

**ANALISIS *TWEET* MASYARAKAT TENTANG KEMANDIRIAN  
ENERGI MENGGUNAKAN *RECURRENT NEURAL NETWORK*  
DAN *NAIVE BAYES***

**(Skripsi)**

**Oleh**

**MOHAMMAD SURYA AKBAR**



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMPUNG  
BANDAR LAMPUNG  
2021**

## **ABSTRAK**

### **ANALISIS *TWEET* MASYARAKAT TENTANG KEMANDIRIAN ENERGI MENGGUNAKAN *RECURRENT NEURAL NETWORK* DAN *NAIVE BAYES***

Oleh

**MOHAMMAD SURYA AKBAR**

Analisis sentimen adalah bagian dari penelitian komputasi yang mengekstrak data tekstual untuk mendapatkan nilai positif, atau negatif, terkait suatu topik. Dalam penelitian terbaru, data biasanya diperoleh dari media sosial, termasuk Twitter, di mana pengguna sering memberikan pendapat pribadi mereka tentang subjek tertentu. Kemandirian energi pernah menjadi trending topic di Indonesia, karena pendapatnya yang beragam, pro dan kontra, menarik untuk dianalisis.

Pembelajaran mendalam adalah cabang pembelajaran mesin yang terdiri dari lapisan tersembunyi jaringan saraf dengan menerapkan transformasi non-linier dan abstraksi model tingkat tinggi dalam database besar. Jaringan saraf berulang (RNN) adalah metode pembelajaran mendalam yang memproses data berulang kali, terutama cocok untuk tulisan tangan, data multi-kata, atau pengenalan suara. Penelitian ini membandingkan tiga algoritma: Simple Neural Network, Bernoulli Naive Bayes, dan Long Short-Term Memory (LSTM) dalam analisis sentimen menggunakan data kemandirian energi dari Twitter. Berdasarkan hasil penelitian, Simple Recurrent Neural Network menunjukkan kinerja terbaik dengan nilai akurasi 78% dibandingkan dengan nilai Bernoulli Naive Bayes 67% dan LSTM dengan nilai akurasi 75%.

Kata kunci: Analisis Sentimen, Simple RNN, LSTM, Bernoulli Naive Bayes, Kemandirian Energi

## **ABSTRACT**

### **COMMUNITY TWEET ANALYSIS ON ENERGY INDEPENDENCE USING RECURRENT NEURAL NETWORK AND NAIVE BAYES**

**By**

**MOHAMMAD SURYA AKBAR**

Sentiment analysis is part of computational research that extracts textual data to obtain positive, or negative values related to a topic. In recent research, data are commonly acquired from social media, including Twitter, where users often provide their personal opinion about a particular subject. Energy independence was once a trending topic discussed in Indonesia, as the opinions are diverse, pros and cons, making it interesting to be analyzed.

Deep learning is a branch of machine learning consisting of hidden layers of neural networks by applying non-linear transformations and high-level model abstractions in large databases. The recurrent neural network (RNN) is a deep learning method that processes data repeatedly, primarily suitable for handwriting, multi-word data, or voice recognition. This study compares three algorithms: Simple Neural Network, Bernoulli Naive Bayes, and Long Short-Term Memory (LSTM) in sentiment analysis using the energy independence data from Twitter. Based on the results, the Simple Recurrent Neural Network shows the best performance with an accuracy value of 78% compared to Bernoulli Naive Bayes value of 67% and LSTM with an accuracy value of 75%.

Key words: Sentiment Analysis, Simple RNN, LSTM, Bernoulli Naive Bayes, Energy Independence.

**ANALISIS *TWEET* MASYARAKAT TENTANG KEMANDIRIAN  
ENERGI MENGGUNAKAN *RECURRENT NEURAL NETWORK* DAN  
*NAIVE BAYES***

Oleh

**MOHAMMAD SURYA AKBAR**

Skripsi

Sebagai Salah Satu Syarat Untuk Mencapai Gelar  
**SARJANA KOMPUTER**

Pada

**Jurusan Ilmu Komputer  
Fakultas Matematika dan Ilmu Pengetahuan Alam**



**JURUSAN ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMPUNG  
2021**

Judul Skripsi : **ANALISIS TWEET MASYARAKAT TENTANG KEMANDIRIAN ENERGI MENGGUNAKAN RECURRENT NEURAL NETWORK DAN NAIVE BAYES**

Nama Mahasiswa : **Mohammad Surya Akbar**

Nomor Pokok Mahasiswa : 1617051077

Jurusan : Ilmu Komputer

Fakultas : Matematika dan Ilmu Pengetahuan Alam



1. Komisi Pembimbing

**Dr. Ir. Kurnia Muludi, M.S.Sc**  
NIP. 19640616 198902 1 001

**Dewi Asiah Shofiana, S.Kom., M.Kom**  
NIP.199509292020122030

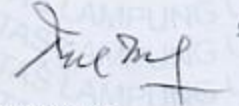
2. Ketua Jurusan Ilmu Komputer

**Didik Kurniawan, S.Si., M.T.**  
NIP. 19800419 200501 1 004

**MENGESAHKAN**

1. Tim Penguji

Ketua : **Dr. Ir. Kurnia Muludi, M.S.Sc**



Penguji I  
Sekretaris : **Dewi Asiah Shofiana, S.Kom., M.Kom**



Penguji II  
Penguji Utama : **Dr. Eng. Admi Syarif**

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



**Dr. Eng. Suropto Dwi Yuwono, S.Si., M.T.**  
NIP 19740705 200003 1 001

Tanggal Lulus Ujian Skripsi : 1 Juli 2021

## PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya yang berjudul **“Analisis *Tweet* Masyarakat Tentang Kemandirian Energi Menggunakan *Recurrent Neural Network* dan *Naive Bayes*”** merupakan karya saya sendiri dan bukan karya orang lain. Semua tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti skripsi saya merupakan hasil salinan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang telah saya terima.

Bandar Lampung, 2 Juli 2021



**MOHAMMAD SURYA AKBAR**  
NPM 1617051077

## **RIWAYAT HIDUP**

Penulis dilahirkan di Palembang pada tanggal 14 Oktober 1998, sebagai anak pertama dari dua bersaudara, dari Ayah Drs. Abdul Gani, dan Ibu Herlina, Amd. Penulis menyelesaikan pendidikan formal pertama kali di Taman Kanan-Kanak (TK) Merpati Pos Palembang pada tahun 2004. Pendidikan Sekolah Dasar (SD) di SD Negeri 17 Palembang diselesaikan pada tahun 2010, Sekolah Menengah Pertama (SMP) di SMP Negeri 17 Palembang pada tahun 2013, dan Sekolah Menengah Atas (SMA) di SMA Negeri 1 Palembang pada tahun 2016.

Pada tahun 2016, penulis terdaftar sebagai mahasiswa jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SBMPTN. Selama menjadi mahasiswa beberapa kegiatan yang dilakukan penulis antara lain.

1. Menjadi anggota Abacus Himpunan Mahasiswa Jurusan Ilmu Komputer pada periode 2016/2017.
2. Pada bulan Desember 2018 sampai dengan bulan Februari 2019 penulis melaksanakan Kerja Praktik di Dinas Kesehatan Provinsi Lampung.
3. Pada bulan Juli 2019 penulis melaksanakan kegiatan Kuliah Kerja Nyata (KKN) di desa Argomulyo, kecamatan Batu Ketulis, Kabupaten Lampung Barat.



## **PERSEMBAHAN**

Puji dan Syukur saya haturkan kepada Tuhan Yang Maha Esa atas segala berkat dan rahmat-Nya sehingga saya dapat menyelesaikan skripsi ini

Skripsi ini saya persembahkan untuk Ayah dan Ibu tercinta yang telah melahirkan, membesarkan, dan memberikan doa, dorongan serta dukungan kepada saya.

Terima kasih telah mendidik dan mendampingi saya dengan penuh kesabaran dan cinta kasih, juga atas segala pengorbanan yang diberikan untuk saya.

Terima kasih juga saya ucapkan kepada:

Ibu Bapak Dosen Pembimbing

Teman-teman Ilmu Komputer 2016

Almamater Tercinta, Universitas Lampung

## **MOTTO**

“Maka jangan sekali-kali membiarkan kehidupan dunia ini memperdayakan  
kamu.”

(Q.S. Fatir: 5)

“Seseorang bertindak tanpa ilmu ibarat bepergian tanpa petunjuk. Dan sudah  
banyak yang tahu kalau orang seperti itu kiranya akan hancur, bukan selamat.”

