

# TUGAS AKHIR

## ANALISIS TEKS BERITA PEMILU DENGAN TEKNOLOGI BIG DATA



Michael Alexander Joeshendrik

NPM: 6181901014

PROGRAM STUDI INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS  
UNIVERSITAS KATOLIK PARAHYANGAN  
2024

**FINAL PROJECT**

**ELECTION NEWS TEXT ANALYSIS WITH BIG DATA  
TECHNOLOGY**



**Michael Alexander Joeshendrik**

**NPM: 6181901014**

**DEPARTMENT OF INFORMATICS  
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES  
PARAHYANGAN CATHOLIC UNIVERSITY  
2024**

**LEMBAR PENGESAHAN**

**ANALISIS TEKS BERITA PEMILU DENGAN TEKNOLOGI BIG  
DATA**

**Michael Alexander Joeshendrik**

**NPM: 6181901014**

**Bandung, 23 Januari 2024**

**Menyetujui,**

**Pembimbing**

**Digitally signed  
by Veronica Sri  
Moertini**

**Dr. Veronica Sri Moertini**

**Ketua Tim Penguji  
Digitally signed  
by Raymond  
Chandra Putra**

**Raymond Chandra Putra, M.T.**

**Anggota Tim Penguji  
Digitally signed  
by Cecilia Esti  
Nugraheni**

**Dr.rer.nat. Cecilia Esti Nugraheni**

**Mengetahui,**

**Ketua Program Studi**

**Digitally signed  
by Lionov**

**Lionov, Ph.D.**

## PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa tugas akhir dengan judul:

### **ANALISIS TEKS BERITA PEMILU DENGAN TEKNOLOGI BIG DATA**

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,  
Tanggal 23 Januari 2024



Michael Alexander Joeshendrik  
NPM: 6181901014

## ABSTRAK

Pemilihan Umum 2024 saat ini menjadi peristiwa yang sangat dinantikan masyarakat Indonesia. Banyak media berita mulai mengeluarkan artikel mengenai kandidat presiden, kandidat wakil presiden, dan partai pemilu Indonesia selanjutnya. Artikel berita tersedia melalui berbagai laman portal yang ada di Indonesia, salah satunya Detik.com yang menjadi salah satu media berita yang orang pakai untuk mengetahui berita terkini. Dengan banyaknya jumlah berita yang ada, serta cepatnya jumlah berita tersebut dikeluarkan, publik dapat kebingungan untuk menilai popularitas dari calon presiden, wakil presiden, dan partai yang didukungnya jika melihat secara satu per satu.

Penelitian ini bertujuan untuk mengatasi masalah yang dibahas pada paragraf pertama dengan menganalisis teks berita terkait Pemilu 2024 menggunakan teknologi Big Data, khususnya dengan memanfaatkan sistem terdistribusi Hadoop dan Spark *Cluster*. Penelitian ini dilakukan dengan langkah-langkah sebagai berikut yaitu menerapkan Tahapan *Data Science*: Implementasi tahapan *data science*, mulai dari pengumpulan, pembersihan, eksplorasi, analisis, hingga visualisasi data teks yang berkaitan dengan Pemilihan Umum 2024 dari portal berita detik.com. Fokus analisis adalah untuk mengukur popularitas Calon Presiden/Wakil Presiden (Capres/Cawapres) dan partai politik baik secara sentimen maupun statistik.

Melakukan eksplorasi dan penyiapan data berita dari detik.com dengan menggunakan query yang mencakup sejumlah Capres, Cawapres dan partai politik yang relevan dengan Pemilu 2024. Analisis dilakukan terhadap data berita dengan tujuan mengukur popularitas Capres/Cawapres dan partai politik. Dalam proses ini, terjadi tiga iterasi: pertama dengan teknik *preprocessing* data per dokumen berita dan penggunaan model *machine learning*, kedua dengan pendekatan per kalimat untuk analisis sentimen menggunakan model *machine learning*, dan ketiga dengan penambahan fitur berupa judul berita untuk meningkatkan akurasi analisis sentimen.

Hasil analisis disajikan dengan metrik statistik dan sentimen yang disusun sedemikian rupa agar menarik dan mudah dipahami. Iterasi ketiga menggunakan polaritas kelas yang didapat dari judul dan teks berita sebagai bobot untuk menciptakan sentimen gabungan yang merepresentasikan berita secara lebih komprehensif. Hasil penelitian ini memberikan wawasan dalam pemahaman popularitas dan sentimen terkait Pemilu 2024 dalam bentuk perangkat lunak berbasis website berupa *Dashboard*, serta menunjukkan upaya iteratif dalam meningkatkan kualitas analisis sentimen menggunakan teknologi Big Data.

**Kata-kata kunci:** Pemilu 2024, Big Data, Hadoop, Klaster Spark, Data Science, Analisis sentimen, Detik.com, Capres, Cawapres, Partai Politik, pembelajaran mesin, popularitas

## ABSTRACT

The upcoming 2024 General Election is eagerly anticipated by the Indonesian public. Numerous news media outlets have begun releasing articles about presidential candidates, vice-presidential candidates, and political parties participating in the upcoming Indonesian elections. News articles are available through various online portals in Indonesia, with Detik.com being one of the widely used news platforms for staying updated on the latest news. With the large amount of news available, and the speed with which the news is released, the public can be confused about assessing the popularity of the presidential, vice presidential candidates and the parties they support if they look at them one by one.

This research aims to overcome the problem discussed in the first paragraph by analyzing news texts related to the 2024 Election using Big Data technology, particularly leveraging the distributed systems of Hadoop and Spark Cluster. The study is conducted through the following steps: implementing Data Science stages, starting from data collection, cleaning, exploration, analysis, to the visualization of text data associated with the 2024 Presidential Election from the news portal detik.com. The focus of the analysis is to measure the popularity of Presidential and Vice Presidential candidates (Capres/Cawapres) and political parties, both in terms of sentiment and statistics.

The research involves exploring and preparing news data from detik.com using queries that encompass various Presidential and Vice Presidential candidates and political parties relevant to the 2024 Election. The analysis is carried out on the news data with the aim of measuring the popularity of Capres/Cawapres and political parties. In this process, three iterations occur: the first involves NLP preprocessing techniques per news document and the use of machine learning models, the second employs a per-sentence approach for sentiment analysis using machine learning models, and the third introduces additional features, namely news headlines, to enhance sentiment analysis accuracy.

The results of the analysis are presented with statistical metrics and sentiments arranged in such a way as to be engaging and easily comprehensible. The third iteration incorporates class polarity obtained from both news headlines and texts as weights to create a combined sentiment representing the news more comprehensively. The results of this research provide insight into understanding the popularity and sentiment related to the 2024 Election in the form of website-based software in the form of *Dashboard*, as well as showing iterative efforts to improve the quality of sentiment analysis using Big Data technology.

