

**ANALISIS SENTIMEN PADA ACARA TELEVISI MENGGUNAKAN  
*IMPROVED K-NEAREST NEIGHBOR***

**SKRIPSI**

**WILLA OKTINAS**

**121402091**



**PROGRAM STUDI TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA**

**MEDAN**

**2017**

**ANALISIS SENTIMEN PADA ACARA TELEVISI MENGGUNAKAN  
*IMPROVED K-NEAREST NEIGHBOR***

**SKRIPSI**

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah  
Sarjana Teknologi Informasi

**WILLA OKTINAS  
121402091**



**PROGRAM STUDI TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA  
MEDAN  
2017**

## PERSETUJUAN

Judul : ANALISIS SENTIMEN PADA ACARA TELEVISI  
MENGUNAKAN *IMPROVED K-NEAREST  
NEIGHBOR*

Kategori : SKRIPSI

Nama : WILLA OKTINAS

Nomor Induk Mahasiswa : 121402091

Program Studi : SARJANA (S1) TEKNOLOGI INFORMASI

Departemen : TEKNOLOGI INFORMASI

Fakultas : ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA

Komisi Pembimbing :

Pembimbing 2 Pembimbing 1

Romi Fadillah Rahmat, B.Comp.Sc., M.Sc  
NIP. 19860303 201012 1 004

Amalia, S.T., M.T  
NIP. 19781221 201404 2 001

Diketahui / Disetujui oleh  
Program Studi S1 Teknologi Informasi  
Ketua,

Romi Fadillah Rahmat, B.Comp.Sc., M.Sc  
NIP. 19860303 201012 1 004

**PERNYATAAN****ANALISIS SENTIMEN PADA ACARA TELEVISI MENGGUNAKAN  
*IMPROVED K-NEAREST NEIGHBOR*****SKRIPSI**

Saya mengakui bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, Oktober 2017

**WILLA OKTINAS**  
121402091

## UCAPAN TERIMA KASIH

Puji dan syukur penulis sampaikan kepada Allah SWT yang telah memberikan rahmat serta restu-Nya sehingga penulis dapat menyelesaikan skripsi ini sebagai syarat untuk memperoleh gelar Sarjana Teknologi Informasi.

Pertama, penulis ingin mengucapkan terima kasih kepada Ibu Amalia,S.T.,M.T selaku pembimbing pertama dan Bapak Romi Fadillah Rahmat, B.Comp.Sc.,M.Sc selaku pembimbing kedua yang telah membimbing penulis dalam penelitian serta penulisan skripsi ini. Tanpa inspirasi serta motivasi yang diberikan dari kedua pembimbing, tentunya penulis tidak dapat menyelesaikan skripsi ini. Penulis juga mengucapkan terima kasih kepada Bapak Dr.Sawaluddin,M.IT sebagai dosen pembimbing pertama dan Bapak Indra Aulia,S.TI.,M.kom sebagai dosen pembimbing kedua yang telah memberikan masukan serta kritik yang bermanfaat dalam penulisan skripsi ini. Ucapan terima kasih juga ditujukan kepada semua dosen serta pegawai pada program studi S1 Teknologi Informasi, yang telah membantu serta membimbing penulis selama proses perkuliahan.

Penulis tentunya tidak lupa berterima kasih kepada kedua orang tua penulis, Bapak M.Nasir dan Ibu Yanti Elfina yang telah membesarkan penulis dengan sabar dan penuh cinta. Terima kasih juga penulis ucapkan kepada kakak penulis Susrianti Pebrinas,Amd,Keb dan adik penulis M.Rabil Septinas. Penulis juga berterima kasih kepada seluruh anggota keluarga penulis yang namanya tidak dapat disebutkan satu per satu.

Terima kasih juga penulis ucapkan kepada seluruh teman-teman angkatan 2012 yang telah bersama-sama dengan penulis melewati perkuliahan pada program studi S1 Teknologi Informasi terutama sahabat penulis yaitu Rosi,Tika, Oan, dan Ipat. Selanjutnya penulis juga mengucapkan terima kasih kepada Hasna, Mayya, Misbah, Ain, Zahara,Ulfa dan Wudda yang telah berbagi ilmu dan memberikan motivasi sehingga penulis dapat menyelesaikan skripsi ini.

## ABSTRAK

Sentimen masyarakat dapat dijadikan sebagai salah satu indikator oleh stasiun televisi untuk menentukan kualitas suatu acara. Pada *twitter* dapat dilakukan proses penggalian informasi mengenai sentimen masyarakat terhadap kualitas acara yang ditayangkan. Salah satu teknik penggalian informasi pada *twitter* adalah analisis sentimen. Pada penelitian ini terdiri dari 3 tahapan proses analisis sentimen. Tahap pertama yaitu proses *pre-processing* yang terdiri dari *cleansing*, *case folding*, *tokenizing*, *stopword removal*, *stemming*, dan *filter reduksi*. Selanjutnya pada tahap kedua yaitu proses perhitungan bobot pada setiap kata menggunakan metode TF-IDF. Tahap terakhir yaitu proses klasifikasi sentimen menjadi 3 kategori yaitu sentimen positif, negatif, dan netral menggunakan metode *improved k-nearest neighbor*. Hasil yang diperoleh dari pengujian analisis sentimen berbahasa Indonesia dengan metode *Improved K-Nearest Neighbor* menghasilkan akurasi tertinggi dengan nilai  $k=10$  sebesar 90%.

Kata kunci : analisis sentimen, tf-idf, *improved k-nearest neighbor*

**SENTIMENT ANALYSIS ON TELEVISION PROGRAMME'S  
BY USING IMPROVED K-NEAREST NEIGHBOR**

**ABSTRACT**

Public sentiment can be used as one of the indicator by tv stations to determine the quality of their tv programme. On *twitter*, information extraction of this public sentiment can be done to determine their tv programme's quality too. One of the method to do the information extraction on *twitter* is by using sentiment analysis method. In this research, sentiment analysis method is applied and it consists of 3 stages. The first stage is *pre-processing* which consists of *cleansing*, *case folding*, *tokenizing*, *stopword removal*, *stemming*, and *redundancy filtering*. The second stage is weighting process for every single word by using TF-IDF method. Then, the last stage is the sentiment classification process which is divided into 3 sentiment category specifically positive, negative and neutral, this process is done using the *improved k-nearest neighbor* method. The result obtained from this research generated the highest accuracy with k=10 as big as 90%.

Keywords : sentiment analysis, tf-idf, *improved k-nearest neighbor*

## DAFTAR ISI

	Hal.
PERSETUJUAN	ii
PERNYATAAN	iii
UCAPAN TERIMA KASIH	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI	vii
DAFTAR TABEL	x
DAFTAR GAMBAR	xi
 BAB 1 PENDAHULUAN	 1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	3
1.3. Tujuan Penelitian	3
1.4. Batasan Masalah	3
1.5. Manfaat Penelitian	4
1.6. Metodologi Penelitian	4
1.7. Sistematika Penulisan	5
 BAB 2 LANDASAN TEORI	 7
2.1. <i>Text Mining</i>	7
2.2. Analisis Sentimen	9
2.3. <i>Preprocessing</i>	10
2.3.1. <i>Cleansing</i>	10
2.3.2. <i>Case folding</i>	10
2.3.3. <i>Tokenizing</i>	10
2.3.4. <i>Stopword removal</i>	11
2.3.5. <i>Stemming</i>	11
2.3.6. <i>Filter redudansi</i>	11
2.4. Algoritma Nazief & Adriani	11



2.5. <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	13
2.6. <i>Improved k- Nearest Neighbor</i>	14
2.7. <i>Penelitian Terdahulu</i>	17
 BAB 3 ANALISIS DAN PERANCANGAN	 21
3.1. <i>Analisis Sistem</i>	21
3.1.1. <i>Pengumpulan dataset</i>	22
3.1.2. <i>Pre-processing</i>	23
3.1.2.1. <i>Cleansing</i>	23
3.1.2.2. <i>Case Folding</i>	24
3.1.2.3. <i>Tokenizing</i>	26
3.1.2.4. <i>Stopword removal</i>	27
3.1.2.5. <i>Steaming</i>	29
3.1.2.6. <i>Filter redudansi</i>	31
3.1.3. <i>Pembobotan kata</i>	33
3.1.4. <i>Klasifikasi dengan algoritma improved k-nearest neighbor</i>	39
3.2. <i>Perancangan Sistem</i>	43
3.2.1. <i>Rancangan tampilan halaman dashboard</i>	43
3.2.2. <i>Rancangan halaman data latih</i>	44
3.2.3. <i>Rancangan halaman data uji</i>	45
3.2.4. <i>Rancangan halaman pembobotan</i>	46
3.2.5. <i>Rancangan halaman klasifikasi</i>	47
3.2.6. <i>Rancangan tampilan halaman visualisasi</i>	58
3.2.7. <i>Rancangan tampilan halaman akurasi</i>	48
 BAB 4 IMPLEMENTASI DAN PENGUJIAN	 49
4.1. <i>Implementasi Sistem</i>	49
4.1.1. <i>Spesifikasi perangkat keras dan perangkat lunak</i>	49
4.1.2. <i>Tampilan halaman dashboard</i>	49
4.1.3. <i>Tampilan halaman data latih</i>	50
4.1.4. <i>Tampilan halaman data uji</i>	51
4.1.5. <i>Tampilan halaman pembobotan</i>	51
4.1.6. <i>Tampilan halaman klasifikasi</i>	52

4.1.7. <i>Tampilan halaman visualisasi</i>	52
4.1.8. <i>Tampilan halaman akurasi</i>	53
4.2. Pengujian Sistem	53
 BAB 5 KESIMPULAN DAN SARAN	 57
5.1. Kesimpulan	57
5.2. Saran	57
 DAFTAR PUSTAKA	 58

## DAFTAR TABEL

Tabel 2.1. Daftar Perfiks yang Meluluh (Nazief & Andriani, 1996)	12
Tabel 2.2 Daftar Kemungkinan Perubahan Perfiks (Nazief & Andriani, 1996)	12
Tabel 2.3. Daftar Kombinasi Prefiks dan Sufiks yang Tidak Diperbolehkan (Nazief & Adriani, 1996)	13
Tabel 2.4. Penelitian Terdahulu	19
Tabel 3.1. Detail Dataset dari Hasil Crawling	23
Tabel 3.2. Hasil <i>Cleansing</i>	24
Tabel 3.3. Hasil <i>Case Folding</i>	25
Tabel 3.4. Hasil <i>Tokenizing</i>	27
Tabel 3.5. Kamus <i>Stopword</i> Tala	27
Tebel 3.6. Daftar Kata yang Tidak Termasuk <i>Stopword</i>	28
Tabel 3.7. Hasil <i>Stopword Removal</i>	29
Tabel 3.8. Hasil <i>Stemming</i>	31
Tabel 3.9. Tabel Kamus Sinonim	31
Tabel 3.10. Hasil Filter Redudansi	31
Tabel 3.11. Contoh Data Latih	33
Tabel 3.12. Contoh Data Uji	34
Tabel 3.13. Perhitungan TF	35
Tabel 3.14. Perhitungan DF	36
Tabel 3.15. Perhitungan IDF	37
Tabel 3.16. Perhitungan TF-IDF	38
Tabel 3.17. Hitung Perkalian Skalar	40
Tabel 3.18. Hitung Panjang Vektor	41
Tabel 3.19. Jumlah Data Latih	42
Tabel 3.20. k-Baru	43
Table 4.1. Porposi Data Latih	53
Tabel 4.2. Pengujian Sistem Berdasarkan Nilai k	53
Tabel 4.3. Hasil Pengujian Sistem	54

## DAFTAR GAMBAR

Gambar 3.1. Arsitektur Umum	22
Gambar 3.2. <i>Flowchart Cleansing</i>	23
Gambar 3.3. <i>Flowchart Case Folding</i>	25
Gambar 3.4. <i>Flowchart Tokenizing</i>	26
Gambar 3.5. <i>Flowchart Stopword Removal</i>	28
Gambar 3.6. <i>Flowchart Stemming</i> Nazief & Andriani	30
Gambar 3.7. <i>Flowchart Filter Redudansi</i>	32
Gambar 3.8. <i>Flowchart TF-IDF</i>	35
Gambar 3.9. <i>Flowchart Improved KNN</i>	39
Gambar 3.10. Rancangan Halaman Dashboard	43
Gambar 3.11. Rancangan Halaman Data Latih	44
Gambar 3.12. Rancangan Halaman Data Uji	45
Gambar 3.13. Rancangan Halaman Pembobotan	46
Gambar 3.14. Rancangan Halaman Klasifikasi	47
Gambar 3.15. Rancangan Halaman Visualisasi	47
Gambar 3.16. Rancangan Halaman Akurasi	48
Gambar 4.1. Tampilan Halaman Dashboard	50
Gambar 4.2. Tampilan Halaman Data Latih	50
Gambar 4.3. Tampilan Halaman Data Uji	51
Gambar 4.4 Tampilan Halaman Pembobotan	51
Gambar 4.5 Tampilan Halaman Klasifikasi	52
Gambar 4.6 Tampilan Halaman Visualisasi	52
Gambar 4.8 Tampilan Halaman Akurasi	53

## **BAB I**

### **PENDAHULUAN**

Pada bab ini dijelaskan secara detail tentang hal yang berkaitan dengan pembuatan tugas akhir. Bab ini dibagi menjadi beberapa bagian yaitu : latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan.

#### **1.1. Latar Belakang**

Dunia pertelevisian Indonesia saat ini sedang berkembang, terbukti dengan semakin banyaknya jumlah stasiun televisi baik yang bersifat nasional maupun lokal. Banyaknya jumlah stasiun televisi tentu berbanding lurus dengan acara yang ditayangkan. Dimana acara tersebut dapat dikelompokkan menjadi beberapa kategori yaitu berita, anak-anak, film/sinetron, dan sebagainya. Beragamnya acara yang ditayangkan, tentunya setiap acara memiliki kualitas yang berdeda. Dimana melalui kualitas acara, stasiun televisi dapat memperitmbangkan berlanjut atau tidaknya penayangan suatu acara. Untuk itu, diperlukan suatu pengetahuan mengenai kualitas suatu acara yang ada. Salah satunya adalah melalui sentimen atau opini masyarakat.

Media sosial dapat digunakan sebagai salah satu wadah untuk menuangkan sentimen atau opini masyarakat, salah satunya adalah *twitter*. *Twitter* sebagai salah satu situs *microblogging* dengan pengguna lebih dari 500 juta dan 400 juta *tweet* perhari (Farber, 2012), dimana *twitter* menyediakan data yang bisa diakses secara bebas dengan menggunakan *twitter* API, mempermudah saat proses pengumpulan *tweets* dalam jumlah yang sangat banyak (Go, 2009).

Pada *twitter* masyarakat sering berbagi sentimen atau opini terhadap suatu acara yang ditayangkan. Melalui *tweet* yang dipublikasi oleh masyarakat tersebut, dapat dilakukan proses penggalian informasi mengenai gambaran sentimen atau opini masyarakat terhadap kualitas acara yang ada. Salah satu teknik penggalian informasi pada *twitter* adalah analisis sentimen.

*Sentiment analysis* atau *opinion mining* adalah studi komputasional dari opini-opini orang, sentimen dan emosi melalui entitas dan atribut yang dimiliki diekspresikan dalam bentuk teks (Liu, 2012). Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral (Pang & Lee, 2008).

Penelitian tentang analisis sentimen pada *twitter* telah banyak dilakukan oleh beberapa peneliti sebelumnya seperti penelitian yang dilakukan oleh Stylios et al. (2010) tentang opini masyarakat terhadap kebijakan pemerintah dengan perbandingan metode *k-Nearest Neighbor*, *Naïve Bayes* dan *Support Vector Machine*. Hasil dari penelitian ini menunjukkan bahwa *performance* metode *Support Vector Machine* lebih baik dibandingkan metode lainya dengan akurasi rata-rata 86%.

Selanjutnya Wang et al. (2012) melakukan penelitian analisis sentimen pada pemilihan presiden Amerika Serikat 2012 dengan menggunakan metode *Naïve Bayes* dan fitur *Unigram*. Hasil penelitian menunjukkan bahwa metode yang digunakan memiliki akurasi rata-rata sebesar 59%. Penelitian yang dilakukan Go et al. (2009) yaitu menentukan kepuasan pelanggan terhadap suatu produk berdasarkan *emoticon* dengan membandingkan tiga metode pembelajaran yaitu *Naïve Bayes*, *Support Vector machine* dan *Maximum Entropy* menggunakan fitur *Unigram* dan *Bigram*. Hasil penelitian menunjukkan bahwa terjadi peningkatan akurasi dengan fitur *Bigram* untuk *Naïve Bayes* dan *Maximum Entropy*.

Selain itu (Jose & Chooralil 2016) melakukan penelitian tentang sentimen masyarakat terhadap pemilihan presiden di India. Penelitian ini menggabungkan dua pendekatan yaitu pembelajaran mesin dan *Lexicon Based*, adapun pendekatan yang digunakan adalah *Naïve Bayes*, *Hidden Markov Model*, dan *SentiWordNet*. Gabungan ketiga metode memberikan akurasi rata-rata sebesar 71.48%. Demikian juga penelitian yang dilakukan oleh Yazdavar et al. (2016) analisis sentimen review obat menggunakan metode fuzzy. Pada penelitian ini akurasi rata-rata yang dicapai sebesar 71%.

