BAB II

LANDASAN TEORI

2.1 Tinjauan Teori

2.1.1 Text Mining

Text mining adalah sebuah proses dalam melakukan sebuah penggalian informasi baru yang belum diketahui berbentuk sebuah, data yang digali biasanya bersifat *unstructured* [8]. Penggunaan dari text mining dapat menghasilkan pola yang menarik saat mengambil informasi ini jika data memiliki nilai.

Implementasi dari *text mining* sangatlah beragam seperti *risk management, fraud detection,business intelligence,*dan juga *media social analyst* [9]. *Text Mining* memiliki beberapa proses tahapan dalam melakukan penerapan nya seperti menurut [10] yaitu :

1) Mendefinisikan Masalah

   Sebelum melakukan text mining,diperlukan pemahaman sebuah masalah untuk mengerti apa yang akan dilakukan pada *text mining*

2) Pengumpulan Data

   Pengumpulan data harus sesusai dengan kebutuhan dari permasalahan yang ada sehingga dapat membantu menyelesaikan masalah tersebut

3) Mendefinisikan fitur

   Melakukan proses pemilihan model apa yang akan cocok untuk menyelesaikan masalah dengan data yang ada .

4) Analisis Data

   Melakukan analisa data dengan menggunakan beberpa model,hasi dari model tersebut akan memunculkan sebuah pola yang terstruktur sehingga masalah dapat di selesaikan.

5) Mendapatkan hasil

   Melakukan penerjemahan dari analisis data menjadi sebuah informasi yang berguna dengan menggunakan visualisasi data

Pada penerapan Text mining khususnya pada analisis sentimen, proses *text preprocessing* merupakan proses yang sangat penting untuk dilakukan karena pada proses ini bertujuan untuk mencegah penurunan performa analisis yang signifikan .Pada uumnya Tahapan yang sering dilakukukan dalam proses *text preprocessing* terdiri dari *casefolding, tokenization, filtering,remove duplicate,stop removal,stemming* [11] .Berikut merupakan penjelasan dari tahapan *text preprocessing* :

1) Remove Duplicate

   Sebelum data di proses, untuk menghindari data yang berulang maka perlu dilakukan *remove duplicate,* data yang memiliki nilai yang sama akan di hapus agar tidak terjadi redudansi.

2) Filtering

   Pada proses *filtering* kata-kata yang dirasa tidak terlalu penting dalam pemprosesan data akan di hapus contoh nya seperti username,url dan query ,kata-kata tersebut akan di hapus agar pemprosesan mendapatkan hasil yang maksimal.

3) Casefolding

   Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Maka dari itu, peran *case folding* diperlukan untuk mengubah keseluruhan Mengonversi teks dalam dokumen ke format standar atau huruf kecil. *Case folding* digunakan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" hingga "z" yang diterima. Fitur Semua huruf kecuali huruf dihilangkan dan diperlakukan sebagai pemisah, misalnya koma, titik koma atau titik dua. Proses case folding tidak memerlukan library tertentu, karena dapat menggunakan fungsi lower() sebagai bawahan Bahasa pemrograman python .

4) Tokenizing

Tokenizing adalah sebuah proses pemecahan sebuah kata yang disebut token. Berikut merupakan contoh penerapan dari tokenizing

Tabel 2.1 *Tokening*

Teks Input	Text Output
ya tuhan semoga sehat selalu kalian semua mari bangun masa depan bangsa kita	'ya', 'tuhan', 'semoga', 'sehat', 'selalu', 'kalian', 'semua', 'mari', 'bangun', 'masa', 'depan', 'bangsa', 'kita'

5) Stopword

Stopword berfungsi untuk meminimalisir jumlah kata dalam suatu dokumen yang dapat mempengaruhi kinerja pemrosesan bahasa alami. Kata-kata yang dibutuhkan pada tahapan *stopword* adalah kata-kata yang penting dan menghapus kata-kata yang tidak penting dari tokenisasi. *Stopwords* adalah kata-kata yang disimpan dalam daftar *stopword* yang diabaikan selama pemrosesan. Ciri utama dari pemilihan stop word adalah seringnya kemunculan kata, seperti konjungsi “dan”, “atau”, “tetapi”, “akan”. Tidak ada aturan yang jelas untuk penentuan stop word yang digunakan, penentuan *stop word* dapat disesuaikan dengan penelitian yang dilakukan.

6) Stemming

Pada tahap stemming berperan untuk mengganti kata awal menjadi kata dasar atau akar dengan menghapus imbuhan, termasuk awalan, kata perantara (*infiks*), kata akhir (*suffix*), dan menghilangkan awalan dan akhiran (*konfiks*) pada kata turunan. *Library Python* digunakan dalam proses stemming bahasa Indonesia.

2.2 Natural Language Processing

Natural Language Processing (NLP) adalah salah satu bidang dari ilmu *artificial intelligence* yang dikembangkan pada interkasi pada manusia dan komputer [12].Fungsi dari NLP yaitu untuk melakukukan perancangan dan membangun

aplikasi untuk memfasilitasi interaksi manusia dengan komputer dan *device* lain dengan penggunaan bahasa alami. NLP berfokus pada pemrosesan bahasa alami, yaitu.bahasa komunikasi manusia yang umum digunakan. Bahasa yang diterima oleh komputer perlu diproses dan dipahami terlebih dahulu sehingga komputer dapat memahami maksud manusia atau pengguna dengan baik.

Masalah yang dapat diselesaikan dengan menggunakan teknologi NLP adalah:

1) Question Answering Systems (QAS)

Merupakan sistem penjawab permasalahan yang mengenakan pengetahuan yang luas untuk membagikan jawaban. Program sanggup berjalan secara otomatis tanpa tertinggal operator.

2) Text Summarization

Suatu sistem peringkasan bacaan yang bisa digunakan untuk meringkas makalah panjang dan ringkasan bacaan dari buku, sembari melindungi inti dari dokumen asli.

3) Machine Translation

Penerjemah mesin ini bisa digunakan buat secara otomatis mengganti bacaan dari satu bahasa ke bahasa lain. Perihal ini membolehkan satu sistem buat bisa menerjemahkan ke dalam sebagian bahasa.