**Keywords:** Election 2024, Big Data, Hadoop, Spark Cluster, Data Science, Sentiment Analysis, Detik.com, Presidential Candidate, Vice Presidential Candidate, Political Parties, Machine Learning, Popularity.

*Skripsi ini dipersembahkan untuk orang tua dan keluarga. . .*

## KATA PENGANTAR

Tugas Akhir ini menggambarkan sebuah perjalanan ilmiah yang dipenuhi dengan berbagai rintangan dan momen berharga yang tak terlupakan. Dengan penuh rasa syukur, Penulis ingin mengucapkan terima kasih kepada Tuhan Yang Maha Esa, atas anugerah, petunjuk, dan keberkahan-Nya yang telah melimpahkan rahmat-Nya sehingga Tugas Akhir ini dapat diselesaikan. Tidak lupa, Penulis juga ingin menyampaikan penghargaan setinggi-tingginya kepada semua pihak yang turut berperan serta dalam mendukung dan membantu Penulis selama perjalanan kuliah dan penulisan Tugas Akhir ini. Adalah suatu kehormatan bagi Penulis untuk dapat mengabadikan rasa terima kasih kepada mereka yang telah menjadi penopang dan sumber inspirasi. Berikut ini adalah ucapan terima kasih Penulis kepada para pihak yang berperan dalam kelancaran penulisan Tugas Akhir ini:

1. Kepada kedua orang tua dan kakak Penulis yang selalu mendukung, menyemangati, dan memberikan doa.
2. Kepada Ibu Prof. Dr. Veronica Sri Moertini selaku dosen pembimbing utama dan Ibu Maria Veronica, M.T. selaku dosen pembimbing serta, yang telah banyak memberikan ilmu, saran, dan bantuan selama proses pengerjaan Tugas Akhir ini.
3. Kepada Bapak Raymond Chandra Putra, M.T. dan Ibu Dr.rer.nat. Cecilia Esti Nugraheni selaku dosen penguji yang telah memberikan masukan dan saran untuk Tugas Akhir ini.
4. Kepada teman-teman seperjuangan kuliah yang menemani Penulis selama masa kuliah, Jeremy, Vincentius Daryl, Vincent Kurniawan, Fabrianus, Axel, Filipus, Aditya, Jose, Daffa, Gasta, Ruben, Faisal, Yoga, Dustin, Gandhi, Michael Sergio, dan Wildan
5. Kepada teman-teman Penulis yang menemani Penulis selama masa penulisan Tugas Akhir ini, Christian, Timothy, Clay, Dewa, Enrico, Erik, Ferdy, Garry, Gheraldy, Glenn, Josephan, Kevin, Michael Evan, Matthew, Pascalis, Okka, Adrian, David, dan John.
6. Kepada seluruh pihak yang mendukung Penulis yang tidak bisa disebutkan satu-satu.

Sebagai akhir dari perjalanan Tugas Akhir ini, Penulis dengan rendah hati mengakui bahwa masih terdapat kekurangan dan peningkatan yang dapat dilakukan pada penelitian ini. Penulis dengan tulus memohon maaf apabila terdapat kesalahan penulisan atau kekurangan dalam penerapan metode pada Tugas Akhir ini.

Terima kasih kepada semua pembaca yang telah meluangkan waktu untuk membaca hasil jerih payah Penulis. Penulis berharap bahwa Tugas Akhir ini dapat memberikan kontribusi positif dan bermanfaat untuk kemajuan ilmu pengetahuan. Semoga karya ini dapat menjadi panduan yang berharga bagi penelitian-penelitian selanjutnya.

Bandung, Januari 2024

Penulis



# DAFTAR ISI

<b>KATA PENGANTAR</b>	<b>xv</b>
<b>DAFTAR ISI</b>	<b>xvii</b>
<b>DAFTAR GAMBAR</b>	<b>xxi</b>
<b>DAFTAR TABEL</b>	<b>xxv</b>
<b>DAFTAR KODE PROGRAM</b>	<b>xxviii</b>
<b>1 PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang . . . . .	1
1.2 Rumusan Masalah . . . . .	5
1.3 Tujuan . . . . .	5
1.4 Batasan Masalah . . . . .	5
1.5 Metodologi . . . . .	5
1.6 Sistematika Pembahasan . . . . .	6
<b>2 LANDASAN TEORI</b>	<b>7</b>
2.1 Pemilu . . . . .	7
2.1.1 Sistem Pemilihan Umum Presiden dan Wakil Presiden . . . . .	7
2.2 Hadoop . . . . .	9
2.2.1 Arsitektur Hadoop . . . . .	9
2.2.2 Hadoop Distributed File System (HDFS) . . . . .	10
2.2.3 Map Reduce . . . . .	12
2.3 Spark . . . . .	13
2.3.1 Application Programming Interface (API) Spark . . . . .	14
2.3.2 Arsitektur Apache Spark . . . . .	15
2.3.3 Library Apache Spark . . . . .	15
2.4 Kafka . . . . .	17
2.4.1 Messages and Batch . . . . .	17
2.4.2 Skema . . . . .	18
2.4.3 Topics . . . . .	18
2.4.4 Brokers . . . . .	18
2.4.5 Producers . . . . .	18
2.4.6 Consumers . . . . .	19
2.4.7 Zookeeper . . . . .	19
2.5 Pengolahan Bahasa Alami ( <i>Natural Language Processing</i> ) . . . . .	20
2.5.1 <i>Teks Preprocessing</i> . . . . .	20
2.5.2 Vektorisasi Teks . . . . .	21
2.5.3 Penggunaan Vektorisasi Teks . . . . .	21
2.5.4 Metode Vektorisasi Teks TF-IDF: . . . . .	21
2.5.5 Teknik Analisis Berbasis Statistik dan Machine Learning . . . . .	22

