Ekpansi Fitur dengan Word2vec dalam klasifikasi Hoax di Twitter

1st Ridho Maulana Cahyudi
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
ridhomcahyudi@students.telkomuniversity.ac.id

2nd Erwin Budi Setiawan
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
erwinbudisetiawan@telkomuniversity.ac.id

Abstrak-Media sosial sekarang sudah banyak digunakan untuk berbagi informasi, dan juga tempat untuk berkomunikasi. Dalam berbagi informasi banyak peluang untuk menyebarkan hoax, contohnya seperti diaplikasi *Twitter*. Terkadang ada ketidaksesuaian kosa kata dalam setiap *tweet*. Oleh karena itu pada penelitian ini dilakukan penerapan metode fitur ekpansi menggunakan *Word2vec* untuk meminimalisir ketidaksesuaian kosakata tersebut. metode klasifikasi yang digunakan adalah *Naive bayes, ANN, Decision Tree*. Hasil dari penelitian ini, nilai tertinggi sebesar 82,44% yang menggunakan ekspansi fitur *Word2vec* pada metode klasifikasi ANN yang meningkat sebesar 1,17%.

Kata kunci - *hoax, fitur ekspansi, twitter.*

Abstract-Social media is now widely used to share information, and also a place to communicate. In sharing information, ther are many opportunities to spead hoaxes, for example, in the application of twiter. Sometimes there is a vocabulary mismatch in every tweet. Therefore, in this study the feature expansion method was applied using Word2vec to minimize the incompatibility of the vocabulary. Classification method used is naïve bayes, ann, decision tree. The results of this study the highest value of 82.44% using the Word2vec feature expansion in the ann classification method which increased by 1,17%.

Keywords-Hoax, Expansion Feature, Twitter.

I. PENDAHULUAN

Hoax adalah informasi palsu yang sering muncul diinternet dan memiliki tujuan untuk menutupi informasi yang sebenarnya dan juga menyebarkan kepanikan atau ketakutan massal, contohnya seperti dibidang ekonomi, politik, kesehatan, teknologi, hingga keamanan. Saat ini sudah ada ratusan hingga ribuan berita yang sudah dimanipulasi isi beritanya yang sudah dan akan menyebabkan kepanikan [1].

Masalah ini tidak dapat dipisahkan dari dampak penggunaan media sosial yang cepat. Akibatnya, setiap hari ada ribuan informasi yang tersebar dimedia sosial, yang belum tentu valid, sehingga orang – orang berpotensi terkena tipuan dimedia sosial. Informasi yang tidak valid *(hoax)* dapat mempengaruhi emosi, perasaan, pikiran, atau bahkan tindakan seseorang atau kelompok [1]. Sangat disayangkan jika informasi tersebut tidak akurat atau bahkan informasi palsu *(hoax)* dengan provokatif judul yang mengarahkan pembaca dan penerima ke opini negatif.

Media sosial adalah platform yang digunakan oleh semua orang untuk berbagi informasi. Orang beralih ke media sosial digunakan untuk tempat komunikasi. Media sosial telah meledak sebagai kategori wacana online di mana orang bisa membuat dan berbagi informasi [2]. Begitu pula dengan berita hoax yang juga berkembang dikalangan pengguna media sosial, contohnya twitter.

Twitter merupakan media sosial bertipe *microblogging* yang didirikan oleh Jack Dorsey pada Maret 2016 dan diluncurkan pada Juli 2006. Keunikan dari twitter adalah mempunyai tweet atau post yang ada ditwitter dengan ukuran maksimum 140 karakter. Pada Twitter juga dapat ditemui berbagai macam pesan positif hingga negatif, seperti *hoax*, gosip, pornografi, penipuan, pencemaran nama baik, bahkan *self- harming*. Pengguna dapat berinteraksi dengan teman diseluruh penjuru dunia melalui pesan singkat yang ditulis. Tidak sedikit dari pesan tersebut sengaja ditulis dengan tujuan untuk menyebarkan *hoax*. *Hoax* menjadi perbincangan panas ditwitter karena dianggap meresahkan publik dengan adanya informasi yang tidak bisa sepenuhnya dipercaya [2].

Menggunakan metode *Word2vec* karena untuk mengurangi ketidaksesuaian kosakata. Penelitian pertama mengenai *Word2Vec* dilakukan oleh Mikolov menghasilkan sebuah model dari representasi kata dan frasa [3]. Dalam sistem pendeteksian *hoax* digunakan cara pengolahan yang didalamnya juga memiliki beberapa tahapan untuk mengolah setiap kata. Crawling data yang mengambil data dari tweet di twitter,

memisahkannya dan membandingkannya dengan kata-kata yang sudah ada sebelumnya. Crawling data otomatis mengambil dari API twitter yang sudah ada di aplikasi web yang di sediakan oleh twitter.

Oleh karena itu, pada penelitian ini penulis akan menambahkan ekspansi fitur *Word2vec* yang dikombinasikan dengan metode klasifikasi seperti *Naïve Bayes, ANN,* dan *Decision tree.* Sebelum melakukan ekspansi fitur, pre-processing data merupakan langkah penting yang harus dilakukan untuk mengurangi jumlah input kata, sehingga data secara efisien dapat digunakan oleh sistem ekspansi fitur. Untuk pembobotan kata, penulis menggunakan algoritma TF-IDF (Term Frequency – Inverse Document Frequency).

Tujuan dari penelitian ini adalah mengetahui pengaruh penerapan metode klasifikasi *Naïve-bayes, ANN,* dan *Decision Tree* dalam mengklasifikasikan topik pada twitter, dan pengaruh penerapan ekspansi fitur pada metode *Naïve-bayes, ANN,* dan *Decision Tree* dalam mengklasifikasikan topik pada twitter.

