LAPORAN PROYEK

11S4037 – Pemrosesan Bahasa Alami

KLASIFIKASI TEKS PADA JUDUL BERITA MENGGUNAKAN NAÏVE BAYES CLASSIFIER

OLEH:

12S16008	**Alfendo S. P. Situmorang**
12S16030	**Boas Demeson Pangaribuan**
12S16050	**Reinheart Christian Simanungkalit**
12S17041	**Dewi Purnama Napitupulu**

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS TEKNOLOGI INFORMATIKA DAN ELEKTRO

INSTITUT TEKNOLOGI DEL

2020

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

Bab pendahuluan berisi penjelasan terkait latar belakang pemilihan topik, tujuan, manfaat yang hendak dicapai, dan ruang lingkup laporan proyek.

1.1 Latar Belakang

Berita adalah sebuah informasi fakta yang sedang dan atau sudah terjadi. Pada tahun 2006 pertumbuhan dan pertukaran informasi sudah mencapai lebih dari 550 triliun dokumen dan 7,3 juta *page* internet baru setiap harinya. Salah satu dampaknya adalah artikel berita yang diunggah di internet sangatlah banyak dalam rentang waktu yang pendek. Selama ini pengkategorian berita masih menggunakan tenaga manusia atau manual. Kategori yang banyak disertai dengan tenggat waktu yang pendek akan mempersulit editor untuk mengkategorikan berita, terutama pada artikel yang tidak memiliki perbedaan yang jelas. [1] Kategori berita pada penelitian ini dapat dibagi menjadi *sport*, *business*, *politics*, *entertainment* dan *tech.* Fokus utama pada penelitian ini adalah berita berbentuk teks. Teks berita akan dipublikasikan dengan judul yang menarik sehingga mengundang minat para pembaca. Namun di balik judul berita yang menarik, terdapat golongan berita yang seharusnya dikategorikan untuk mempermudah pembaca memilih berita yang akan dikonsumsi secara pribadi.

Pada penelitian ini, penulis memanfaatkan kinerja *Teks Mining* dalam pemrosesan bahasa alami untuk mengklasifikasikan berita. *Teks Mining* adalah cara agar teks dapat diolah dengan menggunakan komputer untuk menghasilkan analisis yang bermanfaat. Praproses dalam text mining diantaranya adalah *tokenizing, case folding, stopwords, dan stemming*. Di antara keempat langkah tersebut yang paling penting adalah proses *stemming* yang merupakan proses menghilangkan imbuhan pada suatu kata untuk mendapatkan kata dasar dari kata tersebut.

Salah satu metode statistika yang dapat melakukan pengkategorian adalah klasifikasi. Terdapat banyak metode klasifikasi seperti NBC, SVM, *Decision Tree* dan *Maximum Entropy*. Penelitian ini akan menggunakan metode NBC (*Naïve Bayes Classifier*) yang telah banyak digunakan dalam penelitian *text mining*. Salah satu kelebihan NBC adalah algoritma sederhananya yang memiliki akurasi tinggi. Penelitian ini akan membandingkan

kinerja terbaik dalam mengklasifikasikan berita dari sub-metode yang telah dijabarkan di atas dengan Bernoulli dan Multinomial NBC.

1.2 Tujuan

Tujuan dari proyek adalah sebagai berikut.

a. Mengetahui kinerja metode NBC dalam melakukan klasifikasi dari *dataset* yang disediakan.
b. Mengetahui tingkat akurasi tertinggi dari sub-metode Bernoulli dan Multinomial NBC.

1.3 Manfaat

Manfaat dari pengerjaan proyek adalah sebagai berikut.

a. Membantu para pembaca dalam memilih berita yang akan dikonsumsi sesuai kebutuhan berdasarkan kategori berita yang disediakan.
b. Sebagai referensi pada pengembangan proyek selanjutnya

1.4 Ruang Lingkup

Pada sub-bab ini dijelaskan mengenai batasan penelitian yang akan dilakukan. Adapun batasan ruang lingkup penelitian yang dilaksanakan yaitu:

1. Variabel yang akan digunakan pada proyek sebagai pertimbangan dalam melakukan klasifikasi berita adalah Judul Berita.
2. Dataset yang berisi daftar berita dalam kurun waktu satu tahun terakhir. Dataset bersumber dari BBC dan berformat .csv.
3. Metode Klasifikasi yang digunakan adalah *Naïve Bayes Classifier*.
4. Akurasi yang dibandingkan adalah dari sub-metode Bernoulli dan Multinomial NBC.

BAB 2
ISI

Pada bab ini dijelaskan mengenai tahapan analisis data dan metode, desain pemrosesan bahasa alami, implementasi, dan hasil yang berupa evaluasi dari implementasi yang telah dikerjakan.

2.1 Analisis

Pada bagian ini dijelaskan mengenai berbagai metode yang akan digunakan dan analisis terhadap data. Analisis dilakukan untuk mengenali atau mengetahui struktur dari data dan metode yang menjadi acuan pada tahap implementasi. Analisis yang dilakukan terdiri dari analisis sumber data, analisis bentuk data dan analisis metode yang digunakan.

2.1.1 Analisis Sumber Data

Dataset berita yang didapatkan dari Kaggle melalui tautan https://www.kaggle.com/c/learn-ai-bbc/data.

Kaggle sendiri adalah *platform* standar untuk *dataset*, di mana *platform* ini bersifat *open access* (dapat diakses oleh semua orang). Tahapan Analisis sumber data digunakan untuk mengetahui atribut data relevan yang akan digunakan dalam klasifikasi berita. Data judul berita yang berhasil terkumpul dari *BBC news dataset* sebanyak 1490 judul, dimana judul merupakan atribut yang dipertimbangan dalam pembangunan model.

2.1.2 Analisis Bentuk Data

Analisis terhadap bentuk data yang digunakan pada penelitian mengacu pada data yang telah diambil dari tautan https://www.kaggle.com/c/learn-ai-bbc/data.