(Hasan Al Bashri)

*“Success consists of going from failure to failure without loss of enthusiasm.”*

(Winston Churchill)

## SANCAWANA

Puji syukur kita haturkan pada Allah SWT atas segala berkat-Nya penulis dapat menyelesaikan skripsi yang berjudul “Analisis *Tweet* Masyarakat Tentang Kemandirian Energi Menggunakan *Recurrent Neural Network* dan *Naive Bayes*” dengan Baik.

Penulis menyadari selesainya skripsi ini tidak terlepas dari partisipasi bimbingan serta bantuan dari berbagai pihak baik secara langsung maupun tidak langsung. Maka kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Ayah dan Ibu tercinta, Drs. Abdul Gani, dan Herlina Amd. yang selalu memberi dukungan, memotivasi, dan menyemangati penulis selama proses perkuliahan sampai dengan penyusunan skripsi. Semoga Allah SWT selalu menyertai, memberkati, dan memberi Kesehatan dan kebahagiaan yang berlimpah.
2. Bapak Dr. Ir. Kurnia Muludi, M.S.Sc selaku pembimbing utama yang telah meluangkan banyak waktu dan dengan sabar membimbing penulis, serta memberikan banyak dukungan, motivasi, dan dorongan untuk menyelesaikan skripsi ini. Terimakasih juga penulis ucapkan atas kritik dan saran yang membangun sehingga penulisan skripsi ini dapat diselesaikan.

3. Ibu Dewi Asiah Shofiana, S.Kom., M.Kom selaku pembimbing kedua yang telah meluangkan banyak waktu dan dengan sabar membimbing penulis, serta memberikan ide, kritik dan saran yang membangun sehingga penulisan skripsi ini dapat diselesaikan.
4. Bapak Dr. Eng. Admi Syarif selaku pembahas yang telah memberikan banyak masukan, serta ilmu dan pengetahuan baru yang bermanfaat dalam perbaikan skripsi ini.
5. Bapak Aristoteles, S.Si., M.Si. selaku pembimbing akademik yang telah membimbing penulis selama proses perkuliahan, serta memberikan masukan dan dukungan hingga skripsi ini dapat diselesaikan.
6. Bapak Didik Kurniawan, S.Si., M.T. selaku Ketua Jurusan Ilmu Komputer Universitas Lampung.
7. Bapak Dr. Eng. Suropto Dwi Yuwono, S.Si., M.T. selaku Dekan FMIPA Universitas Lampung.
8. Bapak dan Ibu Dosen Jurusan Ilmu Komputer FMIPA Universitas Lampung yang tak bisa disebutkan satu per satu, atas bimbingan dan pengajarannya selama penulis menjadi mahasiswa FMIPA Universitas Lampung.
9. Seluruh Staf dan karyawan Fakultas MIPA Universitas Lampung: Ibu Ade Nora Maela, Bang Zainuddin, Mas Syam, Mas Ardi Novalia, dan lainnya yang tidak bisa penulis sebutkan satu per satu, yang telah membantu segala urusan administrasi penulis.
10. Ibu Noviani selaku Ka. Seksi Datin dan Litbang Kes. yang telah mengizinkan penulis untuk melakukan Praktik Kerja Lapangan (PKL) di Dinas Kesehatan Provinsi Lampung, selaku pembimbing lapangan yang

telah membimbing dan memberikan banyak ilmu bermanfaat, dan juga para staff Dinas Kesehatan lainnya yang telah menyambut serta memperlakukan penulis dengan sangat baik.

11. Kepala Pekon (Desa) Argomulyo dan aparat desa lainnya yang telah mengizinkan penulis melakukan kegiatan Kuliah Kerja Nyata di Desa Argomulyo, Kecamatan Batu Ketulis, Kabupaten Lampung Barat. Bapak Yadi dan Ibu Yana, para tetangga dan seluruh warga desa Argomulyo yang telah menyambut dan memperlakukan penulis dan teman-teman lainnya dengan sangat baik.
12. Nata, saudara dan keluarga besar yang telah memberi semangat dan dukungan moril selama penyusunan skripsi.
13. Teman-teman terdekat, Aditya Fajrianto, Rachel, Aditya Bimantoro, dan Arief yang telah memberikan banyak dukungan moril, segala bentuk bantuan, dan selalu menemani dari awal perkuliahan.
14. Teman-teman dari komunitas JOS Gaming Community yang telah memberikan semangat, dan dukungan selama pengerjaan skripsi ini.
15. Keluarga Ilmu Komputer 2016 serta kakak tingkat dan adik tingkat yang tidak bisa penulis sebutkan satu per satu.
16. Almamater tercinta, Universitas Lampung yang sudah memberi banyak wawasan dan pengalaman berharga.

Semoga skripsi ini dapat berguna dan bermanfaat bagi agama, masyarakat, bangsa dan negara, para mahasiswa, akademisi, serta pihak-pihak lain yang membutuhkan terutama penulis. Saran dan kritik yang bersifat membangun sangat

diharapkan. Akhir kata penulis ucapkan terima kasih. Semoga Allah SWT senantiasanya memberikan perlindungan dan kebaikan bagi kita semua.

Bandar Lampung, 2 Juli 2021  
Penulis

Mohammad Surya Akbar  
1617051077

## DAFTAR ISI

	Halaman
<b>DAFTAR TABEL .....</b>	<b>iii</b>
<b>DAFTAR GAMBAR.....</b>	<b>iv</b>
<b>I. PENDAHULUAN</b>	
1.1. Latar Belakang .....	1
1.2. Rumusan Masalah .....	4
1.3. Batasan Masalah.....	4
1.4. Tujuan.....	4
1.5. Manfaat.....	5
<b>II. TINJAUAN PUSTAKA</b>	
2.1. <i>Text Mining</i> .....	6
2.2. Analisis sentimen .....	7
2.2.1. Metode dan Fitur .....	8
2.2.2. Penggunaan Analisis Sentimen.....	9
2.3. <i>Deep Learning</i> .....	9
2.4. <i>Recurrent Neural Network</i> .....	11
2.5. Twitter .....	13
2.6. <i>Word2vec</i> .....	14
2.7. <i>POS Tagging</i> .....	15
2.8. Bahasa Pemrograman Python.....	16
2.9. Gensim .....	17
2.10. <i>Naive Bayes Classifier</i> .....	18

2.11. Kemandirian Energi .....	19
2.12. Penelitian Terdahulu .....	20

### **III. METODE PENELITIAN**

3.1. Tempat dan Waktu Penelitian .....	23
3.2. Alat dan Bahan .....	23
3.2.1. Perangkat Keras ( <i>Hardware</i> ) .....	23
3.2.2. Perangkat Lunak ( <i>Software</i> ).....	23
3.3. Tahapan Penelitian .....	24

### **IV. HASIL DAN PEMBAHASAN**

4.1. Pengumpulan Data .....	31
4.2. Pembagian data .....	32
4.3. Normalisasi Teks.....	33
4.4. Tokenisasi.....	33
4.5. Pembuangan <i>Stopword</i> .....	34
4.6. Hasil <i>Word2vec</i> .....	35
4.7. <i>Training</i> Klasifikasi.....	36
4.8. Evaluasi Hasil Klafisifikasi.....	40

### **V. SIMPULAN DAN SARAN**

5.1. Simpulan.....	46
5.2. Saran.....	47

### **DAFTAR PUSTAKA**

### **LAMPIRAN**



**DAFTAR TABEL**

Tabel	Halaman
1. Penelitian terdahulu.....	21
2. Tabel confusion matrix .....	29
3. Pengumpulan data Twitter .....	31
4. Pembagian data Twitter.....	32
5. Hasil pengujian klasifikasi Simple Recurrent Neural Network .....	41
6. Confusion matrix Simple Recurrent Neural Network.....	41
7. Hasil pengujian klasifikasi Long Short-Term Memory .....	42
8. Confusion matrix Long Short-Term Memory.....	42
9. Confusion matrix Bernoulli Naive Bayes .....	43
10. Hasil perbandingan hasil perhitungan klasifikasi .....	43
11. Perbandingan nilai confusion matrix .....	44

## DAFTAR GAMBAR

Gambar	Halaman
1. Simple recurrent neural network .....	12
2. Tahapan penelitian .....	24
3. Dataset yang digunakan untuk penelitian .....	32
4. Hasil normalisasi teks .....	33
5. Hasil dari tahap tokenisasi .....	34
6. Nilai vektor dari hasil latih word2vec .....	35
7. Grafik training Simple Recurrent Neural Network .....	38
8. Grafik training LSTM .....	39