Berdasarkan latar belakang diatas maka pada penelitian ini akan dilakukan penelitian yang berjudul analisis sentimen pada acara televisi menggunakan *Improved k-Nearest Neighbor*. *Improved k-Nearest Neighbor* merupakan modifikasi dari metode *k-Nearest neighbor*. Modifikasi dilakukan dalam penentuan *k-values*, dimana

penetapan *k-values* tetap dilakukan, hanya saja setiap kategori memiliki *k-values* yang berbeda. Perbedaan *k-values* yang dimiliki pada setiap kategori disesuaikan dengan besar-kecilnya jumlah dokumen latih yang dimiliki kategori tersebut.

## 1.2. Rumusan Masalah

Stasiun televisi memerlukan informasi mengenai kualitas suatu acara. Melalui kualitas acara, stasiun televisi dapat memperitmbangkan berlanjut atau tidaknya penayangan suatu acara. Dimana setimen atau opini masyarkat dapat dijadikan sebagai salah satu indikator oleh stasiun televisi untuk menentukan kualitas acara yang ada. *Twitter* dapat dijadikan sebagai salah satu sumber opini atau sentimen masyarakat mengenai suatu acara. Namun, *twitter* tidak mempunyai kemampuan untuk mengagregasi informasi menjadi sebuah kesimpulan. Untuk itu dibutuhkan suatu pendekatan, untuk menarik kesimpulan dari sentimen atau opini masyarakat untuk mendapatkan informasi mengenai kuliatas suatu acara.

## 1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah menentukan kualitas acara yang ditayangkan oleh stasiun televisi berdasarkan sentimen atau opini masyarakat menggunakan metode *Improved k-Nearest Neighbor* .

## 1.4. Batasan Masalah

Dalam penelitian ini, peneliti memberikan batasan ruang masalah agar tidak terjadi kesalahan pada saat penelitian. Batasan masalah dalam melakukan proses penelitian ini yaitu:

1. Penelitian ini hanya menggunakan opini dalam bahasa Indonesia.
2. Data yang digunakan berasal dari *tweets mention* yang ditujukan kepada 4 stasiun televisi yaitu ANTV, RCTI, Global TV, dan MNCTV.
3. Data yang dikumpulkan berupa teks.
4. Klasifikasi opini menjadi 3 kategori yaitu opini positif, negatif, dan netral pada kategori acara berita, anak-anak , dan film /sinetron.

### 1.5. Manfaat Penelitian

Manfaat yang diperoleh pada penelitian ini adalah :

1. Mendapatkan informasi mengenai kualitas suatu acara, dimana informasi tersebut dapat dijadikan sebagai bahan pertimbangan oleh stasiun televisi untuk menentukan berlanjut atau tidaknya penayangan suatu acara.
2. Mengetahui kemampuan metode *improved k- Nearest neighbor* pada bidang analisis sentimen.

### 1.6. Metodologi Penelitian

Tahapan-tahapan yang akan dilakukan pada penulisan skripsi ini adalah sebagai berikut :

#### 1. Studi Literatur

Studi literatur dilakukan dengan cara mengumpulkan bahan referensi yaitu berupa buku, artikel, paper, jurnal, makalah, maupun situs-situs dari internet. Studi literatur yang dilakukan berkaitan dengan analisis sentimen dan algoritma *Improved K-Nearest Neighbor* serta metode TF-IDF.

#### 2. Identifikasi Masalah

Pada tahap ini, dilakukan identifikasi masalah yang akan diselesaikan pada aplikasi yang akan dibangun.

#### 3. Analisis dan Perancangan Sistem

Pada tahap ini dilakukan analisis dan perancangan terhadap permasalahan yang ada dan batasan masalah.

#### 4. Implementasi Sistem

Pada tahap ini dilakukan proses implementasi pengkodean program dalam aplikasi komputer menggunakan bahasa pemrograman yang telah dipilih yang sesuai dengan analisis dan perancangan yang sudah dilakukan.

#### 5. Pengujian Sistem

Pada tahap ini dilakukan proses pengujian dan percobaan terhadap sistem sesuai dengan kebutuhan yang ditentukan sebelumnya serta memastikan program yang dibuat berjalan seperti yang diharapkan.



## 6. Dokumentasi

Pada tahap ini dilakukan pembuatan dokumentasi dalam bentuk laporan tugas akhir.

### 1.7 Sistematika Penulisan

Penulisan skripsi ini terdiri dari lima bab dengan masing-masing bab secara singkat dijelaskan sebagai berikut :

#### **Bab 1 : Pendahuluan**

Bab ini berisikan dari latar belakang yang dilaksanakan, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan.

#### **Bab 2 : Landasan Teori**

Pada bab ini berisikan tentang teori-teori pendukung skripsi yang diperlukan untuk memahami permasalahan yang dibahas pada penelitian ini yaitu Analisis Sentimen, Algoritma *Improved k-Nearest Neighbor*, dan metode TF-IDF (Term Frequency-Inverse Document Frequency).

#### **Bab 3 : Analisis dan Perancangan Sistem**

Pada bab ini berisikan tentang paparan analisis terhadap permasalahan dan penyelesaian persoalan terhadap algoritma *Improved k-Nearest Neighbor* dan metode TF-IDF serta identifikasi kebutuhan perancangan sistem.

#### **Bab 4 : Implementasi dan Pengujian Sistem**

Pada bab ini berisikan implementasi perancangan sistem dari hasil analisis dan perancangan yang sudah dipaparkan pada bab 3, serta menguji sistem untuk menemukan kelebihan dan kekurangan pada sistem yang dibuat.

**Bab 5 : Kesimpulan dan Saran**

Pada bab ini berisikan kesimpulan yang didapatkan terhadap hasil penelitian skripsi dan saran untuk pengembangan lebih lanjut tentang topik terkait pada skripsi ini.

## BAB 2

### LANDASAN TEORI

#### **2.1. Text Mining**

*Text mining* adalah proses mengambil informasi dari teks. Informasi biasanya diperoleh melalui peramalan pola dan kecenderungan pembelajaran pola statistik. *Text mining* yaitu parsing, bersama dengan penambahan beberapa fitur linguistik turunan dan penghilangan beberapa diantaranya, dan penyisipan *subsequent* ke dalam database, menentukan pola dalam data terstruktur, dan akhirnya mengevaluasi dan menginterpretasi output, *text mining* biasanya mengacu ke beberapa kombinasi relevansi, kebaruan, dan *interestingness*.

Kunci dari proses pada *text mining* adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst, 2003). Sedangkan menurut (Harlian, 2006) *text mining* didefinisikan sebagai data yang berupa teks yang biasanya sumber data didapatkan dari dokumen, dengan tujuan adalah mencari kata-kata yang dapat mewakili isi dari dokumen tersebut yang nantinya dapat dilakukan analisa hubungan antar dokumen. Proses *text mining* yang khas meliputi kategorisasi teks, *text clustering*, ekstraksi konsep/entitas, produksi taksonomi granular, *sentiment analysis*, penyimpulan dokumen, dan pemodelan relasi entitas yaitu, pembelajaran hubungan antara entitas (Bridge, 2011).

Pendekatan manual *text mining* secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah memungkinkan ranah tersebut untuk berkembang selama dekade terakhir. *Text mining* adalah bidang interdisipliner yang mengacu pada pencarian informasi, pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistik. Dikarenakan kebanyakan informasi (perkiraan umum mengatakan lebih dari 80%) saat ini disimpan sebagai teks, *text mining* diyakini memiliki potensi nilai komersial tinggi (Bridge, 2011).

Saat ini, *text mining* telah mendapat perhatian dalam berbagai bidang (Sumartini, 2011):

#### 1. Aplikasi keamanan.

Banyak paket perangkat lunak *text mining* dipasarkan terhadap aplikasi keamanan, khususnya analisis *plaintext* seperti berita Internet. Hal ini juga mencakup studi enkripsi teks.

#### 2. Aplikasi biomedis

Berbagai aplikasi *text mining* dalam literatur biomedis telah disusun. Salah satu contohnya adalah PubGene yang mengkombinasikan *text mining* biomedis dengan visualisasi jaringan sebagai sebuah layanan Internet. Contoh lain *textmining* adalah GoPubMed.org. Kesamaan semantik juga telah digunakan oleh sistem *text mining*, yaitu, GOAnnotator.

#### 3. Perangkat Lunak dan Aplikasi

Departemen riset dan pengembangan perusahaan besar, termasuk IBM dan Microsoft, sedang meneliti teknik *text mining* dan mengembangkan program untuk lebih mengotomatisasi proses pertambangan dan analisis. Perangkat lunak *text mining* juga sedang diteliti oleh perusahaan yang berbeda yang bekerja di bidang pencarian dan pengindeksan secara umum sebagai cara untuk meningkatkan performansinya

#### 4. Aplikasi Media Online

Text mining sedang digunakan oleh perusahaan media besar, seperti perusahaan Tribune, untuk menghilangkan ambiguitas informasi dan untuk memberikan pembaca dengan pengalaman pencarian yang lebih baik, yang meningkatkan loyalitas pada site dan pendapatan. Selain itu, editor diuntungkan dengan mampu berbagi, mengasosiasikan dan properti paket berita, secara signifikan meningkatkan peluang untuk menguangkan konten.

## 5. Aplikasi Pemasaran

*Text mining* juga mulai digunakan dalam pemasaran, lebih spesifik dalam analisis manajemen hubungan pelanggan yang menerapkan model analisis prediksi untuk *churn* pelanggan (pengurangan pelanggan).

## 6. Sentiment Analysis

*Sentiment Analysis* mungkin melibatkan analisis dari review film untuk memperkirakan berapa baik review untuk sebuah film. Analisis semacam ini mungkin memerlukan kumpulan data berlabel atau label dari efektifitas kata-kata. Sebuah sumber daya untuk efektifitas kata-kata telah dibuat untuk WordNet.

## 7. Aplikasi Akademik

Masalah *text mining* penting bagi penerbit yang memiliki database besar untuk mendapatkan informasi yang memerlukan pengindeksan untuk pencarian. Hal ini terutama berlaku dalam ilmu sains, di mana informasi yang sangat spesifik sering terkandung dalam teks tertulis. Oleh karena itu, inisiatif telah diambil seperti *Nature's proposal* untuk *Open Text Mining Interface* (OTMI) dan *Health's common Journal Publishing* untuk *Document Type Definition* (DTD) yang akan memberikan isyarat semantik pada mesin untuk menjawab pertanyaan spesifik yang terkandung dalam teks.

## 2.2. Analisis Sentimen

Analisis sentimen adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negatif (Nasukawa & Yi, 2003). Pendapat yang hampir senada dikemukakan oleh (Cvijikj & Michahelles, 2011), di mana analisis sentimen digunakan untuk memahami komentar yang diciptakan oleh pengguna internet dan menjelaskan bagaimana sebuah produk maupun *brand* diterima oleh mereka. Definisi analisis sentimen *twitter* sendiri merupakan bagian dari pendapat pada media *twitter*. Pesan *twitter*, pada kenyataannya, lebih mudah untuk menganalisis karena penulisan yang dibatasi dibanding forum diskusi. Hal ini berbeda pada forum diskusi yang lebih sulit, dikarenakan pengguna dapat mendiskusikan apapun dan berinteraksi satu sama

lain. Kalimat seringkali memuat pendapat tunggal, meskipun tidak bersifat mutlak bahwa setiap kalimat berisi pendapat tunggal. Dalam kasus lain terdapat kalimat dengan pendapat lebih dari satu pada suatu kalimat namun ini hanya sebagian kecil (Liu, 2012).

Pada dasarnya sentimen analisis merupakan tahapan klasifikasi. Namun tahapan klasifikasi sentimen pada *twitter* (tidak terstruktur) sedikit lebih sulit dibanding dengan klasifikasi dokumen terstruktur. Dalam kasus analisis sentiment *twitter* yang merupakan gambaran dari kalimat, langkah pertama (Liu, 2012) adalah untuk mengklasifikasikan apakah kalimat mengungkapkan pendapat atau tidak. Langkah kedua adalah mengklasifikasikan kalimat-kalimat pendapat menjadi positif dan kelas negatif.

### **2.3. Pre-Processing**

*Pre-processing* dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten (Hemalatha et al. 2012). Tahapan dari *pre-processing* adalah sebagai berikut:

#### **2.3.1. Cleansing**

Proses membersihkan dokumen dari kata yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan adalah karakter HTML, kata kunci, ikon emosi, *hashtag* (#), *username* (@username), url (<http://situs.com>), dan email (nama@situs.com).

#### **2.3.2. Case folding**

*Case folding* yaitu perubahan bentuk huruf menjadi huruf kecil.

#### **2.3.3. Tokenizing**

*Tokenizing* adalah proses memecah teks menjadi kata tunggal dan penghapusan tanda baca serta angka, sesuai dengan kamus data yang telah ditentukan. Pada penelitian ini fitur yang digunakan dalam memecah teks adalah *unigram* yaitu token yang terdiri hanya satu kata.

#### 2.3.4. *Stopword removal*

*Stopwords removal* adalah proses menghilangkan kata tidak penting dalam text. Hal ini dilakukan untuk memperbesar akurasi dari pembobotan *term*. Untuk proses ini, diperlukan suatu kamus kata-kata yang bisa dihilangkan. Dalam Bahasa Indonesia, misalnya kata: dan, atau, mungkin, ini, itu, dll adalah kata-kata yang dapat dihilangkan.

#### 2.3.5. *Stemming*

*Stemming* adalah pengubahan kata ke bentuk kata dasar atau penghapusan imbuhan. Stemming disini menggunakan kamus daftar kata berimbuhan yang mempunyai kata dasarnya dengan cara membandingkan kata-kata yang ada di dalam komentar dengan daftar kamus stem.

#### 2.3.6. *Filter redundansi*

*Filter redundansi* merupakan proses mencari sinonim kata pada database sinonim. Untuk mengoptimalkan perhitungan frekuensi kemunculan kata pada proses pembobotan maka diperlukan kamus sinonim untuk mengecek kata yang memiliki makna yang sama. Jika kata tersebut ditemukan didalam kamus sinonim maka kata tersebut diubah ke bentuk sinonimnya.

### 2.4. Algoritma Nazief & Adriani

Algoritma Nazief & Adriani menyimpulkan sebuah kata dasar dapat ditambahkan imbuhan berupa *derivation prefix* (DP) di awal dan/atau diakhiri secara berurutan oleh *derivation suffix* (DS), *possesive pronoun* (PP), dan *particle* (P). Keterangan diatas dirumuskan sebagai berikut :

$$DP + DP + DP + \textit{root word} + DS + PP + P \quad (2.1)$$

Adapun langkah-langkah yang digunakan oleh algoritma Nazief dan Adriani yaitu sebagai berikut: (Nazief & Adriani, 1996)

- a. Kata dicari di dalam daftar kamus. Bila kata tersebut ditemukan di dalam kamus, maka dapat diasumsikan kata tersebut adalah kata dasar sehingga algoritma dihentikan.
- b. Bila kata di dalam langkah pertama tidak ditemukan di dalam kamus, maka diperiksa apakah sufiks tersebut yaitu sebuah partikel (“-lah” atau “-kah”). Bila ditemukan, maka partikel tersebut dihilangkan.
- c. Pemeriksaan dilanjutkan pada kata ganti milik (“-ku”, “-mu”, “-nya”). Bila ditemukan, maka kata ganti tersebut dihilangkan.
- d. Memeriksa akhiran (“-i”, “-an”). Bila ditemukan, maka akhiran tersebut dihilangkan. Hingga langkah ke-4 dibutuhkan ketelitian untuk memeriksa apakah akhiran “-an” merupakan hanya bagian dari akhiran “-kan”, dan memeriksa lagi apakah partikel (“-lah”, “-kah”) dan kata ganti milik (“-ku”, “-mu”, “-nya”) yang telah dihilangkan pada langkah 2 dan 3 bukan merupakan bagian dari kata dasar.
- e. Memeriksa awalan (“se-“, ”ke-“, “di-“, “te-“, “be-“, “pe-“, “me-“). Bila ditemukan, maka awalan tersebut dihilangkan. Pemeriksaan dilakukan dengan berulang mengingat adanya kemungkinan *multi-prefix*. Langkah ke-5 ini juga membutuhkan ketelitian untuk memeriksa kemungkinan peluluhan awalan (Tabel 2.1), perubahan *prefix* yang disesuaikan dengan huruf-awal kata (Tabel 2.2) dan aturan kombinasi *prefix-suffix* yang diperbolehkan (Tabel 2.3).
- f. Setelah menyelesaikan semua langkah dengan sukses, maka algoritma akan mengembalikan kata dasar yang ditemukan.