2.6	Algoritma Klasifikasi . . . . .	23
2.6.1	<i>Multinomial Logistic Regression</i> . . . . .	24
2.6.2	Naive Bayes Classifier . . . . .	24
2.6.3	<i>Decision Tree Classifier</i> . . . . .	25
2.6.4	<i>Random Forest Classifier</i> . . . . .	27
2.6.5	Evaluasi Metrik . . . . .	28
2.7	<i>Web Scraping</i> . . . . .	29
2.7.1	Requests . . . . .	29
2.7.2	Beautiful Soup . . . . .	29
<b>3</b>	<b>STUDI EKSPLORASI</b>	<b>31</b>
3.1	Eksplorasi Instalasi dan Penggunaan Hadoop Sebagai Data Warehouse . . . . .	31
3.2	Instalasi dan Eksplorasi Penggunaan Apache Spark . . . . .	36
3.3	Eksplorasi Instalasi dan Penggunaan Kafka . . . . .	44
3.4	Eksplorasi Penggunaan Web Scraping . . . . .	47
3.5	Eksplorasi Penggunaan Teknik Analisis Berbasis <i>Machine Learning</i> . . . . .	50
3.6	Eksplorasi Berita Pemilu 2024 . . . . .	51
3.6.1	Calon Presiden Nomor Urut 3 Ganjar . . . . .	51
3.6.2	Calon Wakil Presiden Nomor Urut 3 Mahfud MD . . . . .	51
3.6.3	Calon Presiden Nomor Urut 2 Prabowo . . . . .	52
3.6.4	Calon Wakil Presiden Nomor Urut 2 Gibran . . . . .	53
3.6.5	Calon Presiden Nomor Urut 1 Anies . . . . .	55
3.6.6	Calon Wakil Presiden Nomor Urut 1 Muhaimin . . . . .	55
3.6.7	Partai Calon dan Calon Wakil Presiden Pemilu 2024 . . . . .	55
<b>4</b>	<b>PENGUMPULAN, EKSPLORASI, PENYIAPAN DATA, DAN PEMBANGUNAN MO- DEL ANALISIS SENTIMEN</b>	<b>59</b>
4.1	Iterasi Pertama Pembuatan Model Analisis Sentimen . . . . .	59
4.1.1	Pengumpulan Data . . . . .	60
4.1.2	Penyiapan Data . . . . .	60
4.1.3	Analisa Distribusi Panjang Kalimat atau Paragraf . . . . .	63
4.1.4	Analisa Frekuensi Kata . . . . .	63
4.1.5	Analisa N-Gram . . . . .	64
4.1.6	Perubahan format <i>string</i> tanggal menjadi <i>integer</i> . . . . .	65
4.1.7	Pelabelan Data Berita . . . . .	66
4.1.8	Pembuatan Model Analisa Sentimen Iterasi Pertama . . . . .	67
4.2	Iterasi Kedua Pembuatan Model Analisis Sentimen . . . . .	70
4.2.1	Penyiapan Data . . . . .	72
4.2.2	Pemecahan Data Menjadi per Kalimat . . . . .	75
4.2.3	Pembuatan dan Evaluasi Model Machine Learning . . . . .	76
4.2.4	Hasil Prediksi Analisis Sentimen dengan Pendekatan Pemecahan Dokumen menjadi Kalimat . . . . .	79
4.3	Iterasi Ketiga Pembuatan Model Analisis Sentimen . . . . .	81
4.3.1	Pelabelan Data Judul Berita Secara Manual . . . . .	83
4.3.2	Pembuatan Model untuk Analisis Sentimen Judul Berita dan Teks Berita . . . . .	83
4.3.3	Analisis Statistik Teks Berita Pemilu 2024 dan Kerangka ECTL . . . . .	95
4.4	Hasil Analisis Teks Berita Pemilu pada Iterasi Ketiga . . . . .	99
4.4.1	Distribusi Berita Mengenai Masing-masing Calon Presiden, Calon Wakil Presiden, dan Partai Pemilu . . . . .	99
4.4.2	Tren Peningkatan atau Penurunan Jumlah Berita yang Membahas Pemilu 2024 Seiring Berjalannya Waktu . . . . .	102
4.4.3	Sentimen Media Berita Terhadap Setiap Calon Presiden . . . . .	102

4.4.4	Perubahan Dalam Sentimen Media Berita Terhadap calon Presiden Seiring Berjalannya Waktu . . . . .	106
4.4.5	Persebaran Geografis Berita Terkait Pemilu 2024 . . . . .	109
<b>5</b>	<b>PERANCANGAN PERANGKAT LUNAK DAN IMPLEMENTASINYA</b>	<b>113</b>
5.1	Fitur Perangkat Lunak . . . . .	113
5.2	Use Case Diagram . . . . .	113
5.3	Implementasi Perangkat Lunak . . . . .	114
5.3.1	Halaman <i>Insights</i> . . . . .	114
5.3.2	Tampilan Halaman Tren Berita <i>User Input</i> . . . . .	115
<b>6</b>	<b>KESIMPULAN DAN SARAN</b>	<b>117</b>
6.1	Kesimpulan . . . . .	117
6.2	Saran . . . . .	118
	<b>DAFTAR REFERENSI</b>	<b>119</b>
	<b>A KODE PROGRAM</b>	<b>121</b>
	<b>B HASIL EKSPERIMEN</b>	<b>133</b>

## DAFTAR GAMBAR

1.1	Contoh Gambar Berita di Internet <sup>1</sup>	1
1.2	Karakteristik Big Data <sup>2</sup>	2
1.3	Komponen library Spark <sup>3</sup>	2
1.4	Arsitektur penyimpanan file HDFS <sup>4</sup>	3
1.5	Apache Kafka sebagai tempat perantara antara data dan target <sup>5</sup>	3
1.6	Sentimen yang biasa digunakan pada analisis sentimen <sup>6</sup>	4
2.1	Klaster Hadoop <sup>7</sup>	9
2.2	Arsitektur Hadoop <sup>8</sup>	10
2.3	Arsitektur HDFS dan Contoh <i>Replication</i> <sup>9</sup>	11
2.4	Proses MapReduce <sup>10</sup>	12
2.5	Arsitektur MapReduce <sup>11</sup>	12
2.6	Susunan Spark <sup>12</sup>	13
2.7	RDD Spark <sup>13</sup>	14
2.8	Arsitektur Spark <sup>14</sup>	15
2.9	Gambaran Kafka <sup>15</sup>	17
2.10	Gambaran Vektorisasi Teks <sup>16</sup>	21
2.11	Contoh <i>Word Frequency</i> <sup>17</sup>	22
2.12	Contoh <i>Decision Tree</i> <sup>18</sup>	25
2.13	Contoh <i>Pruning</i> <sup>19</sup>	26
2.14	Random Forest Ensemble <sup>20</sup>	27
2.15	Gambaran proses <i>web scraping</i> <sup>21</sup>	29
3.1	Halaman download Apache Hadoop <sup>22</sup>	31
3.2	JPS untuk memastikan Hadoop jalan	32
3.3	Tampilan UI Apache Hadoop	32
3.4	Tampilan Utama UI Hadoop Datanode dan Namenode	33
3.5	Tampilan <i>Datanode</i> UI Hadoop Lokal	33
3.6	Direktori Tempat Penyimpanan Data di HDFS Lokal	33
3.7	Informasi File Data di HDFS Lokal	34
3.8	SSH ke Access Point	35
3.9	Login	35
3.10	Tampilan Hadoop Lab	35
3.11	Informasi dan <i>Replication</i> klaster HDFS Labkom	36
3.12	Tampilan Halaman untuk Mengunduh Spark <sup>23</sup>	37
3.13	Setting variabel lokal untuk Spark	37
3.14	Pyspark berhasil dinyalakan	37
3.15	Halaman UI Master Pengelolaan Pekerjaan Spark	38
3.16	Halaman UI Master Pengelolaan Pekerjaan Spark Ketika Job sudah Jalan	39
3.17	Halaman UI Master Pengelolaan Pekerjaan Spark Bagian DAG	39
3.18	SSH ke Access Point	41
3.19	Login	41
3.20	Spark Cluster Labkom	41