## **I. PENDAHULUAN**

### **1.1. Latar Belakang**

Internet merupakan alat elektronik yang menjadi salah satu kebutuhan utama bagi masyarakat. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) mengungkapkan jumlah pengguna internet di Indonesia mencapai 196.71 juta orang pada tahun 2019 sampai Juni 2020. Internet dapat digunakan sebagai komunikasi, sumber informasi, media sosial, dan sebagainya untuk mengikuti perkembangan zaman. Berbagai opini masyarakat Indonesia dapat ditemukan di media sosial, dan salah satu bahasan masyarakat yang populer yaitu kemandirian energi. Kemandirian energi merupakan salah satu rencana pembangunan berkelanjutan dari pemerintahan Republik Indonesia untuk pemerataan energi di wilayah yang belum menerima sumber energi yang baik di Indonesia (ESDM, 2019)

Twitter merupakan salah satu media sosial yang digunakan sebagai alat komunikasi, media untuk promosi, dan kampanye politik. Twitter merupakan media sosial yang memiliki karakteristik dan format unik dengan simbol ataupun aturan khusus karena pengguna Twitter hanya dapat mengirim dan membaca pesan dengan batasan maksimal 140 karakter yang

diketahui dengan istilah *tweet* (Zhang et al., 2011). *Tweet* itu dapat berupa pendapat, saran, ataupun kritikan mengenai topik tertentu yang beraneka ragam dari pengguna Twitter. Keanekaragaman *tweet* tersebut serta banyaknya penggunaan bahasa yang tidak baku pada *tweet* menjadi alasan diperlukan analisis sentimen.

Analisis sentimen merupakan salah satu cabang dari *text mining* yang melakukan identifikasi teks dan kemudian mengekstrak informasi dari teks yang telah diidentifikasi menjadi informasi subjektif dalam sumber (Soong et al., 2019). Analisis sentimen juga berfokus pada pengelola opini yang mengandung polaritas, yaitu memiliki nilai sentimen positif ataupun negatif (Novantirani et al., 2014). Masalah yang ada dalam analisis sentimen biasanya sulit dalam mendefinisikan, menentukan konsep masalah, sub masalah, dan tujuan yang berfungsi sebagai kerangka kinerja dalam berbagai penelitian (Liu, 2010).

Analisis sentimen dalam pengumpulan data statistik menggunakan berbagai algoritme yang berasal dari cabang ilmu *Artificial Intelligence*, seperti *Deep Learning* dan algoritme *Machine Learning* yang terdiri dari *Naive Bayes*, *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN), regresi, dan sebagainya. Terdapat banyak penelitian terkait analisis sentimen menggunakan algoritme *Naive Bayes*. Keuntungan dari algoritme *Naive Bayes* adalah tidak memerlukan jumlah data yang banyak, tidak perlu memproses data latih yang banyak, serta perhitungannya cepat dan efisien (Saranya dan Jayanthi, 2018).

*Recurrent Neural Network* (RNN) merupakan salah satu algoritme yang masuk dalam kategori *Deep Learning* karena data diproses melalui banyak lapisan (*layer*). RNN telah mengalami kemajuan yang cepat dan merevolusi bidang-bidang seperti pemrosesan bahasa alami, pemrosesan data finansial seri waktu, dan sebagainya (Karpathy, 2015). *Deep learning* dapat digunakan untuk analisis sentimen, karena *Deep Learning* telah diaplikasikan ke dalam *Natural Language Processing* (NLP) (Socher et al., 2013).

Terdapat beberapa penelitian terdahulu mengenai analisis sentimen dengan menerapkan *Deep Learning* yang telah dilakukan. Salah satu penelitian yang telah dilakukan oleh (Zulfa dan Winarko, 2017) melakukan analisis sentimen *tweet* berbahasa Indonesia. Penelitian ini menggunakan algoritme yang termasuk dalam kategori *Deep Learning*, yaitu *Deep Belief Network* (DBN). DBN menggunakan metode lapisan atau tumpukan dari beberapa algoritme dengan *feature extraction* yang memanfaatkan seluruh *resource* semaksimal mungkin. Hasil dari penggunaan algoritme DBN ini menunjukkan nilai akurasi yang tinggi untuk analisis sentimen yaitu 93.31%.

Berdasarkan pemaparan tersebut, maka penelitian ini akan menerapkan *Deep Learning* untuk menganalisis dan mengklasifikasikan data dengan topik *tweet* mengenai kemandirian energi di Indonesia dari media sosial Twitter. Pada penelitian ini juga akan dilakukan perbandingan akurasi,

presisi, *recall*, dan *f1-score* antara dua algoritme, yaitu algoritme *Recurrent Neural Network* (RNN) dengan algoritme *Naive Bayes*.

## 1.2. Rumusan Masalah

Rumusan dalam penelitian ini yaitu:

1. Bagaimana pengklasifikasian sentimen negatif, dan sentimen positif terhadap data yang didapatkan dari Twitter dengan menggunakan algoritme *Recurrent Neural Network*?
2. Bagaimana tingkat akurasi algoritme *Recurrent Neural Network* terhadap data pada analisis sentimen dibandingkan dengan algoritme *Naive Bayes*?

## 1.3. Batasan Masalah

Penelitian ini dibatasi oleh beberapa hal, yang ditentukan sebagai berikut:

1. Menggunakan studi kasus mengenai kemandirian energi di Indonesia.
2. Proses pembuatan model menggunakan data *tweet* menggunakan Twitter API yang merupakan *library* untuk menarik *tweet*.
3. Data *tweet* yang digunakan hanya *tweet* yang menyebutkan kata kunci yang berhubungan dengan kemandirian energi di Indonesia.
4. *Tweet* yang digunakan berbahasa Indonesia.
5. Algoritme klasifikasi yang digunakan adalah *Recurrent Neural Network* dan *Naive Bayes*.

#### 1.4. Tujuan

Tujuan penelitian yang akan dicapai sebagai berikut:

1. Melakukan pengklasifikasian terhadap data dari Twitter dengan menggunakan algoritme *Recurrent Neural Network*.
2. Menguji akurasi, presisi, *recall*, dan *f1-score* algoritme *Recurrent Neural Network* dan *Naive Bayes* pada analisis sentimen dan melakukan perbandingan antara keduanya.

#### 1.5. Manfaat

Manfaat yang ingin dicapai dalam penelitian ini yaitu:

1. Mengetahui cara mengklasifikasikan sentimen pada data *tweet* menggunakan algoritme *Recurrent Neural Network* dalam melakukan analisis sentimen.
2. Mengetahui keunggulan antara dua algoritme (*Recurrent Neural Network* dan *Naive Bayes*) dalam melakukan analisis sentimen.

## II. TINJAUAN PUSTAKA

### 2.1. *Text Mining*

*Text mining* adalah penemuan informasi baru yang sebelumnya tidak diketahui oleh komputer, dengan mengekstraksi informasi dari berbagai sumber tertulis (Hotho et al., 2005). Sumber data dapat berupa situs web, buku, email, ulasan, dan artikel. *Text mining* juga disebut sebagai *text data mining* (penambangan data teks), yaitu proses memperoleh informasi berkualitas tinggi dari teks. Informasi berkualitas tinggi dapat diperoleh melalui perancangan pola dan tren melalui cara-cara seperti pembelajaran pola statistik, dan pengenalan pola. *Text mining* melibatkan proses penataan teks (*parsing*, penambahan beberapa fitur turunan linguistik, serta penyisipan data teks berikutnya ke dalam suatu *database*), memperoleh pola dalam data terstruktur, serta evaluasi dan interpretasi hasil *output*. Tugas *text mining* meliputi pengelompokan teks, ekstraksi konsep / entitas, analisis sentimen, peringkasan dokumen, dan pemodelan hubungan (Hotho et al., 2005).

Analisis teks melibatkan pencarian informasi, analisis makna kata untuk mempelajari distribusi frekuensi kata, pengenalan pola, penandaan / anotasi,



ekstraksi informasi, teknik penambangan data, termasuk analisis tautan dan asosiasi, visualisasi, dan analitik prediktif. Tujuannya adalah untuk mengubah teks menjadi data untuk dianalisis. Untuk data yang dianalisis dapat melalui aplikasi *Natural Language Processing* (NLP), berbagai jenis algoritme, dan metode analitik. Fase penting dari proses ini adalah penafsiran informasi yang dikumpulkan (Feldman dan Sanger, 2007).