**Tabel 2.1. Daftar Perfiks yang Meluluh (Nazief & Andriani, 1996)**

Jenis Prefiks	Huruf	Hasil Peluluhan
pe-/me-	K	-ng-
pe-/me-	P	-m-
pe-/me-	S	-ny-
pe-/me-	T	-n-



Tabel 2.2 Daftar Kemungkinan Perubahan Perfiks (Nazief &amp; Andriani, 1996)

Prefiks	Perubahan
se-	tidak berubah
ke-	tidak berubah
di-	tidak berubah
be-	ber-
te-	ter-
pe-	per-, pen-, pem-, peng-
me-	men-, mem-, meng-

Tabel 2.3. Daftar Kombinasi Prefiks dan Sufiks yang Tidak Diperbolehkan (Nazief &amp; Adriani, 1996)

Prefix	Sufiks yang tidak diperbolehkan
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se	-i, kan
te-	-an
pe-	-kan

### 2.5. Term Frequency-Inverse Document Frequency (TF-IDF)

Metode TF-IDF merupakan metode untuk menghitung bobot dari kata yang digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap *token* (kata) disetiap dokumen dalam korpus.

*Term frequency* (TF) adalah jumlah kemunculan kata pada suatu dokumen. Semakin banyak suatu kata muncul pada dokumen, maka semakin besar kata tersebut

berpengaruh pada dokumen tersebut. Sebaliknya, semakin sedikit suatu kata muncul pada dokumen, maka semakin kecil kata tersebut berpengaruh pada dokumen tersebut.

*Inverse document frequency* (IDF) adalah pembobotan kata yang didasarkan pada banyaknya dokumen yang mengandung kata tertentu. Semakin banyak dokumen yang mengandung suatu kata tertentu, semakin kecil pengaruh kata tersebut pada dokumen. Sebaliknya, semakin sedikit dokumen yang mengandung suatu kata tertentu, semakin besar pengaruh kata tersebut pada dokumen. Pembobotan menggunakan TF-IDF dijelaskan pada Persamaan (2.3) (Feldman & Sanger, 2007).

$$IDF(w) = \log\left(\frac{N}{DF(w)}\right) \quad (2.2)$$

$$TF-IDF(w,d) = TF(w,d) \times IDF(W) \quad (2.3)$$

Keterangan :

TF-IDF (w,d)	: bobot suatu kata dalam keseluruhan dokumen
W	: suatu kata (word)
d	: suatu dokumen
Tf(w,d)	: frekuensi kemunculan sebuah kata w dalam dokumen
IDF(w)	: inverse DF dari kata W
N	: jumlah keseluruhan dokumen
DF (w)	: jumlah dokumen yang mengandung kata w

## 2.6. Improved K-Nearest Neighbor

Penentuan *k-values* yang tepat diperlukan agar didapatkan akurasi yang tinggi dalam proses kategorisasi dokumen uji. Algoritma *Improved k-Nearest Neighbor* melakukan modifikasi dalam penentuan *k-values*. Dimana penetapan *k-values* tetap dilakukan, hanya saja tiap-tiap kategori memiliki *k-values* yang berbeda. Perbedaan *k-values* yang dimiliki pada setiap kategori disesuaikan dengan besar-kecilnya jumlah dokumen latihan yang dimiliki kategori tersebut. Sehingga ketika *k-values* semakin tinggi, hasil kategori tidak terpengaruh pada kategori yang memiliki jumlah dokumen latihan yang lebih besar. Untuk menghitung similaritas antara dua dokumen

menggunakan metode Cosine Similarity (CosSim). Dipandang sebagai pengukuran (*similarity measure*) antara vector dokumen (D) dengan vector query (Q). Semakin sama suatu *vector* dokumen dengan *vector query* maka dokumen dapat dipandang semakin sesuai dengan *query*. Rumus yang digunakan untuk menghitung *cosine similarity* adalah sebagai berikut:

$$\text{cosSim}(X, dj) = \frac{\sum_{i=1}^m x_i \cdot d_{ji}}{\sqrt{(\sum_{i=1}^m x_i)^2} \cdot \sqrt{(\sum_{i=1}^m d_{ji})^2}} \quad (2.4)$$

Dimana  $X$  adalah dokumen uji,  $dj$  dokumen *training*,  $x_i$  dan  $d_{ji}$  adalah nilai bobot yang diberikan pada setiap *term* pada dokumen. Kedekatan *query* dan dokumen diindikasikan dengan sudut yang dibentuk. Nilai *cosinus* yang cenderung besar mengindikasikan bahwa dokumen cenderung sesuai *query*. Dalam proses membandingkan dokumen yang sesuai dengan dokumen yang telah ada atau dokumen lainnya, maka digunakan perhitungan dengan rumus pada persamaan (2.4) untuk mengetahui angka similaritas dari dokumen tersebut.

Perhitungan penetapan *k-values* pada algoritma *Improved k-Nearest Neighbor* dilakukan dengan menggunakan persamaan (2.5) dengan terlebih dahulu mengurutkan secara menurun hasil perhitungan similaritas pada setiap kategori. Selanjutnya pada algoritma *Improved k-Nearest Neighbor*, *k-values* yang baru disebut dengan  $n$ . Persamaan (2.5) menjelaskan mengenai proporsi penetapan *k-values* ( $n$ ) pada setiap kategori

$$n = \left\lceil \frac{k \cdot N(c_m)}{\max\{N(c_m) | j=1 \dots N_C\}} \right\rceil \quad (2.5)$$

Keterangan:

$n$  : *k-values* baru,

$k$ : *k-values* yang ditetapkan,

$N(C_m)$  : jumlah dokumen latih di kategori/kategori  $m$ ,

$\max\{N(C_m) | j=1 \dots N_C\}$  : jumlah dokumen latih terbanyak pada semua kategori .

Dalam menentukan kategori untuk dokumen uji menggunakan algoritma *Improved k-Nearest Neighbors*, maka dilakukan perbandingan kemiripan dokumen uji dengan dokumen latih pada tiap kategori. Persamaan (2.6) menyatakan nilai

maksimum perbandingan antara kemiripan dokumen X dengan dokumen latih  $d_j$  sejumlah top  $n$  tetangga pada suatu kategori dengan kemiripan dokumen X dengan dokumen latih  $d_j$  sejumlah top  $n$  tetangga pada *training set*.

$$p(x, c_m) = \operatorname{argmax}_m \frac{\sum_{d_j \in \text{top } n \text{ k } NN(c_m)} \text{sim}(x, d_j) y(d_j, c_m)}{\sum_{d_j \in \text{top } n \text{ k } NN(c_m)} \text{sim}(x, d_j)} \quad (2.6)$$

dimana:

$p(x, c_m)$  : probabilitas dokumen X menjadi anggota kategori  $c_m$

$\text{sim}(x, d_j)$  : kemiripan antara dokumen X dengan dokumen latih  $d_j$

*top n kNN* : top  $n$  tetangga

$y(d_j, c_m)$  : fungsi atribut dari kategori yang memenuhi persamaan

Adapun langkah-langkah untuk klasifikasi dokumen X menggunakan algoritma *Improved K-Nearest Neighbor* adalah sebagai berikut:

1. Melakukan tahapan *pre-prosesing* sehingga didapatkan representasi dari dokumen X dan semua dokumen latih.
2. Setelah terbentuk vektor, hitung bobot masing-masing dokumen menggunakan TF-IDF.
3. Hitung nilai *cosine similarity* dokumen X dengan semua dokumen latih.
4. Selanjutnya urutkan hasil dari perhitungan nilai *cosine similarity* secara menurun. Nilai yang lebih tinggi menunjukkan bahwa di antara dokumen X dan dokumen latih tersebut memiliki kemiripan.
5. Mengelompokkan hasil dari perhitungan nilai *cosine similarity* berdasarkan kategorinya.
6. Menentukan *k-values* kemudian melakukan perhitungan penetapan *k-values* baru ( $n$ ) pada masing-masing kategori menggunakan persamaan (2.5).
7. Setelah didapatkan nilai  $n$  yang menyatakan sebagai top tetangga dari langkah 6, maka langkah selanjutnya adalah menentukan kategori dokumen X berdasarkan hasil perhitungan menggunakan persamaan (2.6).
8. Berdasarkan perhitungan pada persamaan (2.6), maka dokumen X akan dikategorikan ke dalam kategori yang memiliki  $P(x, c_m)$  terbesar.

## 2.8. Penelitian Terdahulu

Penelitian tentang analisis sentimen sebelumnya yang dilakukan oleh (Stylios, G et al. 2010) tentang opini masyarakat terhadap kebijakan pemerintah dengan perbandingan metode *k-nearest neighbor*, *naïve bayes* dan *support vector machine*. Hasil dari penelitian ini menunjukkan bahwa *performance* metode *support vector machine* lebih baik dibandingkan metode lainya dengan akurasi rata-rata 86% sedangkan untuk *k-nearest neighbor* sebesar 84% dan *Naïve Bayes* sebesar 73%.

Selanjutnya Wang et al. (2012) melakukan penelitian tentang analisis sentimen pada pemilihan presiden Amerika Serikat 2012 dengan menggunakan metode *Naïve Bayes* dan fitur *Unigram* untuk menentukan sentimen positif, negatif, dan netral Hasil penelitian menunjukkan bahwa metode yang digunakan memiliki akurasi rata-rata sebesar 59%.

Penelitian lainnya menggunakan metode pembelajaran mesin untuk menentukan kepuasan pelanggan terhadap suatu produk berdasarkan *emoticon* dengan membandingkan tiga metode pembelajaran yaitu *naïve bayes*, *support vector machine* dan *maximum entropy* untuk menentukan sentiment positif dan negatif. Pada penelitian ini menggunakan dua buah fitur yaitu fitur unigram dan bigram. Hasil penelitian menunjukkan bahwa terjadi peningkatan akurasi dengan fitur bigram untuk metode *Naïve Bayes* dari 81.3% menjadi 82.7% dan *Maxiumum Entropy* dari 80.5% menjadi 82.7% sedangkan untuk metode svm terjadi penurunan akurasi dari 82.2% menjadi 81.6% (Go et al 2009).

Selanjutnya (Jose & Chooralil 2016) melakukan penelitian tentang sentimen masyarakat terhadap pemilihan presiden di India, dimana sentimen tersebut dikelompokkan menjadi dua yaitu positif dan negatif. Penelitian ini menggabungkan dua pendekatan yaitu pembelajaran mesin dan *lexicon based*, adapun pendekatan yang digunakan adalah *naïve bayes*, *hidden markov model*, dan *sentiwordnet*. Penelitian ini menggunakan 12.000 dataset dan hasil penelitian menunjukkan bahwa diperoleh akurasi rata-rata sebesar 71.48%. Demikian juga penelitian yang dilakukan oleh (Yazdavar et al. 2016) analisis sentimen review obat menggunakan metode fuzzy. Pada penelitian ini akurasi rata-rata yang dicapai sebesar 71%.

Dalam penelitian (Altrabsheh et al. 2016) meneliti tentang analisis sentimen siswa terhadap pengajar secara *realtime*. Dalam penelitian ini membanding kinerja metode *naïve bayes* dengan *support vector machine*. Pengujian untuk metode *naïve*

*bayes* tanpa tahap *pre-processing* terjadi peningkatan akurasi sebesar 20%. Dari hasil penelitian didapatkan peningkatan akurasi tertinggi diperoleh sebesar 95% dengan menggunakan metode *support vector machine*.

Pada tahun (2014) Razzaq et al. melakukan penelitian analisis sentimen pada *twitter* mengenai opini masyarakat terhadap pemilihan presiden Pakistan, dimana sentimen tersebut dapat dijadikan sebagai acuan untuk prediksi hasil pemilu. Opini di klasifikasikan kedalam 3 kelas yaitu opini positif, negatif, dan netral. Metode Klasifikasi yang digunakan pada penelitian ini yaitu *naïve bayes* dan *support vector machine*. Akurasi tertinggi diperoleh dengan menggunakan metode *support vector machine* sebesar 70%.

Selanjutnya Mandal et al. (2017) melakukan penelitian analisis sentimen terhadap *review* produk. Sentimen diklasifikasikan menjadi 3 kelas yaitu sentimen positif, negatif, dan netral. Pada penelitian ini menggunakan metode *lexicon based* untuk klasifikasi. Akurasi tertinggi yang diperoleh menggunakan *lexicon based* sebesar 97,1%. Li et al. (2016) meneliti analisis sentimen masyarakat terhadap layanan pemerintah. Dalam penelitian ini Klasifikasi sentimen menggunakan metode *multinomial naïve bayes* dengan fitur *unigram*. Hasil penelitian ini menghasilkan akurasi rata-rata sebesar 72,3%.

Penelitian yang dilakukan oleh (Govindarajan, 2013) yaitu analisis sentimen dari *review* film menggunakan metode *hybrid naïve bayes* dan algoritma genetika. terdiri dari 2000 dataset berbahasa Inggris 1000 untuk label positif dan 1000 untuk label negatif, algoritma *best first search* sebagai filter untuk mengurangi kelebihan data, dalam penggabungan 2 metode *naïve bayes* dan algoritma genetika yang menggunakan  $10 \times 10$ -fold cross-validasi untuk mengevaluasi akurasi. Sehingga menghasilkan akurasi 93,80%. Rangkuman penelitian diatas dapat dilihat pada Tabel 2.4.

**Tabel 2.4. Penelitian Terdahulu**

No	Peneliti	Tahun	Metode	Akurasi
1.	Stylios et al.	2010	<i>k-Nearest Neighbor , Naïve Bayes dan Support Vector Machine</i>	86%
2	Wang et al.	2012	<i>Naïve Bayes</i>	59%
3	Go et al.	2009	<i>Naïve Bayes, Maxiumum Entropy dan Support Vector machine + Unigram</i>	82.2%
			<i>Naïve Bayes, Maxiumum Entropy dan Support Vector machine + Bigram</i>	82.7%,
4	Jose & Chooralil	2016	<i>Hybrid Naïve Bayes, Hidden Markov Model, dan SentiWordNet</i>	71.48%
5	Yazdavar et al.	2016	<i>Fuzzy logic</i>	71%
6	Altrabsheh et al.	2016	<i>Support Vector Machine</i>	95%
7	Razzaq et al.	2014	<i>Support Vector Machine</i>	70%
8	Mandal et al.	2017	<i>Lexicon Based</i>	97,1%.
9	Li et al.	2016	<i>Multinomial Naïve Bayes</i>	72,3%
10	Govindarajan	2013	<i>Hybrid Naive Bayes Dan Algoritma Genetika</i>	93,80%

Pada penelitian sebelumnya dapat diketahui bahwa metode *k-nearest neighbor* dapat digunakan untuk klasifikasi sentimen. Namun metode *k-nearest neighbor* memiliki kelemahan pada tingkat akurasi, karena pada semua kategori memiliki *k-values* yang sama tanpa memperhitungkan jumlah dokumen latih yang dimiliki

masing-masing kategori, sedangkan distribusi dokumen latih dalam data *training* tidak sama. Untuk mengatasi kelemahan *k-nearest neighbor* maka pada penelitian ini penulis mengajukan suatu metode untuk menentukan *rating* acara televisi berdasarkan opini publik menggunakan *improved k-nearest neighbor*.

*Improved k-nearest neighbor* merupakan modifikasi dari *k-nearest neighbor*. Modifikasi dilakukan dalam penetapan *k-values*. Pada *improved k-nearest neighbor* penetapan *k-values* tetap dilakukan, hanya saja setiap kategori memiliki *k-values* yang berbeda. Perbedaan *k-values* yang dimiliki pada setiap kategori disesuaikan dengan besar-kecilnya jumlah dokumen latih yang dimiliki kategori tersebut sehingga dapat meningkatkan hasil akurasi.