3.21	<i>Tasks</i> dikerjakan oleh 2 buah slave . . . . .	42
3.22	Web Interface Kafka Manager . . . . .	45
3.23	Web Interface Kafka Connect . . . . .	45
3.24	Pembuatan Konektor . . . . .	46
3.25	Hasil Consumer . . . . .	47
3.26	Tampilan laman portal Detik.com . . . . .	47
3.27	Laman portal <i>search query</i> . . . . .	47
3.28	Timeline Pemilu 2024 <sup>24</sup> . . . . .	51
3.29	Kontroversi pembatalan piala dunia U-20 oleh Ganjar <sup>25</sup> . . . . .	52
3.30	Berita positif mengenai Ganjar <sup>26</sup> . . . . .	52
3.31	Berita negatif mengenai Mahfud <sup>27</sup> . . . . .	52
3.32	Berita negatif mengenai Prabowo khususnya pemasangannya dengan Gibran <sup>28</sup> . . . . .	53
3.33	Berita negatif mengenai Prabowo <sup>29</sup> . . . . .	53
3.34	Berita positif mengenai Prabowo <sup>30</sup> . . . . .	54
3.35	Berita negatif mengenai Prabowo <sup>31</sup> . . . . .	54
3.36	Tampilan berita-berita mengenai Gibran . . . . .	54
3.37	Berita negatif mengenai Anies . . . . .	55
3.38	Tampilan berita positif mengenai Anies . . . . .	55
3.39	Berita negatif mengenai Muhaimin <sup>32</sup> . . . . .	56
3.40	Berita positif mengenai Muhaimin <sup>33</sup> . . . . .	56
3.41	Berita negatif mengenai partai pemilu <sup>34</sup> . . . . .	56
4.1	Alur Iterasi Pertama . . . . .	59
4.2	Artikel berita tentang pemilu . . . . .	60
4.3	Hasil Scrape teks berita . . . . .	61
4.4	Hasil Pelabelan manual . . . . .	68
4.5	Alur pembuatan model iterasi kedua . . . . .	71
4.6	Spark Lokal membaca csv dari HDFS Labkom FTIS . . . . .	73
4.7	Tampilan hasil <i>write</i> Spark lokal ke HDFS <i>cluster</i> . . . . .	76
4.8	Contoh Pelabelan per kalimat . . . . .	76
4.9	Contoh Alur Pembangunan Model Iterasi Ketiga . . . . .	82
4.10	Contoh Data Judul Berita yang akan dilabelkan . . . . .	84
4.11	Judul Berita dengan sentimen positif yang jelas . . . . .	84
4.12	Isi berita yang judulnya sentimen positif . . . . .	85
4.13	Alur Kerangka ECTL . . . . .	99
4.14	<i>Insights</i> Distribusi Berita Calon Presiden 2024 . . . . .	100
4.15	<i>Insights</i> Distribusi Calon Wakil Presiden 2024 . . . . .	101
4.16	<i>Insights</i> Distribusi Jumlah Berita Mengenai Partai Calon Presiden Pemilu 2024 . . . . .	102
4.17	<i>Insights</i> Tren Peningkatan atau Penurunan Jumlah Berita Pemilu 2024 . . . . .	103
4.18	<i>Insights</i> Distribusi Sentimen Berita Teks Pemilu 2024 Mengenai Calon Presiden . . . . .	103
4.19	<i>Insights</i> Distribusi Sentimen Berita Calon Presiden Nomor Urut 1 Anies Baswedan . . . . .	104
4.20	<i>Insights</i> Distribusi Sentimen Berita Calon Presiden Nomor Urut 2 Prabowo Subianto . . . . .	105
4.21	<i>Insights</i> Distribusi Sentimen Berita Calon Presiden Nomor Urut 3 Ganjar Pranowo . . . . .	106
4.22	<i>Insights</i> Distribusi Sentimen Teks Berita Pemilu 2024 Mengenai Calon Presiden Anies Baswedan . . . . .	107
4.23	<i>Insights</i> Distribusi Sentimen Teks Berita Pemilu 2024 Mengenai Calon Presiden Prabowo Subianto . . . . .	108
4.24	<i>Insights</i> Distribusi Sentimen Teks Berita Pemilu 2024 Mengenai Calon Presiden Ganjar Pranowo . . . . .	109
4.25	<i>Insights</i> Distribusi Geografis Berita terkait Pemilu 2024 dari Bulan Januari hingga November . . . . .	110

5.1	Use Case Diagram . . . . .	113
5.2	Tampilan <i>Home</i> perangkat lunak . . . . .	114
5.3	Tampilan Halaman <i>Insights</i> Perangkat Lunak . . . . .	114
5.4	Tampilan halaman tren berita kandidat per user input dari bulan Januari hingga November . . . . .	115
5.5	Tampilan halaman tren berita kandidat per user input ketika dijalankan, dari bulan Januari hingga November . . . . .	115
B.1	Hasil Analisis Distribusi Berita Kandidat Presiden . . . . .	133
B.2	Hasil Analisis Distribusi Berita Kandidat Wakil Presiden . . . . .	134
B.3	Hasil Analisis Distribusi Berita Partai Pemilu . . . . .	134
B.4	Hasil Analisis Tren Distribusi Berita Partai Pemilu sepanjang Januari hingga November 2023 . . . . .	135
B.5	Hasil Analisis Tren Sentimen Kandidat Presiden Nomor Urut 1 Anies Baswedan . . . . .	135
B.6	Hasil Analisis Tren Sentimen Kandidat Presiden Nomor Urut 2 Prabowo Subianto . . . . .	136
B.7	Hasil Analisis Tren Sentimen Kandidat Presiden Nomor Urut 3 Ganjar Pranowo . . . . .	136
B.8	Hasil Analisis Distribusi Berita berdasarkan Lokasi . . . . .	137
B.9	Hasil Analisis Distribusi Berita berdasarkan Sentimen Kandidat Presiden Nomor Urut 1 Anies Baswedan . . . . .	137
B.10	Hasil Analisis Distribusi Berita berdasarkan Sentimen Kandidat Presiden Nomor Urut 2 Prabowo Subianto . . . . .	138
B.11	Hasil Analisis Distribusi Berita berdasarkan Sentimen Kandidat Presiden Nomor Urut 3 Ganjar Pranowo . . . . .	139

