification value.*

Keywords: *Classification, Hoax article, Naïve Bayes, Preprocessing.*

Received *September 2021 /* ***Revised*** *November 2021 /* ***Accepted*** *December 2021*

PENDAHULUAN

Pengguna layanan internet semakin hari semakin meningkat. Menurut hasil survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) setelah melakukan survei penetrasi dan perilaku pengguna internet di Indonesia, jumlah pengguna internet pada tahun 2016 adalah 132,7 juta jiwa dan jumlah pengguna internet pada tahun 2017 telah mencapai 143,26 juta jiwa atau setara dengan 54,68% dari total jumlah penduduk Indonesia [1]. Pesatnya perkembangan tersebut tidak hanya berdampak positif, terdapat juga dampak negatif yang dihasikan. Setiap informasi yang beredar tanpa melewati penyutingan serta validasi kebenaran yang tidak jelas.

DOI : 10.24014/coreit.v7i2.14290

Fenomena ini sering dimanfaatkan untuk mencari keuntungan dari menyebarkan informasi hoax. Menurut Presiden Direktur VIVA Media Group, Anindya Novyan Bakrie, Persentase berita hoax di media sosial mencapai 92,40%, disusul aplikasi chatting 62,80%, lalu website 34,90%, sementara untuk media yang sudah kurang diminati seperti televisi hanya 8.70%, media cetak 5%, email 3,10% dan radio 1,20% [2]. Tujuan dari berita hoax adalah untuk mempengaruhi opini dan pandangan publik dalam suatu hal tertentu. Mendeteksi berita hoax tidaklah mudah, informasi dicampur dan diolah sedemikian rupa sehingga membuat pembaca terkecoh serta dapat membangkitkan kesan sebagai kebenaran baru dan semua orang harus tahu. Pada penelitian sebelumnya telah diperoleh hasil akurasi yang cukup tinggi, rata-rata hasil akurasi sekitar 78.6% dengan metode naïve bayes untuk klasifikasi artikel [3]. Dengan menggunakan dataset yang sama pada penelitian sebelumnya, penulis ingin meneliti dengan menambahkan 2 atribut dalam klasifikasi apakah akan menambah akurasi dalam mendeteksi berita hoax.

METODE

1. Metode Analisis

Menambahkan atribut untuk mengklasifikasikan atrikel berita valid dan berita hoax sebelumnya yang menggunakan metode klasifikasi *Naïve* Bayes agar lebih akurat [4]. Untuk mengetahuinya perlu dilakukan pengujian atribut terhadap metode klasifikasi naïve bayes. Hal ini dilakukan untuk mengetahui apakah ada pengaruh penambahan atribut terhadap metode klasifikasi naïve bayes. Berdasarkan maksud dan ruang lingkup ini maka menggunakan metode eksperimen. Metode eksperimen ini dilakukan dengan memanipulasi kondisi sesuai dengan kebutuhan permasalahan yang dihadapi di dalam penelitian. Dengan memanipulasi kondisi ini,nantinya hasilnya akan menghasilkan algoritma dengan tingkat akurasi yang lebih tinggi dari penelitian sebelumnya.

2. Metode Pengumpulan Data

Dataset menggunakan data berdasarkan penelitian yang dilakukan Pratiwi [5] yang terdiri dari 250 data artikel dalam bahasa Indonesia. Artikel hoax dan bukan hoax yang terdiri dari 10 topik yang berbeda dan setiap topik terdiri dari 25 berita. Topik berita tersebut adalah "Makan lele menyebabkan sel kanker, Aku puntur dengan jarum menyebabkan stroke, Iphone 6 mudah dibengkokan, Reog Ponorogo dibakar di Filipina, Simpatisan Aksi 212 dilarang masuk masjid Istiqlal, Sikat gigi dari rambut babi, Permen dot mengandung narkoba, Pokemon berarti "Aku Yahudi", Foto Awan Berdoa Di Pemakaman Uje, Munarman pengacara Freeport"