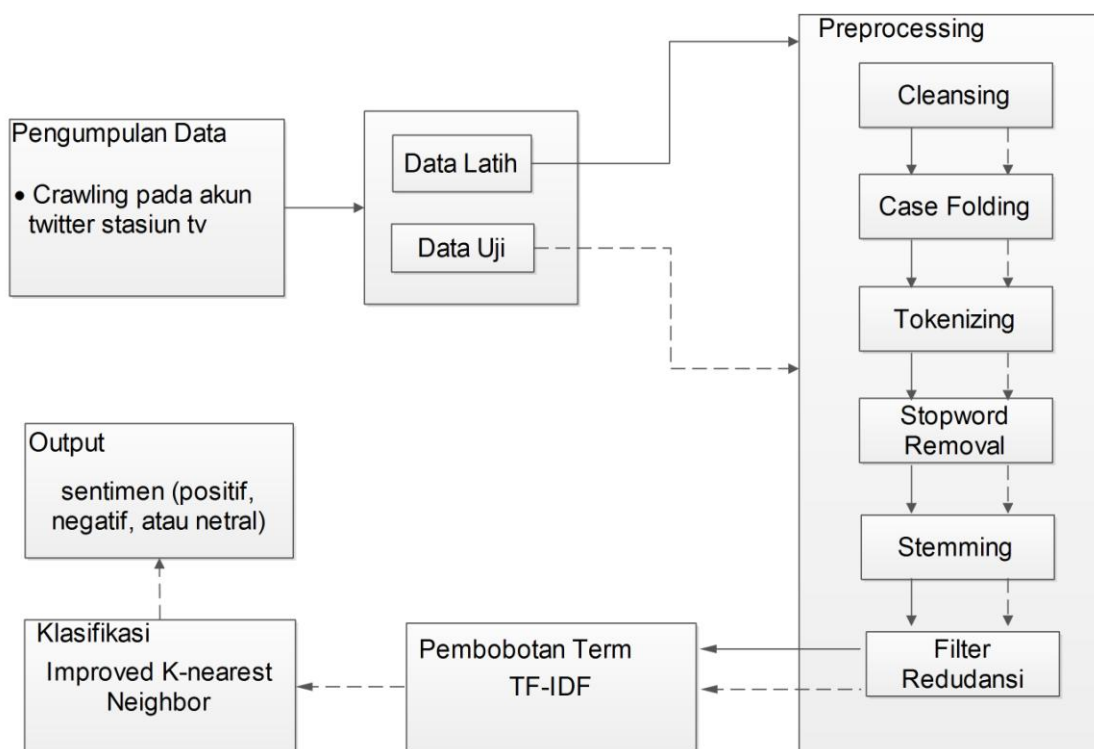
## **BAB 3**

### **ANALISIS DAN PERANCANGAN**

Bab ini akan membahas tentang implementasi metode yang digunakan untuk analisis sentimen pada acara televisi. Adapun dua tahapan yang dibahas pada bab ini yaitu tahap analisis dan tahap perancangan sistem. Analisis terhadap data yang digunakan dan analisis terhadap metode yang digunakan pada setiap langkah pemrosesan data akan dibahas pada tahap analisis. Perancangan tampilan antarmuka sistem akan dibahas pada tahap perancangan sistem.

#### **3.1. Analisis Sistem**

Metode yang diajukan penulis untuk menentukan *rating* stasiun televisi terdiri dari beberapa proses. Proses-proses yang akan dilakukan adalah sebagai berikut: kumpulan opini memasuki proses *preprocessing*, hasil dari *preprocessing* dilakukan perhitungan bobot pada setiap kata menggunakan tf-idf, selanjutnya dilakukan proses klasifikasi dengan algoritma *improved k-nearest neighbor* pada tahap ini dilakukan untuk mengklasifikasikan menjadi 3 kategori yaitu opini positif, negatif, dan netral. Arsitektur umum yang mendeskripsikan setiap metodologi pada penelitian ini ditunjukkan pada Gambar 3.1.



**Gambar 3.1. Arsitektur Umum**

### 3.1.1. Pengumpulan dataset

Data yang digunakan pada penelitian ini diambil dari kumpulan *tweets* bahasa Indonesia yaitu *tweets mention* yang ditujukan kepada 4 akun *twitter* stasiun televisi yaitu @officialRCTI, @WatsonANTV, @Globaltvseru dan @official\_MNCTV melalui *twitter search* API. Data yang digunakan dalam penelitian ini terdiri dari dua jenis, yaitu data latih dan data uji. Data latih yang digunakan diambil dari kumpulan opini yang telah dilabeli dengan kelas sentimennya secara manual yang berjumlah 70% dari keseluruhan data. Data inilah yang digunakan sebagai data latih untuk membentuk model analisis sentimen. 30% dari keseluruhan data, nantinya akan digunakan sebagai data uji. Data uji ini menggunakan kumpulan kata yang belum memiliki label. Detail dari kumpulan *dataset* yang didapat dari hasil *crawling* dapat dilihat pada Tabel 3.1.

**Tabel 3.1. Detail Dataset dari Hasil *Crawling***

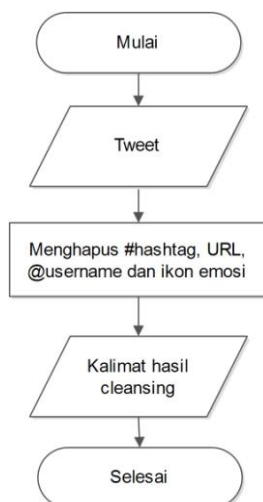
Stasiun TV	Jumlah	
	Data Latih	Data Uji
ANTV	700	300
Global TV	700	300
RCTI	700	300
MNCTV	700	300

### 3.1.2. *Pre-processing*

*Pre-processing* dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten (Hemalatha et al. 2012). Adapun Langkah-langkah *Preprocessing* dalam penelitian ini yaitu *cleansing*, *case folding*, *tokenizing*, *stopword removal*, *stemming*, *filter redudansi*.

#### 3.1.2.1 *Cleansing*

Proses membersihkan dokumen dari kata yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan adalah *hashtag* (#), *username* (@username), url (<http://situs.com>), ikon emosi, dan email (nama@situs.com). Berikut adalah *flowchart* untuk proses cleansing dapat dilihat pada Gambar 3.2.

**Gambar 3.2. *Flowchart Cleansing***

Adapun penjelasan dari flowchart diatas adalah sebagai berikut:

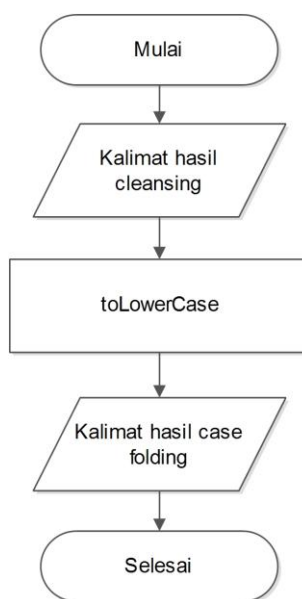
1. Data masukan yang digunakan berupa *tweet*.
2. Tweet akan diperiksa apakah terdapat karakter *hashtag* (#), *username* (@username), url (http://situs.com), dan email (nama@situs.com) maka karakter tersebut akan dihapus. Hasil *cleansing* dapat dilihat pada Tabel 3.2.

**Tabel 3.2. Hasil Cleansing**

Kode	Input Tweet Asli	Output <i>Cleansing</i>
D1	nonton Naruto Shippuden seru tayang di @Globaltvseru	nonton Naruto Shippuden seru tayang di
D2	ntn film di @Bigmovies_GTV kebanyakan kepotong filmnya #globaltvgakseru @Globaltvseru	ntn film di kebanyakan kepotong filmnya
D3	betul Males ah nonton naruto karna di ulang2 terus Benar2 ga seru	betul Males ah nonton naruto karna d iulang2 terus Benar ga seru
D4	katanya #jagonyafilm tapi kok naruto nya di ulang lagi :D bikin males nonton @bigmovie_gtv @globaltvseru	katanya tapi kok naruto nya di ulang lagi bikin males nonton
D5	@Bigmovies_GTV kenapa banyak scence film di potong jauh banget? Ga mutu @Globaltvseru	kenapa banyak scence film di potong jauh banget Ga mutu
D6	@Globaltvseru teruslah tayangkan acara2 berkualitas!!!!	teruslah tayangkan acara2 yang berkualitas

#### .3.1.2.2. Case folding

Pada tahap ini, semua huruf akan diubah menjadi *lowercase* atau huruf kecil. *Flowchart* untuk proses *case folding* dapat dilihat pada Gambar 3.3.



**Gambar 3.3. Flowchart Case Folding**

Adapun penjelasan dari flowchart diatas adalah sebagai berikut:

1. Kalimat yang digunakan adalah kalimat hasil *cleansing*.
2. Hasil dari *cleansing* akan diperiksa apakah karakter yang menggunakan huruf kapital (*uppercase*), maka huruf tersebut akan diubah menjadi huruf kecil (*lowercase*). Hasil *case folding* dapat dilihat pada Tabel 3.3.

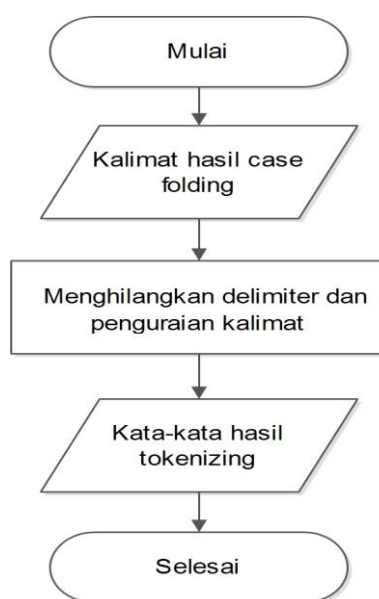
**Tabel 3.3. Hasil Case Folding**

Kode	Input Hasil <i>Cleansing</i>	Output <i>Case Folding</i>
D1	nonton Naruto Shippuden seru tayang di	nonton naruto shippuden seru tayang di
D2	ntn film di kebanyakan kepotong filmnya	ntn film di kebanyakan kepotong filmnya
D3	betul Males ah nonton naruto karna d iulang2 terus Benar ga seru	betul males ah nonton naruto karna di ulang2 terus benar ga seru
D4	katanya tapi kok naruto nya di ulang lagi bikin males nonton	katanya tapi kok naruto nya di ulang lagi bikin males nonton

D5	kenapa banyak scence film di potong jauh banget Ga mutu	kenapa banyak scence film di potong jauh banget ga mutu
D6	teruslah tayangkan acara2 yang berkualitas	teruslah tayangkan acara2 yang berkualitas

### 3.1.2.3.Tokenizing

Pada tahap ini akan dilakukan pemisahan kalimat menjadi kata tunggal dan pengecekan kata dari karakter pertama sampai karakter terakhir. Apabila karakter ke-*i* bukan tanda pemisah kata seperti titik(.), koma(,), spasi dan tanda pemisah lainnya, maka akan digabungkan dengan karakter selanjutnya. *Flowchart* untuk proses *tokenizing* dapat dilihat pada Gmbar 3.4.



**Gambar 3.4. Flowchart Tokenizing**

Adapun penjelasan dari flowchart diatas dalah sebagai berikut:

1. Kalimat yang digunakan adalah kalimat hasil *case folding*.
2. Memotong setiap kata dalam kalimat berdasarkan pemisah kata seperti titik(.), koma(,), dan spasi. Bagian yang hanya memiliki satu karakter *non* alfabet dan angka akan dibuang. Sehingga menghasilkan kata-kata penyusun kalimat. Hasil *tokenizing* dapat dilihat pada Tabel 3.4.

**Tabel 3.4. Hasil *Tokenizing***

Kode	Input Hasil <i>Case Folding</i>	Output <i>Tokenizing</i>
D1	nonton naruto shippuden seru tayang di	[ nonton, naruto, shippuden, seru , tayang, di ]
D2	ntn film di kebanyakan kepotong filmnya	[ ntn, film, di, kebanyakan, kepotong ]
D3	betul males ah nonton naruto karna di ulang2 terus benar ga seru	[ males, ah, nonton, naruto, karna ,di ulang, terus, benar, ga, seru]
D4	katanya tapi kok naruto nya di ulang lagi bikin males nonton	[ kok, narutonya, di, ulang, lagi, bikin, males, nonton ]
D5	kenapa banyak scence film di potong jauh banget ga mutu	[kenapa, banyak, scence, film ,di potong, jauh, banget, ga ,mutu]
D6	teruslah tayangkan acara2 yang berkualitas	[ teruslah, tayangkan, acara, yang, berkualitas]

#### 3.2.1.4. *Stopword removal*

Pada tahap ini, kumpulan kata yang telah melewati tahap *tokenzing* akan melalui tahap *stopword removal*. Setiap kata akan diperiksa. Jika terdapat kata sambung, kata depan, kata ganti atau kata yang tidak ada hubungannya dalam analisis sentimen, maka kata tersebut akan dihilangkan. Pada penelitian ini menggunakan kamus *stopword* Tala (2013) dimana datanya berjumlah 748 data. Kamus *stopword* Tala telah banyak dimanfaatkan oleh peneliti untuk membuang kata-kata yang tidak penting dalam bahasa Indonesia. Contoh *stopword* pada kamus Tala dapat dilihat pada Tabel 3.5.

**Tabel 3.5. Kamus *Stopword* Tala**

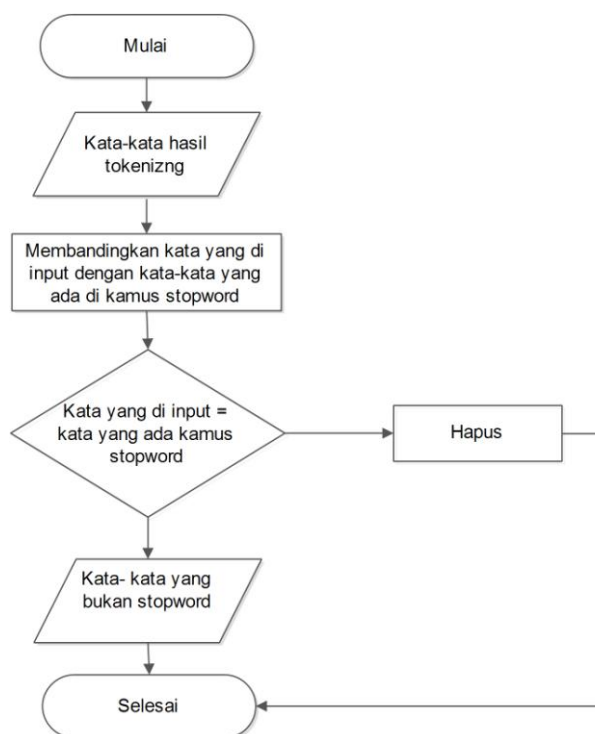
<i>Stopword</i> Tala			
Di	Adalah	Beberapa	Khususnya
Kapan	Adapun	Berupa	Kini
Pula	Aku	Besar	Kok
Dari	Bagaimana	Entah	Oleh
Karena	Bahwasanya	Ketika	Pada

Dalam penelitian ini, tidak semua kata yang ada didalam kamus Tala digunakan sebagai *stopword*. Hal ini dikarenakan kata tersebut berpengaruh terhadap makna atau nilai sentimen khususnya untuk sentimen yang bernilai negatif. Adapun daftar kata didalam kamus Tala yang tidak digunakan sebagai *stopword* dapat dilihat pada Tabel 3.6.

**Tebel 3.6. Daftar Kata yang Tidak Termasuk *Stopword***

Daftar Kata yang Tidak Termasuk <i>Stopword</i>			
Tidak	Enggak	Tambah	Cuma
Tidaklah	Enggaknya	Baik	Hanya
Gak	Bukan	Hampir	Pantas
Nggak	Jangan	Nyaris	Bisa
Ga	Kurang	Lama	Tak

*Flowchart* untuk proses *stopword removal* dapat dilihat pada Gambar 3.5.



**Gambar 3.5. Flowchart Stopword Removal**

Adapun penjelasan *flowchart* diatas adalah sebagai berikut:

1. Kata hasil *tokenizing* akan dibandingkan dengan daftar *stopword*.



2. Dilakukan pengecekan apakah kata sama dengan daftar *stopword* atau tidak.
3. Jika kata sama dengan yang ada pada daftar *stopword*, maka akan dihilangkan.

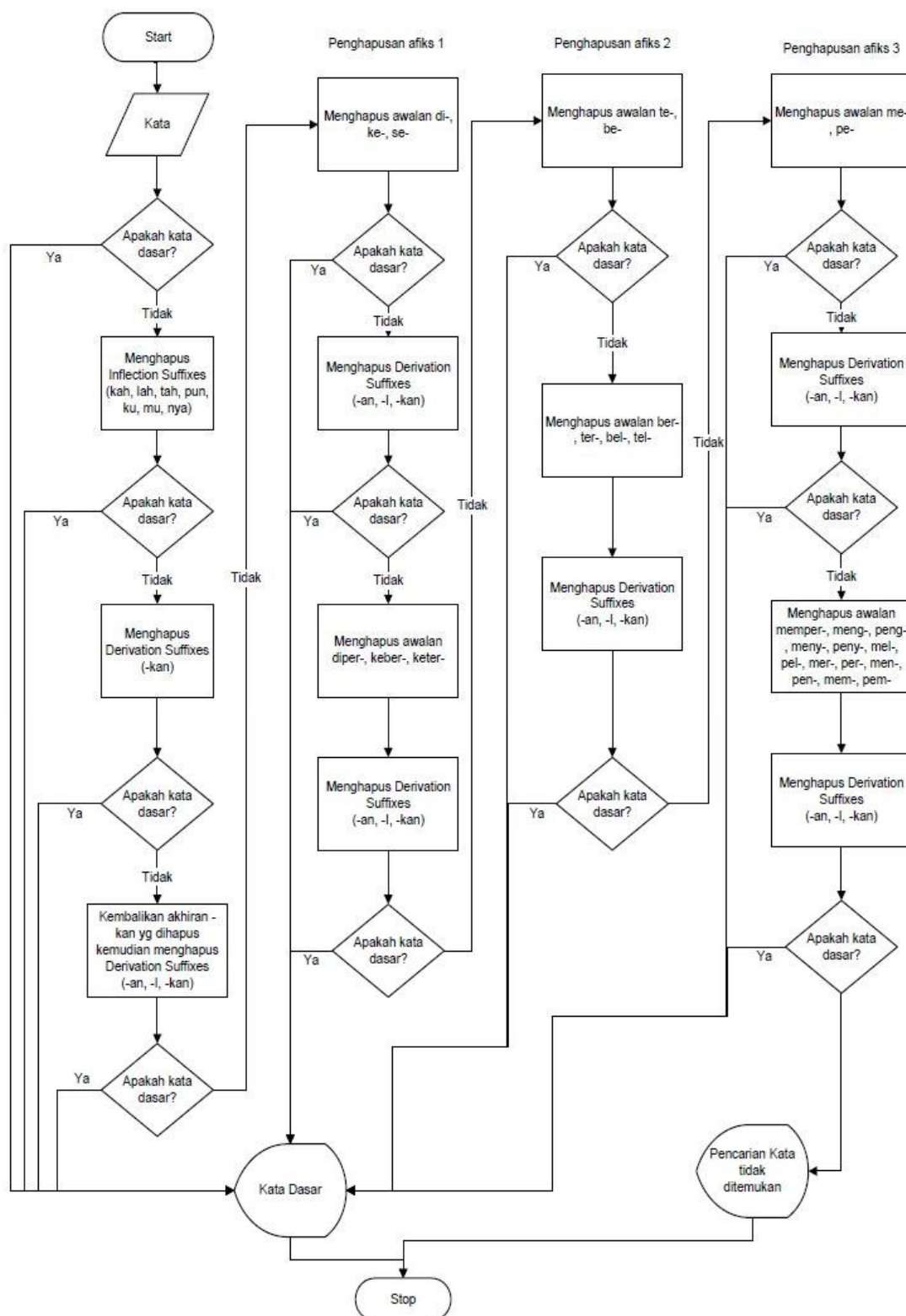
Hasil *stopword removal* dapat dilihat pada Tabel 3.7.