Analisis ini dilakukan dengan tujuan untuk mengetahui karakteristik *dataset* yang akan digunakan pada penelitian guna memberikan gambaran terkait atribut apa saja yang

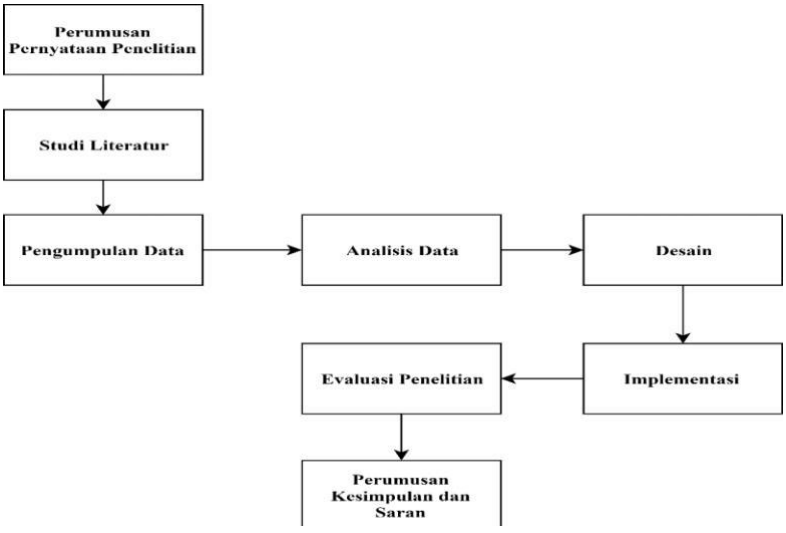
dibutuhkan pada penelitian. Berikut adalah penjelasan lebih rinci terkait bentuk data pada penelitian. Penjelasan atribut, tipe atribut, keterangan dan contoh nilai setiap atribut pada *BBC news dataset* dapat dilihat pada Tabel 1 di bawah ini.

Tabel 1 Atribut pada *BBC News Dataset*

NO	Atribut	Tipe Atribut	Keterangan	Contoh Nilai
1.	Articleid	Numerik	Kode unik berita	1833
2.	Text	Kategorikal	Judul Berita	worldcom boss launches defence lawyers defendi
3.	Category	Kategorikal	Kategori berita	business

2.1.3 Analisis Metode

Subbab ini menjelaskan tentang metodologi penelitian yang digunakan. Metodologi penelitian tersebut adalah rumusan pernyataan masalah, studi literatur, analisis desain, implementasi, evaluasi dan perumusan kesimpulan dan saran. Secara umum, langkah-langkah pengerjaan yang dilakukan dapat dilihat pada gambar berikut ini.



Gambar 1 Metode Pengerjaan Proyek

Gambar 1 merupakan metodologi penelitian yang direpresentasikan dalam bentuk bagan. Berikut adalah penjelasan bagan terkait metode penelitian.

1. **Perumusan Pernyataan Penelitian**

   Pada tahap ini dijelaskan hal yang melatarbelakangi permasalahan terkait dengan pengklasifikasian berita. Hal yang menjadi latar belakang penelitian ini bermaksud untuk menolong para pembaca dalam memilih kategori berita yang akan dibaca. Beberapa faktor yang dipertimbangkan dalam penelitian ini adalah judul dan konten dari suatu berita berita.

2. **Studi Literatur**

   Pada tahap ini dilaksanakan berbagai studi literatur yang akan menjadi informasi untuk penelitian. Studi literatur didapatkan dari berbagai sumber seperti buku dan jurnal penelitian terdahulu yang terkait dengan penelitian yang akan dilakukan. Studi literatur tersebut akan digunakan untuk menjawab pertanyaan penelitian yang akan menghasilkan landasan teoritis untuk penelitian yang akan dilakukan.

   Literatur yang digunakan pada penelitian ini sebagian besar menggunakan bahasa Inggris dan merupakan *paper* yang berasal dari luar Indonesia.

3. **Pengumpulan Data**

   Pada tahap ini dilakukan persiapan data berupa pencarian dan pengumpulan data. Data yang dikumpulkan untuk penelitian ini bersumber dari https://www.kaggle.com/c/learn-ai-bbc/data.

Data yang digunakan adalah data berita yang menggunakan bahasa Inggris. Melalui tahapan ini telah dikumpulkan sebanyak 334.784 untuk data train dan 1490 untuk data test.

4. Analisis

Pada tahap ini akan dilakukan analisis yang dimulai dari analisis data yang mencakup sumber dan bentuk data, pra-proses data, analisis metode klasifikasi dan analisis metode evaluasi terhadap pengerjaan proyek. Tujuan akhir dari analisis adalah untuk mendapatkan informasi dan gambaran mengenai penelitian yang akan dikerjakan.

5. **Desain**

   Pada tahap ini dilakukan perancangan sistem yang tersusun dari beberapa komponen yang terstruktur sebelum masuk ke tahap implementasi. Perancangan dilakukan berdasarkan hasil analisis yang sudah dilakukan dan diperoleh sebelumnya.

6. **Implementasi**

   Pada tahap ini dilakukan implementasi komponen-komponen yang telah dirancang atau didesain pada tahap sebelumnya, untuk menguji ketepatan rancangan yang dibuat. Pada penelitian ini juga dilakukan eksperimen yang mana diharapkan dapat menghasilkan keakurasian yang cukup tinggi melalui pendekatan Bernoulli dan Multinomial NBC.

7. **Evaluasi Penelitian**

   Pada tahap ini dilakukan evaluasi dan pembahasan hasil eksperimen pendeteksi klasifikasi berita. Mengukur kinerja suatu system klasifikasi merupakan hal yang penting, sehingga dapat menggambarkan seberapa baik sistem tersebut untuk mengklasifikasikan data. *Confusion Matrix* merupakan pengukur kinerja klasifikasi yang paling umum. Evaluasi dan pembahasan hasil eksperimen dilakukan untuk mengetahui apakah eksperimen memiliki performa yang cukup baik. *Accuracy* dan *Recall* yang diharapkan berada pada rentang skor tinggi. Apabila masih terdapat *error* maka dilakukan perbaikan pada model proyek.