3. Rancangan Sistem

Tambahan atribut dalam klasifikasi teks berita hoax yang dapat digunakan untuk mendeteksi apakah suatu artikel berita tersebut hoax atau tidak. Atribut yang ditambahkan adalah website yang mempublikasi artikel dan status website yang masih aktif atau tidak, atribut tersebut akan diterapkan pada klasifikasi naïve bayes. Di Indonesia terdapat seitar 43.000 website yang mengklaim sebagai portal berita menurut catatan Dewan Pers. Website yang terverifikasi sebagai portal berita resmi tidak mencapai 300 website [4]. Dari perbandingan jumlah website dan yang website telah terverifikasi dapat disimpulkan terdapat puluhan ribu yang berpotensi menyebarkan berita hoax. Informasi yang berasal dari website tidak terverifikasi seperti blog pribadi perlu diwaspadai karena kemungkinan berita tersebut adalah *hoax*. Sistem yang diusulkan pada penelitian ini dibagi menjadi dua tahap yaitu tahap pelatihan dan pengujian. Pada tahap pelatihan digunakan untuk menciptakan model klasifikasi artikel berita itu hoax atau bukan, sedangkan dalam tahap pengujian untuk mengklasifikasikan apakah artikel atau dokumen masukan tersebut *hoax* atau bukan.

4. Praposes

Teks bisa dalam berbagai bentuk dari daftar kata-kata, hingga kalimat ke beberapa paragraf dengan karakter khusus. Seperti halnya masalah data science, memahami pertanyaan yang ditanyakan akan menginformasikan langkah apa yang dapat digunakan untuk mengubah kata-kata menjadi fitur numerik dengan menggunakan algoritma pembelajaran mesin. Tahap praproses adalah tahapan dimana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen. Praproses digunakan untuk merubah teks yang tidak terstruktur menjadi token representasi yang siap dimodelkan oleh algoritma klasifikasi. Praproses terdiri dari pemrosesan leksikal dan pengubahan kata ke fitur kata. Pemrosesan leksikal meliputi [6] *Case folding, Tokenisasi,* Penghapusan *stopword*,dan *Stemming.*

DOI : 10.24014/coreit.v7i2.14290

5. Ekstraksi Fitur

Ekstraksi fitur adalah proses mengekstrak seluruh fitur kata yang terdapat dalam dokumen latih [7]. Keluaran dari proses ini adalah kumpulan kata yang dijadikan penciri dokumen berita hoax dan bukan hoax. Fitur kata ini di dapatkan dari proses tokenisasi.

6. Seleksi Fitur

Seleksi fitur dilakukan sebelum proses klasifikasi terhadap dataset. Pada penelitian ini penulis menggunakan algoritma *TF-IDF (Term Frequency – Inverse Document Frequency)* yang digabungkan dengan algoritma *SGD (Stochastic Gradient Descent)* karena menurut penelitian dari Agung [8] mengatakan bahwa keakuratan *TF-IDF* lebih baik dikombinasikan dengan *SGD* ketika sampel klasifikasi tidak memiliki ketentuan khusus, sebaliknya, ia memiliki pola unik di penyedia portal berita yang sama.

7. Klasifikasi

Pengkalsifikasian data dilakukan terhadap dataset berdasarkan atribut yang dimiliki oleh dataset berita hoax tersebut. Pada penelitian ini penulis menggunakan klasifikasi *Naïve Bayes*. Untuk merepresentasikan sebuah kelas dokumen, terdapat karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi yang berguna untuk menjelaskan bahwa peluang masuknya sampel karakteristik tertentu kedalam kelas *posterior*. Klasifikasi *Naïve Bayes* diasumsikan bahwa ada atau tidaknya ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Persamaan dari teorema bayes adalah [6]:

$$P(H|X) = (P(X|H)\ P(H)) \ / \ (P(X)) \qquad (1)$$

Dimana nilai X adalah data kelas yang belum diketahui, H adalah hipotesis X pada label tertentu, P(H|X) adalah probabilitas H berdasarkan kondisi X (posteriori), P(H) adalah probabilitas H (prior), P(X|H) adalah probabilitias X.

8. Pengukuran Akurasi

Dalam menguji keefektifan suatu klasifikasi dibutuhkan suatu pengukuran evaluasi. Pengukuran tersebut didapatkan dalam sebuah *set confusion matrix* [9]. Kondisi yang digunakan dalam pengujian adalah sebagai berikut *True Positive* adalah kondisi berita hoax diklasifikasi sebagai hoax, *True Negative* adalah kondisi berita bukan hoax diklasifikasikan sebagai bukan hoax, *False Positive* adalah kondisi berita hoax diklasifikasikan sebagai bukan hoax, *False Negative* adalah kondisi berita bukan hoax diklasifikasikan sebagai berita hoax.

9. Pengujian Sistem

Pengujian sistem yang dilakukan pada penelitian ini untuk melihat performansi atribut tambahan pada klasifikasi *Naïve Bayes* dalam melakukan prediksi terhadap data testing [10]. Performansi diukur dengan melakukan perbandingan antara hasil testing yang diklasifikasikan oleh sistem dengan data testing yang sebelumnya telah diberi label.