## DAFTAR TABEL

3.1	Hasil DataFrame WebScraping berita . . . . .	50
4.1	Atribut Data Teks Berita . . . . .	61
4.2	Hasil Analisa Frekuensi Kata . . . . .	64
4.3	Hasil Ekstraksi Tanggal dan Group per bulan . . . . .	67
4.4	Hasil Evaluasi Model Multinomial Logistic Regression Iterasi Pertama . . . . .	70
4.5	Hasil Evaluasi Model Multinomial Naive Bayes Classifier Iterasi Pertama . . . . .	70
4.6	Hasil Evaluasi Model Decision Tree Iterasi Pertama . . . . .	70
4.7	Hasil Evaluasi Model Random Forest Iterasi Pertama . . . . .	71
4.8	Hasil Evaluasi Model Iterasi Kedua per Kalimat . . . . .	79
4.9	Hasil Evaluasi Model Iterasi Ketiga Teks Berita per Kalimat . . . . .	87
4.10	Hasil Evaluasi Model Iterasi Ketiga Judul Berita . . . . .	87
4.11	Confusion Matrix model Teks Berita per Kalimat . . . . .	87
4.12	Confusion Matrix model Judul Berita . . . . .	88

## DAFTAR KODE PROGRAM

2.1	Contoh penggunaan Spark SQL . . . . .	16
2.2	Contoh Penggunaan MLlib . . . . .	16
2.3	Contoh Penggunaan Requests . . . . .	29
2.4	Contoh Penggunaan Beatiful Soup . . . . .	30
3.1	Kode Sederhana Penghitung Jumlah Kata . . . . .	38
3.2	Pengubahan Konfigurasi <i>Zookeeper.properties</i> . . . . .	44
3.3	Kode untuk mengaktifkan Zookeeper dan Kafka . . . . .	44
3.4	Kode untuk mengaktifkan <i>Web Interface</i> Kafka Manager . . . . .	44
3.5	Menyalakan Kafka Connect . . . . .	45
3.6	Kode Menyalakan <i>Consumer</i> . . . . .	46
3.7	Fungsi untuk mengambil atribut berita . . . . .	48
3.8	Lakukan looping untuk mengambil semua berita dari semua halaman yang didapat . . . . .	48
3.9	Fungsi untuk mengambil teks berita . . . . .	48
3.10	Kode Eksplorasi Teknik Analisis Machine Learning Pyspark . . . . .	50
4.1	Kode Load Data . . . . .	61
4.2	Fungsi Pembersihan Teks . . . . .	62
4.3	Kode Pembuatan UDF CleanExtra . . . . .	62
4.4	Kode perubahan teks berita bersih menjadi <i>bag of words</i> . . . . .	63
4.5	Kode Analisa Distribusi Panjang Kalimat atau Paragraf . . . . .	63
4.6	Kode Analisa Frekuensi Kata . . . . .	64
4.7	Kode Frekuensi Bigram yang paling sering muncul . . . . .	64
4.8	Kode Ekstraksi Tanggal . . . . .	65
4.9	Kode Ekstraksi Tanggal 2 . . . . .	66
4.10	Kode Vektorisasi Teks . . . . .	67
4.11	Kode Pelatihan dan Test Model Analisis Sentimen . . . . .	68
4.12	Kode Evaluasi Model Machine Learning . . . . .	69
4.13	Kode Pembersihan Teks Berita Iterasi Kedua . . . . .	72
4.14	Kode Pemecahan berita menjadi per kalimat . . . . .	75
4.15	Kode Vektorisasi Teks berita per kalimat . . . . .	76
4.16	Kode Pelatihan Model Teks Berita per Kalimat . . . . .	78
4.17	Kode Prediksi Teks Berita per Kalimat . . . . .	79
4.18	Kode Pembuatan Model Iterasi Ketiga . . . . .	86
4.19	Kode Penyiapan data untuk prediksi . . . . .	89
4.20	Kode Prediksi data teks berita untuk Iterasi Ketiga . . . . .	89
4.21	Kode Pengecekan hasil prediksi teks berita dan judul berita . . . . .	92
4.22	Kode Penggabungan hasil prediksi teks berita dan judul berita . . . . .	93
4.23	Kode Analisis Distribusi Berita . . . . .	95
4.24	Kode Analisis Tren Berita Pemilu . . . . .	96
4.25	Kode Analisis Sentimen Berita Pemilu . . . . .	96
4.26	Kode Analisis Tren Sentimen Berita . . . . .	97



4.27	Kode Analisis Tren Lokasi Berita . . . . .	98
4.28	Kode Analisis Tren Berita Sentimen Berdasarkan User Input . . . . .	98
A.1	Kode Web Scraping . . . . .	121
A.2	Kode Preprocess Data . . . . .	122
A.3	Kode Pembuatan Model Machine Learning Analisis Sentimen . . . . .	123
A.4	Kode Transformasi Atribut . . . . .	125
A.5	Kode Prediksi dan Penyiapan Data untuk Analisis . . . . .	126
A.6	Kode Analisis Statistik dan Dump Data . . . . .	130

# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang

Pemilu adalah singkatan dari pemilihan umum, yang digunakan sebagai suatu sarana untuk mengisi jabatan-jabatan politik di pemerintahan suatu negara demokrasi, berdasarkan pilihan formal dari warga negaranya yang memenuhi syarat. Salah satu negara yang rutin melakukan pemilihan umum yaitu negara Indonesia, yang melakukan pemilu selanjutnya pada tahun 2024. Dalam melakukan pemilihan umum tentunya para kandidat, wakil kandidat, dan partainya melakukan kampanye untuk mendorong popularitas kandidat dan partai tersebut sehingga kemungkinan kandidat terpilih lebih besar. Hasil dari kampanye-kampanye yang dilakukan menghasilkan opini publik yang tersebar dalam media sosial, seperti Twitter, Tiktok, Instagram dan juga berita-berita pada Internet.