8. **Perumusan Kesimpulan dan Saran**

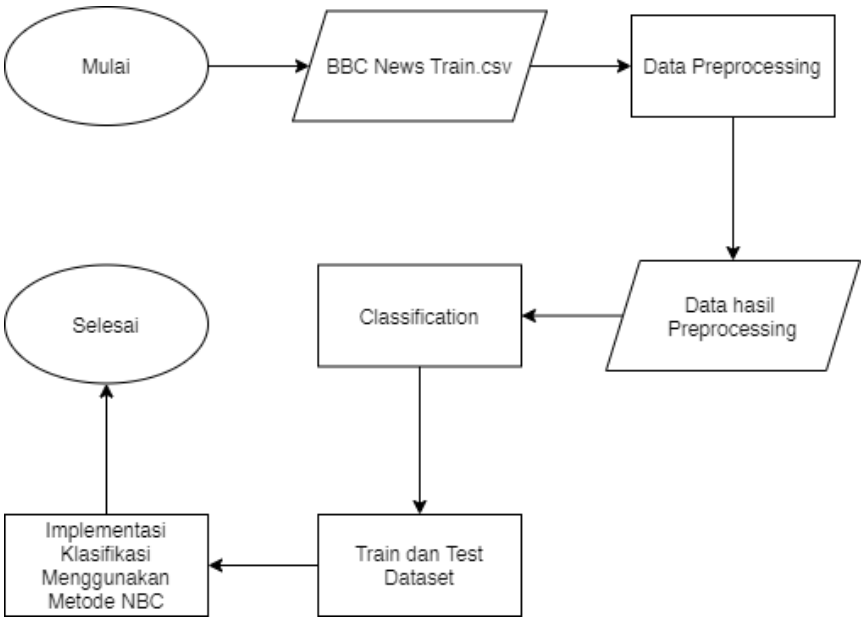
   Pada akhir penelitian, dilakukan perumusan kesimpulan dan saran terkait hasil penelitian pendeteksi berita palsu yang dicapai serta saran yang diberikan untuk pengembangan proyek selanjutnya.

2.2 Desain

Pada sub-bab ini dijelaskan desain yang kemudian diimplementasikan oleh penulis. Desain yang ada pada bab ini mencakup desain umum sistem dan desain pemrosesan data.

2.2.1 Desain Umum Sistem

Pada sub-bab ini dijelaskan mengenai gambaran umum dari sistem yang diimplementasikan. Desain umum tersebut dapat dilihat pada Gambar 2.



Gambar 2 Desain Umum Sistem

Adapun proses yang akan dilakukan pada gambar diatas adalah sebagai berikut.

1. *Input Dataset*

   Input Dataset merupakan proses di mana *dataset* yang sudah didapatkan akan digunakan sebagai *input* demi pelaksanaan pemrosesan hingga implementasi.
2. *Data Preprocessing*

   Data Preprocessing merupakan tahapan untuk membersihkan data sehingga data tersebut dapat digunakan sesuai kebutuhan yang diperlukan dalam melakukan implementasi.
3. *Classification*

   Pada tahapan ini, terdapat proses di mana data diklasifikasikan ke dalam label/kelas yang sesuai dengan ketentuan.
4. *Train* dan *Test Dataset*

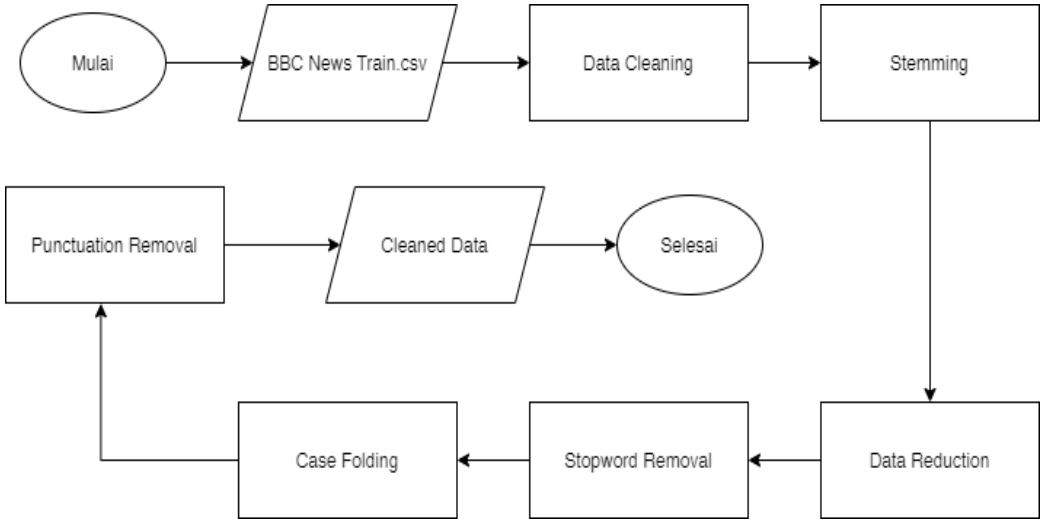
Pada tahapan ini, *dataset* dipilah untuk menentukan alokasi *training dataset* dan *testing dataset*.

5. Implementasi Klasifikasi Menggunakan Metode NBC

   Pada tahap ini, dilakukan implementasi antara 2 metode NBC untuk menentukan metode yang terbaik dalam mengklasifikasi berita.

2.2.2 Desain Prapemrosesan

Setelah *dataset* sudah didapat, maka data tersebut harus diolah terlebih dahulu. Adapun yang menjadi tahapan dalam prapemrosesan data yang akan digunakan adalah, Data Cleaning, Stemming, Data Reduction, Stopword removal, Case Folding, dan Punctuation Removal. Gambaran prapemrosesan dapat dilihat pada Gambar 3.