HASIL DAN PEMBAHASAN

1. Analisa Sistem

Proses yang akan dilakukan untuk membuat sistem prediksi adalah proses pelatihan yang akan dilanjutkan dengan proses pengujian untuk mendapatkan hasil prediksinya. Data berita yang digunakan pada sistem ini terbatas yaitu terdiri dari 250 data artikel dalam bahasa Indonesia. Artikel hoax dan bukan hoax yang terdiri dari 10 topik yang berbeda dan setiap topik terdiri dari 25 berita. Pada dataset yang akan digunakan terdapat beberapa perbandingan dari proses tagging manual dan memeriksa status website. Proses tagging valid atau hoax dilakukan dengan menggunakan sistem voting 3 reviewer. Hasil dari manual tagging dari 250 dataset artikel yang digunakan dalam penelitian terdapat Artikel Valid adalah 159 dan Artikel Hoax adalah 91. Untuk mengetahui status website aktif atau tidak, penulis mengunjungi link website yang ada dalam dataset pada penelitian sebelumnya satu per satu secara manual. terdapat website 192 aktif dan 58 website mati.
Data yang digunakan dapat dilihat pada Tabel 1 dibawah ini:

DOI : 10.24014/coreit.v7i2.14290

Tabel 1. Data Yang Digunakan

ID	Articles	Tagging	Website	Status
1	Isu bahwa ikan lele mengandung sel kanker di jejaring sosial dan berita dari mulut ke mulut terus menyebar. Dampak dari isu tersebut para ibu-ibu enggan membeli ikan lele. Waspada Online berhasil merangkum komentar ibu-ibu yang biasanya membeli ikan lele untuk konsumsi rutin	*valid*	waspada.co.id	aktif
2	Ikan lele merupakan salah satu makanan favorit di Indonesia. Selain harganya murah, rasanya juga sangat enak. Meski demikian, ada sebagian masyarakat yang takut menikmati masakan dari ikan air tawar tersebut. Mereka beranggapan jika ikan lele penyebab kanker dan penyakit lainnya. Namun apakah anggapan yang menyatakan lele mengandung kanker tersebut benar?. Berikut ini penjelasannya. Habitat dan Kehidupan Ikan Lele.....	*valid*	transferfactorformula.com	aktif
3	Kepala Bagian Penerangan Umum (Kabagpenum) Polri Kombes Pol Martinus Sitompul memantah isu terjadi pelemparan Alquran oleh petugas jaga di Rutan Mako Brimob Cabang Salemba, Jakarta Pusat, Jumat (10/11/2017). Rutan Mako Brimob Cabang Salemba kerusuhan……	*valid*	bacaberitaupdate.com	mati
4	Rumah Tahanan (Rutan) Mako Brimob, Kelapa Dua, Depok, Jawa Barat, dikabarkan rusuh, Jumat (10/11/2017) sore. Kabar ini ramai dibicarakan di media sosial. "Mako Brimob rusuh tadi sore. Kabarnya 3 blok ruang tahanan terbakar & pintu2nya hancur. Menurut sumber saya, sebabnya karena sekelompok Polisi di sana melempar Al-Qur'an dan buku2 hadits milik tahanan," kicau pemilik akun Twitter @CondetWarrior…….	*hoax*	harianumum.com	aktif
5	Malam peringatan ulang tahun ke-90 Kolese Kanisius di Hall D JIExpo Kemayoran, Jakarta Utara, Sabtu (11/11/2017) lalu terus menjadi perbincangan hangat beberapa hari terakhir……	*hoax*	netralnews.com	aktif
6	Tahun 2012 kita dihebohkan dengan munculnya kuas atau sikat yang bebahan bulu babi sampai-sampai MUI mengeluarkan fatwa haram menggunakan sikat bulu babi, namun baru-baru ini muncul kembali info sikat gigi yang berbahan bulu babi. Informasi bagi anda kaum Muslim agar lebih teliti dalam memilih sikat gigi, karena ada produk pembersih gigi yang menggunakan bahan dari yang tidak halal….	*hoax*	pelangimuslim.com	mati

Pada dataset yang akan digunakan terdapat beberapa perbandingan dari proses tagging manual dan memeriksa status website. Proses tagging valid atau hoax dilakukan dengan menggunakan sistem voting 3 reviewer dapat dilihat pada Tabel 2.