Gambar 1.1: Contoh Gambar Berita di Internet<sup>1</sup>

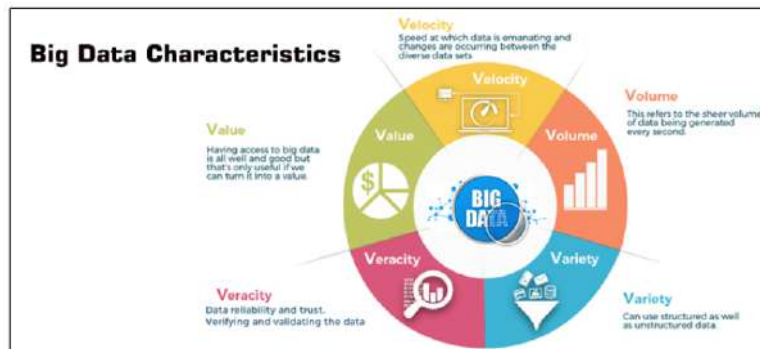
Selain media sosial, wadah yang dapat publik gunakan untuk melihat dan mendapatkan informasi tentang calon dan partai yang terkait yaitu berita-berita dari Internet dengan sumber yang jelas seperti pada gambar 1.1. Berita-berita di internet sangat berperan bagi publik, opini dan pandangan publik akan terhasut dengan banyaknya sumber berita mengenai topik yang mereka cari. Dengan banyaknya berita yang bersebaran di Internet mengenai kandidat dan partai yang terkait, publik dapat kebingungan untuk membaca dan beropini sendiri tentang kandidat dan partai yang didukung.

Oleh karena itu solusi yang ingin ditawarkan pada tugas akhir ini adalah menganalisis popularitas dan sentimen dari teks-teks yang berkaitan tentang pemilu berupa kandidat-kandidat dan partainya pada berita di Internet. Teknologi Big Data dapat digunakan karena data-data pada berita di Internet memenuhi karakteristik dari 5V pada Big Data.

Karakteristik pada 5V yaitu *volume*, *veracity*, *velocity*, *variety*, dan *value* seperti yang tertera pada Gambar 1.2. *Volume* adalah sekumpulan data dalam jumlah dan ukuran yang sangat besar. *Veracity* adalah seberapa akurat dan dapat dipercaya suatu data. *Velocity* adalah kecepatan pertambahan data dalam kurun waktu tertentu. *Variety* adalah berbagai jenis data yang ada, mulai

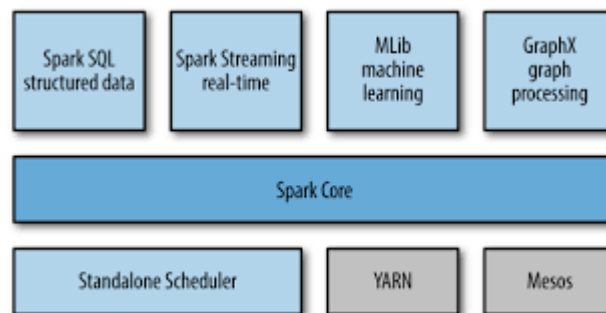
<sup>1</sup>Sumber:<https://news.detik.com/pemilu/d-6760967/analisis-pakar-soal-teori-jalan-tengah-dan-elektabilitas-tinggi-prabowo>

dari data terstruktur seperti tabel basis data, data semi-terstruktur seperti XML dan JSON, dan data tidak terstruktur seperti citra, suara, dan teks. Yang terakhir yaitu *Value* adalah seberapa bernilainya suatu data yang digunakan.



Gambar 1.2: Karakteristik Big Data<sup>2</sup>

Dengan penggunaan teknologi Big Data seperti Spark, Hadoop, dan Kafka dapat menangani karakteristik-karakteristik Big Data yang tidak bisa ditangani dengan teknik penanganan data secara biasa. Spark adalah platform komputasi distribusi *open-source* yang dirancang untuk mengolah dan menganalisis data secara efisien. Pada gambar 1.3 Spark menyediakan antarmuka pemrograman yang kaya, pemrosesan data yang cepat, dan dukungan untuk berbagai jenis tugas pemrosesan data, termasuk pemrosesan batch, pemrosesan aliran (*streaming*), pembelajaran mesin (*machine learning*), dan pemrosesan grafik.



Gambar 1.3: Komponen library Spark<sup>3</sup>

Hadoop adalah kerangka kerja open-source yang digunakan untuk menyimpan, mengelola, dan menganalisis data secara terdistribusi. Hadoop menggunakan sistem file terdistribusi (*Hadoop Distributed File System - HDFS*) untuk menyimpan data, terlihat flow penyimpanan datanya pada gambar 1.4 dan menggunakan model pemrosesan paralel (MapReduce) untuk memproses data secara efisien di dalam *cluster* komputasi.

Kafka merupakan sistem pengumpul data dan sistem penyederhanaan data stream. Kafka menyederhanakan data stream dengan cara menjadi perantara antara sumber data stream dan target data stream (*client*). Hubungan antara target dan sumber harus disederhanakan karena sumber dari data stream sangat banyak. Kafka juga memiliki fungsi lain yaitu sebagai tempat transit data sebelum dikirim ke sistem lain, terlihat pada gambar 1.5.

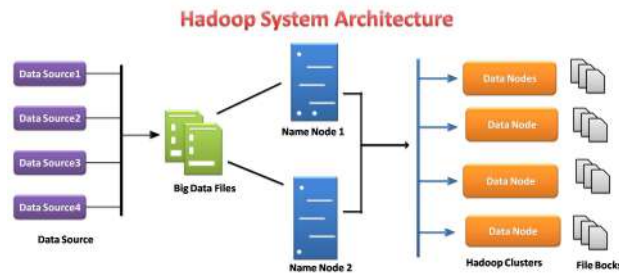
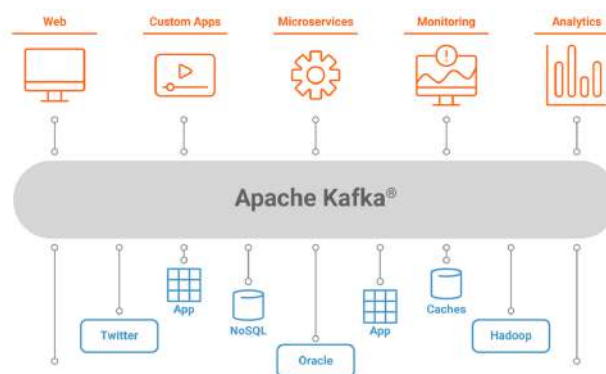
Spark memiliki integrasi yang baik dengan Hadoop. Spark dapat membaca dan menulis data langsung dari HDFS, memungkinkan pemrosesan data yang efisien. Dengan menggunakan Hadoop

<sup>2</sup>Sumber: <https://www.simpliaxis.com/resources/big-data-characteristics>

<sup>3</sup>Sumber: <https://blog.knoldus.com/things-to-know-about-spark-rdd/>

<sup>4</sup>Sumber : <https://discuss.boardinfinity.com/t/hdfs-hadoop-distributed-file-system/6571>

<sup>5</sup>Sumber : <https://halovina.com/apache-kafka-platform-terdistribusi-untuk-data-streaming/>

Gambar 1.4: Arsitektur penyimpanan file HDFS<sup>4</sup>Gambar 1.5: Apache Kafka sebagai tempat perantara antara data dan target<sup>5</sup>

dan Spark bersama-sama, dapat memanfaatkan kecepatan pemrosesan dan kemampuan analisis Spark dengan memanfaatkan sistem penyimpanan terdistribusi dan skalabilitas Hadoop.