II. KAJIAN TEORI

Penelitian yang dilakukan oleh Fitri dkk, berisikan tentang Perancangan algoritme *Word2vec* untuk membuat automated lexicon menggunakan 150 data latih dengan pembagian 75 data sentimen positif dan 75 sentimen negatif. Proses perhitungan *Word2Vec* ini dilakukan pada tiap kata dalam data latih secara berurutan. Kemudian, bobot akhir dari *Word2Vec* digunakan untuk mencari kemiripan kata dan membuat automated lexicon berdasarkan peluang terbesar dari kelasnya. Pengujian terhadap data uji dilakukan dengan menggunakan metode Naive-Bayes Lexicon Based yang mana leksikon merupakan hasil dari automated lexicon yang telah dibuat sebelumnya. Hasil klasifikasi yang didapatkan berupa sentimen positif dan sentimen negatif. Penggunaan metode *Naive-Bayes* untuk melakukan klasifikasi kelas sentimen memiliki nilai *precision* sebesar 0,36, *recall* sebesar 0,818, *f-measure* sebesar 0,5 dan akurasi sebesar 64%. Dari analisis yang telah dilakukan sebelumnya, pembentukan *automated lexicon* memiliki pengaruh terhadap hasildari data uji tersebut. Semakin banyak variasi dan jumlah data latih yang digunakan untuk membuat *automated lexicon,* maka akan semakin akurat juga hasil prediksi kelas data uji [4].

Penelitian yang dilakukan oleh Nurirwan Saputra dkk, berisikan tentang analisis setimen analisis sentimen data *Presiden Joko Widodo* dengan preprocessing normalisasi dan stemming menggunakan metode naive bayes, akurasi yang dihasilkan ketika data dilakukan stemming, terdapat peningkatan rata-rata sebesar 0,85% untuk metode Naive Bayes [5].

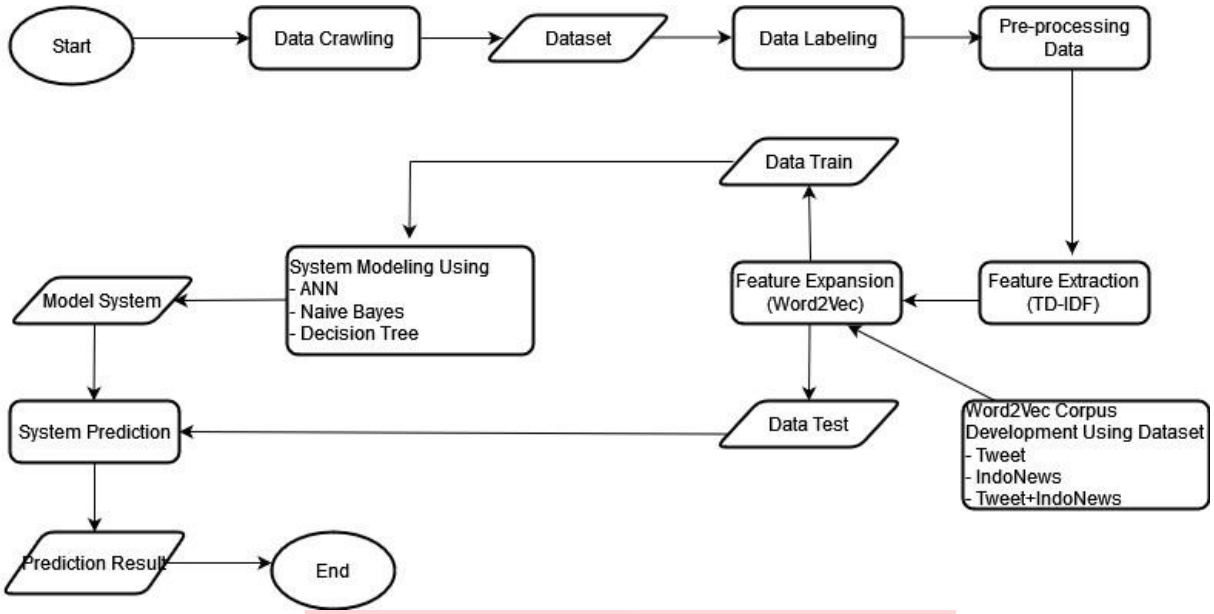
Penelitian yang dilakukan oleh Ari Muzakir dkk, berisikan tentang data mining sebagai prediksi penyakit hipertensi kehamilan dengan Teknik decision tree, Implementasi data mining dengan teknikdecision tree menggunakan algoritma C4.5 dapat menghasilkan informasi berupa prediksi penyakit hipertensi dalam kehamilan, dimana dari data training yang digunakan dengan jumlah 286 instance dapat dibangun sebuah decision tree yang menghasilkan rules yang bisa digunakan untuk memprediksi penyakit hipertensi dalam kehamilan. Dari decision tree yang dibangun, menunjukkan bahwa atribut yang menjadi faktor pendukung seorang ibu hamil bisa menderita penyakit hipertensi dalam kehamilannya, yaitu berdasarkan usia, berat badan, riwayat hipertensi, dan paritas. Setelah mendapatkan decision tree dan rules yang dapat memprediksi penyakit hipertensi dalam kehamilan, dilakukan evaluasi dengan supplied test set menggunakan WEKA dihasilkan kesalahan (error) 7.3427% dan tingkat akurasi 92.6573%. Data training yang berjumlah 286 instances, hal ini menunjukkan bahwa terdapat 265 instances yang akurat dan 21 instances yang error atau prediksinya salah [6].

Penilitian yang dilakukan oleh Diah Wahyuningsih dkk, berisikan tetang prediksi Indonesia dengan model artificial neural network, Hasil prediksi inflasi dengan menggunakan analisis ANN lebih baik dibandingkan dengan analisis regresi linear. Hal ini dapat dilihat dari korelasi keempat variabel terhadap inflasi, dimana pada ANN sebesar 0,83 sedangkan pada regresi linear hanya 0,16 [7].