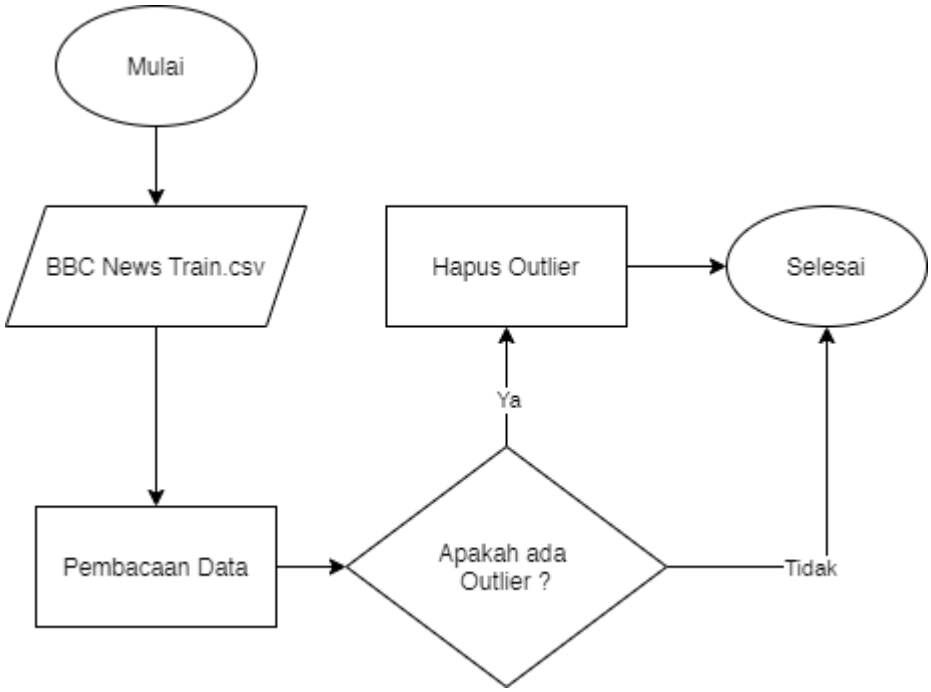


Gambar 3 Desain Prapemrosesan

1. *Data Cleaning* digunakan untuk menghapus *outlier* pada *dataset*.
2. *Stemming* digunakan untuk menghapus imbuhan terhadap kata dalam *dataset*.
3. *Data Reduction* digunakan untuk menghapus kata yang hanya mengandung karakter yang lebih sedikit dari dua dan lebih banyak dari 21.
4. *Stopword Removal* digunakan untuk menghilangkan kata yang dianggap tidak memiliki makna.
5. *Case Folding* digunakan untuk mengubah setiap kata ke dalam bentuk *lowercase*.
6. *Punctuation Removal* digunakan untuk menghapus tanda baca pada *dataset*.

2.2.2.1 Data Cleaning

Gambaran proses yang ada pada *data cleaning* dapat dilihat pada Gambar 4.

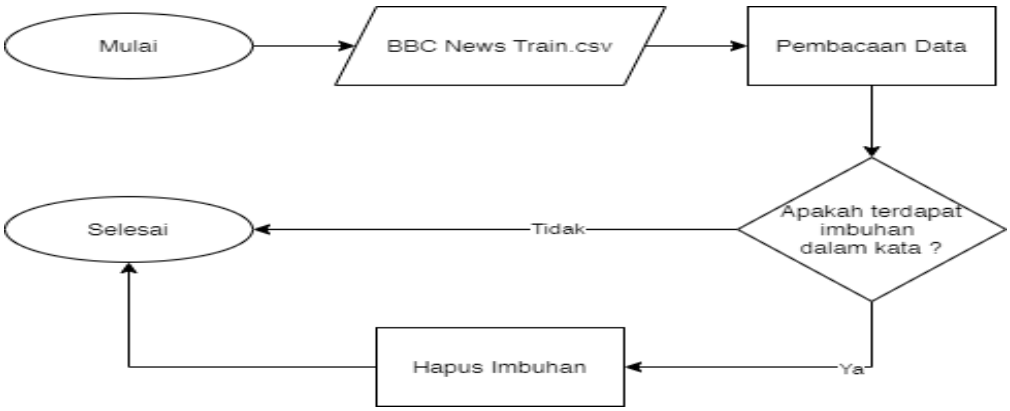


Gambar 4 *Data Cleaning*

Dataset akan dibuat sebagai *input* dalam proses ini, kemudian *dataset* tersebut akan dibaca terlebih dahulu. Dalam pembacaan *dataset* maka akan diperiksa apakah data mengandung *outlier*. Jika ada, maka *outlier* akan dihapus.

2.2.2.2 Stemming

Gambaran proses yang ada pada *stemming* dapat dilihat pada Gambar 5.

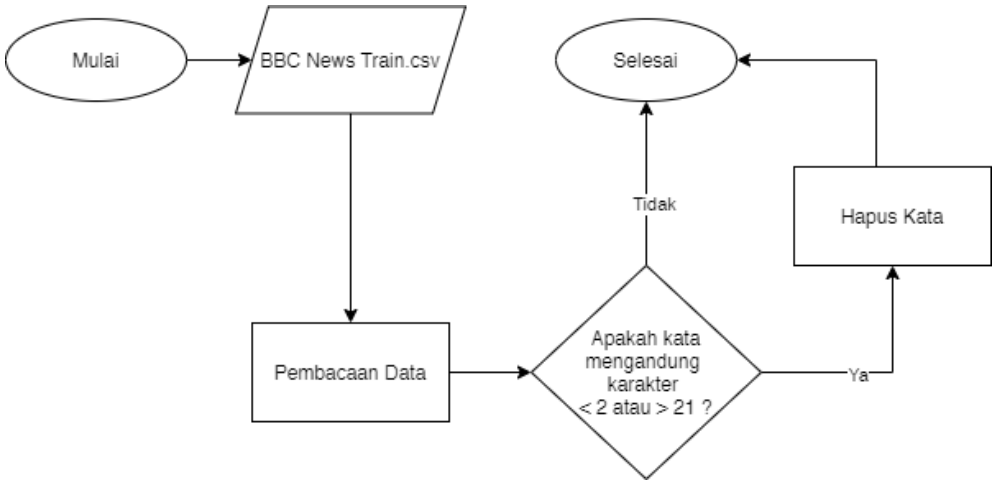


Gambar 5 *Stemming*

Dataset akan digunakan sebagai *input*, dan kemudian akan dilakukan pembacaan pada data tersebut. Setelah itu akan diperiksa apakah terdapat kata yeng berimbuhan. Jika ada imbuhan pada kata, maka imbuhan akan dihapus.

2.2.2.2 Data Reduction

Gambaran proses yang ada pada *data reduction* dapat dilihat pada Gambar 6.