**Tabel 3.7. Hasil *Stopword Removal***

Kode	Input Hasil <i>Tokenizing</i>	Output <i>Stopword Removal</i>
D1	[ nonton, naruto, shippuden, seru , tayang, di ]	[ nonton, naruto, shippuden, seru, tayang, ]
D2	[ ntn, film, di, kebanyakan, kepotong ]	[ ntn, film, kebanyakan, kepotong]
D3	[ males, ah, nonton, naruto, karna ,di ulang, terus, benar, ga, seru]	[ males, nonton, naruto, ulang, terus, benar, ga, seru]
D4	[ kok, narutonya, di, ulang, lagi, bikin, males, nonton ]	[ narutonya, ulang, bikin, males, nonton ]
D5	[kenapa, banyak, scence, film ,di potong, jauh, banget, ga ,mutu]	[ banyak, scence, film , potong, jauh, banget, ga ,mutu]
D6	[teruslah, tayangkan, acara, yang, berkualitas]	[ teruslah, tayangkan, acara, berkualitas ]

#### 3.1.2.5. *Stemming*

*Stemming* adalah pengubahan kata ke bentuk kata dasar atau penghapusan imbuhan. *Stemming* disini menggunakan kamus daftar kata berimbuhan yang mempunyai kata dasarnya dengan cara membandingkan kata-kata yang ada di dalam komentar dengan daftar kamus stem. Data kata dasar didapatkan dari kamus bahasa Indonesia online dimana datanya berjumlah 29932 data. Algoritma *stemming* yang digunakan adalah algoritma *stemming* Nazief & Andriani yang telah dijelaskan pada bagian 2.5. *Flowchart* untuk proses *stemming* dapat dilihat pada Gambar 3.6.



**Gambar 3.6. *Flowchart Stemming* Nazief & Andriani**

Hasil *stemming* dapat dilihat pada Tabel 3.8

**Tabel 3.8. Hasil *Stemming***

Kode	Input Hasil <i>Stopword Removal</i>	Output <i>Stemming</i>
D1	[ nonton, naruto, shippuden, seru, tayang, ]	[ nonton, naruto, shippuden, seru, tayang, ]
D2	[ ntn, film, kebanyakan, kepotong]	[ntn, film, banyak, potong ]
D3	[ males, nonton, naruto, ulang, terus, benar, ga, seru]	[ males, nonton, naruto , ulang, terus, benar, ga, seru]
D4	[ narutonya, ulang, bikin, males, nonton ]	[ naruto, ulang, bikin, males, nonton ]
D5	[ banyak, scence, film, potong, jauh, banget, ga ,mutu]	[ banyak, scence, film, potong, jauh, banget, ga ,seru].
D6	[ teruslah, tayangkan, acara, berkualitas ]	[ terus, tayang, acara, mutu ,kualitas]

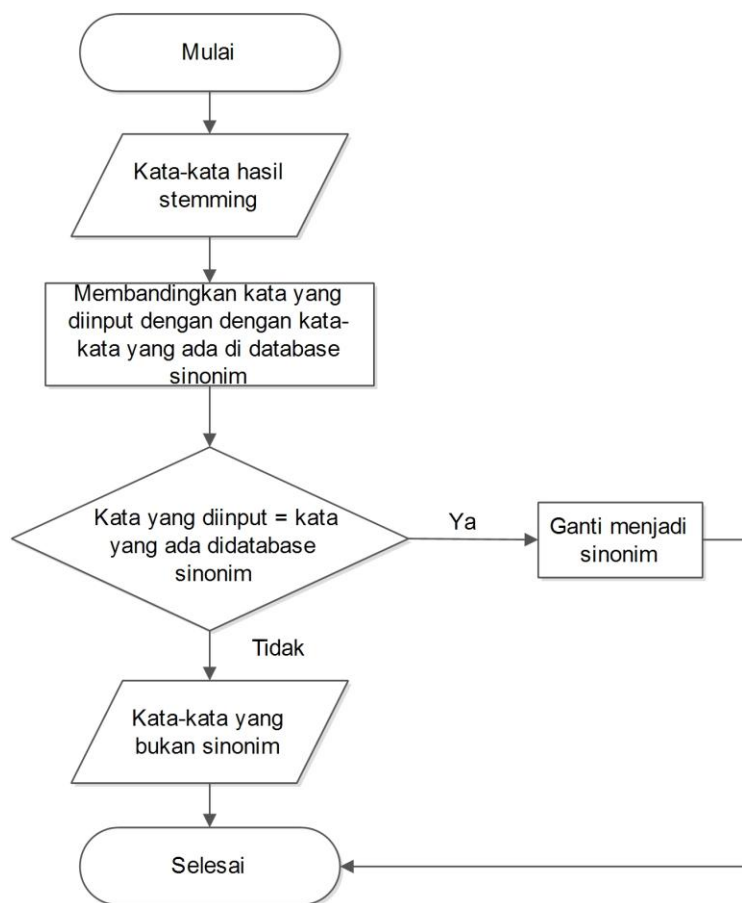
#### 3.1.2.6. *Filter redudansi*

Pada tahap ini, kumpulan kata yang telah melewati tahap *stemming* akan melalui tahap *filter redudansi*. *Filter redudansi* disini menggunakan kamus sinonim yang ada di database. Untuk mengoptimalkan perhitungan frekuensi kemunculan kata pada proses pembobotan maka diperlukan kamus sinonim untuk mengecek kata yang memiliki makna yang sama. Dengan cara kata yang ada didalam dokumen dibandingkan dengan kamus sinonim yang ada di database. Jika kata tersebut ditemukan didalam kamus sinonim maka kata tersebut diubah ke bentuk sinonimnya. Adapun contoh daftar sinonim yang digunakan dapat dilihat pada Tabel 3.9.

**Tabel 3.9. Tabel Kamus Sinonim**

Id	Daftar Kata	Sinonim
1.	ntn	nonton
2.	males	malas
3.	bosen	bosan
4.	bagus	mantap

Flowchart untuk proses *filter redudansi* dapat dilihat pada Gambar 3.7.



**Gambar 3.7. Flowchart Filter Redudansi**

Adapun penjelasan dari *flowchart* diatas adalah sebagai berikut:

1. Kata hasil *stemming* akan dibandingkan dengan daftar sinonim.
2. Dilakukan pengecekan apakah kata sama dengan daftar sinonim atau tidak.
3. Jika kata sama dengan yang ada pada daftar sinonim, maka akan diganti ke bentuk sinonimnya. Hasil proses *filter redudansi* dapat dilihat pada Tabel 3.10.

**Tabel 3.10. Hasil Filter Redudansi**

Kode	Input Hasil <i>Stemming</i>	Output <i>FILTER Redudansi</i>
D1	[ nonton, naruto, shippuden, seru, tayang, ]	[ nonton, naruto, shippuden, seru, tayang, ]
D2	[ntn, film, banyak, potong ]	[nonton, film, banyak, potong ]

D3	[ males, nonton, naruto , ulang, terus, benar, ga, seru]	[ males, nonton, naruto , ulang, terus, benar, ga, seru]
D4	[ naruto, ulang, bikin, males, nonton ]	[ naruto, ulang, bikin, males, nonton ]
D5	[ banyak, scence, film ,potong, jauh, banget, ga ,seru].	[ banyak, scence, film ,potong, jauh, banget, ga ,seru].
D6	[ terus, tayang, acara, mutu ,kualitas]	[ terus, tayang, acara, mutu ,kualitas]

### 3.1.3. Pembobotan kata

Pembobotan adalah proses pemberian nilai terhadap setiap kata yang ada pada setiap opini yang sudah melewati proses *pre-processing*. Pada penelitian ini digunakan metode TF-IDF persamaan (2.2) dan (2.3) sebagai proses pembobotan, yaitu akan dilakukan pembobotan pada tiap kata berdasarkan tingkat kepentingan tersebut didalam sekumpulan dokumen masukan. Pembobotan ini bertujuan untuk memberikan nilai kepada suatu kata yang dimana nilai dari kata tersebut akan dijadikan sebagai input pada proses klasifikasi. Detail dari metode TF-IDF ini dapat dilihat pada bagian 2.5. Adapun contoh data latih dan data uji dapat dilihat di Tabel 3.11. dan Tabel 3.12.

**Tabel 3.11. Contoh Data Latih**

Kode	Sebelum Proses <i>Pre - Processing</i>	Setelah Proses <i>Pre - Processing</i>	Kelas
D1	nonton "Naruto Shippuden" seru tayang di @Globaltvseru	[ nonton, naruto, shippuden, seru, tayang, ]	Positif
D2	ntn film di @Bigmovies_GTV kebanyakan kepotong filmnya #globaltvgakseru @Globaltvseru	[ nonton, film, banyak, potong ]	Negatif
D3	betul. Males ah nonton naruto karna di ulang2 terus. Benar2 ga seru	[ males, nonton, naruto , ulang, terus, benar, ga, seru]	Negatif
D4	katanya #jagonyafilm tapi kok	[ naruto, ulang, bikin,	Negatif

---

	naruto nya di ulang lagi :D, males, nonton ]
	bikin males nonton
	@bigmovie_gtv @globaltvseru
<b>D5</b>	@Bigmovies_GTV kenapa [ terus, tayang, acara, mutu Positif banyak scence film di potong ,kualitas] jauh banget? Ga mutu @Globaltvseru

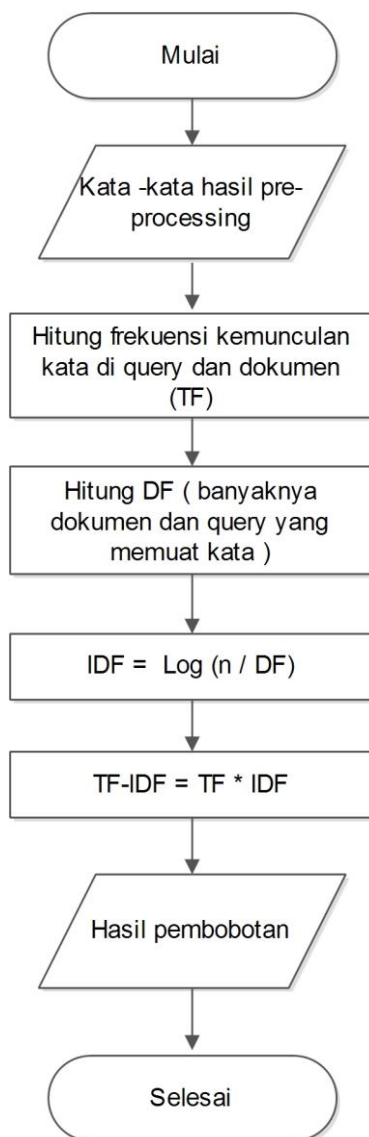
---

Tabel 3.12. Contoh Data Uji

Kode	Sebelum Proses <i>Pre – Processing</i>	Setelah Proses <i>Pre-Processing</i>	Kelas
<b>D6</b>	@Globaltvseru tayangkan acara2 yang berkualitas!!!!	teruslah [banyak, scence, film ? ,potong, jauh, banget, ga ,seru]	

---

Berdasarkan Tabel 3.11 dan Tabel 3.12, D1 sampai D6 merupakan data yang akan diuji bobot dokumennya. D1 sampai D5 merupakan data yang sudah diketahui kategorinya, sedangkan D6 data yang belum diketahui kategori sentimennya dan yang akan diuji. Untuk menentukan masuk ke kelas manakah D6. Pertama hitung bobot setiap kata dengan metode TF-IDF. *Flowchart* untuk proses TF-IDF dapat dilihat pada Gambar 3.8.



**Gambar 3.8. Flowchart TF-IDF**

Sehingga proses pembobotan kata sebagai berikut :

Proses I : menghitung jumlah frekuensi tiap kata pada tiap dokumen (TF)

**Tabel 3.13. Perhitungan TF**

Kata	TF					
	D6	D1	D2	D3	D4	D5
nonton	1	1	1	1	1	
naruto	1			1	1	
shippuden		1				
seru	1	1		1		

<b>tayang</b>	1			1		
<b>film</b>	1		1			
<b>banyak</b>	1		1			
<b>potong</b>	1		1			
<b>males</b>				1	1	
<b>ulang</b>				1	1	
<b>terus</b>				1		1
<b>benar</b>				1		
<b>ga</b>	1			1		
<b>bikin</b>					1	
<b>acara</b>						1
<b>mutu</b>						1
<b>kualitas</b>						1
<b>science</b>	1					
<b>jauh</b>	1					
<b>banget</b>	1					

Proses II : Menghitung DF

**Tabel 3.14. Perhitungan DF**

<b>Kata</b>	<b>DF</b>						<b>DF</b>
	<b>D6</b>	<b>D1</b>	<b>D2</b>	<b>D3</b>	<b>D4</b>	<b>D5</b>	
<b>nonton</b>		1	1	1	1		4
<b>naruto</b>		1		1	1		3
<b>shippuden</b>		1					1
<b>seru</b>	1	1		1			3
<b>tayang</b>		1				1	2
<b>film</b>	1		1				2
<b>banyak</b>	1		1				2
<b>potong</b>	1		1				2
<b>males</b>				1	1		2
<b>ulang</b>				1	1		2



<b>terus</b>		1	1	2
<b>benar</b>		1		1
<b>ga</b>	1	1		2
<b>bikin</b>			1	1
<b>acara</b>			1	1
<b>mutu</b>			1	1
<b>kualitas</b>			1	1
<b>science</b>	1			1
<b>jauh</b>	1			1
<b>banget</b>	1			1

Proses III : Menghitung IDF menggunakan persamaan (2.4)

**Tabel 3.15. Perhitungan IDF**

<b>Kata</b>	<b>IDF</b>
	<b>LOG (n/df)</b>
<b>nonton</b>	$^{10}\text{Log } (6/4) = 0.176$
<b>naruto</b>	$^{10}\text{Log } (6/3) = 0.301$
<b>shippuden</b>	$^{10}\text{Log } (6/1) = 0.778$
<b>seru</b>	$^{10}\text{Log } (6/3) = 0.301$
<b>tayang</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>film</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>banyak</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>potong</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>males</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>ulang</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>terus</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>benar</b>	$^{10}\text{Log } (6/1) = 0.778$
<b>ga</b>	$^{10}\text{Log } (6/2) = 0.477$
<b>bikin</b>	$^{10}\text{Log } (6/1) = 0.778$
<b>acara</b>	$^{10}\text{Log } (6/1) = 0.778$
<b>mutu</b>	$^{10}\text{Log } (6/1) = 0.778$

<b>kualitas</b>	$^{10}\text{Log}(6/1) = 0.778$
<b>science</b>	$^{10}\text{Log}(6/1) = 0.778$
<b>jauh</b>	$^{10}\text{Log}(6/1) = 0.778$
<b>banget</b>	$^{10}\text{Log}(6/1) = 0.778$

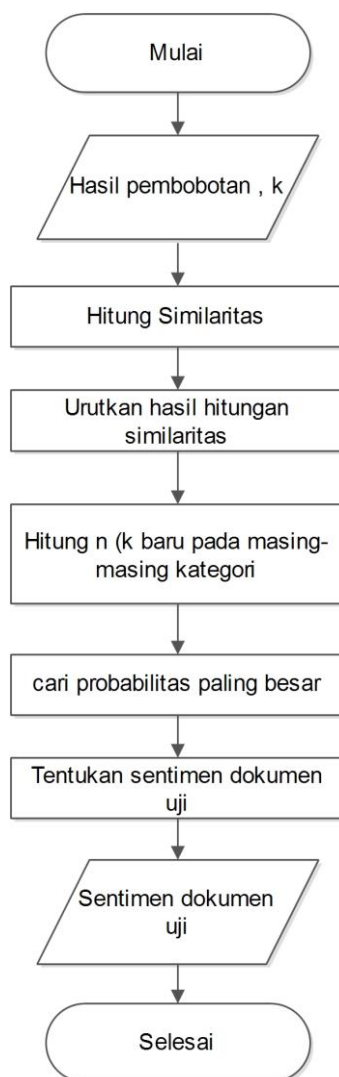
Proses IV : Menghitung TF-IDF menggunakan persamaan (2.3).