Setelah melakukan penelitian, belum ada penggunaan fitur ekspansi Word2Vec dengan menggunakan 3 metode klasifikasi. Maka dari itu penulis melakukan pengabungan ekpansi fitur Word2vec dengan menggunakan metode klasifikasi Naïve Bayes, Decision Tree, ANN.

III. METODE

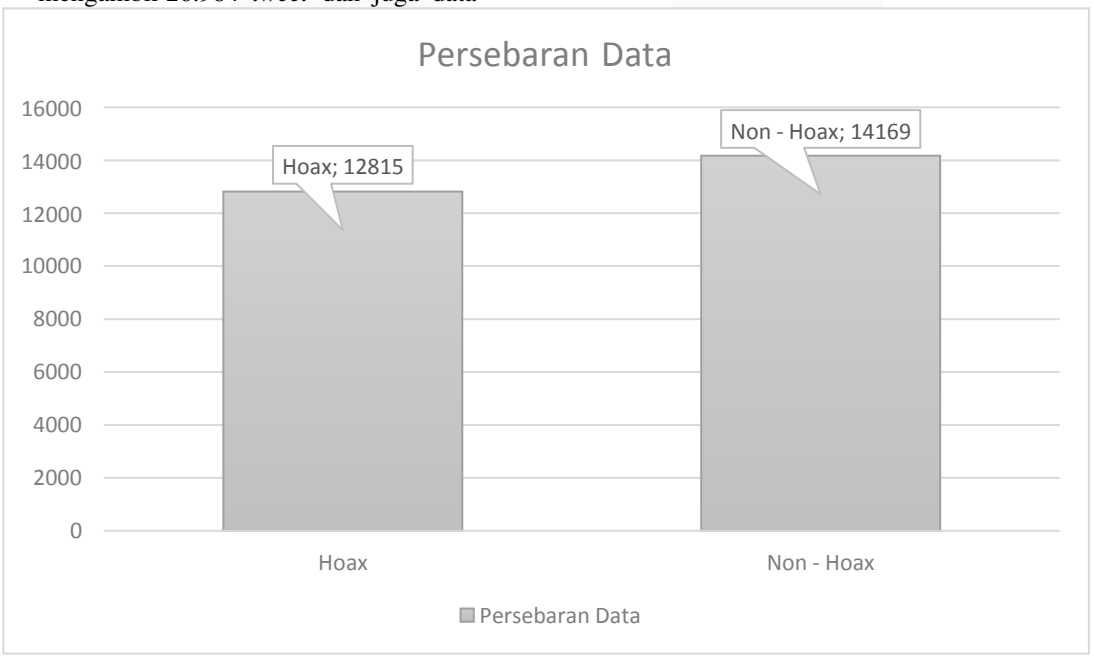
Gambaran umum untuk sistem deteksi *hoax* direpresentasikan dengan rancangan diagram alur berikut ini.



GAMBAR 1
SISTEM DETEKSI *HOAX*

A. Crawling Data

Pengumpulan data ini dilakukan secara otomatis menggunakan aplikasi yang sudah diberikan yaitu *crawling* data yang bersumber dari Twitter menggunakan API (*Aplication Program Interface)* yang sudah ada dari aplikasi Twitter. Untuk pengambilan dari *crawling data* twitter tersebut hanya mengambil 26.984 *tweet* dan juga data yang rentang waktunya dari tanggal 26 Februari 2021 sampai dengan 28 April 2021 untuk mengambil data dari twitter. Data ini terdiri dari beberapa keyword berbeda yang berkaitan dengan hastag, #vaksin, #presiden, #nkri, #mudik, #paspampres, #wakilbupati sangihe, #gajijakarta. Jumlah persebaran datanya untuk hoax 47,50% dan untuk non – hoax 52,50%



GAMBAR 2
PERSEBARAN DATA

Kemudian, terdapat data untuk memenuhi pembuatan kamus kata menggunakan data yang telah diambil dari beberapa media seperti CNN Indonesia, Sindonews, Kompas, Tempo, Detik.com, Liputan6, dan Republika sebanyak 142.533 data. Komposisi data yang digunakan dalam pembuatan kamus *similarity* dengan *word embedding Word2Vec* yang dapat dilihat pada tabel 1.

TABLE 1
SEBARAN DATA UNTUK KORPUS WORD2VEC

Nama Redaksi	Jumlah Data
CNN Indonesia	29349
Republika	53812
Kompas	15055
Tempo	15055
SindoNews	13702
Detik.com	7974
Liputan6	251
TOTAL	**142544**

B. Labeling Data

Sebelum melakukan labeling, pelabelan dilakukan secara manual, dikerjakan oleh 3 orang untuk 1 tweet dengan prinsip *majority vote.* Ciri – ciri umum berita hoax yang digunakan dalam penelitian ini adalah Hoax pencemaran nama, Hoax tentang kisah menyedihkan, Hoax pengalihan isu, Hoax pemicu kepanikan publik. Contoh pelabelan bisa dilihat pada tabel 1.

TABLE 2
CONTOH *HOAX DAN NON - HOAX*

Tweet	Label
Vaksin Covid-19 Sinovac Ilegal karena Tak Bersertifikasi WHO	Hoax
Secara keseluruhan sebanyak 15.500 orang ditargetkan untukmemperoleh suntikan dosis vaksin pada penyelenggaraan vaksinasi pada Rabu, 19 Mei 2021.	Non-Hoax

C. Pre-processing

Jika mengambil data dari Twitter maka akan terdapat noise juga tidak terstruktur dan banyak sekali karakter yang tidak diperlukan. Maka dari itu diperlukan *Pre-processing data.* Tahap pre-processing atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Praproses sangat penting dalam melakukan analisis sentimen, terutama untuk media sosial yang sebagian besar berisi kata-kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki noise yang besar [8]. Ini adalah langkah yang sangat penting sebelum masuk ke proses klasifikasi agar di proses klasifikasi data yang digunakan sudah sangatberkualitas. Berikut 5 proses dalam melakukan *pre-processing* data.