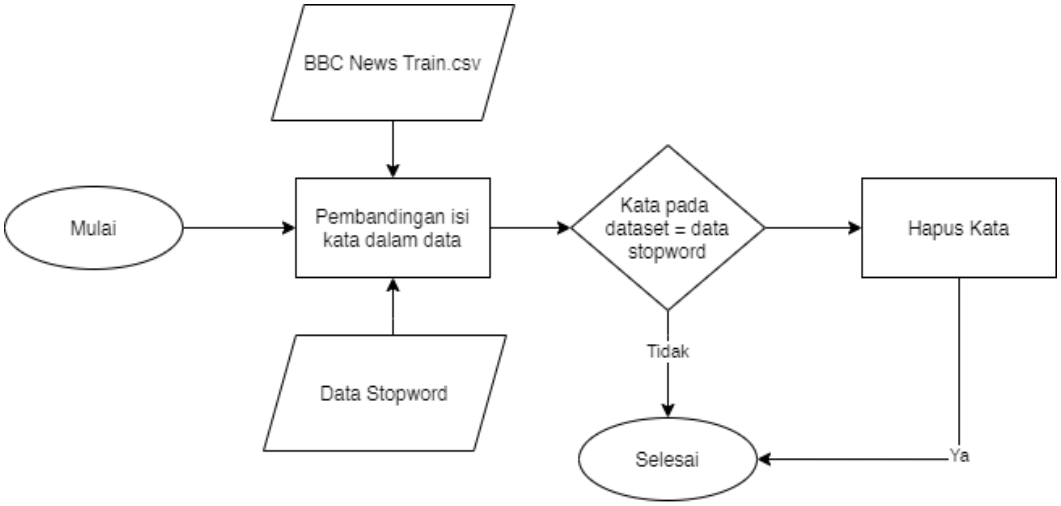


Gambar 6 *Data Reduction*

Pada tahapan ini, *dataset* akan dibaca terlebih dahulu dan kemudian akan diperiksa apakah terdapat kata yang mengandung karakter yang lebih kecil daripada 2 atau lebih besar daripada 21. Jika ada, maka kata tersebut akan dihapus.

2.2.2.2 Stopword Removal

Gambaran proses yang ada pada *stopword removal* dapat dilihat pada Gambar 7.

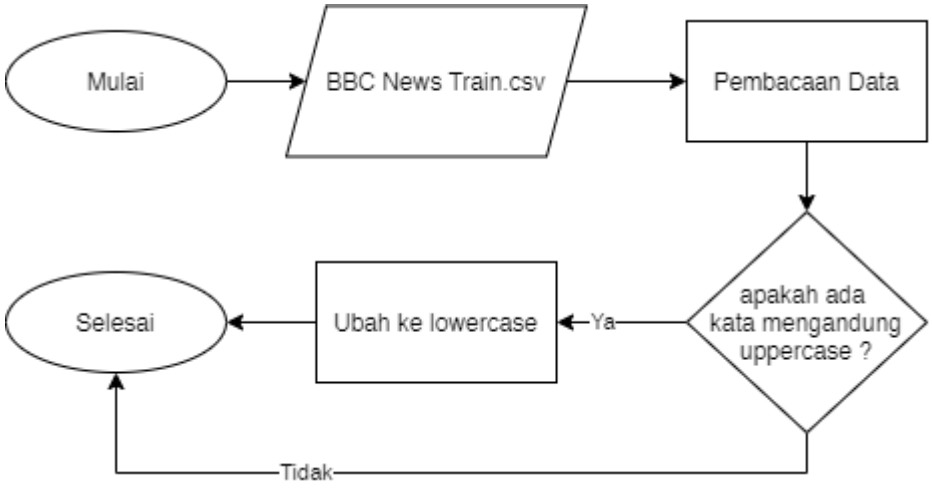


Gambar 7 *Stopword Removal*

Pada tahap ini yang menjadi input adalah *dataset* berita dan data daftar *stopword*. Kedua data itu akan dibandingkan kata demi kata. Jika kata pada data daftar *stopword* terdapat pada *dataset* berita, maka, kata tersebut akan dihapus.

2.2.2.2 Case Folding

Gambaran proses yang ada pada *case folding* dapat dilihat pada Gambar 8.

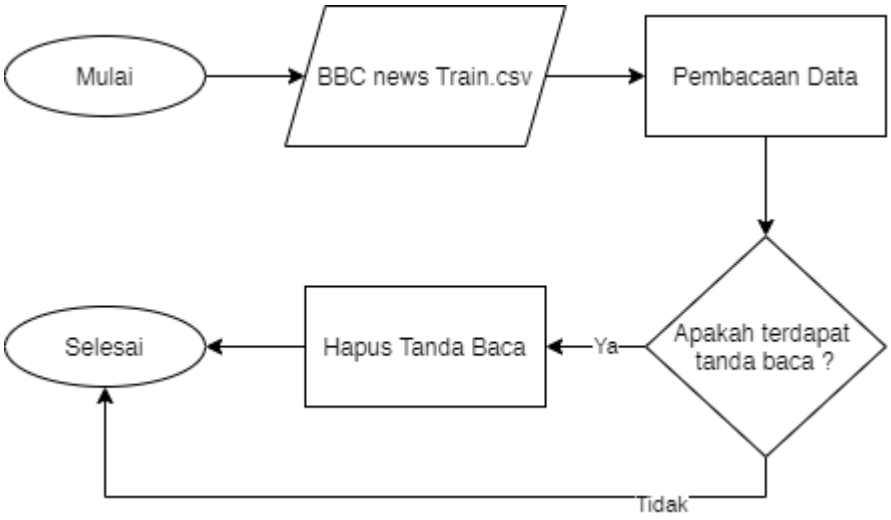


Gambar 8 *Case Folding*

Pada tahap ini, *dataset* akan menjadi *input*. Setelah itu data tersebut akan dibaca dan diperiksa apakah ada kata yang mengandung *uppercase*. Kemudian setiap *uppercase* akan diubah ke dalam *lowercase*.

2.2.2.2 Punctuation Removal

Gambaran proses yang ada pada punctuation removal dapat dilihat pada Gambar 9.



Gambar 9 *Punctuation Removal*

Pada tahap ini, *dataset* akan dimasukkan sebagai *input*, lalu dilakukan pembacaan pada *dataset* tersebut. Setelah itu akan diperiksa apakah terdapat tanda baca pada data tersebut. Jika ada, maka setiap tanda baca akan dihapus.

2.3 Implementasi dan Hasil

Pada tahap ini dilakukan evaluasi dan pembahasan hasil eksperimen klasifikasi berita berdasarkan *text-based feature.* Mengukur kinerja suatu sistem klasifikasi merupakan hal yang penting, sehingga dapat menggambarkan seberapa baik sistem tersebut untuk mengklasifikasikan data. Confusion Matrix merupakan pengukur kinerja klasifikasi yang paling umum.

Evaluasi dan pembahasan hasil eksperimen dilakukan untuk mengetahui apakah eksperimen memiliki performansi yang cukup baik untuk menentukan apakah suatu berita dikategorikan palsu atau tidak, berdasarkan jumlah (*number of neuron Accuracy*), *Precision*, *Recall* dan *Computational time* yang diharapkan tinggi. Apabila masih terdapat *error* maka dilakukan perbaikan pada penelitian.