## **2.2. Analisis sentimen**

Analisis sentimen (juga dikenal sebagai *opinion mining*) mengacu pada penggunaan pemrosesan bahasa alami, analisis teks, linguistik komputasi, dan biometrik untuk mengidentifikasi, mengekstraksi, mengukur, dan mempelajari status afektif dan informasi subjektif secara sistematis (Saranya dan Jayanthi, 2018). Analisis sentimen secara luas diterapkan untuk memberikan informasi seperti ulasan dan tanggapan survei, media *online* dan sosial, dan bahan perawatan kesehatan untuk aplikasi yang berkisar dari pemasaran, layanan pelanggan hingga kedokteran klinis (Taboada et al., 2011).

Tugas dasar dalam analisis sentimen adalah mengklasifikasikan polaritas teks yang diberikan pada tingkat dokumen, kalimat, atau fitur apakah opini yang diungkapkan dalam dokumen, kalimat atau fitur entitas adalah positif, negatif, atau netral (Koppel dan Schler, 2006). Ada berbagai jenis analisis sentimen seperti analisis sentimen berbasis aspek, analisis sentimen

penilaian (positif, negatif, netral), serta analisis sentimen multibahasa dan deteksi emosi.

### 2.2.1. Metode dan Fitur

Pendekatan yang ada untuk analisis sentimen dapat dikelompokkan ke dalam tiga kategori utama, yaitu *knowledge-based techniques*, *statistical methods*, dan *hybrid approaches* (Cambria et al., 2013). *Knowledge-based techniques* mengklasifikasikan teks dengan mempengaruhi kategori berdasarkan pada kehadiran kata-kata yang mempengaruhi seperti senang, sedih, takut, dan bosan (Ortony et al., 2011). Beberapa *knowledge-based* tidak hanya mencantumkan kata-kata yang mempengaruhi secara jelas, tetapi juga menetapkan kata-kata yang mungkin berhubungan dengan emosi tertentu.

*Statistical method* memanfaatkan elemen dari *machine learning* seperti *support vector machines* (SVM), *bag-of-words*, *latent semantic analysis*, *pointwise mutual information* untuk *semantic operations*, dan *deep learning*. Untuk menambang pendapat dalam konteks dan mendapatkan fitur tentang pendapat pembicara, hubungan tata bahasa kata digunakan. Hubungan ketergantungan *grammatical* diperoleh oleh *deep parsing* teks (Dey dan Haque, 2009). *Parsing* atau *syntax analysis* adalah proses menganalisis serangkaian simbol, baik dalam bahasa alami, bahasa komputer, atau struktur data, sesuai dengan aturan tata bahasa.

*Hybrid approaches* memanfaatkan *machine learning* dan elemen dari representasi pengetahuan seperti ontologi dan *semantic network* untuk mendeteksi makna kata yang diekspresikan secara halus. Misalnya melalui analisis dapat mengetahui konsep penyampaian informasi yang tidak relevan (Chaturvedi et al., 2015).

### 2.2.2. Penggunaan Analisis Sentimen

Penggunaan analisis sentimen dalam kehidupan sehari-hari yang sering digunakan adalah sistem rekomendasi (Hu dan Liu, 2004). Penggunaan rekomendasi banyak digunakan dalam beberapa aplikasi seperti untuk rekomendasi film, restoran, tempat liburan, dan sebagainya. Sistem ini menggunakan analisis sentimen dengan memanfaatkan komentar-komentar dari pengguna aplikasi atau pendapat orang dari berbagai sumber seperti media sosial dan internet.

Pada penelitian ini, *knowledge-based techniques* dan *statistical methods* akan digunakan untuk membuat model sentimen karena dapat melakukan kategorisasi sentimen. Teknik klasifikasi *knowledge-based* dan *statistical methods* dapat melakukan kategori berdasarkan kemunculan kata, dan menganalisis struktur bahasa.

## 2.3. *Deep Learning*

*Deep learning* adalah bidang ilmu yang muncul dari pembelajaran mesin (*machine learning*), yang terdiri dari beberapa lapisan tersembunyi dari jaringan saraf tiruan. Metodologi *deep learning* menerapkan transformasi

non-linear dan abstraksi model tingkat tinggi dalam basis data besar. Kemajuan terbaru dalam arsitektur *deep learning* di berbagai bidang telah memberikan kontribusi signifikan dalam kecerdasan buatan (*artificial intelligence*). Kecerdasan buatan sebagai kecerdasan yang dimiliki oleh mesin telah menjadi pendekatan yang efektif untuk pembelajaran dan penalaran manusia. Pada tahun 1950, "The Turing Test" diusulkan sebagai penjelasan yang memuaskan tentang bagaimana komputer dapat melakukan penalaran kognitif manusia (Vargas dan Ruiz, 2018).

Konsep *deep learning* muncul untuk pertama kalinya pada tahun 2006 sebagai bidang penelitian baru dalam pembelajaran mesin. *Deep learning* pertama kali dikenal sebagai pembelajaran hierarkis, dan biasanya melibatkan banyak bidang penelitian yang berkaitan dengan pengenalan pola (Lecun et al., 2015). *Deep learning* mempertimbangkan dua faktor utama: pemrosesan non-linear dalam berbagai lapisan atau tahapan dan pembelajaran yang diawasi atau tidak diawasi. Hierarki dibangun di antara lapisan-lapisan untuk mengatur pentingnya data agar dianggap berguna atau tidak (Bengio, 2009).

Untuk menggunakan analisis sentimen, aplikasi *deep learning* yang akan digunakan adalah *Natural Language Processing* (NLP). *Neural networks* telah digunakan untuk mengimplementasikan model bahasa sejak awal tahun 2000. Teknik penyisipan kata seperti word2vec dapat dianggap sebagai lapisan representasi dalam arsitektur *deep learning* yang mengubah kata atom menjadi representasi posisi kata relatif terhadap kata lain dalam

*dataset* dan posisi direpresentasikan sebagai titik dalam ruang vektor. Penggunaan vektor sebagai lapisan input RNN memungkinkan jaringan untuk mengurai kalimat dan frasa menggunakan tata bahasa vektor komposisi yang efektif. Tata bahasa vektor komposisi dapat dianggap sebagai tata bahasa bebas konteks probabilistik atau *probabilistic context free grammar* (PCFG) yang diterapkan oleh RNN.

Pada penelitian ini, NLP akan digunakan untuk dasar implementasi dari analisis sentimen. NLP dapat digunakan untuk penandaan kelas kata (*part-of-speech tagging*) dan disambiguasi makna kata (*word sense disambiguation*).

#### **2.4. Recurrent Neural Network**

*Recurrent neural network* (RNN) adalah kelas *artificial neural network* yang menunjukkan koneksi antara titik yang membentuk grafik yang diarahkan sepanjang urutan. RNN dapat menggunakan status internal (memori) untuk memproses urutan panjang variabel input. Oleh karena itu, RNN dapat digunakan untuk menyelesaikan masalah yang berkaitan dengan *natural language processing* (NLP) seperti pengenalan tulisan dan pengenalan ucapan (Graves et al., 2009).

RNN menggunakan arsitektur yang biasa disebut *simple recurrent neural network* atau *Elman network*. Arsitektur ini merupakan versi sederhana dari *recurrent neural network*, dan sangat mudah diimplementasikan dan dilatih. *Network* ini memiliki lapisan input  $x$ , lapisan *context*  $s$ , dan lapisan *output*  $y$ .

Input *network* dalam *timestep*  $t$  adalah  $x(t)$ , *output network* dilambangkan dengan  $y(t)$ , dan  $s(t)$  adalah *state network (hidden layer)*. Vektor input  $x(t)$  dibentuk oleh vektor  $w$  yang mewakili kalimat atau kata yang dilatih, dan *output* dari *neuron* di lapisan *context*  $s$  pada saat  $t - 1$ . Lapisan input, *context*, dan *output* dihitung dengan persamaan (Mikolov et al., 2010):

$$x(t) = w(t) + s(t - 1) \quad (1)$$

$$s_j(t) = f \left( \sum_i x_i(t) u_{ji} \right) \quad (2)$$

$$y_k(t) = g \left( \sum_j s_j(t) u_{kj} \right) \quad (3)$$

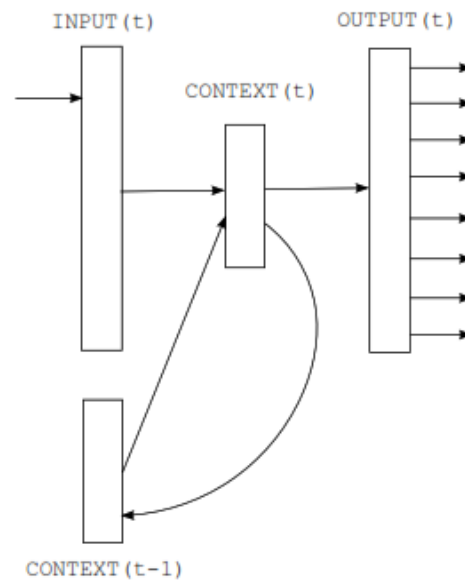
dengan  $f(z)$  adalah fungsi aktivasi *sigmoid*:

$$f(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

dan  $g(z)$  adalah fungsi *softmax*:

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \quad (5)$$

Berdasarkan persamaan (1), (2), dan (3), maka cara kerja dari *simple recurrent neural network* diilustrasikan pada Gambar 1.