Tabel 2. Proses Voting Label Valid atau Hoax

No.	Artikel	Rev-1	Rev-2	Rev-3	Hasil
1	Isu bahwa ikan lele mengandung sel kanker di jejaring sosial dan berita dari mulut ke mulut terus menyebar. Dampak dari isu tersebut para ibu-ibu enggan membeli ikan lele. Waspada Online berhasil merangkum komentar ibu-ibu yang biasanya membeli ikan lele untuk konsumsi rutin	*Valid*	*Valid*	*Valid*	*Valid*
2	Ikan lele merupakan salah satu makanan favorit di Indonesia. Selain harganya murah, rasanya juga sangat enak. Meski demikian, ada sebagian masyarakat yang takut menikmati masakan dari ikan air tawar tersebut. Mereka beranggapan jika ikan lele penyebab kanker dan penyakit lainnya. Namun apakah anggapan yang menyatakan lele mengandung kanker tersebut benar?. Berikut ini penjelasannya. Habitat dan Kehidupan Ikan Lele.....	*Valid*	*Valid*	*Hoax*	*Valid*

DOI : 10.24014/coreit.v7i2.14290

3	Kepala Bagian Penerangan Umum (Kabagpenum) Polri Kombes Pol Martinus Sitompul memantah isu terjadi pelemparan Alquran oleh petugas jaga di Rutan Mako Brimob Cabang Salemba, Jakarta Pusat, Jumat (10/11/2017). Rutan Mako Brimob Cabang Salemba kerusuhan…….	*Valid*	*Valid*	*Valid*	*Valid*
4	Rumah Tahanan (Rutan) Mako Brimob, Kelapa Dua, Depok, Jawa Barat, dikabarkan rusuh, Jumat (10/11/2017) sore.Kabar ini ramai dibicarakan di media sosial."Mako Brimob rusuh tadi sore. Kabarnya 3 blok ruang tahanan terbakar & pintu2nya hancur. Menurut sumber saya, sebabnya karena sekelompok Polisi di sana melempar Al-Qur'an dan buku2 hadits milik tahanan," kicau pemilik akun Twitter @CondetWarrior……..	*Hoax*	*Hoax*	*Hoax*	*Hoax*
5	Malam peringatan ulang tahun ke-90 Kolese Kanisius di Hall D JIExpo Kemayoran, Jakarta Utara, Sabtu (11/11/2017) lalu terus menjadi perbincangan hangat beberapa hari terakhir…….	*Hoax*	*Hoax*	*Hoax*	*Hoax*
6	Tahun 2012 kita dihebohkan dengan munculnya kuas atau sikat yang bebahan bulu babi sampai-sampai MUI mengeluarkan fatwa haram menggunakan sikat bulu babi, namun baru-baru ini muncul kembali info sikat gigi yang berbahan bulu babi. Informasi bagi anda kaum Muslim agar lebih teliti dalam memilih sikat gigi, karena ada produk pembersih gigi yang menggunakan bahan dari yang tidak halal……	*Hoax*	*Hoax*	*Hoax*	*Hoax*

Hasil dari manual tagging dari 250 dataset artikel yang digunakan dalam penelitian dapat dilihat pada Tabel 3 dibawah ini :

Tabel 3. Perbandingan Valid dan Hoax

Tagging	**Total**
Artikel Valid	159
Artikel Hoax	91

Untuk mengetahui status website aktif atau tidak, penulis mengunjungi link website yang ada dalam dataset pada penelitian sebelumnya satu per satu secara manual dapat dilihat pada Tabel 4 dibawah ini:

Tabel 4. Perbandingan Status Website

Status	**Total**
Aktif	192
Mati	58

Terdapat banyak jenis website yang telah tidak bisa di akses atau mati. terdapat beberapa website yang diblock oleh google, webserver mati dan juga domain yang sudah habis masa berlakunya dapat dilihat pada Gambar 1.

Gambar 1. Contoh Website yang Telah Mati

DOI : 10.24014/coreit.v7i2.14290

Artikel pada dataset tersebut akan di-import ke dalam prototype dengan menggunakan libraries pandas pada pyhton untuk dilakukan proses selanjutnya seperti preprocessing dan klasifikasi dapat dilihat pada Gambar 2.