2.3.1 Hasil Prapemrosesan Data

Tahapan dalam melakukan *text preprocessing* dilakukan dalam 7 tahapan, yaitu *data cleaning*, *data transformation*, *data reduction*, *punctuation removal*, *tokenization*, *stopword removal*, dan *lemmatization*.

2.3.1.1 *Data Cleaning*

Tahapan *data cleaning* diimplementasikan untuk mengatasi adanya *missing value* maupun *noisy value*, dengan menghapus baris data yang memiliki nilai yang kosong. Hasil dari penerapan *data cleaning* pada tahapan prapemrosesan data, dapat dilihat pada Gambar 10.

```
In [5]: df["text"].isnull().sum()
Out[5]: 0

In [6]: df["category"].isnull().sum()
Out[6]: 0
```

Gambar 10 Cuplikan Hasil *Data Cleaning*

Dari gambar tersebut, dapat dilihat bahwa tidak ada lagi adanya nilai yang kosong pada atribut *text* dan *category* yang merupakan *attribute* yang digunakan untuk melakukan klasifikasi.

2.3.1.2 *Data Reduction*

Data reduction dilakukan untuk mengurangi atribut dengan cara menghapus atribut yang tidak diperlukan dalam membangun sebuah model. Hasil dari implementasi *data reduction* terhadap data *train* dan data *test* adalah atribut *text* dan *category* dengan menghapus data yang duplikat. Penghapusan data duplikat dapat dilihat pada Gambar 11.

```
In [7]: df = df.drop_duplicates()
        df.head(10)
```

Out[7]:

	articleid	text	category
0	1833	worldcom ex-boss launches defence lawyers defe...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll indicates economic gloom citizens in ...	business
3	1976	lifestyle governs mobile choice faster bett...	tech
4	917	enron bosses in $168m payout eighteen former e...	business
5	1582	howard truanted to play snooker conservative...	politics
6	651	wales silent on grand slam talk rhys williams ...	sport
7	1797	french honour for director parker british film...	entertainment
8	2034	car giant hit by mercedes slump a slump in pro...	business
9	1866	fockers fuel festive film chart comedy meet th...	entertainment

Gambar 11 Cuplikan Hasil *Data Reduction*

2.3.1.3 *Punctuation Removal*

Punctuation Removal dilakukan untuk memperoleh data yang bersih dari tanda baca dan siap untuk diproses pada tahapan selanjutnya. Hasil yang telah diperoleh dari tahapan implementasi *punctuation removal* terhadap data *train* ditampilkan pada Gambar 12.

Out[13]:

	articleid	text	category
0	1833	worldcom boss launches defence lawyers defendi...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll economic gloom citizens majority nati...	business
3	1976	lifestyle governs mobile choice faster funkier...	tech
4	917	enron bosses payout eighteen enron directors a...	business
5	1582	howard truanted play snooker conservative lead...	politics
6	651	wales silent grand slam talk rhys williams wal...	sport
7	1797	french honour director parker british film dir...	entertainment
8	2034	car giant hit mercedes slump slump profitabili...	business
9	1866	fockers fuel festive film chart comedy meet fo...	entertainment

Gambar 12 Cuplikan Hasil *Punctuation Removal*

Berdasarkan hasil implementasi yang telah diperoleh, dapat dilihat bahwa penggunaan tanda baca pada data telah dihapus. Berdasarkan hasil implementasi *punctuation removal* dapat diperoleh dengan baik, yang kemudian akan lebih mudah untuk ditokenisasi.

2.3.1.4 *Tokenization*

Tahapan tokenisasi dilakukan untuk memecah setiap teks berita menjadi token-token agar lebih mudah untuk diproses pada tahapan *stopword removal*. Hasil yang diperoleh pada tahapan implementasi tokenisasi terhadap data *train* ditampilan pada Gambar 13.

```
In [28]: language = 'english'
         corpus = processCorpus(corpus, language)
         corpus

Out[28]: [['worldcom',
           'boss',
           'launches',
           'defence',
           'lawyers',
           'defending',
           'worldcom',
           'chief',
           'bernie',
           'ebbers',
           'battery',
           'fraud',
           'charges',
           'called',
           'company',
           'whistleblower',
           'witness',
           'cynthia',
           'cooper',
```

Gambar 13 Cuplikan Hasil *Tokenization*

Berdasarkan hasil implementasi Tokenisasi yang telah dilakukan, dapat dilihat bahwa setiap teks berita yang dimiliki atribut *text* telah berubah menjadi token.

2.3.1.5 *Stopword Removal*

Setelah dilakukan tahapan implementasi *stopwords removal*, data yang tersedia hanyalah kata-kata yang penting (*wordlist*) dari hasil token, karena pada tahapan implementasi, kata-kata yang kurang bermakna (*stoplist*) telah dibuang. Gambar 14 menunjukkan hasil dari tahapan implementasi *stopword removal* terhadap data *train.*

Out[12]:

	articleid	text	category
0	1833	worldcom boss launches defence lawyers defendi...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll economic gloom citizens majority nati...	business
3	1976	lifestyle governs mobile choice faster funkier...	tech
4	917	enron bosses payout eighteen enron directors a...	business
5	1582	howard truanted play snooker conservative lead...	politics
6	651	wales silent grand slam talk rhys williams wal...	sport
7	1797	french honour director parker british film dir...	entertainment
8	2034	car giant hit mercedes slump slump profitabili...	business
9	1866	fockers fuel festive film chart comedy meet fo...	entertainment

Gambar 14 Cuplikan Hasil *Stopword Removal*

2.3.1.6 *Stemming*

Pada tahapan implementasi *Stemming*, bentuk data akan direduksi sehingga hanya akan menyisakan bentuk dasar dari data. Gambar 15 dan Gambar 16 menunjukkan cuplikan kode dan hasil keluaran proses *stemming* pada data *train.*

```
# applies stemming to a list of tokenized words
def applyStemming(listOfTokens, stemmer):
    return [stemmer.stem(token) for token in listOfTokens]
```

Gambar 15 Cuplikan Kode Fungsi *Stemming*

	articleid	text	category
0	1833	worldcom boss launches defence lawyers defendi...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll economic gloom citizens majority nati...	business
3	1976	lifestyle governs mobile choice faster funkier...	tech
4	917	enron bosses payout eighteen enron directors a...	business
5	1582	howard truanted play snooker conservative lead...	politics
6	651	wales silent grand slam talk rhys williams wal...	sport
7	1797	french honour director parker british film dir...	entertainment
8	2034	car giant hit mercedes slump slump profitabili...	business
9	1866	fockers fuel festive film chart comedy meet fo...	entertainment

Gambar 16 Cuplikan Hasil *Stemming*

Berdasarkan gambar tersebut, dapat dilihat bahwa bentuk data telah berubah menjadi bentuk dasar.