4) Analisis Sentimen

Suatu sistem untuk mengenali serta mengekstrak data subjektif dari sumber. Data bisa diperoleh dari pembahasan produk ataupun pesan media sosial, serta tugasnya merupakan buat mengenali apakah sentimen tersebut berpolaritas positif, netral, ataupun negatif. Ini menunjang pebisnis menguasai sentimen sosial dari sesuatu merk, produk, ataupun layanan sembari memantau obrolan online di media sosial.

5) Speech Recognition

Suatu sistem yang membolehkan komputer untuk menerima masukan dalam wujud lisan. Teknologi ini membolehkan fitur buat mengidentifikasi serta menguasai perkata yang diucapkan dengan mendigitalkan perkata serta mencocokkan sinyal digital dengan pola tertentu yang tersimpan di fitur.

2.1.3 Sentimen Analisis

Sentimen analisis adalah proses memahami dan mengkategorikan emosi (baik positif maupun negatif) yang terkandung dalam tulisan menggunakan teknik analisis teks. Fungsi dari analisis sentimen adalah untuk menentukan nilai suatu opini terhadap suatu topik tertentu. Oleh karena itu, hasil analisis sentimen dapat berupa evaluasi pengenalan emosi.

Analisis sentimen diperlukan atas dasar menemukan sentimen seseorang terhadap suatu topik atau polaritas kontekstual dari keseluruhan dokumen [13]. Dengan analisis sentimen, kita dapat mengetahui bagaimana orang bereaksi terhadap pendapat mereka tentang sebuah topik tertentu. Ini akan memudahkan manusia untuk memahami sentimen akhir dari keseluruhan konteks topik

2.1.4 Python

Python adalah sebuah Bahasa pemograman yang bertujuan untuk memudahkan dalam pengembangan aplikasi secara cepat. *Python* merupakan Bahasa pemograman yang mudah dipahami karena memiliki bentuk yang jelas untu digunakan [14].

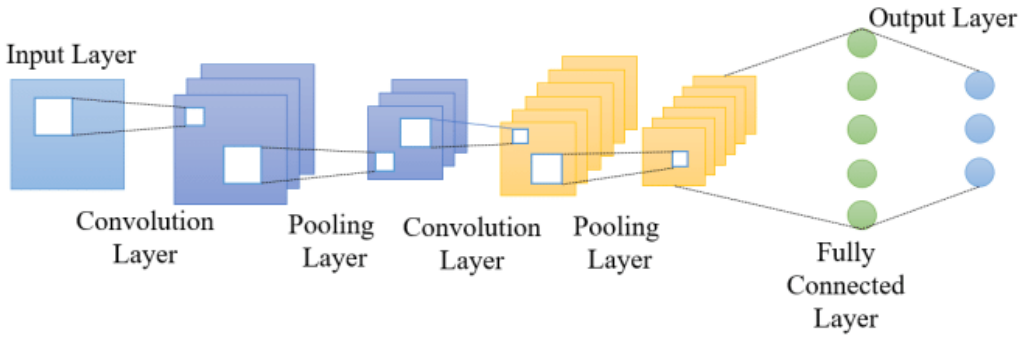
Python juga bisa digunakan dalam melaksanakan analisis statistik, pembangunan model *machine learning* maupun tentang yang berhubungan dengan data *science* dan lainnya,terdapat berbagai macam *library open-source* yang dapat dipakai dalam pemrograman Pyhton. Sebagian package yang popular di *Python* antara lain:

1) Spicy dan Scikit, *library* untuk membuat model *Artificial Intellegence* dan *machine learning*.

2) Matplotlib, *library* dapat digunakan dalam membuat visualisasi data seperti grafik. Setiap sumbu memiliki sumbu horizontal (x) dan sumbu vertikal (y)
3) OpenCV Python, *library* untuk membuat aplikasi *Computer vision*.
4) TensorFlow, *library* untuk membuat model dalam implementasi deep learning, dan beberapa lainnya

2.1.5 Convolutional Neural Network

Convolutional Neural Network atau yang sering sebut dengan CNN adalah sebuah model algoritma *neural network multilayer* jenis *feed forward network* yang terdiri dari 2 *deep layer* dan digabungkan dengan *fully connected layer*. Tujuan asli dari desain *Convolutional Neural Network* adalah pengenalan gambar. Namun, telah digunakan sebagai berbagai model tujuan seperti mengenali informasi prediktif dari sebuah objek seperti text,suara, *Convolutional Neural Network* dapat membedakan antara fitur lokal dan rendering bidang multidimensi. pengurangan ukuran keluaran Kemudian lolos ke lapisan koneksi. *Convolutional Neural Network* Arsitektur mendefinisikan tiga lapisan - lapisan input, lapisan ekstraksi fitur, dan lapisan klasifikasi, seperti yang ditunjukkan pada gambar di gambar. 2.1. Arsitektur model dapat dimodifikasi menjadi: Perlu untuk mencapai akurasi yang lebih baik [15]



Gambar 2.1 *Architectture CNN*

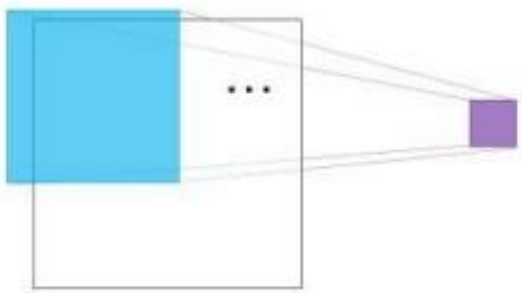
Convolutional Neural Network sendiri adalah salah satu metode penerapan dari *deep learning* berbeda dengan model *machine learning* ,dalam segi waktu proses penerapan dari *deep learning* memekan waktu yang lebih lama dibandingkan dengan penggunaan *machine learning* dikarenakan proses pembobotan yang dilakukan besar dan adanya beberapa penambahan dari parameter.maka penerapan dari *machine learning* lebih baik untuk ukuran data yang lebih kecil sedangkan untuk *deep learning* lebih baik untuk penerapan data yang lebih besar