1. *Cleaning*

Cleaning atau proses pembersihan data mengisi nilai yang hilang, menghilangkan data yang bersifat noise, identifikasi atau hapus outliers dan atasi ketidak konsistenan. Meskipun dalam penulisan komentar selalu menyertakan sebuah angka di setiap awal atau akhir kalimat untuk menunjukkan bahwa kalimat tersebut diulang – ulang. Begitu pula dengan *url,* satu huruf, maupun symbol –symbol seperti @atau tanda kurung, tanda kutip dan sebagainya.

TABLE 3
CLEANING

Sebelum	Sesudah
@SaveMoslem1 @didienAZHAR Keliatannya habibakan diputuskan 4 tahun penjara perkara berita bohong	Keliatannya habib akan di putuskan tahunpenjara perkara berita bohong

2. *Case Folding*

Case Folding atau proses yang dilakukan untuk merubah setiap kata menjadi huruf kecil, agarsetiap kata kapital akan menjadi huruf kecil. Berikut contoh dari *Case Folding*:

TABLE 4
CASE FOLDING

Sebelum	Sesudah
Keliatannya habib akan di putuskan tahun penjara perkara berita bohong	keliatannya habib akan di putuskan tahun penjara perkara berita bohong

.

3. *Tokenizing*

Tokenizing atau proses pemecahan kalimat menjadi kata – kata yang dipisahkan oleh spasi.

Berikut Contoh dari Tokenizing:

TABLE 5
TOKENIZING

Sebelum	Sesudah
keliatannya habib akan di putuskan tahun penjaraperkara berita bohong	“keliatannya” “habib” “akan” “di” “putuskan” “tahun” “penjara” “perkara”“berita” “bohong”

4. *Stopword Removal*

Stop Removal atau proses penghapusan kata yang tidak memiliki pengaruh penting atau tidak sama sekali berhubungan. Berikut contoh dari *Stopword Removal*:

TABLE 6
STOPWORLD REMOVAL

Sebelum	Sesudah
“keliatannya” “habib” “akan” “di” “putuskan” “tahun” “penjara” “perkara” “berita” “bohong”	“keliatannya” “habib” “penjara” “penjara” “perkara “berita” “bohong”

5. *Stemming*

Stemming atau proses pengubahan kata – kata yang ada menjadi kata mendasar, dengan cara menghapus imbuhan dari sebuah kata. Berikut contoh dari *Stemming*:

TABLE 7
STEMMING

Sebelum	Sesudah
“keliatannya” “habib” “penjara” “penjara” “perkara “berita” “bohong”	“lihat” “habib” “penjara” “perkara” “berita” “bohong”

D. TF-IDF

Setelah melewati proses *pre-processing* data, data yang sudah berkualitas di tahap ini akan dilakukan pengubahan atau pengekstrakan data berupa dokumen teks menjadi bentuk vector. *TF-IDF* adalah metode untuk menghitung *term, inverse document frequency,* dan *term weighting* pada dokumen dataset. *Term Frequency* (TF) adalah kemunculan suatu *term* terhadap dokumen atau berapa banyaknya jumlah suatu *term* dalam satu dokumen. *Document Frequency* (DF) adalah banyaknya dokumen yang mengandung suatu *term* [9]. Pembobotan global digunakan untuk memberikan tekanan terhadap *term* yang mengakibatkan perbedaan dan berdasarkan pada penyebaran dari *term* tertentu di seluruh dokumen. Banyak skema didasarkan pada pertimbangan bahwa semakin jarang suatu *term* muncul di dalam total koleksi maka *term* tersebut menjadi semakin berbeda. Pemanfaatan pembobotan ini dapat menghilangkan kebutuhan *stop word removal* karena *stop word* mempunyai bobot global yang sangat kecil. Namun pada prakteknya lebih baik menghilangkan *stop word* di dalam fase *pre-processing* sehingga semakin sedikit *term* yang harus ditangani [10]. Data yang telah melalui tahap *pre-processing* akan melalui tahap *TF-IDF* ini. *TF – IDF* adalah proses yang digunakan untuk menentukan seberapa jauh keterhubungan kata terhadap dokumen dengan memberikan bobot setiap kata. Metode *TF-IDF* ini menggabungkan dua konsep yaitu frekuensi kemunculan dari sebuah kata di dalam dokumen dan inverse frekuensi dokumen yang mengandung kata tersebut. Dalam perhitungan bobot menggunakan *TF-IDF*, dihitung terlebih dahulu nilai *TF* perkata dengan bobot masing-masing kata adalah 1.

IDF (word) adalah nilai *IDF* dari setiap kata yang akan dicari, n adalah jumlah keseluruhan dokumen yang ada, *DF* jumlah kemunculan kata pada semua dokumen [11].

E. Ekspansi Fitur dengan Word2Vec

Word2Vec didasarkan pada ide deep learning di mana kata direpresentasikan dalam vektor. *Word2Vec* mentransformasikan operasi dokumen menjadi perhitungan vektor dalam ruang vektor kata. Relasi semantik pada dokumen dapat dikarakterisasi berdasarkan kesamaan kata di dalam ruang vektor. Tahap awal pada proses *Word2Vec* yaitu membangun kosakata dari data teks pelatihan dan kemudian mempelajari representasi vektor dari kumpulan kata. Vektor yang dihasilkan dapat digunakan sebagai fitur untuk penerapan dalam kasus natural language processing dan machine learning [12]. *Word2Vec* adalah salah satu teknik *word embedding* (mengubah kata menjadi vektor yang terdiri dari kumpulan angka). Kata pada sebuah kalimat bisa merepresentasikan makna kata itu sendiri dan konteks kata (kalimat) merepresentasikan makna secara keseluruhan sebagai hasil dari gabungan setiap kata yang menyusun kalimat tersebut. Cara kerja *Word2Vec* yaitu dengan mengambil *corpus* data sebagai input yang sudah melalui tahap praproses dan *one hot encoding* (membuat variabel *binary code* sebanyak jumlah teks yang ada dalam kalimat). Kemudian akan menghasilkan nilai vektor dari setiap kata yang ada ada *corpus* data. *Word2Vec* mempunyai dua jenis model arsitektur untuk merepresentasikan vektor kata yaitu, *Continous bag-of-word* (CBOW) dan *Skip-gram* [13].