Berita berdasarkan karakteristiknya, dilihat dari jumlah popularitasnya pun dapat berubah dari waktu ke waktu. Dilihat dari sentimennya yang bernada positif, negatif, dan netral juga dapat berubah waktu ke waktu. Oleh karena itu dengan bantuan teknologi Big Data, permasalahan tersebut bisa diselesaikan dengan adanya program yang dibuat pada tugas akhir ini untuk memvisualisasikan hasil analisis yang didapat mengenai capres dan cawapres pemilu 2024

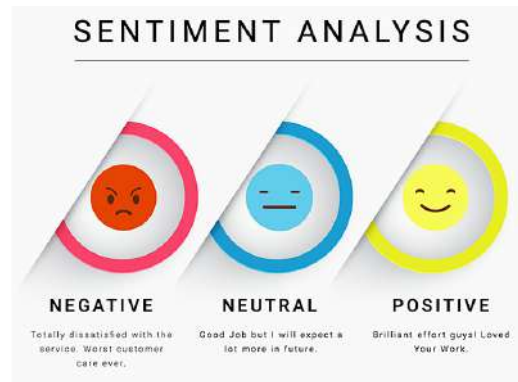
Penelitian dengan melibatkan analisis sentimen pada data sosial media dan artikel berita sudah pernah dilakukan, tetapi penggunaan Big Data dan teknologi pemrosesan data terdistribusi masih terbatas dalam konteks pemilihan umum. Oleh karena itu, tugas akhir ini bertujuan untuk mengisi kesenjangan ini dengan menerapkan pendekatan Big Data yang melibatkan Spark, dan Hadoop untuk menganalisis sentimen pada teks artikel berita selama periode pemilihan umum.

Dengan menganalisis sentimen dalam artikel berita, tugas akhir ini berpotensi memberikan wawasan yang lebih akurat dan komprehensif tentang opini dan sentimen publik terkait dengan pemilihan umum. Hasil tugas akhir ini dapat membantu dalam pemahaman isu-isu yang relevan, dukungan untuk kandidat atau partai politik, serta evaluasi respons publik terhadap peristiwa-peristiwa yang terjadi selama pemilihan umum.

Analisis yang dilakukan berupa analisis sentimen dan analisis statistik untuk mengukur popularitas capres dan cawapres. Analisis sentimen adalah proses untuk mengidentifikasi, memahami, dan mengekstraksi sentimen atau pendapat yang terkandung dalam teks, baik itu positif, negatif, atau netral dapat dilihat pada gambar 1.6. Dalam konteks pemilihan umum, analisis sentimen digunakan untuk memahami sikap dan pendapat publik terhadap kandidat, partai politik, isu politik, dan peristiwa terkait pemilihan.

Dalam tugas akhir ini, analisis sentimen dilakukan pada artikel berita yang terkait dengan kandidat presiden, kandidat wakil presiden, dan partai pemilu pemilihan umum tahun 2024. Dimana

<sup>6</sup>Sumber: <https://blog.gopenai.com/sentiment-analysis-on-healthcare-reviews-2ec229d04e69>



Gambar 1.6: Sentimen yang biasa digunakan pada analisis sentimen<sup>6</sup>

hal tersebut merupakan wadah yang dapat publik gunakan untuk melihat dan mendapatkan informasi tentang calon dan partai yang terkait yaitu berita-berita dari Internet dengan sumber yang belum tentu jelas.

Analisis teks *tweet* dari media sosial seperti Twitter sudah pernah dilakukan sebelumnya, salah satunya dilakukan oleh Hamid Bagheri dan MD Johirul Islam dari Iowa State University yang meneliti sentimen dari teks *tweet* Twitter, tertera pada Link <https://arxiv.org/vc/arxiv/papers/1711/1711.10377v1.pdf>.

Mereka mengekstrak sentimen dengan cara :

1. Mengunduh diksi sentimen
2. Mengunduh data test dan memasukkannya ke dalam program
3. Membersihkan *tweets* dengan menghilangkan *stopwords*
4. Melakukan tokenisasi
5. Setiap kata yang ada dibandingkan dengan sentimennya pada diksi
6. Di akhir mendapatkan persentase sentimen dan mendapatkan *polarity*.

Sedangkan penelitian analisis sentimen teks berita, juga pernah dilakukan oleh Antony Samuels dan John Mcgonical dari University of Southern California, Caltech tertera pada Link <https://arxiv.org/ftp/arxiv/papers/2007/2007.02238.pdf>. Pada penelitian tersebut dilakukan analisis sentimen dengan menggunakan korpus berbasis *Lexicon-based*. Dimana kata-kata dalam dokumen berita dihitung nilai *polaritynya* dan diambil kelas sentimen yang terbesar untuk merepresentasikan dokumen tersebut.

Dari kedua eksperimen tersebut, penggunaan model *machine learning* masih sangat jarang, hasil penelitian belum berbentuk perangkat lunak dan hanya berupa nilai polaritas. Oleh karena itu dengan dilakukan penelitian ini, permasalahan tersebut bisa diselesaikan dengan adanya program yang dibuat pada penelitian ini untuk memvisualisasikan hasil analisis yang didapat mengenai capres dan cawapres pemilu 2024. Program yang dibuat menggunakan teknologi Big Data *Hadoop*, dan *Spark*. Dimana Hadoop digunakan sebagai tempat penyimpanan data yang digunakan, dan Spark digunakan untuk komputasi kode yang menjalankan semuanya.

Data yang digunakan pada eksperimen yaitu data artikel berita yang diambil dengan teknik *web scraping* yang diambil secara bertahap. Data berita yang dipakai atributnya yaitu atribut teks.