2.3.2 Hasil TF-IDF

Metode pembobotan yang digunakan yaitu *Term frequency-Inverse document frequency (Tf-Idf)* yang mempertimbangkan seringnya kemunculan *term* dalam dokumen dan rasio panjang dokumen tersebut di dalam korpus. TF-IDF diperoleh dengan menggunakan *library Tfidf transformer* pada *Python library*. Hasil TF-IDF dapat dilihat pada Gambar 17.

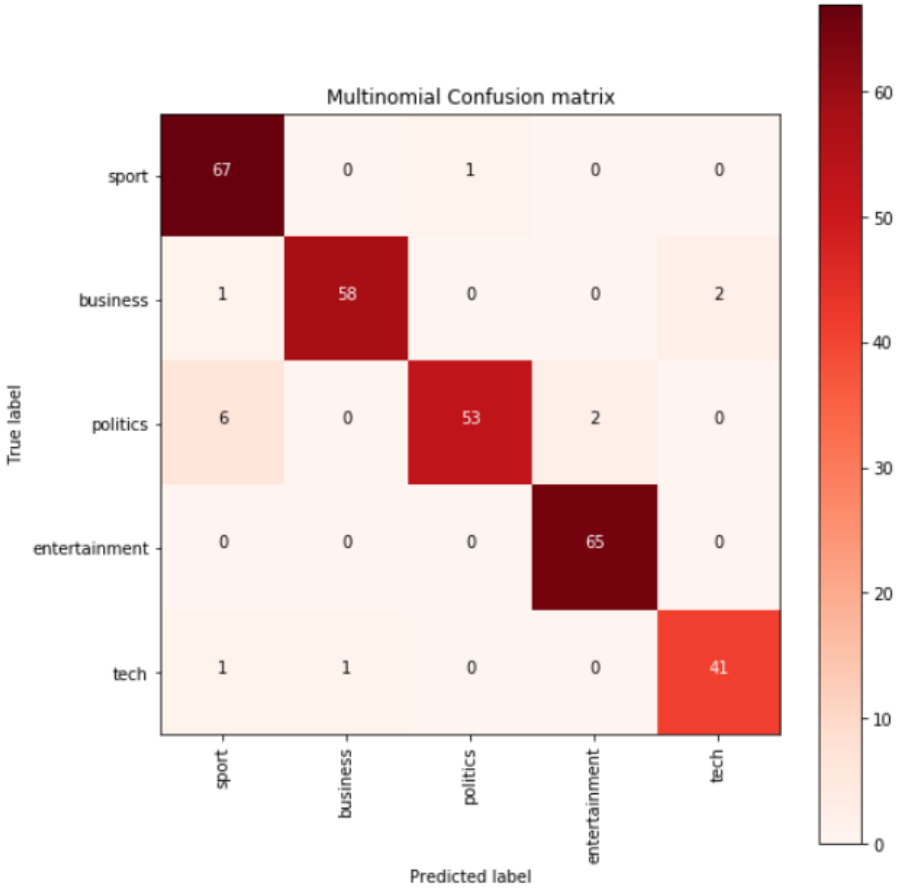
```
In [28]: tfidf = TfidfTransformer()
tfidf.fit(train_mat)
train_tfmat = tfidf.transform(train_mat)
print (train_tfmat.shape)
test_tfmat = tfidf.transform(test_mat)
print (test_tfmat.shape)

(1192, 7529)
(298, 7529)
```

Gambar 17 Cuplikan Hasil TF-IDF

2.4.3 Hasil Klasifikasi *Naïve Bayes*

Model yang telah dibangun pada tahapan implementasi. Model klasifikasi dibangun dengan menggunakan *Naïve Bayes classifier* menggunakan *Multinomial Naïve Bayes.* Pada klasifikasi dihasilkan model klasifikasi yang dapat mengelompokkan berita berdasarakan *category* dan dikelompokkan berdasarakan *text-based feature* pada dataset. Hasil klasifikasi dapat dilihat pada *confusion matrix* yang ditunjukkan pada Gambar 18.



Gambar 18 Confusion Matrix Hasil Klasifikasi *Naive Bayes Multinomial*

Dapat kita lihat pada *confusion matrix,* hasil klasifikasi bekerja dengan baik. Pada kategori *sport* hanya ada kesalahan pengelompokan berjumlah 1 butir data di mana data yang seharusnya ada di kelompok *sport* ditempatkan pada kelompok dengan kategori *politics.* Data yang seharusnya dikelompokkan pada kategori *business,* 1 butir dikelompokkan menjadi kategori *sport* dan 2 butir data dikelompokkan menjadi kategori *tech.* Data yang seharusnya dikelompokkan pada kategori *politics,* 6 butir ditempatkan pada kategori sport dan 2 butir data ditempatkan pada kategori *entertainment*. Sementara data dengan kategori *entertainment* tidak memiliki kesalahan klasifikasi dan data dengan kategori *tech* memiliki kesalahan pengelompokan data, di mana 1 butir pada kategori *sport* dan 1 butir data pada kategori *business*.

2.4.3 Hasil Evaluasi Model

Evaluasi pada model klasifikasi menggunakan beberapa metric, yaitu *precision, recall, f1-score* dan *support. Precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. *F1-score* merupakan perbandingan rata-rata presisi dan recall yang dibobotkan dimana hasil skor menunjukkan bahwa model klasifikasi memiliki kualitas yang baik.

Hasil Evaluasi

```
In [40]: # Multinomial Naïve Bayes - tf-idf
         from sklearn.metrics import classification_report

         print(classification_report(test_lbl, ypredMnb, target_names=class_names, zero_division=1))
```

	precision	recall	f1-score	support
sport	0.89	0.99	0.94	68
business	0.98	0.95	0.97	61
politics	0.98	0.87	0.92	61
entertainment	0.97	1.00	0.98	65
tech	0.95	0.95	0.95	43
accuracy			0.95	298
macro avg	0.96	0.95	0.95	298
weighted avg	0.96	0.95	0.95	298

Gambar 19 Cuplikan Hasil Evaluasi Model

BAB 3
PENUTUP

3.1 Pembagian Tugas dan Tanggungjawab

Pada subbab ini akan dipaparkan pembagian tugas dan tanggung jawab dari setiap anggota kelompok yang tampak pada table berikut.