**Tabel 3.16. Perhitungan TF-IDF**

<b>Wdt = TF .IDF</b>					
<b>D6</b>	<b>D1</b>	<b>D2</b>	<b>D3</b>	<b>D4</b>	<b>D5</b>
0	0.176	0.176	0.176	0.176	0
0	0.301	0	0.301	0.301	0
0	0.778	0	0	0	0
0.301	0.301	0	0.301	0	0
0	0.477	0	0	0	0.477
0.477	0	0.477	0	0	0
0.477	0	0.477	0	0	0
0.477	0	0.477	0	0	0
0	0	0	0.477	0.477	0
0	0	0	0.477	0.477	
0	0	0	0.477	0	0.477
0	0	0	0.778	0	0
0.477	0	0	0.477	0	0
0	0	0	0	0.778	0
0	0	0	0	0	0.778
0	0	0	0	0	0.778
0	0	0	0	0	0.778
0.778	0	0	0	0	0
0.778	0	0	0	0	0
0.778	0	0	0	0	0

### 3.1.4. Klasifikasi dengan algoritma *improved k-nearest neighbor*

Setelah melalui proses pembobotan dokumen akan melalui tahap pengklasifikasian, pada proses ini akan digunakan algoritma *improved k-nearest neighbor*. Adapun langkah langkahnya adalah sebagai berikut: Menghitung similaritas antara dua dokumen menggunakan metode *Cosine Similarity* (*CosSim*). Hitung kemiripan vektor dokumen D6 dengan setiap dokumen yang telah terklasifikasi (D1, D2, D3, D4, dan D5). Kemiripan antar dokumen dapat menggunakan persamaan (2.4). Hasil dari perhitungan kemiripan diurutkan kemudian akan disimpan untuk proses selanjutnya yaitu klasifikasi dengan menggunakan *improved k-nearest neighbor*. Gambar 3.9 menunjukkan *flowchart improved k-nearest neighbor*.



**Gambar 3.9. Flowchart Improved KNN**

Proses I : hitung *similarity* vektor [dokumen] query dengan setiap dokumen yang ada, yaitu hitung hasil perkalian skalar antar dokumen. Hasilnya perkalian dari setiap dokumen dijumlahkan, sesuai dengan pembilang pada persamaan (2.4). Kemudian, hitung panjang setiap dokumen. Caranya, kuadratkan bobot setiap kata pada Tabel 3.16 dalam setiap dokumen, jumlahkan nilai kuadrat dan terakhir akarkan. Hasilnya dapat dilihat pada Tabel 3.18.

**Tabel 3.17. Hitung Perkalian Skalar**

Kata	WD6 *WDi				
	D6* D1	D6*D2	D6*D3	D6*D4	D6*D5
nonton	0	0	0	0	0
naruto	0	0	0	0	0
shippuden	0	0	0	0	0
Seru	0.0906	0	0.0906	0	0
tayang	0	0	0	0	0
Film	0	0.2276	0	0	0
banyak	0	0.2276	0	0	0
potong	0	0.2276	0	0	0
males	0	0	0	0	0
ulang	0	0	0	0	0
terus	0	0	0	0	0
benar	0	0	0	0	0
Ga	0	0	0.2276	0	0
bikin	0	0	0	0	0
acara	0	0	0	0	0
mutu	0	0	0	0	0
kualitas	0	0	0	0	0
science	0	0	0	0	0
Jauh	0	0	0	0	0
banget	0	0	0	0	0
	0.0906	0.6828	0.3182	0	0

Tabel 3.18. Hitung Panjang Vektor

Kata	Panjang Vektor					
	D6	D1	D2	D3	D4	D5
nonton	0	0.031	0.031	0.031	0.031	0
naruto	0	0.0906	0	0.0906	0.0906	0
shippuden	0	0.606	0	0	0	0
Seru	0.0906	0.0906	0	0.0906	0	0
tayang	0	0.2276	0	0	0	0.2276
Film	0.2276	0	0.2276	0	0	0
banyak	0.2276	0	0.2276	0	0	0
potong	0.2276	0	0.2276	0	0	0
Males	0	0	0	0.2276	0.2276	0
Ulang	0	0	0	0.2276	0.2276	
Terus	0	0	0	0.2276	0	0.2276
benar	0	0	0	0.606	0	0
Ga	0.2276	0	0	0.2276	0	0
Bikin	0	0	0	0	0.606	0
Acara	0	0	0	0	0	0.606
Mutu	0	0	0	0	0	0.606
kualitas	0	0	0	0	0	0.606
science	0.606	0	0	0	0	0
Jauh	0.606	0	0	0	0	0
banget	0.606	0	0	0	0	0
	2.819	1.046	0.714	1.729	1.183	2.273
	1.67	1.02	0.84	1.31	1.08	1.50

Kemudian hitung *similarity* Dokumen 6 dengan Dokumen 1, 2, 3, 4, 5.

$$\text{Cos (D6,D1)} = 0.0906 / (1.67 * 1.02) = 0.053$$

$$\text{Cos (D6,D2)} = 0.6828 / (1.67 * 0.84) = 0.486$$

$$\text{Cos (D6,D3)} = 0.3182 / (1.67 * 1.31) = 0.145$$

$$\text{Cos (D6,D4)} = 0 / (1.67 * 1.08) = 0$$

$$\text{Cos (D6,D5)} = 0 / (1.67 * 1.50) = 0$$

Hasil perhitungan adalah sebagai berikut:

D1	D2	D3	D4	D5
0.053	0.486	0.145	0	0

Proses II : urutkan hasil perhitungan similarity sebagai berikut:

1	2	3	4	5
D2	D3	D1	D4	D5

Proses III : Hitung nilai  $n$  (*k-values baru*) pada masing-masing kategori menggunakan persamaan (2.5). Hasil perhitungan  $n$  dapat dilihat pada Tabel 3.20.

**Tabel 3.19. Jumlah Data Latih**

Data Latih		
Positif	Negatif	Jumlah
2	3	5

$$n = 3 * 2 / 3 = 2 \text{ Positif}$$

$$n = 3 * 3 / 3 = 3 \text{ Negatif}$$

**Tabel 3.20. k-Baru**

Nilai K	n (k-Baru)	
	Positif	Negatif
3	2	3

Proses IV : Hitung perbandingan kemiripan dokumen uji dengan dokumen latih pada tiap kategori menggunakan persamaan (2.6).

Jumlahkan nilai similaritas sebanyak top  $n$  tetangga yang termasuk dalam suatu kategori.

$$\sum \text{CosSim positif} = D1 + D5 = 0.053 + 0 = 0.053$$

$$\sum \text{CosSim negatif} = D2 + D3 + D4 = 0.486 + 0.145 + 0 = 0.631$$

Selanjutnya hitung penjumlahan nilai similaritas sebanyak top n tetangga pada data latih.

$$\begin{aligned}\sum \text{CosSim data latih} &= D2 + D3 + D1 \\ &= 0.486 + 0.145 + 0.053 \\ &= 0.684\end{aligned}$$

Proses V : hitung nilai maksimum perbandingan antara kemiripan D6 dengan dokumen latih sebanyak top n tetangga pada suatu kategori dengan kemiripan D6 dengan dokumen latih sebanyak top n tetangga pada data latih.

$$P_{(x,cm)} \text{ positif} = 0.053 / 0.684 = 0.07$$

$$P_{(x,cm)} \text{ negatif} = 0.631 / 0.684 = 0.922$$

Nilai maksimum merupakan kategori dari D6, sehingga D6 terklasifikasi sebagai kategori negatif.

### 3.2. Perancangan Sistem

Perancangan pada sistem akan dilakukan perancangan antarmuka dari analisis sentimen pada acara televisi.

#### 3.2.1. Rancangan tampilan halaman dashboard

Rancangan halaman dashboard berisikan informasi mengenai sistem. Rancangan tampilan halaman dashboard dapat dilihat pada Gambar 3.10.

MENU		
Dashboard	[1]	
Data Latih	[2]	
Data Uji	[3]	
Pembobotan	[4]	
Klasifikasi	[5]	
Visualisasi	[6]	
Akurasi	[7]	

**Gambar 3.10. Rancangan Halaman Dashboard**

Keterangan :

1. Menu dashboard merupakan halaman awal ketika sistem dijalankan.
2. Menu data latih berfungsi untuk menampilkan data latih yang sudah di *crawling* yang tersimpan di database.
3. Menu data uji berfungsi untuk menampilkan data uji yang sudah di *crawling* yang tersimpan di database.
4. Menu pembobotan berfungsi untuk melakukan proses pembobotan.
5. Menu klasifikasi berfungsi untuk melakukan proses klasifikasi data uji.
6. Menu visualisasi berfungsi untuk menampilkan persentasi kualitas acara pada masing-masing stasiun televisi.
7. Menu akurasi berfungsi untuk menampilkan akurasi algoritma.

### 3.2.2 Rancangan halaman data latih

Halaman ini berfungsi untuk menampilkan data latih yang sudah di *crawling* dari tweet mention yang ditujukan pada masing-masing stasiun televisi. Rancangan halaman data latih dapat dilihat pada Gambar 3.11.

MENU								
	<div> <div>ANTV [2]</div> <div>Global TV [3]</div> <div>RCTI [4]</div> <div>MNCTV [5]</div> </div>							
Data Latih [1]	<table border="1"> <tr><td></td><td></td></tr> <tr><td></td><td></td></tr> <tr><td></td><td></td></tr> </table>							

**Gambar 3.11. Rancangan Halaman Data Latih**

Keterangan:

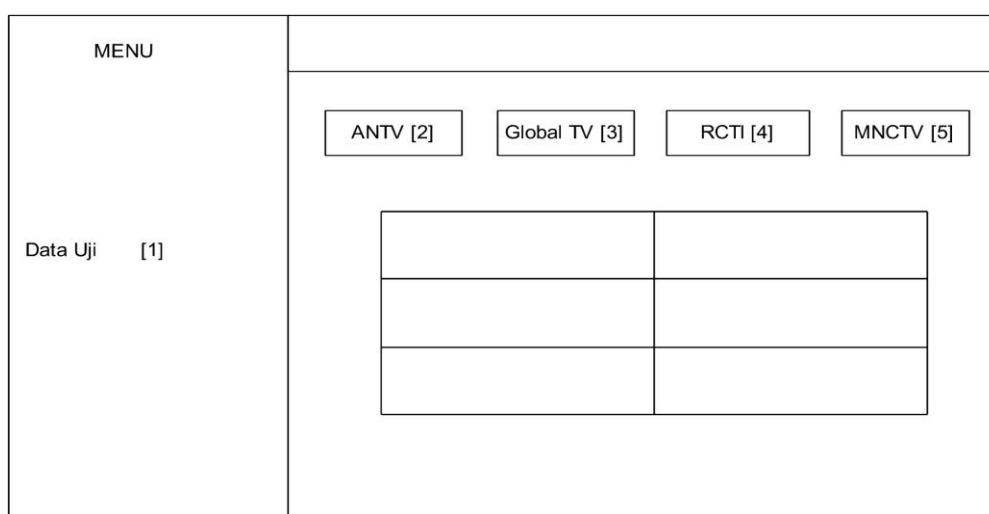
1. Menu data latih berfungsi untuk menampilkan data latih yang sebelumnya sudah di *crawling*. Pada menu ini terdapat 4 *button* yaitu “ANTV”, “Global tv”, “RCTI”, dan “MNCTV”.



2. *Button* “ANTV” berfungsi untuk mengambil data latih ANTV yang sudah di *crawling* dari database.
3. *Button* “Global TV” berfungsi untuk mengambil data latih Global TV yang sudah di *crawling* dari database.
4. *Button* “RCTI” berfungsi untuk mengambil data latih RCTI yang sudah di *crawling* dari database.
5. *Button* “MNCTV” berfungsi untuk mengambil data latih MNCTV yang sudah di *crawling* dari database.

### 3.2.3 Rancangan halaman data uji

Halaman ini berfungsi untuk menampilkan data uji yang sudah di *crawling* dari tweet mention yang ditujukan pada masing-masing stasiun televisi. Rancangan halaman data uji dapat dilihat pada Gambar 3.12.



**Gambar 3.12. Rancangan Halaman Data Uji**

Keterangan:

1. Menu data uji berfungsi untuk menampilkan data uji yang sebelumnya sudah di *crawling*. Pada menu ini terdapat 4 *button* yaitu “ANTV”, “Global tv”, “RCTI”, dan “MNCTV”.
2. *Button* “ANTV” berfungsi untuk mengambil data uji ANTV yang sudah di *crawling* dari database.

3. *Button* “Global TV” berfungsi untuk mengambil data uji Global TV yang sudah di *crawling* dari database.
4. *Button* “RCTI” berfungsi untuk mengambil data uji RCTI yang sudah di *crawling* dari database.
5. *Button* “MNCTV” berfungsi untuk mengambil data uji MNCTV yang sudah di *crawling* dari database.

### 3.2.4 Rancangan halaman pembobotan

Halaman ini berfungsi untuk menampilkan melakukan proses pembobotan pada masing-masing stasiun televisi. Rancangan halaman pembobotan dapat dilihat pada Gambar 3.13.

<p style="text-align: center;">MENU</p>          <p>Pembobotan [1]</p>	<div style="display: flex; justify-content: space-around; margin-bottom: 20px;"> <div style="border: 1px solid black; padding: 5px 10px;">ANTV [2]</div> <div style="border: 1px solid black; padding: 5px 10px;">Global TV [3]</div> <div style="border: 1px solid black; padding: 5px 10px;">RCTI [4]</div> <div style="border: 1px solid black; padding: 5px 10px;">MNCTV [5]</div> </div> <table border="1" style="margin: 0 auto; border-collapse: collapse; text-align: center;"> <tr><td style="width: 50px; height: 30px;"></td><td style="width: 50px; height: 30px;"></td></tr> <tr><td style="width: 50px; height: 30px;"></td><td style="width: 50px; height: 30px;"></td></tr> <tr><td style="width: 50px; height: 30px;"></td><td style="width: 50px; height: 30px;"></td></tr> </table>						

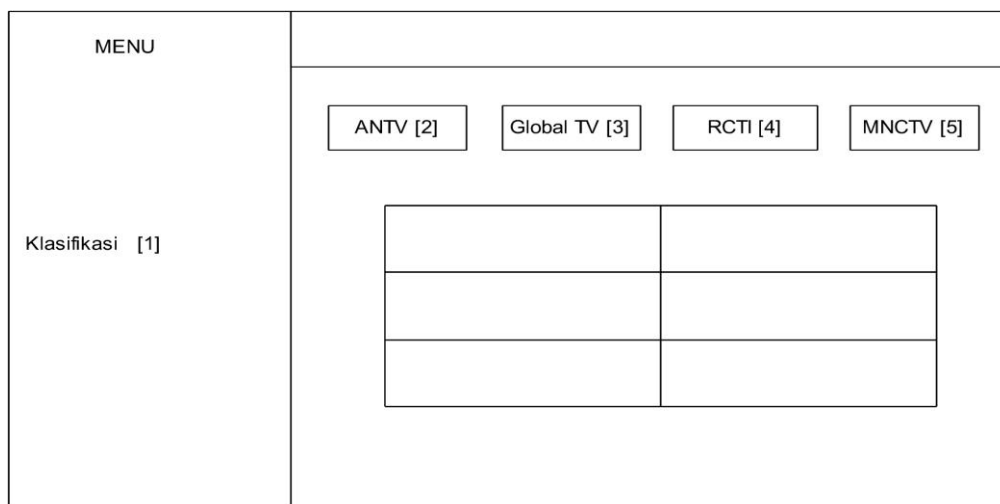
**Gambar 3.13. Rancangan Halaman Pembobotan**

Keterangan:

1. Menu pembobotan terdapat 4 *button* yaitu “ANTV”, “Global TV”, “RCTI”, dan “MNCTV”.
2. *Button* “ANTV” berfungsi untuk melakukan proses pembobotan data ANTV.
3. *Button* “Global TV” berfungsi untuk melakukan proses pembobotan data Global TV.
4. *Button* “RCTI” berfungsi untuk melakukan proses pembobotan data RCTI.
5. *Button* “MNCTV” berfungsi untuk melakukan proses pembobotan data MNCTV.

### 3.2.5 Rancangan halaman klasifikasi

Halaman ini berfungsi untuk menampilkan hasil klasifikasi sentimen. Rancangan halaman klasifikasi dapat dilihat pada Gambar 3.14.