Perangkat lunak yang dibuat berbentuk dashboard berbasis website sebagai hasil visualisasi dari analisis sentimen dan popularitas. Pengguna dapat melihat jumlah sentimen negatif dan positif setiap calon, cawapres, dan partainya. Pengguna pun dapat melihat hasil analisis popularitas statistik yang ditampilkan berupa *insights*, pengguna dapat melihat tren perkembangan sentimen kandidat, wakil kandidat, dan partainya sesuai dengan input pengguna.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang permasalahan yang dijelaskan pada Bagian latar belakang, berikut merupakan masalah–masalah yang hendak diselesaikan oleh tugas akhir ini:

1. Mengumpulkan data teks dari media (berita) yg terkait dengan pemilu 2024 secara waktu nyata dengan memanfaatkan teknologi big data (Hadoop dan Spark).
2. Menganalisis dan menampilkan hasil analisis yang menjadi ukuran popularitas dari capres/ca-wapres dan partai politik, secara visual dan menarik.
3. Mendeteksi sentimen terhadap capres/cawapres dan partai politik (yang berhubungan dengan pilpres) secara periodik.
4. Menampilkan hasil analisis sentimen dan statistik dengan perangkat lunak.

## 1.3 Tujuan

Berdasarkan rumusan masalah yang dijelaskan pada Bagian rumusan masalah, berikut tujuan dari tugas akhir ini:

1. Menerapkan tahapan data science (mengumpulkan data, membersihkan data, mempelajari dan eksplorasi data, menganalisis data, visualisasi hasil) untuk menganalisis data teks yang terkait dengan pilpres 2024 (media berita) untuk menjadi ukuran popularitas dari capres/cawapres dan partai politik.
2. Melakukan eksplorasi dan penyiapan data berita yang sudah dikumpulkan.
3. Melakukan analisis popularitas dan sentimen pada data Twitter dan berita untuk menjadi ukuran popularitas dari capres/cawapres dan partai politik.
4. Meluncurkan hasil analisis (metrik statistik maupun sentiment) sedemikian rupa sehingga menarik dan mudah dipahami.

## 1.4 Batasan Masalah

Batasan masalah untuk penelitian ini adalah :

1. Dalam pengambilan data, data berita yang diambil berasal dari Detik.com karena penggunaan class html yang dipakai untuk mengekstrak paragraf, hanya Detik.com yang konsisten memakai class yang sama sehingga bisa diambil secara konsisten tanpa adanya *error*.
2. Pembangunan proses penyiapan, eksplorasi, analisis data, dan pembangunan model tidak sepenuhnya dilakukan pada *cluster* Spark dan Hadoop yang terdistribusi karena adanya batasan *hardware* pada *worker* berupa modul-modul yang tidak terinstall. Sehingga pembuatan kerangka ECTL untuk *batch processing* yang sudah dibuat tidak dapat dicoba.
3. *Stemming* tidak bisa digunakan untuk *preprocessing* data, dikarenakan penggunaan *library Stemming* hanya bisa dilakukan pada Spark lokal sehingga dengan jumlah data yang besar, proses Spark tidak kuat menanganinya.
4. Dalam menganalisis sentimen teks berita, teks berita tersebut dapat berisi berbagai calon presiden, calon wakil presiden, dan partai pemilunya, tidak dapat dipisah per calon.

## 1.5 Metodologi

Berikut langkah-langkah yang dilakukan untuk mencapai tujuan pada tugas akhir ini:

1. Melakukan studi literatur mengenai Pemilu 2024
2. Melakukan studi literatur mengenai Kafka
3. Melakukan studi literatur mengenai Hadoop
4. Melakukan studi literatur dan eksplorasi mengenai Spark dan Library untuk analisis data teks yang dibutuhkan
5. Melakukan studi literatur dan eksplorasi mengenai Web Scraping

6. Melakukan studi literatur dan eksplorasi mengenai *Text Preprocessing*, dan *Text Vectorization* untuk transformasi data teks dan Pengolahan Bahasa Alami yang relevan untuk analisis teks pemilu
7. Melakukan studi literatur dan eksplorasi mengenai teknik analisis statistik dan *Machine Learning* yang bermanfaat bagi analisis teks pemilu
8. Membuat program untuk mengumpulkan data dari berbagai berita di media internet dengan *web scraping*.
9. Menyiapkan data dengan melakukan: pembersihan data (*tokenization, text cleaning, stop words removal, stemming*), transformasi ke data terstruktur.
10. Melakukan eksplorasi terhadap data dengan memanfaatkan statistika deskriptif maupun visualisasi dengan menggunakan teknologi Big Data.
11. Melakukan analisis data untuk mengolah teks dengan memanfaatkan berbagai library yang disediakan *Spark*.
12. Melakukan analisis popularitas terhadap capres dan partai dengan memanfaatkan teknologi big data dan pengolahan bahasa alami.
13. Melakukan analisis sentimen terhadap kandidat, wakil kandidat, dan partai dengan pendekatan NLP/Machine Learning dengan teknologi Big Data seperti *SparkMLlib*, dan *NLTK*
14. Merancang dan mengimplementasikan program untuk meluncurkan hasil analisis popularitas maupun sentimen
15. Menulis Dokumen Tugas Akhir

## 1.6 Sistematika Pembahasan

Berikut susunan yang digunakan dalam memaparkan isi tugas akhir ini:

1. Bab 1 Pendahuluan  
Dijelaskan tentang latar belakang masalah terkait pemilu dan Big Data, pengambilan data terkait pemilu, dan pencarian masalahnya. Pada Bagian pendahuluan, dijelaskan juga terkait rumusan masalah, tujuan penelitian, batasan masalah, metodologi, dan sistematika pembahasan pada penelitian ini.
2. Bab 2 Landasan Teori  
Pada Bagian landasan teori dijelaskan landasan teori terkait pemilu, teknologi big data berupa Hadoop, Spark dan Kafka dan konsep lainnya yang digunakan pada penelitian ini
3. Bab 3 Studi Eksplorasi  
Pada Bab 3 membahas tentang langkah-langkah dan eksperimen untuk melakukan konfigurasi pada teknologi yang digunakan serta teknik analisis yang dipakai.
4. Bab 4 Pengumpulan, Eksplorasi, Penyiapan Data, dan Pembangunan Model Analisis Sentimen  
Pada Bab 4 membahas tentang langkah pendekatan pembangunan model analisis sentimen dan hasil analisis statistik yang dilakukan
5. Bab 5 Perancangan Perangkat Lunak dan Implementasinya  
Pada Bab 5 membahas perancangan perangkat lunak *Dashboard* dan implementasinya tentang fitur yang tersedia pada perangkat lunak dan juga memperlihatkan cara kerja dari perangkat lunak yang dibangun beserta hasilnya.
6. Bab 6 Kesimpulan dan Saran  
Pada Bab 6 membahas tentang kesimpulan dari keseluruhan penelitian yang sudah dilakukan dan saran yang membangun agar bisa diterapkan di pengembangan selanjutnya