Adapun proses pemodelan *Word2vec*, yaitu:

I. Membaca seluruh isi dari data korpus yang sudah dilakukan proses preprocessing. Dimana data yang dibaca yaitu berupa kata-kata pada suatu kalimat yang telah diubah kedalam bentuk array.
II. Pembuatan Model
    a. Membangun konteks pasangan kata dari data korpus dengan berdasarkan jumlah window size. Apabila ditemukan konteks pasangan kata pada window size tersebut maka frekuensi kata ditambah 1.
    b. Setelah itu melakukan training untuk mengubah data menjadi bentuk one-hot-vector. Hal ini dilakukan untuk mengubah bentuk dari setiap kata pada dataset menjadi bentuk binary vector.
    c. Langkah selanjutnya yaitu sistem melatih model untuk memprediksi vektor kata input berdasarkan konteks kata disekitarnya dengan satu hidden layer.
    d. Dari hidden layer

dihasilkan matriks output, kemudian matriks tersebut diubah dengan Softmax fun- ction untuk mendapatkan Word Vector.

III. Word Vector

Setelah proses pembuatan model selesai, maka sistem menghasilkan vektor-vektor dari setiap kata dari data korpus. Di dalam Word2vec, setiap satu kata bisa memiliki lebih dari satu vektor hal ini dikarenakan setiap kata pada sebuah kalimat memiliki konteks yang berbeda[14].

F. Metode Klasifikasi

1. Naïve Bayes

Naive Bayes adalah sebuah algoritma analisa statistik, yang melakukan pengolahan data terhadap data numerik menggunakan probabilitas Bayesian. Klasifikasi–klasifikasi Bayes adalah klasifikasi statistik yang dapat memprediksi kelas suatu anggota probabilitas. Untuk klasifikasi Bayes sederhana yang lebih dikenal sebagai naïve Bayesian Classifier dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas tidak dipengaruhi atau mempengaruhi nilai dari atribut lainnya. Asumsi ini disebut class conditional independence yang diciptakan untuk memudahkan perhitungan, pengertian ini dianggap "naive", dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kata-kata yang lain, sehingga dalam analisis sentimen kata yang muncul memiliki bobot masing-masing yang kemudian dihitung total bobot seluruhnya apakah kalimat tersebut termasuk positif, netral ataupun negatif [5].

2. Artificial Neural Network (ANN)

Merupakan suatu metode *artificial intelligence* yang konsepnya meniru sistem jaringan syaraf yang ada pada tubuh manusia, dimana dibangun *node-node* yang saling berhubungan satu dengan yang lainnya. *Node* tersebut terhubung melalui suatu *link* yang biasa disebut dengan istilah *weight* atau bobot. *Neural network* pertama kali dirancang oleh *Warren McCulloch* dan *Walter Pitts* pada tahun 1943 yang dikenal dengan *McCulloch-Pitts neurons* tentang dua neuron aktif secara bersamaan kemudian kekuatan tersebut terkoneksi antara neuron yang seharusnya bertambah. Kemudian pada tahun 1957, Frank Rosenblatt mengenalkan dan mengembangkan sekumpulan besar jaringan saraf tiruan yang disebut *perceptrons* [15]. ANN lahir dari usaha memodelkan otak manusia karena manusia dianggap sebagai system yang paling sempurna. Berbagai usaha memodelkan otak manusia telah dilakukan dan memunculkan tiga golongan model. Pertama, golongan pertama meniru pola manusia dalam mengambil keputusan. Seperangkat diinputkan dalam otak mesin atau komputer, sehingga komputer dapat mengambil keputusan sesuai dengan pengetahuan yang sesuai dengan input *("pengetahuan")* yang diberikan. Golongan ini disebut sebagai sistem pakar *(expert system)*. Kedua, golongan berikutnya menirukan cara kerja manusia yang tidak pernah dilakukan dalam variabel tegas *(crisp)*. Semua wariabel yang yang diolah dalam otak manusia bersifat samar *(fuzzy)*. Dengan menggabungkan variabel samar dengan sistem pakar maka lahirlah *fuzzy logic*. Ketiga, golongan berikutnya lahir dari usaha memodelkan sel syaraf. Oleh karena itu disebut sebagai ANN *(artificial neural network)* [7].

3. Decision Tree

Decision tree merupakan proses menemukan kumpulan pola atau fungsi-fungsi yang mendeskripsikan dan memisahkan kelas data satu dengan yang lainnya. Metode *decision tree* mengubah fakta yang sangat besar menjadi *Decision Tree* yang mempresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk basis data seperti *Structure Query Language* (SQL) untuk mencari *record* pada data tertentu. Sebuah *decision tree* adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Pada *decision tree* setiap simpul daun menandai label kelas. Simpul yang bukan simpul akhir terdiri dari akar dan simpul internal yang terdiri dari kondisi tes atribut pada sebagian *record* yang mempunyai karakteristik yang berbeda. Simpul akar dan simpul internal ditandai dengan bentuk oval dan

simpul daun ditandai dengan bentuk segi empat [6].