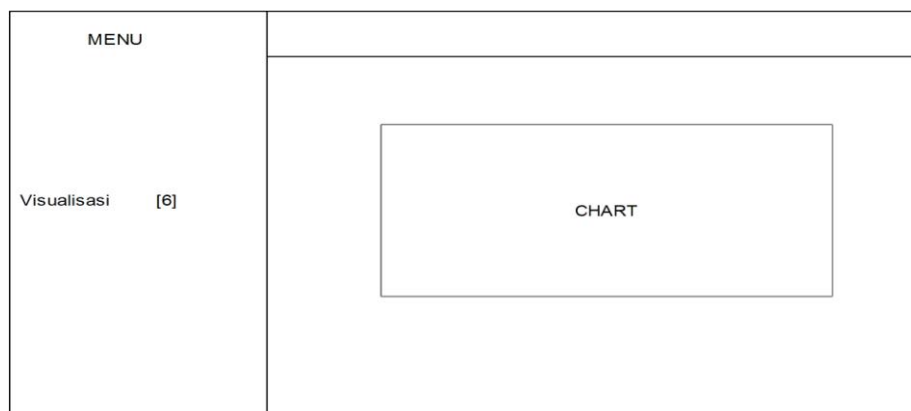
**Gambar 3.14. Rancangan Halaman Klasifikasi**

Keterangan:

1. Menu Klasifikasi terdapat 4 *button* yaitu “ANTV”, “Global TV”, “RCTI”, dan “MNCTV”.
2. *Button* “ANTV” berfungsi untuk melakukan proses klasifikasi data uji ANTV.
3. *Button* “Global TV” berfungsi untuk melakukan proses klasifikasi data uji Global TV.
4. *Button* “RCTI” berfungsi untuk melakukan proses klasifikasi data uji RCTI.
5. *Button* “MNCTV” berfungsi untuk melakukan proses klasifikasi data uji MNCTV.

### 3.2.6. Rancangan tampilan halaman visualisasi

Rancangan halaman ini berfungsi untuk menampilkan presentasi kualitas acara berdasarkan sentimen yang diperoleh pada masing-masing stasiun televisi. Rancangan tampilan halaman dashboard dapat dilihat pada Gambar 3.15.



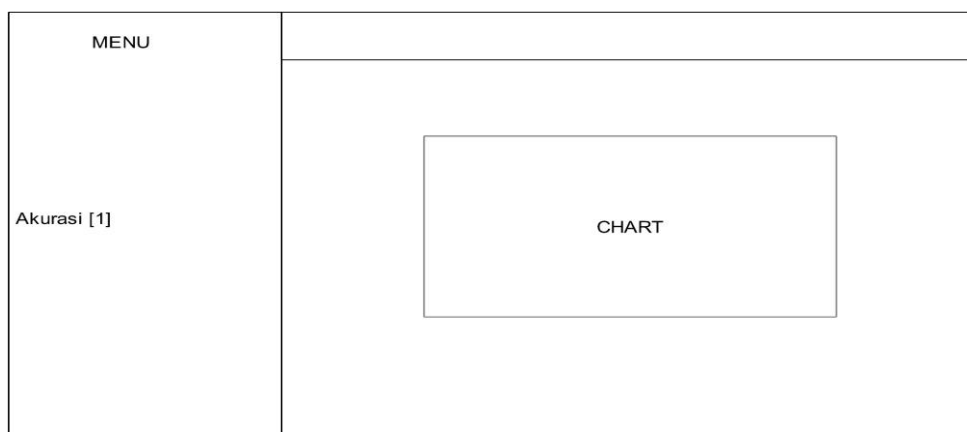
**Gambar 3.15. Rancangan Halaman Visualisasi**

Keterangan:

1. Menu visualisasi berfungsi untuk menampilkan presentasi kualitas acara pada masing-masing stasiun televisi dalam bentuk *chart*.

### *3.2.7. Rancangan tampilan halaman akurasi*

Rancangan halaman ini berfungsi untuk menampilkan akurasi sistem. Rancangan tampilan halaman akurasi dapat dilihat pada Gambar 3.16.



**Gambar 3.16. Rancangan Halaman Akurasi**

Keterangan:

1. Menu akurasi berfungsi untuk menampilkan akurasi yang diperoleh algoritma.

## **BAB 4**

### **IMPLEMENTASI DAN PENGUJIAN**

Bab ini membahas hasil yang didapatkan dari implementasi metode *improved k-nearest neighbor* untuk melakukan klasifikasi sentimen dan pengujian sistem sesuai dengan analisis dan perancangan yang ada pada bab 3.

#### **4.1 Implementasi Sistem**

Berdasarkan analisis dan perancangan yang telah dibuat, sistem ini akan di buat menggunakan bahasa pemrograman php dan javascript.

##### *4.1.1 Spesifikasi perangkat keras dan perangkat lunak*

Adapun spesifikasi perangkat keras yang digunakan dalam pembuatan sistem ini adalah sebagai berikut:

1. Processor Inter(R) Core (TM) i3 CPU M330 @ 2.13GHz
2. Memory (RAM) 2.00 GB
3. Kapasitas Hardisk 250 GB

Selain perangkat keras, sistem juga dibuat dalam lingkungan spesifikasi perangkat lunak sebagai berikut:

1. Sistem operasi Windows 7 Ultimate
2. Composer
3. Php dan javascript
4. Database MySql

##### *4.1.2 Tampilan halaman dashboard*

Tampilan dashboard merupakan halaman awal ketika sistem dijalankan. Tampilan halaman Dashboard dapat dilihat pada Gambar 4.1



**Gambar 4.1. Tampilan Halaman Dashboard**

#### 4.1.3. Tampilan halaman data latih

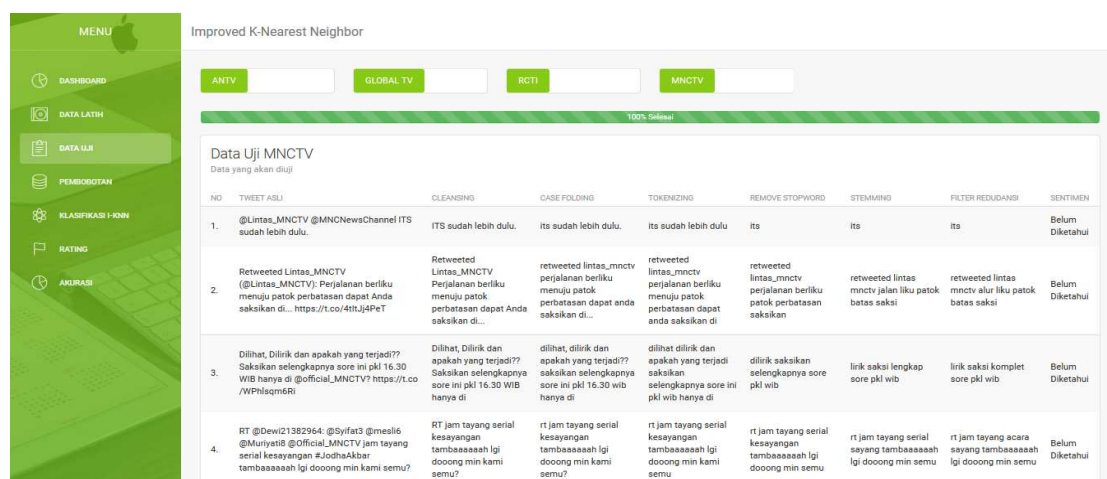
Pada halaman ini terdapat empat tombol diantaranya, tombol “ANTV” ,”Global TV”, “RCTI” dan “MNCTV” .Dimana tombol tersebut digunakan untuk mengambil data latih yang sudah dicrawling sebelumnya dari tweet mention yang ditujukan pada masing-masing stasiun tv. Data tersebut sebelumnya sudah disimpan di database. Pada saat tombol di klik semua tahapan preprocessing dilakukan secara bersamaan. Halaman data latih dapat dilihat pada Gambar 4.2.

MENU		Improved K-Nearest Neighbor							
		ANTV	GLOBAL TV	RCTI	MNCTV	100% Selesai			
Data Latih MNCTV Database Data Latih Yang Berasal dari API Twitter									
NO	TWEET ASLI	CLEANSING	CASE FOLDING	TOKENIZING	REMOVE STOPWORD	STEMMING	FILTER REDUNDANSI	SENTIMEN	
1.	@Official_MNCTV Mantap tu	Mantap tu	mantap tu	mantap tu	mantap	mantap	bagus	Positif :)	
2.	@agnes_korea22 @Official_MNCTV sama say di primetime gak ada yg seru m moga ratingnya nampol. gosip di ig pak otis https://t.co/5hVyoYh1bu	sama say di primetime gak ada yg seru m moga ratingnya nampol. gosip di ig pak otis	sama say di primetime gak ada yg seru m moga ratingnya nampol. gosip di ig pak otis	sama say di primetime gak ada yg seru moga ratingnya nampol gosip di ig pak otis	primetime gak ada yg seru moga ratingnya nampol gosip di ig otis	primetime gak ada yg seru moga ratingnya nampol gosip di otis	primetime tidak ada yg seru moga ratingnya nampol gosip di otis	Negatif :(	
3.	Min @Official_MNCTV tayangin Film layar kaca Dil To Pagel Hai dong min, keren tuh film nya ?	Min tayangin Film layar kaca Dil To Pagel Hai dong min, keren tuh film nya ?	min tayangin film layar kaca dil to pagel hai dong min, keren tuh film nya ?	min tayangin film layar kaca dil to pagel hai dong min keren tuh film nya	min tayangin film layar kaca dil to pagel hai min keren tuh film nya	min tayangin film layar kaca dil to pagel hai min keren tuh film nya	min tayang film layar kaca dil to pagel hai min keren tuh film nya	Positif :)	
4.	@Official_MNCTV Drama favorit mah #MahabharataRajaBaru	Drama favorit mah	drama favorit mah	drama favorit mah	drama favorit	drama favorit	drama favorit	Positif :)	

**Gambar 4.2. Tampilan Halaman Data Latih**

#### 4.1.4. Tampilan halaman data uji

Pada halaman ini terdapat empat tombol diantaranya, tombol “ANTV”, “Global TV”, “RCTI” dan “MNCTV”. Dimana tombol tersebut digunakan untuk mengambil data latih yang sudah dicrawling sebelumnya dari tweet mention yang ditujukan pada masing-masing stasiun tv. Data tersebut sebelumnya sudah disimpan di database. Halaman data uji dapat dilihat pada Gambar 4.3.

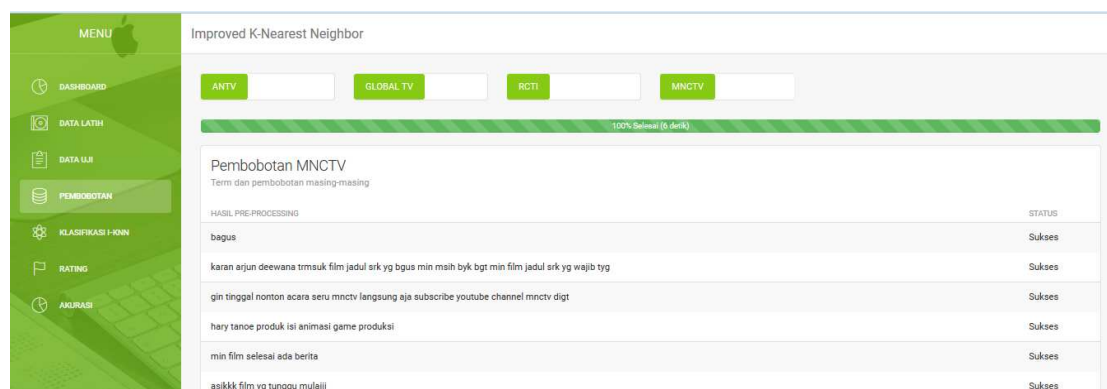


NO	TWEET ASLI	CLEANSING	CASE FOLDING	TOKENIZING	REMOVE STOPWORD	STEMMING	FILTER REDUNDANSI	SENTIMEN
1.	@Lintas_MNCTV @MNCNewsChannel ITS sudah lebih dulu.	ITS sudah lebih dulu.	its sudah lebih dulu.	its sudah lebih dulu	its	its	its	Belum Diketahui
2.	Retweeted Lintas_MNCTV (@Lintas_MNCTV): Perjalanan berliku menuju patok perbatasan dapat anda saksikan di... https://t.co/4tUj4PeT	Retweeted Lintas_MNCTV Perjalanan berliku menuju patok perbatasan dapat anda saksikan di...	retweeted lintas_mnctv perjalanan berliku menuju patok perbatasan dapat anda saksikan di...	retweeted lintas_mnctv perjalanan berliku menuju patok perbatasan dapat anda saksikan di	retweeted lintas_mnctv perjalanan berliku patok perbatasan saksikan	retweeted lintas mnctv jalan liku patok batas saksi	retweeted lintas mnctv alur liku patok batas saksi	Belum Diketahui
3.	Dilihat, Dilirik dan apakah yang terjadi?? Saksikan selengkapnya sore ini pkl 16.30 WIB hanya di @Official_MNCTV? https://t.co/WP9lsqm6Rl	Dilihat, Dilirik dan apakah yang terjadi?? Saksikan selengkapnya sore ini pkl 16.30 WIB hanya di	dilihat, dilirik dan apakah yang terjadi?? saksikan selengkapnya sore ini pkl 16.30 wib hanya di	dilihat dilirik dan apakah yang terjadi?? saksikan selengkapnya sore ini pkl wib hanya di	dirlik saksikan selengkapnya sore pkl wib	lirik saksi lengkap sore pkl wib	lirik saksi komplet sore pkl wib	Belum Diketahui
4.	RT @Dewi21382964: @Syfat3 @mesil6 @Muriyati8 @Official_MNCTV jam tayang serial kesayangan #JodhaAkar tambaaaaah lgi dooong min kami semu?	RT jam tayang serial kesayangan tambaaaaah lgi dooong min kami semu?	rt jam tayang serial kesayangan tambaaaaah lgi dooong min kami semu?	rt jam tayang serial kesayangan tambaaaaah lgi dooong min kami semu	rt jam tayang serial kesayangan tambaaaaah lgi dooong min semu	rt jam tayang serial sayang tambaaaaah lgi dooong min semu	rt jam tayang acara sayang tambaaaaah lgi dooong min semu	Belum Diketahui

**Gambar 4.3. Tampilan Halaman Data Uji**

#### 4.1.6. Tampilan halaman pembobotan

Halaman ini bertujuan untuk melakukan proses pembobotan terhadap data uji dan latih. Pada halaman ini terdapat empat tombol diantaranya, “tombol “ANTV”, “GlobalTV”, “RCTI” dan “MNCTV”. Pada saat tombol tersebut diklik proses pembobotan dilakukan. Halaman pembobotan dapat dilihat pada Gambar 4.4.



HASIL PRE-PROCESSING	STATUS
bagus	Sukses
karan arjun deewana trmsuk film judul srk yg bgus min masih byk bgt min film judul srk yg wajib tyg	Sukses
gin tinggal nonton acara seru mnctv langsung aja subscribe youtube channel mnctv digt	Sukses
hary tanoe produk isi animasi game produksi	Sukses
min film selesai ada berita	Sukses
asikkk film yg tunggu mulaili	Sukses

**Gambar 4.4 Tampilan Halaman Pembobotan**

#### 4.1.7. Tampilan halaman klasifikasi

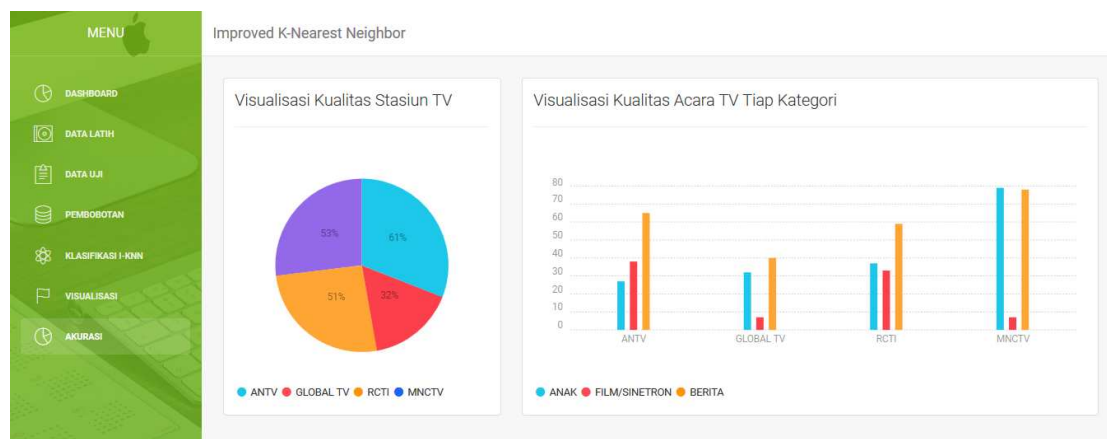
Halaman ini bertujuan untuk melakukan proses klasifikasi terhadap data uji. Pada halaman ini terdapat empat tombol diantaranya , “tombol “ANTV”, “GlobalTV”, “RCTI” dan “MNCTV”. Pada saat tombol tersebut diklik maka proses klasifikasi dilakukan. Halaman klasifikasi dapat dilihat pada Gambar 4.5.