Gambar 1. *Simple recurrent neural network*

Jumlah *context* atau *hidden layer* harus sama dengan jumlah data latih. Untuk data dalam jumlah besar, *hidden layer* dalam jumlah besar sangat diperlukan. *Output layer*  $y(t)$  merupakan distribusi probabilitas kata berikutnya dari kata sebelumnya  $w(t)$  dan *context*  $s(t - 1)$ . *Softmax* memastikan bahwa distribusi probabilitas ini valid, yaitu  $y_m(t) > 0$  untuk kata apapun dari  $m$  dan  $\sum_k y_k(t) = 1$ .

## 2.5. Twitter

Twitter adalah layanan mikroblog dan jejaring sosial Amerika tempat pengguna mengunggah dan berinteraksi dengan pesan yang dikenal sebagai "*tweets*". *Tweet* awalnya dibatasi hingga 140 karakter, tetapi pada tanggal 7 November 2017, batas ini digandakan menjadi 280 untuk semua bahasa kecuali Cina, Jepang, dan Korea. Pengguna terdaftar dapat mengunggah, menyukai, dan me-*retweet tweet*, tetapi pengguna yang tidak terdaftar hanya

dapat membacanya. Pengguna dapat mengakses Twitter melalui antarmuka situs webnya, melalui *Short Message Service* (SMS) atau perangkat lunak aplikasi perangkat selulernya. Twitter, Inc. berbasis di San Francisco, California, dan memiliki lebih dari 25 kantor di seluruh dunia (Kelly, 2009).

Twitter didirikan pada Maret 2006 oleh Jack Dorsey, Noah Glass, Biz Stone, dan Evan Williams, dan diluncurkan pada bulan Juli di tahun yang sama. Layanan ini dengan cepat mendapatkan popularitas di seluruh dunia. Pada tahun 2012, lebih dari 100 juta pengguna mengunggah 340 juta *tweet* per hari, dan Twitter telah menangani rata-rata 1,6 miliar permintaan pencarian per hari. Pada tahun 2013, Twitter menjadi salah satu dari sepuluh situs web yang paling banyak dikunjungi dan digambarkan sebagai "SMS Internet". Pada tahun 2018, Twitter memiliki lebih dari 321 juta pengguna (Holton et al., 2014).

Pada penelitian ini, Twitter digunakan sebagai media tempat pengunduhan data yang akan digunakan. Untuk pengambilan data akan menggunakan Twitter *API* yang merupakan *library* Twitter untuk pemrograman yang dapat mengambil data dari Twitter.

## 2.6. *Word2vec*

*Word2vec* adalah sekelompok model yang saling terkait yang digunakan untuk menghasilkan vektor kata. *Word2vec* mengambil kata-kata dari dokumen atau kalimat sebagai masukan kumpulan teks yang banyak dan menghasilkan ruang vektor, biasanya beberapa ratus dimensi, dengan setiap



kata unik dalam kumpulan tulisan ditugaskan vektor yang sesuai dalam ruang. Vektor kata diposisikan dalam ruang vektor sehingga kata-kata yang berbagi konteks umum dalam kumpulan tulisan terletak berdekatan satu sama lain dalam ruang tersebut (Mikolov et al., 2013).

*Word2vec* dibuat dan diterbitkan pada tahun 2013 oleh tim peneliti yang dipimpin oleh Tomas Mikolov di Google dan dipatenkan. Algoritme ini kemudian dianalisis dan dijelaskan oleh peneliti lain. Vektor kata yang dibuat menggunakan algoritme *Word2vec* memiliki banyak keunggulan dibandingkan dengan algoritme sebelumnya seperti *latent semantic analysis*. Keunggulan dari *word2vec* adalah memiliki ekstensi untuk menyimpan kata-kata dari seluruh kalimat yang disebut dengan *paragraf2vec* atau *doc2vec*. *Word2vec* juga telah diimplementasikan dalam beberapa bahasa pemrograman yaitu C, Python, dan Java yang dapat mendukung inferensi penyimpanan dokumen pada dokumen tak terlihat (Le dan Mikolov, 2014).

*Word2vec* dapat menggunakan salah satu dari dua model arsitektur untuk menghasilkan representasi kata-kata yang terdistribusi: bagan kata-kata kontinu (CBOW) atau skip-gram kontinu. Dalam arsitektur *bag-of-words* yang berkelanjutan, model memprediksi kata saat ini dari jendela kata konteks sekitarnya. Urutan kata konteks tidak memengaruhi prediksi (asumsi *bag-of-words*). Dalam arsitektur *skip-gram*, model menggunakan kata saat ini untuk memprediksi jendela kata konteks di sekitarnya.

Arsitektur *skip-gram* menimbang kata-kata konteks terdekat lebih berat daripada kata-kata konteks lebih jauh (Mikolov et al., 2013).

Pada penelitian ini, *Word2vec* digunakan untuk penghitungan kemunculan kata dan menyimpan kata yang muncul dari keseluruhan data. *Word2vec* hanya melakukan penghitungan kata yang muncul dari setiap data dan yang paling sering muncul. *Word2vec* dapat membantu dalam pengkategorian nilai sentimen dari data.

## **2.7. POS Tagging**

Dalam *corpus linguistics*, *part-of-speech tagging* (POS tagging atau PoS tagging atau POST), juga disebut penandaan gramatikal atau disambiguasi kata-kategori, adalah proses menandai kata dalam teks (*corpus*) sesuai dengan bagian tertentu pidato, berdasarkan definisi dan konteksnya yaitu hubungannya dengan kata-kata yang berdekatan dan terkait dalam frasa, kalimat, atau paragraf. Bentuk yang disederhanakan dari ini umumnya diajarkan kepada anak-anak usia sekolah, dalam mengidentifikasi kata-kata sebagai kata benda, kata kerja, kata sifat, kata keterangan, dan lain-lain (Güngör, 2010).

Pada penelitian ini, POS Tagging digunakan untuk melakukan penandaan kata yang paling sering muncul. Dari kata yang paling sering muncul dari setiap data, kata tersebut akan ditandai sebagai kata yang paling sering muncul dari fungsi *Word2vec* dan akan menjadi acuan untuk nilai sentimen.

## 2.8. Bahasa Pemrograman Python

Python adalah bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang mengacu pada tingkat keterbukaan kode. Python diciptakan oleh Guido van Rossum dan dirilis pertama kali pada tahun 1991. Filosofi desain Python menekankan pembacaan kode dengan penggunaan spasi yang signifikan. Bahasa dan pendekatan berorientasi objek Python bertujuan untuk membantu *programmer* menulis kode yang jelas dan logis untuk proyek skala kecil dan besar (Gutttag, 2016).

Python diketik dan dikumpulkan secara dinamis. Hal ini mendukung beberapa paradigma pemrograman, termasuk pemrograman prosedural, berorientasi objek, dan fungsional. Python sering digambarkan sebagai bahasa "termasuk baterai" karena pustaka standarnya yang komprehensif (Kuhlman, 2012).

Python merupakan bahasa pemrograman *multi-paradigm* dengan maksud untuk meraih berbagai jenis tujuan dengan metode yang beragam. Pemrograman berorientasi objek dan pemrograman terstruktur sepenuhnya didukung, dan banyak fitur-fiturnya mendukung pemrograman fungsional dan pemrograman berorientasi aspek termasuk oleh *metaprogramming* dan *metaobjects*. Banyak paradigma lain didukung melalui ekstensi, termasuk desain berdasarkan kontrak dan pemrograman logika (Sarkar, 2019).