G. Ukuran Performansi Sistem

1. *Confusion Matrix*

Confusion Matrix merupakan konsep *machine learning* yang berisi fakta mengenai prediksi & klasifikasi aktual menurut suatu sistem klasifikasi [16]. *Confusion Matrix* memiliki dua dimensi, satu dimensi ditandai oleh *actual class* dari suatu objek, yang lain ditandai oleh kelas yang diprediksi oleh pengklasifikasi [17]. Kinerja sistem biasanya dievaluasi menggunakan data dalam matriks. Tabel 8 menunjukkan *Confusion Matrix* untuk dua pengklasifikasi[18].

TABLE 8
CONFUSION MATRIX

		Prediksi	
		Hoax	Non-Hoax
Aktual	Hoax	a	b
	Non-hoax	c	d

Arti dari entri *Confusion Matrix* adalah sebagai berikut[18]:

1. a adalah jumlah prediksi yang benar bahwa suatu pernyataan tersebut bernilai hoax,
2. b adalah jumlah prediksi yang salah bahwa sebuah pernyataan tersebut bernilai non-hoax,
3. c adalah banyaknya prediksi yang salah pada sebuah pernyataan tersebut bernilai hoax, dan
4. d adalah jumlah prediksi yang benar bahwa suatu pernyataan tersebut bernilai non-hoax. Istilah standar telah ditetapkan untuk matriks [18]:

a. Akurasi

Akurasi adalah jumlah persentase dari total prediksi yang benar, di tentukan oleh persamaan

$$\text{Akurasi} = \frac{\square+}{\square+\square+\square+\square} \quad (1)$$

b. Recall

Recall atau true positive rate (TP) adalah persentase kasus positif yang teridentifikasi dengan benar dan dihitung menggunakan rumus berikut:

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

c. *Precision*

Presisi (P) adalah proporsi kasus positif yang diprediksi yang benar, yang dihitung menggunakan persamaan:

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

d. F1-Measure

Pertimbangan rata – rata presisi dan recall yang dibobotkan

$$\text{F1} - \text{Measure} = 2\,x\,\frac{(precision\ x\ recall)}{(precision+recall)} \quad (4)$$

[19].

IV. HASIL DAN PEMBAHASAN

Evaluasi sistem ini memiliki 3 skenario yang bertujuan untuk mengetahui hasil dari tujuan awal yaitu penerapan metode klasifikasi *Naïve-bayes, ANN,* dan *Decision Tree* dalam mengklasifikasikan topik pada twitter, ingin mengetahui pengaruh penerapan pembobotan *TF-IDF* pada metode *Naïve-bayes, ANN,* dan *Decision Tree* dalam mengklasifikasikan topik pada twitter, mengetahui pengaruh penerapan ekspansi fitur pada metode *Naïve-bayes, ANN,* dan *Decision Tree* dalam mengklasifikasikan topik pada twitter.

A. Hasil Pengujian Skenario 1

Pengujian Skenario 1 ini yang bertujuan penerapan metode klasifikasi *Naïve-bayes, ANN,* dan *Decision Tree* dalam mengklasifikasikan topik pada twitter, melakukan tanpa menggunakan pembobotan *TF-IDF*. Dalam pengujian ini menentukan rasio yang akan digunakan yaitu 70:30, 80:20, 90:10 untuk perbandingan data latih dan data uji yang selanjutnya akan digunakan sebagai *baseline*. Pengujian ini dilakukan sebanyak 5 kali lalu diambil rata – rata dari akurasi dan *F1-Measure*, dengan hasil seperti tertera pada table 9.

TABLE 9
HASIL PENGUJIAN SKENARIO 1 MENENTUKAN RASIO DAN *BASELINE*

Metode	Akurasi(%)			F1-Measure		
	Rasio 70:30	**Rasio 80:20**	**Rasio 90:10**	**Rasio 70:30**	**Rasio 80:20**	**Rasio 90:10**
Naïve Bayes	78,89	**79,60**	78,66	77,60	**77,89**	77,13
ANN	79,26	**81,27**	79,62	79,54	**80,31**	78,58
Decision Tree	79,66	**81,00**	79,19	79,48	**79,83**	78,34

Dari pengujian pada table 9, dapat disimpulkan bahwa rasio 80:20 memiliki performansi tertinggi pada perbandingan data latih dan data ujinya. Jadi, pada pengujian selanjutnya akan menggunakan rasio dengan perbandingan 80:20 untuk data latih dan data ujinya, dan juga data tersebut akan digunakan sebagai *baseline*.

B. Hasil Pengujian Skenario 2

Pengujian Skenario 2 ini melakukan penerapan ekspansi fitur yang sudah dilakukan pembobotan *TF-IDF* pada *baseline Naïve Bayes, ANN,* dan *Decision Tree.* Ekspansi fitur menggunakan 3 jenis *Corpus Word2vec,* yaitu *corpus* data tweet, *corpus* data berita, *corpus* data tweet dan data berita. pengujian ini menggunakan perbandingan rasio 80:20 untuk data latih dan data ujinya. Pada tahap ini pengujian dilakukan sebanyak 5 kali lalu diambil hasil nilai rata – rata dari akurasi.

Hasil nilai dari akurasi terhadap pengujian ekpansi fitur menggunakan algoritma klasifikasi *Naïve Bayes, ANN* dan *Decision Tree.* Kolom dari *baseline* menunjukan hasil tanpa menggunakan pembobotan *TF-IDF* dan ekpansi fitur. Sedangkan kolom *corpus* tweet, *corpus* berita, *corpus* tweet dan berita menunjukkan hasil dengan pembobotan *TF-IDF* yang kemudian dilakukan penerapan metode menggunakan ekspansi fitur sesuai dengan *corpus*nya masing – masing.

A. Klasifikasi Naïve Bayes

Pada bagian ini adalah pengujian dengan penerapan ekspansi fitur pada klasifikasi *Naïve Bayes* yang hasilnya dapat dilihat pada tabel 10.