NO	TWEET ASLI	HASIL PRE-PROCESSING	KLASIFIKASI SISTEM	KLASIFIKASI MANUAL
1.	@Lintas_MNCTV Perjalanan berliku menuju patok perbatasan dapat Anda saksikan di... <a href="https://t.co/4tUJ4PeT">https://t.co/4tUJ4PeT</a>	jalan liku patok batas saksi	Netral :)	Netral :)
2.	RT @Lintas_MNCTV: @Hary_Tanoe mendampingi penandatanganan MoU @MNCKapital dengan Bank Banten #LintasSiang <a href="https://t.co/utZHQOfusm">https://t.co/utZHQOfusm</a>	dampingi penandatanganan mou bank banten	Netral :)	Netral :)
3.	@Official_MNCTV Saya suka...saya suka...	sukesaya suka	Positif :)	Positif :)
4.	Lagi nonton film animasi Upin Ipin dan Kawann-kawan di @Official_MNCTV. Tonton ya !!! <a href="https://t.co/M6iyOsH01L">https://t.co/M6iyOsH01L</a>	nonton film animasi upin ipin kawanikawan nonton	Netral :)	Netral :)
5.	@Official_MNCTV tolong scene nasihat" wejangan Krishna jng d cut justru itu yg mngandung pengajaran #MahabharataRajaBaru	tolong scene nasihat wejang krishna jangan potong mngandung didik	Positif :)	Negatif :(

**Gambar 4.5 Tampilan Halaman Klasifikasi**

#### 4.1.8. Tampilan halaman visualisasi

Halaman ini akan menampilkan persentasi kualitas perkategori acara berdasarkan hasil klasifikasi sentimen pada empat stasiun tv nasional yaitu ANTV, GlobalTV, RCTI dan MNCTV .Halaman visualisasi dapat dilihat pada Gambar 4.6.

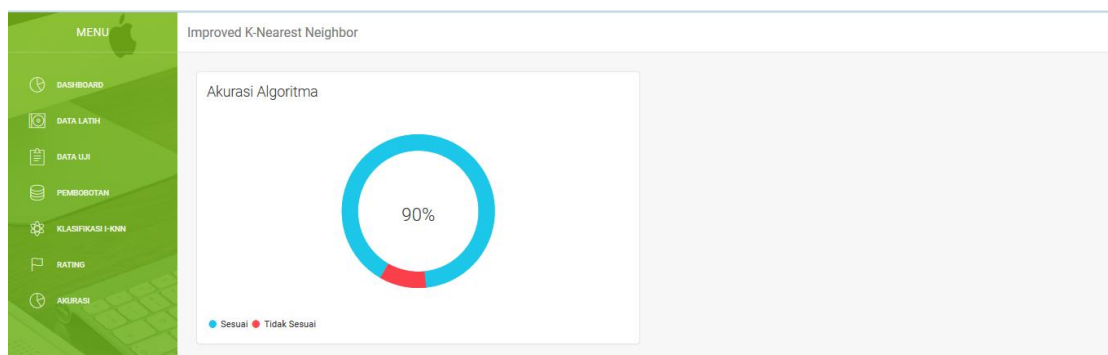


**Gambar 4.6 Tampilan Halaman Visualisasi**



#### 4.1.9. Tampilan halaman akurasi

Halaman ini berfungsi untuk menampilkan akurasi yang diperoleh sistem dengan menggunakan metode *improved k-nearest neighbor*.



**Gambar 4.8 Tampilan Halaman Akurasi**

## 4.2 Pengujian Sistem

Pada tahap ini akan dilakukan pengujian sistem untuk mengetahui kemampuan dari sistem. Untuk mengetahui pengaruh proporsi data latih pada setiap kategori dan nilai  $k$  terhadap efektivitas sistem klasifikasi maka dilakukan 10 kali uji coba. Dimana jumlah data uji yang digunakan sebanyak 1200 data uji dengan nilai  $k$  yang berbeda.

**Table 4.1. Porposi Data Latih**

Porposi Data Latih			
Positif	Negatif	Netral	Jumlah
1385	700	724	2800

Pada Tabel 4.1. menunjukan perbedaan proporsi jumlah data latih yang digunakan dalam melakukan pengklasifikasian.

**Tabel 4.2. Pengujian Sistem Berdasarkan Nilai  $k$**

k	n ( $k$ -values baru)			Akurasi
	Positif	Negatif	Netral	
5	2	1	1	89%
10	5	2	2	90%

15	7	4	4	88%
20	10	5	5	88%
25	12	6	6	88%
30	15	8	8	87%
35	17	9	9	87%
40	20	10	10	87%
50	25	13	13	86%
60	30	15	15	85%

Dari pengujian sistem dengan menggunakan 1200 data uji pada Tabel 4.2 dapat dilihat nilai akurasi rata-rata dari 10 kali pengujian. Dimana 2800 data latih dengan proporsi jumlah 1385 positif, 700 negatif dan 724 netral. Dari tabel diatas dapat diketahui bahwa akurasi tertinggi berada pada nilai k=10 yaitu 90% dan terendah pada nilai k=60 yaitu 85%

**Tabel 4.3. Hasil Pengujian Sistem**

No	Data Uji	Klasifikasi	
		Sistem	Manual
1.	@whatsonANTV wajib di tonton Karena Alur ceritanya pasti bikin Baper dan selalu bikin penasaran, Jadi selalu tak in <del>in</del> <a href="https://t.co/Fql8XWcyYv">https://t.co/Fql8XWcyYv</a>	Positif	Positif
2.	Arigatou buat yang udah nonton @merlynasun di @LensorANTV edisi pagi ini, wassalamu'alaikum dan BRAVO OLAHRAGA!!! <a href="https://t.co/wiQv6VHCSP">https://t.co/wiQv6VHCSP</a>	Positif	Positif
3.	@whatsonANTV Apa ANTV kehabisan stok film Indonesia sampe impor dari india? Ternyata bukan hanya pangan, film sj impor.	Negatif	Negatif
4.	Tontonan @SeriesANTV Hari Ini #MASHAANDTHEBEAR (Episode Terbaru) @whatsonANTV :).	Netral	Netral

---

5.	Agak prihatin nih liat Shiva yg ditayangin @whatsonANTV soalnya bnyak adegan kekerasan, gk cocok jadi tontonannya anak kecil,,	Positif	Negatif
6.	Isi liburanmu dengan movie petualangan seru satu ini! "SPY KID" Pkl 14.30 WIB @Globaltvseru #JagonyaFilm <a href="https://t.co/4pVcG0cKtF">https://t.co/4pVcG0cKtF</a>	Netral	Netral
7.	@Globaltvseru ah elaaah lg nonton film tommorow land iklannya banyak bgt mbok ya tolong yg sesuai masak filmnya 30detik iklannya 5 menit	Positif	Negatif
8.	@Globaltvseru awas kau ya globaltv sial yg trlalu banyak nyensor dulu seru skrng udh gk trlalu prcy lg sensor BAGIAN WANITA AJA JGN YG LAIN	Negatif	Negatif
9.	@Globaltvseru Malah gak seru,kalau narutonya mainnya hannya sebentar. Apalagi dari dulu hanya sampai film naruto saja. Malah diulang2 terus	Negatif	Negatif
10.	Waaauuw Inuyasha di global tv keren euuyy lagunya ngangenin banget I want to change the world ? Terimakasih @globaltvseru	Positif	Positif
11.	@OfficialRCTI om mau kartun jepang dong om. Bosen sinetron	Negatif	Negatif
12.	Saksikan Big Office Indonesia Malam ini #Pupus Jam:01.45wib @officialRCTI @DonitaCantik @marcel_fc <a href="https://t.co/toKegXOzTx">https://t.co/toKegXOzTx</a>	Netral	Netral
13.	RT @MNCNewsChannel: Retweeted SeputarIndonesiaRCTI (@SindoRCTI): Ketiga berita tersebut dapat anda saksikan di #SindoSiang di @OfficialRCTI	Positif	Netral
14.	@OfficialRCTI Ini filem kesukaan saya.. Stay trus buat doraemon	Positif	Positif
15.	Iklan sampe 10 menit belum kelar... Luar biasa ya #DuniaTerbalik #rcti @OfficialRCTI ... Suruh nonton sinetron apa iklan ini?	Negatif	Negatif

---

16.	ai Hai! Merapat yuk ke @Official_MNCTV banyak keseruan deh sama si kembar #UpinIpin <a href="https://t.co/WAptULnLZE">https://t.co/WAptULnLZE</a>	Netral	Netral
17.	Saksikan kisah perjuangan Wa Opah, #PahlawanTenunIndonesia dari Masalili dalam @Pahlawan_MNC Minggu (16/7) pk1 15:30 <a href="https://t.co/KNw4knMYj6">https://t.co/KNw4knMYj6</a>	Netral	Netral
18.	@Official_MNCTV Saya sekeluarga tak ada bosan nonton ni kartun,Best punya????tq MNCtv.klo bisa yng Boboiboy di tyngkn lg stiap harinya.	Positif	Positif
...	...	..	..
1200	Mantap jiwa dubbingnya keren. Walau tayangnya malem kalo nonton Jodha Akbar mata tetep meleak @Official_MNCTV #JodhaAkbarMNCTV	Positif	Positif

Berdasarkan Tabel 4.2 dapat diketahui bahwa akurasi tertinggi diperoleh dengan nilai  $k = 10$ . Pada Tabel 4.3 terdiri dari 1200 data uji dengan nilai  $k = 10$ . Untuk menghitung akurasi yang diperoleh sistem maka digunakan persamaan (4.1).

$$\text{Akurasi} = \frac{\text{Jumlah data yang benar}}{\text{Jumlah data keseluruhan}} \times 100\% \quad (4.1)$$

Dari tabel diatas maka dapat dihitung akurasi sistem dengan menggunakan persamaan (4.1)

$$\begin{aligned} \text{Akurasi} &= \frac{1080}{1200} \times 100\% \\ &= 90\% \end{aligned}$$

Berdasarkan persamaan (4.1) maka dapat diketahui bahwa akurasi sistem untuk klasifikasi sentimen menggunakan metode *Improved K-Nearest Neighbor* sebesar 90%.

## **BAB 5**

### **KESIMPULAN DAN SARAN**

Bab ini akan membahas tentang kesimpulan dari penerapan metode yang diajukan untuk klasifikasi sentimen dan saran untuk pengembangan yang dapat dilakukan pada penelitian selanjutnya.

#### **5.1. Kesimpulan**

Adapun kesimpulan yang diperoleh dari penelitian ini, yaitu:

1. Metode *Improved K-Nearest Neighbor* dapat digunakan untuk klasifikasi sentimen berbahasa Indonesia menjadi 3 kategori yaitu sentimen positif, negatif, dan netral.
2. Dalam mengatasi permasalahan penurunan akurasi, metode *Improved K-Nearest Neighbor* setiap kategorinya memiliki *k-values* yang berbeda. Dimana *k-values* disesuaikan dengan besar-kecilnya jumlah data latih pada setiap kategori. Sehingga ketika *k-values* ditetapkan semakin tinggi, hasil akurasi tidak terpengaruh oleh kategori yang memiliki jumlah data latih yang lebih besar.
3. Pengujian analisis sentimen berbahasa Indonesia dengan metode *Improved K-Nearest Neighbor* menghasilkan akurasi tertinggi dengan nilai  $k=10$  sebesar 90%.

#### **5.2. Saran**

Adapun saran peneliti untuk penelitian selanjutnya adalah sistem diharapkan dapat mengatasi kesalahan penulisan sehingga didapatkan hasil yang lebih optimal

## DAFTAR PUSTAKA

- Alhumoud, S. O., & Altuwaijri, M. I. 2015. Survey on Arabic sentiment analysis in Twitter. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* 9(1): 364-368.
- Altrabsheh, N., Cocea, M. & Fallahkhair, S. 2014. Sentiment analysis: towards a tool for analysing real-time students feedback. *Proceedings 26<sup>th</sup> International Conference on Tools with Artificial Intelligence*, pp. 419-423.
- Baoli, Li., Shiwen, Yu., dan Qin, Lu. 2003. An Improved k-Nearest Neighbors for Text Categorization. *Proceedings of the 20th International Conference of Computer Processing of Oriental Language*.
- Chen, I.-L., Pai, K. C., Kuo, B. c., & Li, C. H. 2010 . An adaptive rule based on unknown pattern for improving k-nearest neighbor classifier. *International Conference on Technologies and Applications of Artificial Intelligence*, pp. 331-334.
- Cvijikj, I.P. & Michahelles, F., 2011. Understanding social media marketing: a case study on topics, categories and sentiment on a facebook brand page.
- Darma, I. M. 2017. Penerapan Sentimen Analisis Acara Televisi Pada Twitter Menggunakan Support Vector Machine dan Algoritma Genetika. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , 998-1007.
- Farber. 2012. *Twitter hits 400 million tweets per day, mostly mobile*. <http://www.cnet.com/news/twitter-hits-400-million-tweets-per-day-mostly-mobile/>. (13 januari 2017).
- Feldman, R & Sanger, J. 2007. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press : New York.
- Go, A., Huang, L., & Bhayani, R. 2009. Twitter sentiment analysis. *Final Projects from CS224N* , 17.

- Go, A., Bhayani, R., & Huang, L. 2009. Twitter sentiment classification using distant supervision.
- Govindarajan, M. 2013. Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm. *International Journal of Advanced Computer Research* (ISSN (print): 2249-7277. (Online): 2277-7970 (30 Agustus 2017).
- Harlian, Milka. 2006. *Machine Learning Text Kategorization*. Austin : University of Texas.
- Hearst, Marti. 2003. *What Is Text Mining*. SIMS, UC Berkeley. <http://www.sims.berkeley.edu/~hearst/text.mining.html> .(5 Juni 2017).
- Jose, R., & Chooralil, V. S. 2015. Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach. *International Conference on Data Mining and Advanced Computing*, pp. 64-67.
- Koncz, P. & Paralic, J., 2011. An approach to feature selection for sentiment analysis. *International Conference on Intelligent Engineering Systems*, pp.357–362.
- Li, M., Ch'ng, E. & See, S. 2016. The New Eye Of Smart City: Novel Citizen Sentiment Analysis In Twitter. *IEEE*, pp. 557-562.
- Liu, Bing. 2012. *Sentiment Analysis And Opinion Mining*. Chicago: Morgan & Claypool Publisher. <http://www.dcc.ufrj.br/~valeriab/DTMSentiment-AnalysisAndOpinionMining-BingLiu.pdf>. ( 13 Januari 2017).
- Mandal, S. & Gupta, S. 2017. A lexicon-Based Text Classification Model to Analyse and Predict Sentiments from Online Review. *IEEE*.
- Medhat, W., Hassan, A. & Korashy, H., 2014. Sentiment analysis algorithms and applications. *Ain Shams Engineering Journal* 5(4): 1093–1113.
- Nargund, K., & S, N. 2016. Public health allergy surveillance using microblogs. *Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1429-1433.

- Nazief, B. A. A. & Adriani, M. (1996), Con\_xstripping: Approach to Stemming Algorithm for Bahasa Indonesia. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta
- Nasukawa, T. & Yi, J., 2003. Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture*. pp. 70-77.
- Pang, B., Lee, L., & Vithyanathan, S. (2002). Thumbs Up ? SentimentClassification Using Machine Learning Techniques. Dalam Proceedings of The ACL-02 conference on Empirical methods in natural language processing, pp. 79-86. Stroudsburg: Association for computational Linguistic.
- Qiao, Y.-L., Pan, J. S., & He Sun, S. 2004. Improved K Nearest Neighbor Classification Algorithm. *Asia-Pacific Conference on Circuits and Systems*, pp. 1101-1104.
- Razzaq, M.A., Qamar, A.M & Bilal, H.S.M .2014. Prediction and Analysis of Pakistan Election 2013 based on Sentiment Analysis. *Proceedings International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 700-703.
- Stylios, G, Christodoulakis, D, Besharat, J, Vonitsanou, M, Kotrotsos, I, Koumpouri, A & Stamou, S. 2010. Public opinion mining for governmental decisions. *Electronic Journal of e-Government* 8 (2): 203-214.
- Tala, F. Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, Institute for Logic, Language and Computation, Universiteit van Amsterdam.
- T.Joachims. 1997. A probabilistic analysis of the Rochhio algorithm with TFIDF for text categorization. *Proceedings of the fourteenth international conference on machine learning*.
- Wang, H., Can, D., & Kazemzadeh, A. 2012. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 115-120.



- Yazdavar, A. H., Ebrahimi, M., & Salim, N. 2016. Fuzzy based implicit sentiment analysis on quantitative sentences. *Journal of Soft Computing and Decision Support Systems* 3 (4): 7-18.
- Yong Z, Youwen L, Xhixion X. 2009. An Improved kNN Text Classification Algotihm based on Clustering. *Journal of Computers* 4(3).