2.1.5.1 Layer pada Convolutional Neural Network

Pada penggunaan *Convolutional Neural Network* terdapap beberapa lapisan yang yang ada pada model *Convolutional Neural Network*, secara garis besar lapisan tersebut adalah *feature extraction layer* dan *fully-connected layer* .pada bagian *feature extraction layer* terdiri dari 2 bagian yaitu *pooling layer* dan *convolutional layer* berikut merupakan penjelasan lebih detail mengenai lapiasan dari *Convolutional Neural Network.*

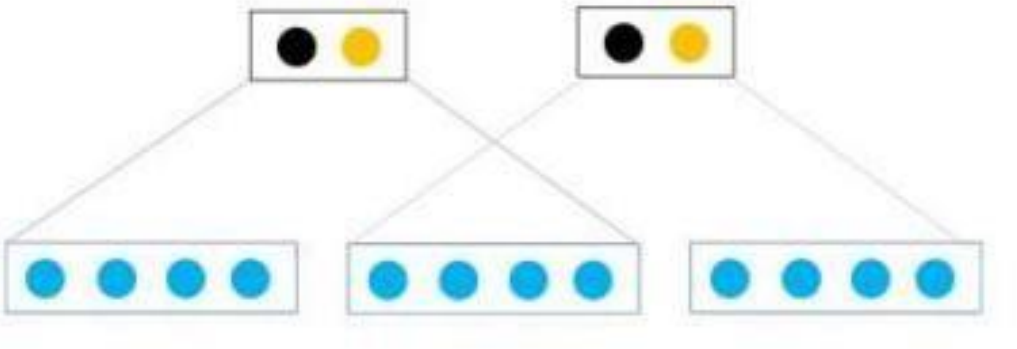
1) Convolutional Layer

Konsep dari *convolutional layer* merupakan memanfaatkan prinsip dari *sliding window* serta *weigh sharing* yang mempunyai tujuan untuk mengurangi dari suatu kompleksitas perhitungan. Pelaksanaan pada *window* dicoba untuk memandang aspek yang sangat informatif ataupun yang diketahui dengan *filter* yang dapat dikenali oleh *window* semacam yan tertera pada gambar 2. 2



Gambar 2.2 *Sliding Window*

Warna biru mewakili dari *window* dan *filter* diwakili pada warna ungu. *Window* akan bergeser sebanyak N yang dan menghasilkan *vector* dengan ukuran tertentu ,setelahnya *window* akan ditrasnformasikan menjadi nilai numerik

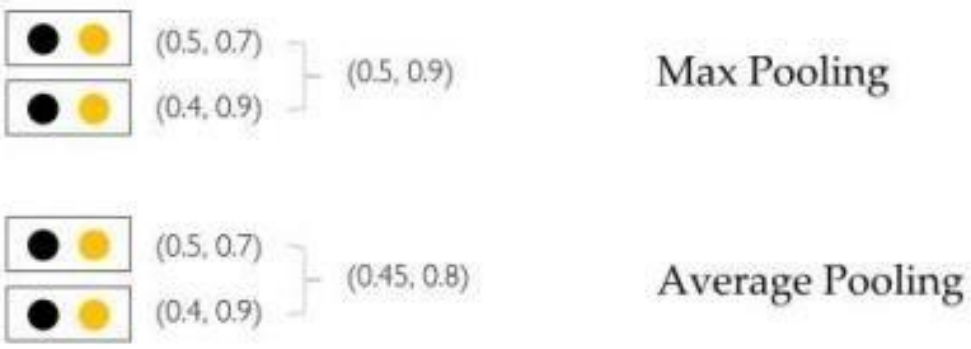


Gambar 2.3 *1D Convolution*

Lalu pada gambar 2.3 warna biru mewaliki dari *feature vector* dari suatu input, setiap 2 input yang akan ditransformasi menjadi (2) dua dimensi yang dapat menghasilkan *vector* 4 dimensi, jika suatu input x dapat mengunakan *stride* sebesar s untuk menentukan seberapa banyak data yang dapat digeser untuk *window* baru

2) *Pooling layer*

Proses dari *pooling layer* berada pada setelah dari proses *convolutional layer*, pada tahp ini, vektor-vektor yang dihasilkan akan dikombinasikan dengan *pooling layer* menjadi sebuah *vector* yag baru. Proses *pooling* yang umum dilakukan ada 2 yaitu max pooling dan average pooling seperti pada gambar 2.3. Proses *Max pooling* yaitu melakukan pengambilan nilai terbesar dari lingkaran hitam dan lingkaran kuning . sedangkan proses average pooling yaitu mengambil nilai rata rata dari lingkaran hitam dan kuning dan menghasilkan suatu nilai yang baru



Gambar 2.3 *Max Pooling dan Average Pooling*

Tujuan dari penggunaan *pooling layer* adalah untuk meminimalisir demensi dari *feature map*, sehingga mempercepat komputasi karena parameter yang harus di *update* semakin mengecil

3) Fully Connected Layer

Pada proses *fully connected layer* akan melanjutkan proses dari sebelum nya yaitu *convolutional layer* dan *pooling layer*, pada proses sebelumnya output yang dihasilkan masih berupa *multidimentional array*, maka diperlukan tahapan *flatten* atau *reshape feature map* menjadi sebuah *vector* untuk bisa diunakan sebaai *input* dari *fully connected layer Fully connected layer* 35 sama dengan *multi-layer perceptron* yang memiliki *hidden layer, activation function, output layer,* dan *loss function*.