TABLE 10
HASIL AKURASI FITUR EKSPANSI PADA METODE *NAÏVE BAYES*

Top Similarity	Akurasi (%)			
	Baseline	Corpus Tweet	Corpus Berita	Corpus Tweet + berita
1	79,6	80,59(+0,99)	79,94(+0,34)	79,29(-0,31)
5	79,6	80,11(+0,51)	79,55(-0,05)	79,9(+0,3)
10	79,6	80,89(+1,29)	79,85(+0,25)	79,66(+0,06)

Dari table diatas, terlihat bahwa peningkatan nilai akurasi terjadi pada setiap fitur, kecuali pada bagian Top Similarity 5 terjadi penurunan sebesar 0,05% pada *corpus* berita. Nilai akurasi tertinggi dengan klasifikasi *Naïve Bayes* terdapat pada *top similarity* 10 yang menggunakan *corpus* tweet yaitu sebesar 80,89% atau mengalami kenaikan nilai sebesar 1,29% sedangkan nilai akurasi terendah sebesar 79,55% yang terdapat pada *top similarity* 5 yang menggunakan *corpus* berita atau mengalami penurunan nilai sebesar 0,05%.

B. Klasifikasi ANN (Artificial Neural Network)

Performansi untuk hasil pengujian ekspansi fitur menggunakan algoritma klasifikasi *ANN (Artificial*

Neural Network) dapat dilihat pada tabel

TABLE 11
HASIL AKURASI FITUR EKSPANSI PADA METODE *ANN(ARTIFICIAL NEURAL NETWORK)*

Top Similarity	Akurasi (%)			
	Baseline	Corpus Tweet	Corpus Berita	Corpus Tweet + berita
1	81,27	80,74(-0,53)	81(-0,27)	81,81(+0,54)
5	81,27	81,63(+0,36)	81,52(+0,25)	81,11(-0,16)
10	81,27	80,96(-0,31)	80,81(-0,46)	82,44(+1,17)

Penurunan nilai akurasi pada bagian *top similarity* 1 dan 10 dibagian *corpus* tweet sebesar 0,53% dan 0,31%, sedangkan pada bagian *top similarity* 5 mengalami peningkatan sebesar 0,36%. Demikian juga dibagian *corpus* berita yaitu mengalami terdapat peningkatan nilai akurasi pada bagian *top similarity* 1 dan 10 yaitu sebesar 0,27% dan 0,46%, dan pada bagian *top similarity* 5 mengalami peningkatan sebesar 0,25%. Sedangkan pada bagian *corpus* tweet + berita bagian mengalami peningkatan akurasi adalah pada bagian *top similarity* 1 dan 10 yaitu sebesar 0,54% dan 1,17%, selain itu pada bagian *top similarity* 5 terjadi penurunan sebesar 0,16%.

Untuk nilai akurasi tertinggi pada klasifikasi *ANN (Artificial Neural Network)* adalah 82,44% pada *top similarity* 10 yang menggunakan *corpus* Tweet+Berita dan nilai akurasi terendahnya adalah 80,74% pada *top similarity* 1 yang menggunakan *corpus* tweet.

C. Klasifikasi Decision Tree

Performansi untuk hasil pengujian ekspansi fitur menggunakan algoritma klasifikasi *Decision Tree* dapat dilihat pada tabel 12.

TABLE 12
HASIL AKURASI FITUR EKSPANSI PADA METODE *DECISION TREE*

Top Similarity	Akurasi (%)			
	Baseline	Corpus Tweet	Corpus Berita	Corpus Tweet + berita
1	81	81,07(+0,07)	80,63(-0,37)	80,24(-0,76)
5	81	80,81(-0,19)	80,18(-0,82)	81,02(+0,02)
10	81	81,63(+0,63)	80,4(-0,6)	80,68(-0,32)

Berdasarkan table diatas terlihat ada peningkatan akurasi pada *corpus* tweet khususnya pada bagian top similarity 1 dan 10 yaitu sebesar 0,07% dan 0,63%, sedangkan untuk *top similarity 5* sendiri mengalami penurunan sebanyak 0,19%. Untuk *corpus* berita terjadi penurunan sebesar 0,37%, 0,82% dan 0,6% pada setiap bagian top similarity. Untuk *corpus* tweet + berita terdapat penurunan pada bagian *top similarity* 1 dan 10 yaitu sebesar 0,76% dan 0,32% lalu *top similarity* 5 mengalami peningkatan sebanyak 0,02%.

Pada klasifikasi *Decision Tree* ini nilai akurasi tertinggi sebesar 81,63% yaitu pada *top similarity* 10 yang menggunakan *corpus* tweet dan nilai akurasi terendah sebesar 80,18% pada *top similarity* 5 yang menggunakan *corpus* berita.

V. KESIMPULAN

Pada penelitian ini, telah dilakukan pembuatan deteksi hoax dengan menggunakan ekpansi fitur metode *Word2Vec* dengan klasifikasi *Naïve bayes, ANN, Decision Tree*. Ekpansi fitur metode *Word2vec* yang digunakan pada system pendeteksi hoax bertujuan untuk mengurangi ketidaksesuaian kosakata pada kalimat tweet tersebut. Ekpansi fitur dilakukan terhadap 3 jenis *corpus Word2vec* (tweet, berita, dan tweet + berita) dan juga variasi ekspansi fitur (Top Similarity 1, Top Similarity 5, Top Similarity 10) untuk mencari model terbaik. Penggunaan ekpansi fitur 3 jenis corpus *Word2vec* (tweet, berita, dan tweet + berita) berpengaruh pada nilai akurasi pada setiap klasifikasi *Naïve Bayes, ANN, Decision Tree, yaitu:*