Out[140]:

	Articles	Tagging	Website	Status
0	Jakarta, Di jejaring sosial, banyak beredar in...	valid	health.detik.com	aktif
1	Isu bahwa ikan lele mengandung sel kanker di j...	valid	waspada.co.id	aktif
2	Bagi penikmat kuliner dengan bahan dasar ikan ...	valid	tongkatmaduraasli.com	mati
3	Ikan lele merupakan salah satu makanan favorit...	valid	transferfactorformula.com	mati
4	Ikan lele merupakan bahan makanan yang cukup p...	valid	tribunnews.com	aktif
5	"Dalam sesuap daging ikan lele, terkandung 3.0...	hoax	regional.kompas.com	aktif
6	Bahaya Mengkonsumsi Ikan Lele Yang Mengandung ...	hoax	mediamasha.com	aktif
7	Di jejaring sosial banyak beredar informasi y...	hoax	beritalive.com	aktif
8	Jakarta Sebuah artikel yang cukup viral di in...	valid	crystalsweb.com	aktif
9	Pada dasarnya tidak ada makanan yang membawa s...	valid	deherba.com	aktif

Gambar 2. Import Dataset ke Prototype

A. Implementasi

Pada tahap ini berisi hasil implementasi dari proses yang telah dirancang sebelumnya yaitu preprocessing, Ekstraksi fitur, seleksi fitur, dan implemenasi proses klasifikasi untuk prediksi berita *hoax*.

1) Proses *Stemming*

Pada proses s*temming* hasilnya dapat dilihat Gambar 3 dibawah ini :

```
Out[20]: ['Ikan => Ikan',
          'lele => lele',
          'merupakan => rupa',
          'salah => salah',
          'satu => satu',
          'makanan => makan',
          'favorit => favorit',
          'di => di',
          'Indonesia => Indonesia',
          '. => .',
          'Selain => Selain',
          'harganya => harganya',
          'murah => murah',
          '  =>  ',
          'rasanya => rasa',
```

Gambar 3. Stemming

Dari Gambar 3 diatas dapat dilihat bahwa hasil dari proses *stemming* menggunakan *stemmer spacy* berjalan dengan baik. Hal ini dibuktikan dengan kata "merupakan" berubah menjadi "rupa" yang memang merupakan kata dasar dari kata "merupakan".

2) Proses *Case Folding*

Pada proses *case folding*, variabel hasil stemming akan dilakukan proses *case folding* sehingga semua kata dalam dataset berubah menjadi kata dengan huruf kecil. Hasil casefolding dapat dilihat pada Gambar 4.

DOI : 10.24014/coreit.v7i2.14290

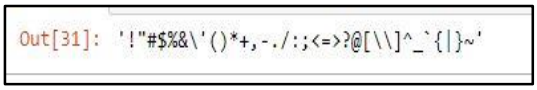
Gambar 4 . *Case Folding*

3) Proses Penghilangan *Stopword* dan Tanda Baca

Pada proses penghilangan *stopword*, dibutuhkan modul *Spacy*. Modul *Spacy* menyediakan beberapa *corpora teks*, salah satunya adalah *Stopwords Corpus.* Selain kata-kata umum, ada juga kelompok kata *stopword* yang memiliki posisi penting dalam morfologi dan tidak bisa berdiri sendiri. *Stopword* yang akan dihilangkan dapat dilihat pada Gambar 5.

Gambar 5. *Stopword*

Jumlah keseluruhan stopword bahasa Indonesia ada 757 kata. Setelah menghilangkan *stopword* pada kalimat maka proses selanjutnya adalah mengilangkan tanda baca pada kalimat. Tanda baca yang akan dihilangkan dapat dilihat pada Gambar 6.



Gambar 6. Tanda Baca

B. Proses Tokenisasi

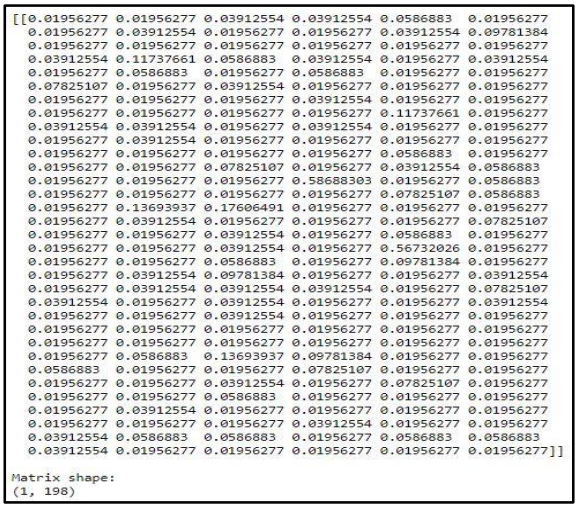
Pada proses tokenisasi adalah proses dimana mengubah sebuah kalimat menjadi *unigram* kata [11]. Dalam proses ini dibutuhkan modul *spacy NLP*. Meski Python memiliki kemampuan untuk melakukan tugas-tugas *Natural Language Processing* dasar, namun tidak cukup powerful untuk melakukan tugas-tugas standar NLP, maka dari itu muncullah modul *spacy*. Modul ini menyediakan berbagai fungsi dan *wrapper*, serta *corpora* standar baik itu mentah atau *pre-processed*. Pada proses ini pula penulis mencoba untuk menggabungkan beberapa implementasi sebelumnya seperti *stemming, case folding, stopword removal* menjadi satu fungsi. Untuk proses tokenisasi adalah seperti terlihat pada Gambar 7.