2.1.6 K-Nearest Neighbor

K- Nearest Neighbor(KNN) ialah salah satu tata cara machine learning yang mengklasifikasikan objek menurut informasi pembelajaran yang sangat dekat dengan objek tersebut. Tata cara ini sangat simpel, mudah direpresentasikan, mempunyai ketangguhan untuk melatih informasi yang mempunyai banyak noise, serta efisien untuk proses pengelompokan. Tujuan dari algoritma ini mengklasifikasikan objek baru, atribut serta pelatihan ilustrasi Nilai k terbaik untuk algoritma ini bergantung pada informasi. Terutama, nilai k yang besar

akan mengurangi dampak noise pada klasifikasi, namun membuat Batas antara tiap klasifikasi terus menjadi kabur. Nilai k yang baik bisa diseleksi berlandaskan parameter optimasi, misalnya dengan memanfaatkan cross-validation[16]. Permasalahan khusus dimana klasifikasi diprediksi bersumber pada informasi pelatihan terdekat(dengan kata lain, k= 1) disebut algoritma tetangga terdekat.

Keakuratan algoritma KNN dipengaruhi oleh ada ataupun tidak terdapatnya fitur yang tidak relevan ataupun apabila nilai fitur tersebut tidak setara dengan relevansinya untuk klasifikasi. Sebagian besar penelitian tentang algoritme ini mangulas metode memilih dan menimbang fitur sehinggabahwa kinerja klasifikasi lebih baik

2.1.7 Media Sosial

Media sosial adalah sebuat wadah yang memungkina pengguna didalam nya dapat melakukan interaksi sosial tanpa melakuka pertemuan secara langsung, melalui media sosial dengan mudah terciptaka interaksi seperti komunikasi dengan pengguna lainya [17]. Dengan media sosial, kemudahan diberikan bagi orang-orang bertemu dengan orang baru yang belum pernah ditemui sebelumnya .

Seiring kemajuannya media sosial tidak lagi hanya menjadi sebuah wadah untuk berkomunikasi dengan pengguna lainnya melainkan jua dapat digunakan sebagai tempat untuk mengekspresikan diri,mengemukakan pendapat mengenai sebuah hal yang ramai diperbincangkan [18]

Dengan pemahaman mengenai media sosial tersebut, media sosial dapat menjadi sebuah *platform* yang menjembatani para pengguna dalam hal interaksi sosial dan mengekspresikan diri ,hal tersebut dapat membuka kesempatan baru untuk orang orang dalam mengemukakan pendapat dan bisa didengar oleh orang lain ,maka dari itu dengan media sosial orang-orang dapat beropini dan berinterkasi tanpa batas.

2.1.8 Twitter

Twitter adalah merupakan platform media sosisal yang dapat menghubungkan pengguna di seluruh dunia, Twitter menjadi wadah untuk penguna dalam beropini,mengukapkan pendapat melalui sebuah *text* ,twitter menyebutnya dengan istilah *tweet* [19],dalam aktifitas dalam memposting sebuah *tweet*, Twitter telah membatasi jumlah *tweet* sebesar 140 kata.

Twitter memberikan layanan komunikasi yang membuat pengguna dapat menuis sebuah opini yang dapat dijangkau secara luas[20].dengan banyak nya pengguna yang menggunakan Twitter maka Twitter pun menggunakan sistem yang berbais *real-time* dalam layanan komunikasinya.

Bedasarkan penelitian Statista [4] Indonesia menduduki peringkat kelima pengguna aktif Twitter terbanyak di dunia. banyaknya pengguna Twitter yang ada di Indonesia maka twitter dapat mewadahi opini masyarakat terhadap isu yang yang sedang hangat

2.1.9 Tweepy

Tweepy adalah sebuah library yang digunakan untuk dapat bisa mengakses API dari Twitter, proses ini biasa dilakukan dengan menggunakan Bahasa pemgoraman python, dengan menggunakan tweepy penarikan data twitter dapat dilakukan dengan cepat ,untuk bisa menggunakan Tweepy diperlukan untuk mendapatkan akses token terlebih dahulu.akses token bisa didapatkan jika melakukan pendaftaran pada Twitter Development ,akses token yang diberikan biasanya berupa *consumer key, consumer secret, access token, access token secret*.

2.1.10 TF-IDF

TF-IDF adalah sebuah algoritma pembobotan dalam *text information processing,* TF (term frequency) berfungsi untuk menampilakan banyaknya kata yang sering muncul. sedangkan IDF(inverse document frecuency) lebih berfokus pada pengukuran terhadap kata yang muncul pada dataset

TFIDF adalah metode yang mengintegrasikan *frequency term* (TF) serta *inverse document frequency* (IDF). *Term Frequency* dihitung memakai persamaan(2. 1), dan frekuensi suku ke- i yaitu frekuensi kemunculan suku ke- i pada dokumen ke- j. *Inverse document frequency*(IDF) yaitu logaritma dari perbandingan jumlah total dokumen dalam korpus dengan jumlah dokumen dengan pertanyaan, seperti yang ditunjukkan oleh rumus matematika pada Persamaan (2.2). Nilai ini diperoleh dengan mengalikan kedua rumus pada Persamaan (2.3).[21]

$$tf_i = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (2.1)$$

$$idf_i = log \frac{|D|}{|\{d : t_i \epsilon d\}} \quad (2.2)$$

$$(tf - idf)_{ij} = tf_i(d_j) * idf_j \quad (2.3)$$

2.1.11 Confusion Matrix

Confusion Matrix merupakan suatu tabel yang kerap digunakan dalam mengukur kinerja dari model klasifikasi di *machine learning* .Tabel ini menggambarkan lebih perinci tentang jumlah informasi yang diklasifikasikan dengan benar ataupun salah.Dalam bentuknya terdapat beberapa faktor dalam menentukan peforma klasifikasi. Berikut merupakan 4 faktor dalam klasifikasi *confusion matrix* bedasarkan [19]:

1. *True Positive*: Memprediksi Positif yang benar
2. *False Positive*: Memprediksi Positif yang salah (eror klasifikasi)
3. *True Negative*: Memprediksi Negatif yang benar
4. *False Negative*: Memprediksi Negatif yang salah (eror klasifikasi)

Rumus dalam menghitung performa yang umum digunakan dalam *confusion matrix* [22] adalah:

1. *Accuracy* = Keakurasi klasifikasi

$$\frac{\text{TP+TN}}{TP+FN+FP+TN} \quad (2.4)$$

2. *Precission* = perhitungan benar positif dengan keseluruhan yang diprediksikan positif

$$\frac{TP}{TP+FP} \quad (2.5)$$

3. *Recall* = perhitungan prediksi benar positif dengan keseluruhan data yang benar positif

$$\frac{TP}{TP+\mathrm{F}N} \quad (2.6)$$

Berikut merupakan contoh dari visualisasi tabel *confusion matrix:*

Tabel 2.2 Contoh Tabel Confusion Matrix

	Positive	***Negative***
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

Confusion matrix sering digunakan untuk mengevaluasi dari hasil sebuah klasifikasi. Dimana hasil akurasi tersebut berpengaruh dalam hasil dari sebuah klasifikasi [23]. *confusion matrix* digunakan untuk menganalisa performa dari hasil kalsifikasi dari kelas yang berbeda.