Klasifikasi *Naïve Bayes* mengalami kenaikan pada 8 data dari 9 data yang diperoleh. Nilai akurasi tertinggi terjadi pada *Top Similarity 10* dengan *corpus* tweet sebesar 80,89% atau mengalami peningkatan sebesar 1,29% dan nilai akurasi terendah terjadi pada *Top Similarit*y 5 sebesar 79,55% atau mengalami penurunan sebesar 0,05%. Klasifikasi *ANN* terdapat kenaikan nilai hanya pada 4 data dari 9 data yang diperoleh. Nilai akurasi tertinggi pada *Top Similarity 10*

dengan *corpus* tweet+berita sebesar 82,44% yang mengalami peningkatan sebesar 1,17% dan nilai akurasi terendah pada *top similarity* 1 dengan *corpus* tweet yaitu sebesar 80,74% atau penurunan sebesar 0,53%. Klasifikasi *Decision Tree* mengalami kenaikan nilai akurasi hanya pada 3 data dari 9 data yang diperoleh. Nilai akurasi tertinggi pada *top similarity* 10 dengan *corpus* tweet sebesar 81,63% yang mengalami peningkatan sebesar 0,63%, dan nilai akurasi terendah pada *top similarity 5* dengan *corpus* bertia sebesar 80,18% mengalami penurusan 0,82%.

Pengaruh dari ekpansi fitur pada model *ANN* dengan nilai tertinggi dari setiap klasifikasinya yaitu sebesar 82,44% yang dimana mengalami peningkatan sebesar 1,17% dari nilai akurasi baseline.

REFERENSI

[1] S. Sucipto, A. G. Tammam, and R. Indriati, "Hoax Detection at Social Media With Text Mining Clarification System-Based," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 3, no. 2, pp. 94– 100, 2018, doi: 10.29100/jipi.v3i2.837.

[2] R. Sistem, "JURNAL RESTI Hoax Detection on Twitter using Feed-forward and Back-propagation,"vol. 1, no. 10, pp. 655–663, 2021.

[3] A. N. Laili, P. P. Adikara, and S. Adinugroho, "Rekomendasi Film Berdasarkan Sinopsis Menggunakan Metode Word2Vec," vol. 3, no. 6, pp. 6035–6043, 2019.

[4] A. Fitri Niasita, P. P. Adikara, and S. Adinugroho, "Analisis Sentimen Pembangunan Infrastruktur di Indonesia dengan Automated Lexicon Word2Vec dan Naive-Bayes," *J-Ptiik*, vol. 3, no. 3, pp. 2673– 2679, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[5] N. Saputra, T. B. Adji, and A. E. Permanasari, "Analisis Sentimen Data Presiden Jokowi dengan Preprocessing Normalisasi dan Stemming Menggunakan Metode Naive Bayes dan SVM," *J. Din. Inform.*, vol. 5, no. November, p. 12, 2015, [Online]. Available: http://ojs.upy.ac.id/ojs/index.php/dinf/article/view/113

[6] A. Muzakir and R. A. Wulandari, "Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree," *Sci. J. Informatics*, vol. 3, no. 1, pp. 19–26, 2016, doi: 10.15294/sji.v3i1.4610.

[7] D. Wahyuningsih, I. Zuhroh, and - Zainuri, "Prediksi Inflasi Indonesia Dengan Model Artificial Neural Network," *J. Indones. Appl. Econ.*, vol. 2, no. 2, pp. 2–2008, 2008, doi: 10.21776/ub.jiae.2008.002.02.7.

[8] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[9] F. W. Nurwariz and Y. Sibaroni, "Analisis Sentimen Review Game pada Steam Menggunakan Metode Support Vector Machine dengan Information Gain," 2019.

[10] H. W. A. Kesuma, "Penerapan Metode TF-IDF dan Cosine Similarity dalam Aplikasi Kitab Undang- Undang Hukum Dagang," 2016.

[11] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp.306–312, 2018.

[12] F. W. KURNIAWAN, "Analisis Sentimen Twitter Bahasa Indonesia dengan Word2Vec," 2020, [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/159923/slug/analisis-sentimen-twitter-bahasa-indonesia-dengan-word2vec.html%0A/home/catalog/id/159923/slug/analisis- sentimen-twitter-bahasa-indonesia-dengan-word2vec.html

[13] J. Nurjaman, R. Ilyas, F. Kasyidi, J. Informatika, U. Jenderal, and A. Yani, "Pengukuran Kesamaan Semantik Pasangan Kalimat Sitasi Menggunakan Convolutional Neural Network," pp. 26–27, 2020.

[14] I. L. S. Nabila Nanda Widyastuti, Arif Arif Bijaksana, "Analisis Word2vec untuk Perhitungan Kesamaan Semantik antar Kata | Widyastuti | eProceedings of Engineering," vol. 5, no. 3, pp. 7603–7612, 2018, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/7263

[15] S. R. DEWI, "Deep Learning Object Detection Pada Video," *Deep Learn. Object Detect. Pada Video Menggunakan Tensorflow Dan Convolutional Neural Netw.*, pp. 1–60, 2018, [Online]. Available: https://dspace.uii.ac.id/bitstream/handle/123456789/7762/14611242_Syarifah Rosita Dewi_Statistika.pdf?sequence=1

[16] C. Sammut and G. I. Webb, "Encyclopedia of Machine Learning".

[17] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf. Sci. (Ny).*, vol. 340–341, pp. 250–261, 2016, doi:

10.1016/j.ins.2016.01.033.

[18] a. K. Santra and C. J. Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering," *Int. J. Comput. Sci.*, vol. 9, no. 1, pp. 322–328, 2012, [Online]. Available: http://ijcsi.org/papers/IJCSI-9-1-2- 322-328.pdf

[19] C. Kim, V. Zhu, J. Obeid, and L. Lenert, "Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke," *PLoS One*, vol. 14, no. 2, pp. 1–13, 2019, doi: 10.1371/journal.pone.0212778.