## 2.9. Gensim

Gensim adalah *open-source library* untuk *unsupervised topic modeling* dan *natural language processing* menggunakan *modern statistical machine learning*. Gensim diimplementasikan dalam Python dan Cython, dan dirancang untuk menangani koleksi teks besar menggunakan *streaming data* dan algoritme *online* tambahan, yang membedakannya dari sebagian besar paket perangkat lunak *machine learning* lainnya yang hanya menargetkan pemrosesan dalam memori (Rehurek dan Sojka, 2010). Gensim mencakup implementasi paralel dari algoritme *FastText*, *Word2vec* dan *Doc2vec*, yang sebaik *Latent Semantic Analysis (LSA)*, *non-negative matrix factorization (NMF)*, *Latent Dirichlet Allocation (LDA)*, *tf-idf*, dan *random projections* (Řehůřek, 2011).

Pada penelitian ini, Gensim akan diimplementasikan dalam pemrograman Python dan akan digunakan untuk mengatasi data yang memiliki nilai teks yang besar dan juga untuk membantu penggunaan *word2vec*.

## 2.10. Naive Bayes Classifier

*Naive Bayes Classifier (NBC)* merupakan salah satu metode pembelajaran mesin yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris bernama Thomas Bayes. Metode ini memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. *Naive Bayes Classifier* mengasumsikan bahwa nilai fitur

tertentu tidak tergantung pada nilai fitur lain, karena variabel kelas (Singh et al., 2019).

Algoritme *Naive Bayes* yang digunakan adalah model *Bernoulli Naive Bayes*. Model *Bernoulli* ini menggunakan fitur *Boolean* (variabel biner) untuk nilai input. Model ini banyak digunakan untuk tugas klasifikasi dokumen dengan fitur kemunculan istilah biner digunakan untuk melakukan klasifikasi kalimat yang ada di dalam suatu dokumen. Persamaan *Bernoulli* dapat dihitung dengan persamaan (Singh et al., 2019):

$$p(X|C_k) = \prod_{i=1}^n p_{ki}^{x_i} (1 - p_{ki})^{1-x_i} \quad (6)$$

Dengan adanya persamaan (6),  $p(x|C_k)$  merupakan peluang dokumen  $X$  nilai vektor dari setiap *class*  $C_k$  yang merupakan *document class*. Kemudian  $x_i$  adalah nilai *boolean* dari kemunculan nilai  $i$  dari kosakata dengan  $C_k$ , dan  $p_{ki}$  adalah nilai *probability* dari *class*  $C_k$  dan menghasilkan nilai  $x_i$ . Model ini sering digunakan untuk mengklasifikasikan data teks pendek.

## 2.11. Kemandirian Energi

Kemandirian energi adalah energi yang tidak memanfaatkan energi alam yang disediakan oleh bumi, tetapi energi buatan yang dikembangkan oleh manusia. Sama seperti energi alternatif, kemandirian energi adalah istilah yang merujuk kepada semua sumber energi yang dapat digunakan yang bertujuan untuk menggantikan bahan bakar konvensional tanpa akibat yang tidak diharapkan dari hal tersebut. Kemandirian energi digunakan untuk

mengurangi penggunaan bahan bakar hidrokarbon yang mengakibatkan kerusakan lingkungan akibat emisi karbon dioksida yang tinggi, yang berkontribusi besar terhadap pemanasan global (UNDESA, 2019).

Kemandirian energi ini adalah energi yang berkelanjutan yang memenuhi kebutuhan saat ini tanpa mengorbankan kemampuan generasi mendatang untuk memenuhi kebutuhan. Energi berkelanjutan yang termasuk sumber energi terbarukan, seperti pembangkit listrik tenaga air, energi surya, energi angin, tenaga ombak, energi panas bumi, fotosintesis buatan, tenaga pasang surut, dan juga teknologi yang dirancang untuk meningkatkan efisiensi energi. Kontribusi penting untuk menuju kemandirian energi adalah efisiensi energi karena penggunaan energi yang efisien dapat dibangun berdasarkan upaya individu dalam penghematan daya yang tidak harus bergantung pada infrastruktur skala besar yang mahal.

Di Eropa, terdapat harapan untuk lebih mandiri dan tidak bergantung lagi terhadap suplai energi (minyak dan gas) dari Rusia. Begitu juga di Amerika Serikat yang berharap terbebas dari impor minyak yang diproduksi oleh negara lain. Berdasarkan sudut pandang ini, gas alam, dan bahan bakar fosil adalah energi alternatif terhadap bahan bakar yang diimpor dari luar. Ini adalah sudut pandang T. Boone Pickens yang menjelaskan Pickens Plan untuk kemandirian energi, dan merefleksikan undang-undang di Negara Bagian Florida, Amerika Serikat. Meski gas alam tidaklah dapat diperbarui, tetapi dalam sudut pandang ini, hal tersebut adalah energi alternatif (Bluszcz, 2017).

Pada penelitian ini, kemandirian energi digunakan sebagai topik dari penelitian analisis sentimen ini. Data dari Twitter yang berkaitan dengan kemandirian energi akan digunakan untuk melakukan klasifikasi analisis sentimen.

### **2.12. Penelitian Terdahulu**

Penelitian terkait dengan *text mining* untuk masalah analisis sentimen telah banyak dilakukan sebelumnya dengan metode yang berbeda-beda. Berikut ini adalah penelitian terdahulu yang berkaitan dengan analisis sentimen, yang dapat dilihat pada Tabel 1.

Tabel 1. Penelitian terdahulu

Perbandingan teori	Peneliti (tahun penelitian)	Tujuan	Langkah-langkah	Metode	Output
<i>Twitter Sentiment Analysis Terhadap Brand Reputation: Studi Kasus PT XL Axiata Tbk</i>	(Vidya, 2015)	Mencari algoritme terbaik dari <i>Twitter sentiment analysis</i> dalam mengkategorisasi sentimen positif dan negatif yang akan digunakan sebagai variabel perhitungan <i>Net Brand Reputation</i>	<ol style="list-style-type: none"> <li>Pengumpulan awal data</li> <li>Pengolahan <i>dataset</i> dan data <i>training</i></li> <li>Kategorisasi data <i>testing</i></li> <li>Perhitungan reputasi <i>brand</i></li> </ol>	Algoritme klasifikasi <i>Naive Bayes</i> , <i>support vector machine</i> dan <i>decision tree</i> dengan perhitungan <i>net sentiment</i> dan <i>net brand reputation</i>	<ul style="list-style-type: none"> <li><i>Naive Bayes</i>: 78.90%</li> <li>SVM: 82.40%</li> <li><i>Decision Tree</i>: 72.90%</li> </ul>
Analisis Sentimen Twitter dengan Klasifikasi <i>Naive Bayes</i> Menggunakan Seleksi Fitur <i>Mutual Information</i> Dan <i>Inverse Document Frequency</i>	(Putra, 2017)	Mengklasifikasikan sentimen pada data <i>tweet</i> menggunakan seleksi fitur MI dan IDF dengan metode <i>Multinomial Naive Bayes</i> dan membandingkan hasil akurasi tersebut.	<ol style="list-style-type: none"> <li>Pengumpulan data <i>tweet</i></li> <li>Penentuan sentimen secara manual</li> <li>Pembagian data</li> <li><i>Indexing</i> data latih dan data uji</li> <li>Seleksi fitur</li> <li>Fungsi klasifikasi <i>multinomial Naive Bayes</i></li> <li>Evaluasi</li> </ol>	<i>Mutual Information</i> (MI) dan <i>Inverse Document Frequency</i> (IDF) dengan metode <i>Multinomial Naive Bayes</i>	<ul style="list-style-type: none"> <li>MI: 71.89%</li> <li>IDF: 60.67%</li> </ul>



Tabel 1. Penelitian terdahulu (lanjutan)

<b>Perbandingan teori</b>	<b>Peneliti (tahun penelitian)</b>	<b>Tujuan</b>	<b>Langkah-langkah</b>	<b>Metode</b>	<b>Output</b>
Sentimen Analisis Tweet Berbahasa Indonesia dengan <i>Deep Belief Network</i>	(Zulfa dan Winarko, 2017)	Melakukan pengklasifikasian terhadap sentimen positif, negatif, dan netral terhadap data uji untuk mengetahui akurasi model klasifikasi dengan menggunakan metode <i>Deep Belief Network</i> ketika diaplikasikan klasifikasi <i>tweet</i> untuk menandai kelas sentimen <i>data training</i> <i>tweet</i> berbahasa Indonesia	<ol style="list-style-type: none"> <li>a. Pengumpulan data</li> <li>b. <i>Preprocessing</i> data</li> <li>c. Pelabelan data</li> <li>d. Klasifikasi sentimen</li> <li>e. Pengujian akurasi</li> </ol>	Menggunakan metode <i>Deep Belief Network</i> yang kemudian dibandingkan dengan algoritme klasifikasi <i>Naive Bayes</i> dan <i>Support Vector Machine</i>	Nilai akurasi DBN: 93.31% <i>Naive Bayes</i> : 79.10% SVM: 92.18%

### **III. METODE PENELITIAN**

#### **3.1. Tempat dan Waktu Penelitian**

Penelitian ini dilakukan di Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung. Penelitian dilaksanakan mulai bulan Juli sampai bulan Desember tahun 2020.