DOI : 10.24014/coreit.v7i2.14290

Gambar 7. Hasil *Tokenisasi* dan Penggabungan Algoritma*Preprocessing*

C. Proses Pembobotan TF-IDF

Pada proses pembobotan *TF-IDF* digunakan modul pustaka *scikit-learn* untuk mengekstraksi teks dengan menggunakan *TfidfVectorizer*. Hasil dari pembobotan TF-IDF berupa matriks. Untuk hasil proses pembobotan *TF-IDF* adalah seperti terlihat pada Gambar 8 dibawah ini :



Gambar 8. Matriks pembobotan *TF-IDF*

Gambar 8 diatas menunjukkan hasil *TF-IDF* yang telah dilakukan. Matriks *TF- IDF* yang dihasilkan berukuran 1 x 198 dengan hanya menggunakan satu artikel contoh. Matriks dari total keseluruhan data artikel adalah 250 x 5853, data website adalah 250 x 205 dan data status adalah 250 x 2. Setelah memanggil nilai label atau tagging dan data matriks *TF-IDF* selanjutnya dilakukan proses klasifikasi.

D. Klasifikasi Berita Hoax

Pada proses ini penulis akan menggabungkan semua fitur *preprocessing text*, tokenisasi dan ekstrasi fitur ke dalam sebuah fungsi. Fungsi tersebut akan digunakan bersamaan dengan proses klasifikasi menggunakan *pipeline*. Fitur-fitur dan fungsi yang telah dibuat sebelumnya akan dimasukan ke dalam *pipeline* ini. *Pipeline* pertama berisi fitur hasil dari *preprocessing* artikel, *Pipeline* kedua berisi fitur hasil dari *preprocessing website* dan ketiga berisi fitur hasil dari *preprocessing* status dari usulan atribut penulis. Sebelum melakukan proses klasifikasi, ketiga *pipeline* tersebut akan di-*union* dengan menggunakan fungsi *union pipeline* pada *pipeline* gabungan dan hasil *union* akan digunakan untuk proses klasifikasi pada *pipeline* gabungan dapat dilihat pada Gambar 9 dibawah ini :

DOI : 10.24014/coreit.v7i2.14290

```
pipe = Pipeline([("cleaner", predictors()),
                 ('vectorizer', tfvectorizer),
                 ('to_dense', DenseTransformer())])
pipeZ1 = Pipeline([("cleaner", predictorsZ1()),
                 ('vectorizer', vectorizer),
                 ('to_dense', DenseTransformerZ1())])
pipeZ2 = Pipeline([("cleaner", predictorsZ2()),
                 ('vectorizer', vectorizer),
                 ('to_dense', DenseTransformerZ2())])
customPipeline = Pipeline([
    ('feat_union', FeatureUnion(transformer_list=[
        ('pl1', pipe),
        ('pl2', pipeZ1),
        ('pl3', pipeZ2)
     ]))
    ])
```

Gambar 9. Implementasi Pembuatan *Pipeline*

Pada Gambar 9 diatas terlihat bahwa pada *pipeline* terdapat beberapa langkah berurutan yang akan dieksekusi oleh sistem. Urutan pertama ialah *cleanner* yang akan berfungsi untuk *text cleaning* seperti *case folding*, yang kedua ialah ekstraksi fitur menggunakan TF-IDF atau tokenisasi, dalam proses ini terdapat proses *stemming*, *stopword removal*, *string punctuation*. Pada proses kedua ini akan menghasilkan matriks TF-IDF yang akan ditransformasikan ke dalam *dense matriks* pada urutan ketiga agar dapat dilakukan perhitungan *FeatureUnion*.