Confusion matrix digunakan dalam menentukan *object* yang benar atau salah [24]. Dalam pengerjaannya *confusion matrix* bekerja dengan membandingkan hasil prediktif dengan informasi kelas asli,lalu menampilkan perbandingan tersebut ke dalam sebuah matrix .

2.1.2 Keras

Keras adalah library deep learning jaringan saraf tiruan tingkat teratas yang

ditulis dengan Bahasa python dan dapat berjalan diatas TensorFlow, CNTK dan Theano. Library ini mempunyai fitur yang bisa digunakan untuk memudahkan pengembangan yang lebih dalam mengenai deep learning. Library ini dibesarkan untuk mengizinkan pengujian yang cepat pada CPU(Central Processing Unit) dan GPU(Graphic Processing Unit) dan menunjang algoritma convolutional neural network dan recurrent neural network maupun gabungan dari keduanya[25]. Sebagian fitur yang menojol dari Keras ialah:

1) Keras merupakan antarmuka tingkatan atas untuk TensorFlow dan Theano selaku backend.
2) Keras dapat berjalan mudah di CPU dan GPU.
3) Keras menunjang hampir seluruh model jaringan saraf– seluruhnya Concatenation, convolution, pooling, looping, embedding, dll. Berikutnya, model- model ini bisa digabungkan untuk membuat model yang lebih kompleks.
4) Keras merupakan framework berbasis Python yang mudah digunakan dideteksi serta dieksplorasi ataupun dipelajari.

2.1.13 Sastrawi

Sastrwi adalah sebuah *library* yang digunakan untuk melakukan sebuah prose untuk menemukan kata dasar dari sebuah kata, sastrawi juga merupakan bagian dari proses *stemming.* Stemming digunakan untuk menentukan hasil dari sebuah *text mining* [26]. Python ini adalah bagian dari proyek asli Sastrawi ditulis dalam bahasa pemrograman PHP. gunakan modul literatur ini itu diinstall di python dan akan digunakan untuk proses klasifikasi teks dan dapat dijelaskan pada subbab *Text Classification* selanjutnya.

2.1.14 ROC AUC

ROC *(Receiver Operating Characteristics)* adalah sebuah *tools* untuk melakukan pengukuran *performance* pada masalah klasifikasi untuk menentukan *threshold* dari sebuah model. Kurva ROC juga sering digunakan

untuk melakukan penilaian dari hasil prediksi bedasarkan kinerja model. [27]

Dalam praktik nya ROC dipaki untuk melakukan visualisasi data dari *confusion matrix,* dan manghasilkan garis biru dan merah yang dinamakan AUC (*Area Under Curve*). AUC sendiri memiliki fungsi untuk membentuk batasan antara kurva garis biru yang menandakan *performance* algoritma dalam membandingkan sebuah *object* ke dalam kategori

Dalam penilaian nya,AUC terbagi menjadi beberapa kolompok yaitu:

Tabel 2.3 Kategori Klasifikasi *AUC*

Nilai AUC	Kategori Klasifikasi
0.90-1.00	Excellent Classification
0.80 -0.90	Good Clasification
0.70-0.80	Fair Clasification
0.60- 0.70	Poor Clasification
0.50-0.60	Fail Clasification

2.1.5 Word2Vec

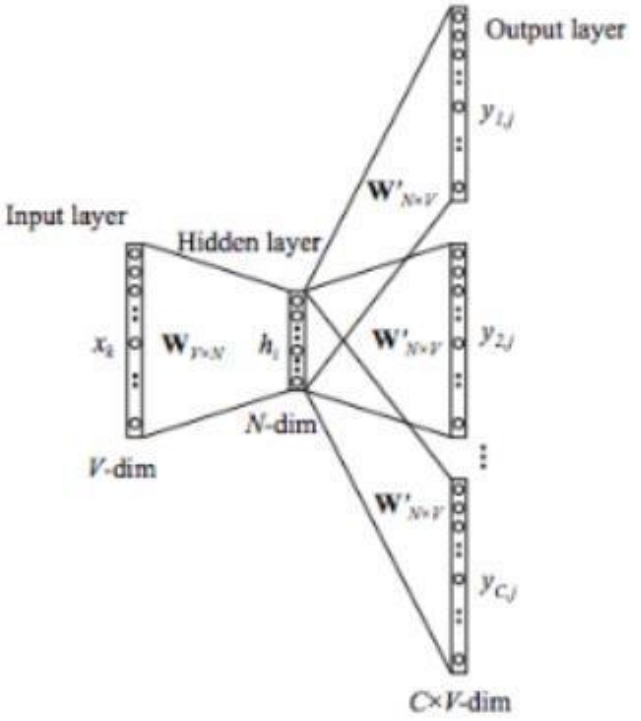
Word2vec merupakan salah satu tata cara *embedding word* yang bermanfaat untuk merepresentasikan kata jadi suatu vektor dengan panjang N, Vektor tersebut tidak cuma merepresentasikan kata secara sintaktik tetapi pula secara semantik ataupun secara makna

Word2Vec memakai neural network untuk memperoleh vektor tersebut. Arsitektur Word2vec hanya terdiri dari 3 layer ialah *Input, Projection(Hidden Layer)*, serta *Output*. *Input* pada Word2vec berupa *one- hot encoded vector* dengan panjang= jumlah kata unik pada informasi training..