#### **3.2. Alat dan Bahan**

Penelitian ini dilakukan dengan menggunakan alat untuk mendukung dan menunjang pelaksanaan penelitian, yaitu sebagai berikut.

##### **3.2.1. Perangkat Keras (*Hardware*)**

Perangkat keras yang digunakan dalam penelitian analisis sentimen ini adalah satu unit laptop dengan spesifikasi:

- Prosesor: Intel(R) Core™ i3-5005U CPU @ 2.00GHz (4 CPUs), ~2.0GHz
- RAM: 8192 MB

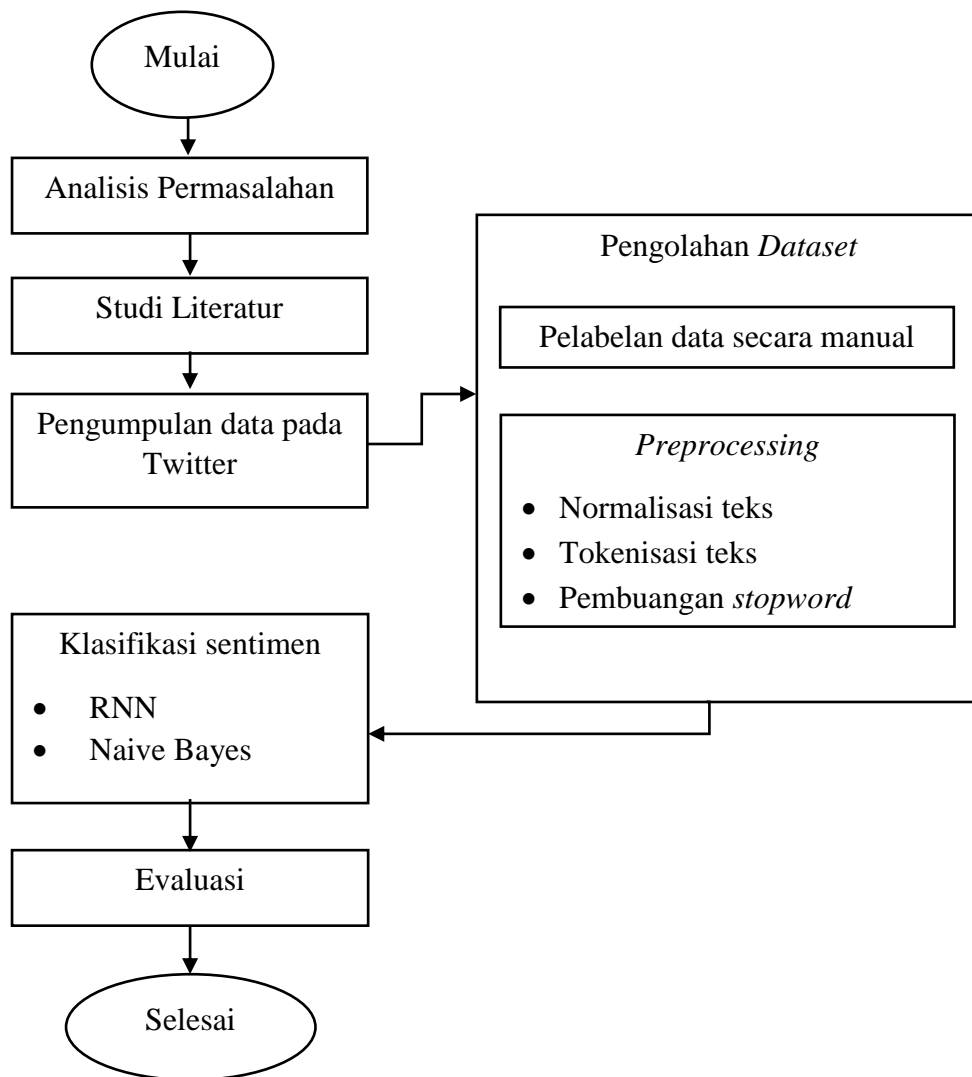
##### **3.2.2. Perangkat Lunak (*Software*)**

Perangkat lunak yang digunakan dalam analisis sentimen ini adalah sebagai berikut:

- *Operating System*: Windows 10
- *Web Browser*: Google Chrome, digunakan untuk pengunduhan data.
- Python 3, digunakan sebagai bahasa pemrograman.
- Jupyter Notebook, digunakan untuk membuat *script* program.

### **3.3. Tahapan Penelitian**

Penelitian ini melalui beberapa tahapan yang diilustrasikan pada Gambar 2.



Gambar 2. Tahapan penelitian

Gambar 2 menunjukkan tahapan-tahapan yang dilalui dalam penelitian. Berikut ini adalah penjabaran dari masing-masing tahapan penelitian yang dilakukan.

#### 1. Analisis Permasalahan

Pada tahap ini dilakukan analisis permasalahan yang ada yaitu mengenai metode algoritme yang mendapatkan nilai akurasi yang terbaik mengenai analisis sentimen. Topik penelitian ini adalah analisis

sentimen untuk membandingkan pendapat masyarakat mengenai kemandirian energi dengan *deep learning* menggunakan algoritme *Recurrent Neural Network* (RNN) dan *Naïve Bayes*. Untuk metode yang digunakan dari algoritme *Recurrent Neural Network* yaitu metode *Long Short-Term Memory* (LSTM) dan *Simple Recurrent Neural Network*. Untuk metode yang digunakan dari algoritme *Naive Bayes* yaitu metode *Bernoulli Naive Bayes*. Kedua algoritme ini merupakan salah satu algoritme yang paling banyak digunakan untuk melakukan analisis sentimen dan pada penelitian ini akan dilakukan perbandingan antara algoritma *Recurrent Neural Network* dan *Naive Bayes*.

## 2. Studi Literatur

Pada tahap ini dilakukan pengumpulan informasi dan penelitian terdahulu yang berkaitan dengan topik permasalahan atau penelitian yang sama dan mempelajari berbagai metode penelitian. Untuk penelitian terdahulu mengenai analisis sentimen, terdapat tiga penelitian terdahulu dengan topik yang berbeda yang dijadikan sebagai studi literatur untuk penelitian ini. Algoritme yang digunakan dari ketiga penelitian terdahulu berbeda-beda.

## 3. Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data *tweet* dari sosial media Twitter menggunakan API key yang disediakan oleh Twitter yaitu *tweepy*. Untuk menggunakan *tweepy* ini, diperlukan akun Twitter yang telah didaftarkan menjadi akun *developer* Twitter. Data yang

dikumpulkan dari Twitter merupakan data teks yang tidak memiliki nilai sentiment positif ataupun negatif. Kata kunci untuk pengumpulan data dari Twitter menggunakan kata “Kemandirian Energi”. Untuk pengumpulan datanya sendiri termasuk lama karena dilakukan hampir per minggu karena data di Twitter yang selalu update. Untuk pengumpulan datanya sendiri menggunakan program yang berbeda dengan program klasifikasi.

#### 4. Pengolahan *Dataset*

Tahap ini mengolah *dataset* yang telah diperoleh dari Twitter dengan melakukan *preprocessing* data. Tahapan yang dilakukan saat melakukan *preprocessing* yaitu tokenisasi teks, normalisasi teks, dan pembuangan stopword. Fungsi dari tahapan *preprocessing* adalah sebagai berikut:

##### a. Normalisasi Teks

Normalisasi teks berfungsi untuk melakukan penormalan kata atau kalimat dalam suatu data. Untuk normalisasi teks ini tujuannya adalah untuk memperbaiki struktur kata dan kosakata dalam kalimat. Pada penelitian ini normalisasi akan dilakukan pada seluruh data teks penelitian yang digunakan. Normalisasi teks yang dilakukan adalah penghapusan URL, penghapusan tanda baca, karakter-karakter khusus, dan simbol-simbol dengan bentuk huruf alphabet sehingga tidak terjadi ambuigitas dalam data teks.