E. Implementasi Klasifikasi Dengan Algoritma *Naive Bayes*.

Implementasi proses klasifikasi berita dengan algoritma *Naïve Bayes* ialah dengan cara menambahkan fungsi classifier pada *pipeline* dengan fungsi algoritma *Naïve Bayes* [12] seperti yang dapat dilihat pada potongan *script* pada Gambar 10 dibawah ini.

```
customPipeline = Pipeline([
    ('feat_union', FeatureUnion(transformer_list=[
        ('pl1', pipe),
        ('pl2', pipeZ1),
        ('pl3', pipeZ2)
     ])),
    ('classify', GaussianNB())
    ])
```

Gambar 10. Implementasi klasifikasi dengan *Naïve Bayes*

F. Implementasi Pembentukan Model Klasifikasi.

Implementasi klasifikasi berita adalah proses dimana fungsi – fungsi yang sebelumnya telah dibangun digunakan untuk melatih model menggunakan data berita. Pada proses ini semua proses *preprocessing*, seleksi fitur dan pembelajaran itu sendiri dijalankan. Hasil keluaran dari proses ini ialah suatu model yang dapat kita simpan dan kita gunakan untuk klasifikasi berita. *Script* untuk menjalankan semua proses atau fungsi tersebut dapat dilihat pada Gambar 11 dibawah ini :

```
X_train, X_test, y_train, y_test = train_test_split(X, ylabels, test_size=0.3, random_state=42)
trainProcess = customPipeline.fit(X_train, y_train)
trainProcess.score(X_test,y_test)
```

Gambar 11. Implementasi Pembentukan Model Klasifikasi

G. Pengujian Confusion Matrix

Pada proses pembobotan *TF-IDF* digunakan modul pustaka *scikit-learn* untuk mengekstraksi teks dengan menggunakan *TfidfVectorizer* [13]. Hasil dari pembobotan *TF-IDF* berupa matriks. Pengujian *confusion matrix* akan dilakukan uji coba dengan merubah perbandingan antara *data test* dan *data train*. Pada pengujian terdapat beberapa kondisi sebagai berikut: *True Positive* adalah kondisi berita *hoax* diklasifikasi sebagai *hoax*, *True Negative* adalah kondisi berita bukan *hoax* diklasifikasikan sebagai bukan *hoax*, *False Positive* adalah kondisi berita *hoax* diklasifikasikan sebagai bukan *hoax*, *False Negative* adalah kondisi berita bukan *hoax* diklasifikasikan sebagai berita *hoax* [14].

DOI : 10.24014/coreit.v7i2.14290

Pengujian *confusion matrix* yang menggunakan perbandingan 10:90 menghasilkan akurasi 68% dengan nilai *True Positive* (TP) adalah 11, *True Negative* (TN) adalah 6, *False Positive* (FP) adalah 4, *False Negative* (FN) adalah 4. Untuk pengujian *confusion matrix* menggunakan perbandingan 20:80 menghasilkan akurasi 72% dengan nilai *True Positive (TP)* adalah 25, *True Negative (TN)* adalah 11, *False Positive (FP)* adalah 7, *False Negative (FN)* adalah 7. Pengujian *confusion matrix* menggunakan perbandingan 30:70 menghasilkan akurasi 72% dengan nilai *True Positive* (TP) adalah 41, *True Negative* (TN) adalah 13, *False Positive* (FP) adalah 15, *False Negative* (FN) adalah 6. Pengujian *confusion matrix* menggunakan perbandingan 40:60 menghasilkan akurasi 71% dengan nilai *True Positive* (TP) adalah 55, *True Negative* (TN) adalah 16, *False Positive* (FP) adalah 22, *False Negative* (FN) adalah 7. Pengujian *confusion matrix* menggunakan perbandingan 50:50 menghasilkan akurasi 71,2% dengan nilai *True Positive (TP)* adalah 68, *True Negative (TN*) adalah 21, *False Positive (FP)* adalah 27, *False Negative (FN)* adalah 9. Berikut pengujian perbandingan yang dilakukan dapat dilihat pada Tabel 5 dibawah ini :

Tabel 5. Hasil Pengujian Perbandingan

Test : Train	Akurasi	TP	TN	FP	FN
10% : 90%	68%	11	6	4	4
20% : 80%	72%	25	11	7	7
30% : 70%	72%	41	13	15	6
40% : 60%	71%	55	16	22	7
50% : 50%	71,2%	68	31	27	9

Dari hasi Pengujian Perbandingan pada Tabel 5 tersebut dapat disimpulkan bahwa perbandingan *data test* dan *data train* yang menghasilkan nilai terbaik adalah 20:80 dan 30:70 dengan tingkat akurasi 72%.