Terdapat 2 jenis arsitektur neural network dari Word2Vec yaitu "*Skip-gram*" dan "*Continous Bag of Word*" (CBOW).[28]

a. Skip-Gram

Model ini memanfaatkan suatu kata guna memprediksi arah konteks. *Skip-Gram* bekerja dengan baik dengan informasi pelatihan yang jumlahnya sedikit serta bisa merepresentasikan kata- kata yang diduga langka.



Gambar 2.2 Skip-Gram

Berikut merupakan ilustrasi arsitektur *skip- gram* dengan *window size* merupakan 2 serta *current word* ataupun *input*“ universitas”

“Jurusan Sistem Informasi Universitas Multimedia Nusantara”

Jurusan : [1,0,0,0,0,0]

Sistem : [0,1,0,0,0,0]

Informasi : [0,0,1,0,0,0]

Universitas : [0,0,0,1,0,0]

Multimedia : [0,0,0,0,1,0]

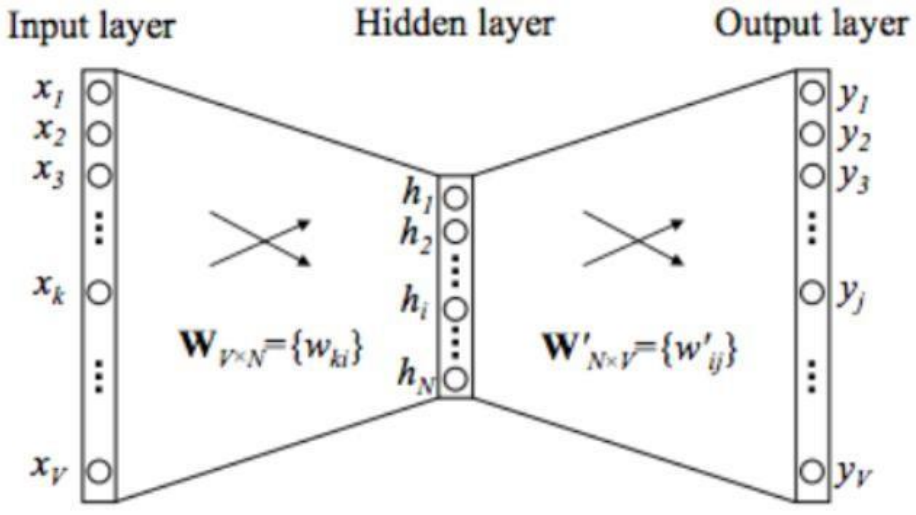
Nusantara : [0,0,0,0,0,1]

Data input berbentuk *one-hot encoded vector*, berikut sebagai proses dari *forwardbackward* proses *training* arsitektur *skip-gram*:

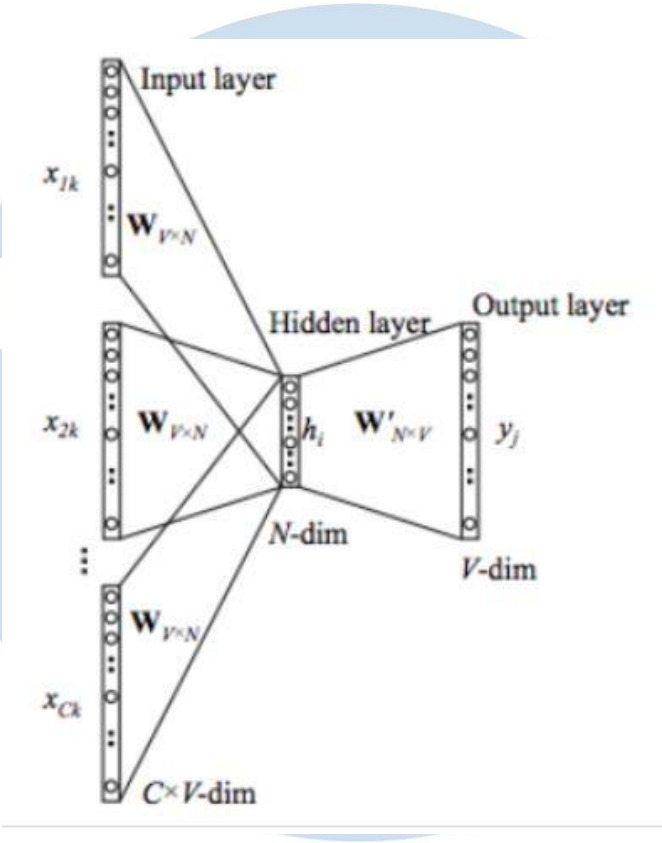
a) Jurusan Sistem Informasi Universitas Multimedia Nusantara
   Input ,target = (jurusan,sistem);(jurusan,informasi)
b) Jurusan Sistem Informasi Universitas Multimedia Nusantara
   Input ,target = (sistem,jurusan);(sistem, informasi);(sistem,universitas)
c) Jurusan Sistem Informasi Universitas Multimedia Nusantara
   Input ,target = (informasi,jurusan);(informasi,sistem);
   (informasi,universitas); (informasi,multimedia)
d) Jurusan Sistem Informasi Universitas Multimedia Nusantara
   Input ,target = (universitas,sistem);(universitas,informasi);
   (universitas,multimedia);(universitas,nusantara)

b. Continous Bag of Word

Model ini memakai konteks untuk memprediksi sasaran kata. CBOW mempunyai waktu *training* lebih kilat serta mempunyai akurasi yang sedikit lebih baik buat *frequent words*.