Tabel 2 Pembagian Tugas dan Tanggungjawab

Nama	Tugas dan Tanggung Jawab
Alfendo S. P. Situmorang	1. Pencarian *dataset* 2. Hasil evaluasi implementasi 3. Slide presentasi 4. Praprosesing tahap II dataset
Boas Demeson Pangaribuan	1. Pencarian *dataset* 2. Desain *flowchart* proyek 3. Kesimpulan dan saran proyek 4. *Slide* Presentasi
Reinheart Christian Simanungkalit	1. Pencarian *dataset* 2. Praprosesing *dataset* 3. Implementasi klasifikasi menggunakan metode NBC (Bernoulli dan Multinomial NBC)
Dewi Purnama Napitupulu	1. Pencarian *dataset* 2. Bab Pendahuluan Laporan Akhir 3. Analisis Data dan Metode 4. Praprosesing tahap I dataset 5. *Exploratory Data Analysis* (EDA) *dataset*

3.2 Kesimpulan dan Saran

Pada bab ini akan dipaparkan apa yang menjadi kesimpulan dari proyek ini dan saran untuk proyek kedepannya

3.2.1 Kesimpulan

Pada proyek ini yang menjadi fokus adalah melakukan klasifikasi pada *dataset* teks berita. Data yang akan diolah harus melewati berbagai tahap pemrosesan terlebih dahulu sebelum dapat digunakan untuk pengembangan model klasifikasi. Klasifikasi yang digunakan pada pelaksanaan proyek ini adalah *Naïve Bayes classifier* dengan *sub-classifier Multinomial Naïve Bayes* dikarenakan tingkat akurasinya yang sangat baik dalam memprediksi teks. Setelah semua tahapan dilakukan, maka akan dilakukan evaluasi. Evaluasi pada model klasifikasi menggunakan beberapa metrik, yaitu *precision*, *recall*, *f1-score* dan *support*.

3.2.2 Saran

Saran yang dapat disampaikan kepada partisipan proyek di masa yang akan dating adalah sebagai berikut.

- Partisipan disarankan untuk lebih memperkuat analisis dokumen pelaksanaan proyek dan meneliti *dataset* secara lebih mendalam sebelum menjalankan implementasi proyek.
- Partisipan disarankan untuk lebih memperkuat komunikasi dan koordinasi, serta terus menginformasikan perkembangan komponen-komponen proyek kepada tim pengembangan proyek.

DAFTAR PUSTAKA

[1] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *JURNAL SAINS DAN SENI ITS*, vol. 4, pp. 2337-3520, 2015.

[2] Erwin, "Analisis Markset Basket Dengan Algoritma Apriori dan FP-Growth," *Jurnal Generic*, vol. 4, p. 26, 2009.

[3] a. M. K. Han Jiawei, Data Mining: Concepts and Techniques, USA: Morgan Kaufmann, 2006.

[4] R.Banik, Hand-On Recommendation System with Python, Birmingham: Packt Publishing, 2018.

[5] P.Chapman, CRISP_DM 1.0 Step by Step Data Mining Guide, SPSS Inc, 2019.

[6] M. K. a. J. P. J. Han, Data Mining: Concepts and Techniques, Third Edition, Vols. pp. 459-461, Waltham,USA, 2012.

[7] S. D, "Natural Language Proccesing," *Binus Universitu*, 2013.

[8] R. R, R. Rainer and R. Potter, "Introduction to Information Technology, Second Edition," *New York: John Wiley & Sons*, 2003.

[9] A. P. Widyassar, . S. Rustad, . G. F. Shidik, E. Noersasongko, . A. Syukur , A. Affandy and D. R. I. M. Setiadi, "Review of automatic text summarization techniques & methods," *Journal of King Saud University –Computer and Information Sciences*, p. 8, 2020.

[10] P. P. A. M. A. F. Alvandi Fadhil Sabily, "Analisis Sentimen Pemilihan Presiden 2019 pada Twitter menggunakan Metode Maximum Entropy," *Jurnal*

Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 5, p. 4205, 2019.

[11] P. P. A. S. A. Albert Bill Alroy, "Klasifikasi Hoaks Menggunakan Metode Maximum Entropy Dengan Seleksi Fitur Information Gain," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 3, no. 9, p. 9292, 2019.

[12] E. W. Ira Zulfa, "Sentimen Analisis Tweet Berbahasa Indonesia dengan Deep Belief Network," *IJCCS,* vol. 11, no. 2, pp. 187-198, 2017.

[13] Z.-H. Y. X. C. K. C. ,. X. L. Yu-An Huang, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics,* 2016.

[14] B. A. S. A. A. B. Z. A. Arliyanti Nurdin, "Perbandingan Kinerja Word Embedding Word2vec,Glove, dan FastText pada Klasifikasi Teks," *Jurnal TEKNOKOMPAK,* vol. 14, no. 2, pp. 74-79, 2020.

[15] S. Z, S. Q, Z. X, S. H, X. B and Y. , "Pattern Recognit," vol. 4, pp. 1623-1637, 2015.

[16] D. S. Pulkit Mehndiratta, "Identification of Sarcasm in Textual Data: A Comparative Study," *Journal of Data and Information Science,* vol. 4, no. 4, pp. 56-83, 2019.

[17] R. Suman and J. SIngh, "Sentimen Analysis of Tweets Using Support Vector Machine," *International Journal of Computer Science and Mobile Applications,* vol. 5, no. 10, pp. 83-91, October 2017.

[18] A. M. B. A. G. B. P. A. M. Nethravathi B., "Study of Techniques Used in Sentiment Analysis of Social Media Data," *MAT Journals,* vol. 5, no. 3, pp. 21-28, 2019.

[19] T. W. H. &. S. P. Perkasa, "RANCANG BANGUN PENDETEKSI GERAK MENGGUNAKAN," *Journal of Control and Network Systems,* vol. 3, no. 2, p. 92, 2014.

[20] Luthfi E. T, Suryono S and Utami E, "Analisis Sentiment Pada Twitter Dengan Menggunakan Metode Naïve Bayes Classifier," *Seminar Nasional Geotik,* pp. 9-15, 2018.