KESIMPULAN

Beberapa kesimpulan yang dapat diperoleh dari penelitian yang telah dilakukan adalah sebagai berikut:

1. Akurasi terbaik dari percobaan di hasilkan dari data train 80%, data test 20% dan data train 70%, data test 30%.
2. Akurasi yang dihasilkan sistem pada proses klasifikasi dengan penambahan atribut adalah 72% dibandingkan dengan penelitian sebelumnya terjadi penurunan 6.6% dari hasil sebelumnya yaitu 78.6%.
3. Turunnya akurasi dikarenakan penurunan bobot klasifikasi dari *website*. setiap *website* penyebar *hoax* hanya mempublikasi satu topik dan satu berita dan kemudian domain website tersebut dibiarkan *expired* atau mati. Berita hoax mengikuti tren terkini, penyebar *hoax* hanya akan menggunakan satu website untuk mempublikasi sedikit artikel hoax.

REFERENSI

[1]. APJII. Asosiasi Penyelenggara Jasa Internet Indonesia. 2017, https://www.apjii.or.id/content/read/39/342/Hasil-Survei-Penetrasi-dan-Perilaku-Pengguna-Internet-Indonesia-2017.

[2]. Tim VIVA. Anindya Bakrie: Penyebar Hoax Terbanyak Itu Media Sosial – VIVA. 2018, https://www.viva.co.id/berita/nasional/1005218-anindya-bakrie-penyebar-hoax-terbanyak-itu-media-sosial.

[3]. Rasywir, Errissya, and Ayu Purwarianti. "Eksperimen Pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin." Jurnal Cybermatika, vol. 3, no. 2, 2015, pp. 1–8.

[4]. Yunita. Ini Cara Mengatasi Berita "Hoax" di Dunia Maya. 2017, https://kominfo.go.id/content/detail/8949/ini-cara-mengatasi-berita-hoax-di-dunia-maya/0/sorotan_media.

[5]. Pratiwi, Inggrid Yanuar Risca, et al. "Study of Hoax News Detection Using Naïve Bayes Classifier in Indonesian Language." 2017 11th International Conference on Information & Communication Technology and System (ICTS), no. February, 2017, pp. 73–78, doi:10.1109/ICTS.2017.8265649.

[6]. Adetunji, A B, Oguntoye, J P, Fenwa, O D, Akande, N O. "Web Document Classification Using Naïve Bayes". Journal of Advances in Mathematics and Computer Science 29(6): 1-11, 2018; Article no.JAMCS.34128 ISSN: 2456-9968, 2017

DOI : 10.24014/coreit.v7i2.14290

[7]. Wang, Yong Hodges, Julia Tang, Bo. "Classification of Web Documents Using a Naive Bayes Method" Department of Computer Science & Engineering, Mississippi State University Mississippi State, MS 39762-9637, 2015
[8]. Prasetijo, Agung B., et al. "Hoax Detection System on Indonesian News Sites Based on Text Classification Using SVM and SGD." Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017, vol. 2018–Janua, 2018, pp. 45–49, doi:10.1109/ICITACEE.2017.8257673.
[9]. Sarkar, Dipanjan. Text Analytics with Python. 2016, doi:10.1007/978-1-4842-2388-8.
[10]. Zhang, Yunan, et al. "Using Multi-Features and Ensemble Learning Method for Imbalanced Malware Classification." Proceedings - 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 10th IEEE International Conference on Big Data Science and Engineering and 14th IEEE International Symposium on Parallel and Distributed Proce, 2016, pp. 965–73, doi:10.1109/TrustCom.2016.0163.
[11]. Chen, Yoke Yie, et al. "Email Hoax Detection System Using Levenshtein Distance Method." Journal of Computers, vol. 9, no. 2, 2014, pp. 441–46, doi:10.4304/jcp.9.2.441-446.
[12]. Kusrini, luthfi taufiq Emha. Algoritma Data Mining. Andi, 2009.
[13]. Informatikalogi. Pembobotan Kata atau Term Weighting TF-IDF. 2016, https://informatikalogi.com/term-weighting-tf-idf/#1.
[14]. Kuliahkomputer. Pengujian Dengan Confusion Matrix. 2018, http://www.kuliahkomputer.com/2018/07/pengujian-dengan-confusion-matrix.html.

DOI : 10.24014/coreit.v7i2.14290