Gambar 2.3 One Word Context CBOW



Gambar 2.4 Multiple Context Word CBOW

Berikut merupkan ilustrasi dari arsitektur CBOW yang merupakan dari arsitektur skip-gram dengan memprediksi kata yang kosong :

a) Jurusan Sistem informasi Universitas Multimedia Nusantara
   Input,target = (sistem ,jurusan);(informasi,jurusan)
b) Jurusan Sistem informasi Universitas Multimedia Nusantara
   Input,target = (jurusan,sistem);(informasi,sistem);(universitas,sistem)
c) Jurusan Sistem informasi Universitas Multimedia Nusantara
   Input,target = (informasi ,jurusan);(sistem,informasi);
   (universitas,informasi);(multimedia,informasi)
d) Jurusan Sistem informasi Universitas Multimedia Nusantara
   Input,target = (sistem ,universitas);(informasi,universitas);
   (multimedia,universitas);(nusantara,universitas)

2.2 Penelitian Terdahulu

2.2.1 Penelitian Terdahulu

Tabel 2.4 Penelitian -Penelitian Terdahulu

1	**Penulis**	Dhaifa Farah Zhafira, Bayu Rahayudi, Indriati
	Nama jurnal	Jurnal Sistem Informasi, Teknologi Informasi, dan Edukasi Sistem Informasi (JUST-SI) Vol. 2, No. 1, Agustus 2021, hlm. 55-63
	Judul	Analisis sentimen kebijakan Kampus Merdeka menggunakan naïve bayes dan pembobotan TF-IDF bedasarkan komentator pada youtube
	Tahun	2021
	Metode	Naïve bayes, TF-IDF, k-fold cross validation
	Kesimpulan	Hasil akurasi terbaik sebesar 97% yang didapat dengan memanfaatkan 900 data latih, 100 data uji, menjalankan pembobotan TF- IDF, serta 10- fold cross validation. Rata- rata akurasi yang didapat dari 10 iterasi pada k- fold cross validation ialah sebesar 91. 8% dengan nilai precision, recall, f- measure sebesar 90. 35%, 93. 6%, 91. 95%. Berlandaskan hasil tersebut, Naive Bayes Classifier lumayan baik selaku alternatif untuk analisis sentimen.

	Adopsi	Topik yang diangkat serupa pada penelitian ini mengangkat topik mengenai Kampus Merdeka
2	**Penulis**	Hans Juwiantho, Esther Irawati Setiawan, Joan Santoso, Mauridhi Hery Purnomo
	Nama jurnal	Jurnal Teknologi Informasi Dan Ilmu Komputer
	Judul	Sentiment Analysis Twitter Bahasa Indonesia Berbasis WORD2VEC Menggunakan Deep Convolutional Neural Network
	Tahun	2020
	Metode	Convolutional Neural Network, Word2vec
	Kesimpulan	Riset tersebut memakai model Word2Vec buat Bahasa Indonesia selaku representasi kata dalam wujud vektor. Pemakaian Word2Vec pula memesatkan proses pelatihan serta tingkatkan akurasi algoritma Deep Convolutional Neural Network. Hasil percobaan yang sudah dilakukan menciptakan akurasi sebesar 76, 40%.
	Adopsi	Penggunaan Word2vec untuk merubah kata ke vector dan penerapan algoritma Convolutional Neural Network
3	**Penulis**	Akhmad Deviyanto , M. Didik R. Wahyud

	Nama jurnal	JISKa (Jurnal Informatika Sunan Kalijaga), Vol. 3, No. 1, MEI, 2018, Pp. 1–13
	Judul	Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor
	Tahun	2018
	Metode	K-Nearest Neighbor, TF-IDF
	Kesimpulan	Pada sentimen analisis yang dilakukan pada topik pilkadan DKI 2017 metode yang dilakukan menggunakan pembobotan TF-IDF dan algoritma KNN mendapatkan hasil akurasi sebesar 78% pada K-5
	Adopsi	Penggunaan model algoritma K-Nearest Neighbor dan pembobotan TF-IDF
4	**Penulis**	Sindhy Genjang Setyorini, Mustakim
	Nama jurnal	Journal of Physics: Conference Series
	Judul	Application of The Nearest Neighbor Algorithm for Classification of Online Taxibike Sentimens In Indonesia In The Google Playstore Application
	Tahun	2021
	Metode	KNN
	Kesimpulan	K- Nearest Neighbor bisa digunakan buat melaksanakan klasifikasi serta mempunyai akurasi yang bagus dalam melaksanakan komparasi pada aplikasi taxibike
	Adopsi	Penggunaan algoritma KNN.

5	**Penulis**	Auliya Rahman Isnain, Jepi Supriyanto, Muhammad Pajar Kharisma
	Nama jurnal	Indonesian Journal of Computing and Cybernetics Systems
	Judul	Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentimen Analysis of Online Learning
	Tahun	2021
	Metode	KNN, TF-IDF
	Kesimpulan	Riset ini mempraktikkan algoritma KNN buat analisis sentimen. Analisis sentimen memakai informasi Twitter dengan kata kunci“ belajar daring” dalam bahasa Indonesia. Riset ini menampilkan 84, 65% sentimen positif, Anggapan positif dihasilkan sebab kepuasan warga terhadap pendidikan daring
	Adopsi	Penerapan algoritma KNN
6	**Penulis**	Siti Saidah,Joanna Mayary
	Nama jurnal	Jurnal Ilmiah Informatika Komputer, Vol 25, No 1
	Judul	Analisis Sentimen Pengguna Twitter Terhadap Dompet Elektronik dengan metode Lexicon Based dan K-Nearest Neighbor
	Tahun	2020
	Metode	Lexicon Based ,K-Nearest Neighbor
	Kesimpulan	Riset yang diakukan yaitu melakukan analisa sentimen pada dompet digital yang ada seperti GOPAY,LinkAja dan OVO twitter

		menggunakan Lexicon Based untuk mementukan sentimen positif dan negatif dan model KNN dilakukan untuk perhitungan akurasi model, dari hasil yang penelitian tersebut menghasilkan nilai sebanyak 86,91% untuk OVO ,94% untuk LinkAja ,dan Gopay sebesar 76,31%
	Adopsi	Pengambilan metode K-Nearest Neighbor
7	**Penulis**	Siti Ernawati, Risa Wati
	Nama jurnal	JURNAL KHATULISTIWA INFORMATIKA, VOL. VI, NO. 1 JUNI 2018
	Judul	Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel
	Tahun	2018
	Metode	K-NN
	Kesimpulan	Dalam penelitian ini melakukan analisa sentimen review pada agen travel ,proses ini dilakukan menggunakan review pelanggan yang sudah menggunakan jasa agen travel,penelitian yang digunakan menggunakan model algoritma KNN ,dari penelitian yang telah dilakukan mendapatkan hasil mencapai 87% dan untuk pengukuran *performance* menggunakan perhitungan ROC AUC sebesar 0,916
	Adopsi	Penggunaan ROC AUC dan algortima K-NN

8	**Penulis**	Jeremy Andre Septian, Tresna Maulana Fahrudin, Aryo Nugroho
	Nama jurnal	JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION
	Judul	Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor
	Tahun	2021
	Metode	TF-IDF ,KNN
	Kesimpulan	Pada penelitian yang dibuat menganalisa sentimen mengenai polemic pesepakbolaan Indonesia dengan menggunakan TF-IDF dan KNN ,dengan mengam data melalui media sosial Twtter sebagai platform yang diteliti,dari hasil penelitian tersebut mendapatkan hasil nilai akurasi sebesar 79,99% dengan akurasi optimal pada nilai k=23
	Adopsi	Pengambilan penerapan yang sama yaitu meotde K-Nearest Neighbor dan TF-IDF
9	**Penulis**	Ahmed Sulaiman M.Alharbi
	Nama jurnal	Cognitive Systems Research, Volume 54, May 2019, Pages 50-61
	Judul	Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information
	Tahun	2019
	Metode	Support Vector Machine, Naïve Bayes, K-Nearest Neighbor,CNN

	Kesimpulan	Sentimen analisis mengenai informasi prilaku pengguna,menggunakan beberapa model algoritma SVM,Naïve Bayes,KNN dan CNN ,dan hasil menunjukan bahwa model CNN mendaptkan hasil yang jauh lebih baik dari algoritma lainya.
	Adopsi	Penerapan algoritma CNN
10	**Penulis**	S. Rani, Parteek Kumar
	Nama jurnal	Arabian Journal for Science and Engineering
	Judul	Deep Learning Based Sentiment Analysis Using Convolution Neural Network
	Tahun	2018
	Metode	Convolutiona Neural Network,Naïve bayes
	Kesimpulan	Riset analisis sentimen berbahasa India dengan informasi yang diperoleh dari pesan berita daring dan web. Tata cara classical machine learning semacam Naïve Bayes, K- Nearest Neighbor, Maximum Entropy, dan Support Vector Machine dikomparasikan dengan prosedur deep learning yaitu Covolutional Neural Network(CNN). Algoritma CNN diberikan pengaturan konfigurasi berbeda seperti banyaknya convolutional layer, hidden layer, serta filter size buat 12 kali percobaan. Sementara itu ukuran output, regularizer, dropout, serta jumlah epoch tidak diperhitungkan karena tidak menampilkan kenaikan akurasi yang signifikan. Hasil yang diperoleh algoritma Covolutional Neural Network menggapai akurasi 95, 4% dibanding tata cara classical machine learning dengan hasil akurasi paling tinggi pada algoritma Naïve Bayes sebesar 90%

	Adopsi	Penerapan algortima Convolutional Neural Network
11	**Penulis**	Abas Sunarya , Rina Refianti , Achmad Benny Mutiara , Wiranti Octaviani
	Nama jurnal	International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019
	Judul	Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter
	Tahun	2019
	Metode	Convolutional neural network (CNN), naïve bayes
	Kesimpulan	Penggunaan algoritma CNN mendapatkan peningkatan tingkat akurasi dibandingkan dengan naïve bayes. Dimana tingkat nilai akurasi yang memiliki sebesar 88% untuk melakukan klasifikasi teks.
	Adopsi	Penggunaan metode algoritma yang Digunakan serupa dengan penelitian yaitu CNN
12	**Penulis**	Ong Jun Ying , Muhammad Mun'im Ahmad Zabidi , Norhafizah Ramli , Usman Ullah Sheikh
	Nama jurnal	IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 9, No. 2, June 2020, pp. 212~220
	Judul	Sentiment analysis of informal Malay tweets with deep learning

	Tahun	2020
	Metode	Naïve Bayes,SVM,CNN
	Kesimpulan	Penggunaan algoritma cnn untuk melakukan analisis kemiripan Bahasa Indonesia dengan Bahasa Malaysia memiliki tinggat akurasi yang lebih tinggi dari algoritma yang lainya sebesar 77%
	Adopsi	Penerapan algoritma CNN.

Bedasarkan dari hasil jurnal terdahulu yang dilampirkan dapat disimpulkan bahwa pada artikel jurnal 1 hingga 3 merupakan jurnal utama acuan dari penelitian ini dimana pada artikel jurnal 1 pengadopsian yang diambil adalah topik mengenai kampus merdeka. Kemudian pada jurnal ke-2 pengadopsian yang diambil adalah penggunaan implementasi dari pembobotan Word2Vec dan algoritma CNN. Pada artikel jurnal ke-3 pengadopsian yang diambil adalah penggunaan dari pembobotan TF-IDF dan algoritma KNN.

Pada artikel jurnal ke-4 hingga ke-8 yang sudah dilampirkan merupakan jurnal acuan dari algoritma K-Nearest Neighbor. Bedasarkan dari beberapa jurnal tersebut dapat disimpulkan bahwa algoritma K-Nearest Neighbor memiliki akurasi yang lebih baik dibandingkan dengan algortima yang lainnya . Dengan begitu maka algoritma yang akan diadopsi pada penelitian kali ini adalah algoritma K-Nearest Neighbour (KNN).

Pada artikel jurnal ke-9 hingga ke-12 yang sudah dilampirkan merupakan jurnal acuan dari algoritma Convolutional Neural Network. Bedasarkan dari beberapa jurnal tersebut dapat disimpulkan bahwa algoritma Convolutional Neural Network memiliki akurasi yang lebih baik dibandingkan dengan algortima yang lainnya. Dengan begitu maka algoritma yang akan diadopsi pada penelitian kali ini adalah algoritma Convolutional Neural Network (CNN).