##### b. Tokenisasi Teks

Tahap tokenisasi teks adalah tahap pemotongan *string* atau kalimat input berdasarkan tiap kata yang menyusun kalimat dalam data teks. Tokenisasi secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata, bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan.

c. Pembuangan *Stopword*

*Stopwords* merupakan kumpulan kata-kata yang sering muncul tetapi jika dihapus tidak mengubah makna dari *tweet* tersebut. Pembuangan *stopword* dimaksudkan untuk mengetahui suatu kata yang tidak memiliki arti atau tidak relevan. Kata yang diperoleh dari tahap tokenisasi di periksa dalam suatu *stopword*, apabila ada sebuah kata masuk di dalam *stopword* maka kata tersebut tidak akan diproses lebih lanjut.

Untuk tahapan preprocessing akan dilakukan disaat melakukan klasifikasi sentimen dengan masing-masing algoritme. Selain tahapan *preprocessing* data, data-data yang telah dikumpulkan akan dilakukan pelabelan nilai sentimen pada data secara manual apakah data tersebut memiliki nilai positif atau negatif. Tahapan pelabelan data ini dilakukan karena pada saat pengumpulan data, data-data yang telah didapatkan tidak memiliki nilai sentimen positif ataupun negatif.

## 5. Klasifikasi Sentimen

Pada tahap ini akan dilakukan klasifikasi sentimen menggunakan dua algoritme yaitu algoritme *Recurrent Neural Network* (RNN) dan *Naive Bayes*. Data yang akan diklasifikasi adalah data *tweet* yang telah melewati tahap pengolahan data. Data-data tersebut akan diklasifikasi dan analisis menggunakan metode *Simple Recurrent Neural Network*, *Long Short-Term Memory*, dan *Bernoulli Naive Bayes*. Sebelum itu dilakukan pembagian data menjadi dua, yaitu sebagai data latih, dan data uji. Data latih digunakan untuk melatih algoritme atau metode untuk, sedangkan data uji digunakan untuk mengetahui performa algoritme yang sudah dilatih sebelumnya ketika menemukan data baru yang belum pernah ditemui sebelumnya. Setelah data dibagi menjadi data latih, dan data uji, melakukan latih data menggunakan data latih yang telah dibagi menggunakan *word2vec*. Selanjutnya adalah melakukan klasifikasi dengan *Simple Recurrent Neural Network*, *Long Short-Term Memory*, dan *Bernoulli Naive Bayes* yang program klasifikasinya telah dibuat. Hasil akhir yang didapatkan dari klasifikasi akan berupa prediksi sentimen positif dan negatif dalam bentuk *confusion matrix* dan *classification report* yang isinya berupa akurasi, presisi, *recall*, dan *f1-score* hasil dari melakukan klasifikasi.

## 6. Evaluasi

Evaluasi yang dilakukan adalah melakukan perbandingan nilai sentimen dari pelabelan manual dan analisis sentimen menggunakan kedua algoritme tersebut. Setelah itu akan dilakukan perbandingan antara hasil dari algoritme RNN dan algoritme *Naive Bayes*. Hasil klasifikasi yang



didapatkan adalah *confusion matrix* dan *classification report* berupa akurasi, presisi, *recall*, dan *f1-score*.

Untuk *confusion matrix* sendiri berisi informasi hasil prediksi klasifikasi dari data aktual yang dilakukan oleh sistem klasifikasi. Kinerja sistem klasifikasi umumnya dihitung menggunakan data dalam tabel *confusion matrix* pada Tabel 2.

Tabel 2. Tabel *confusion matrix*

Fakta	Prediksi	
	Negatif	Positif
Negatif	TN ( <i>True Negative</i> )	FP ( <i>False Positive</i> )
Positif	FN ( <i>False Negative</i> )	TP ( <i>True Positive</i> )

Nilai *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Negative* (FN) merupakan data positif yang terdeteksi negatif. Nilai *True Positive* (TP) merupakan data positif yang terdeteksi benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi positif. Tabel *confusion matrix* digunakan untuk mengukur kinerja suatu metode klasifikasi dengan menghitung nilai akurasi, presisi, *recall*, dan *f1-score*.

- a. Akurasi (*accuracy*) merupakan rasio prediksi benar (positif, dan negatif) dengan keseluruhan data.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7)$$

- b. Presisi (*precision*) merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (8)$$

- c. *Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (9)$$

- d. *F1-Score* merupakan perbandingan rata-rata presisi dan *recall* yang dibobotkan.

$$f1\ score = 2 \times \frac{(recall \times precision)}{(recall + precision)} \quad (10)$$

## V. SIMPULAN DAN SARAN

### 5.1. Simpulan

Pada penelitian ini dapat disimpulkan bahwa:

1. Hasil klasifikasi dari metode *Simple Recurrent Neural Network* menunjukkan nilai klasifikasi yang cukup baik dengan akurasi 78%, presisi 78%, *recall* 78%, dan *f1-score* 77%. Untuk hasil *confusion matrix* dari *Simple Recurrent Neural Network* memiliki banyak nilai 377 data diprediksi *true positive* dari 405 data uji sebenarnya positif dan 96 *true negative* dari 195 data uji sebenarnya negatif, ini menunjukkan bahwa tidak terlalu banyak nilai prediksi yang error pada metode *Simple Recurrent Neural Network*.
2. Untuk perbandingan klasifikasi antara metode *Simple Recurrent Neural Network* dengan LSTM yang termasuk dalam algoritme RNN dan *Bernoulli Naive Bayes* memiliki nilai klasifikasi yang berbeda. *Simple Recurrent Neural Network* lebih unggul dibandingkan dengan LSTM dengan akurasi 75%, presisi 74%, *recall* 75%, dan *f1-score* 73% dan *Bernoulli Naive Bayes* dengan akurasi 68%, presisi 58%, *recall* 67%, dan *f1-score* 56%. Dari hasil ketiga klasifikasi tersebut dapat

disimpulkan bahwa algoritme RNN memiliki hasil klasifikasi analisis sentimen yang lebih baik dibandingkan dengan algoritme *Naive Bayes*.

## 5.2. Saran

Terdapat beberapa hal yang dapat ditambahkan atau diperbaiki untuk penelitian selanjutnya, yaitu:

1. Data *tweet* yang digunakan dapat diperbanyak jumlahnya, agar data latih yang digunakan dapat lebih banyak dan menyetarakan kelas data positif dan negatif sehingga sistem klasifikasi dapat memiliki nilai akurasi yang lebih baik.
2. Dapat mengembangkan analisis sentimen ini dengan mengkombinasikan metode *Deep Learning* yang lain selain RNN seperti *Deep Neural Networks*, dan *Convolutional Neural Networks* untuk dilakukan pengujian, dan mendapatkan nilai akurasi yang lebih baik.

## DAFTAR PUSTAKA

- Bengio, Y. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*. 2(1), hal. 1–127. <https://doi.org/10.1561/22000000006>.
- Bluszcz, A. 2017. European economies in terms of energy dependence, *Quality & Quantity*. Springer Nature. 51(4), hal. 1531–1548. <https://doi.org/10.1007/s11135-016-0350-1>.
- Cambria, E., Schuller, B., Xia, Y., dan Havasi, C. 2013. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*. 28(2). hal. 15-21. <https://doi: 10.1109/MIS.2013.30>.
- Chai, K. M. A., Ng, H. T., dan Chieu, H. L. 2002. Bayesian online classifiers for text classification and filtering. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*. hal. 97–104. <https://doi.org/10.1145/564392.564395>.
- Chaturvedi, I., Cambria, E. dan Qiu, L. 2015. Multilingual Subjectivity Detection Using Deep Multiple Kernel Learning. *4th International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'15)*, (October). Tersedia pada <http://www.sentic.net/multilingual-subjectivity-detection.pdf>.
- Dey, L. dan Haque, S. K. M. 2009. Opinion Mining From Noisy Text Data. In *International Journal on Document Analysis and Recognition*. Vol. 12, hal. 205–226. <https://doi.org/10.1007/s10032-009-0090-z>.
- ESDM. 2019. Indonesia Energy Out Look 2019. *Journal of Chemical Information and Modeling*. <https://doi.org/10.1017/CBO9781107415324.004>.

- Feldman, R. dan Sanger, J. 2007. The text mining handbook: Advanced approaches in analyzing unstructured data. *Review of Social Economy*. 34(1), hal. 126–127. <https://doi.org/10.1080/00346768200000022>.
- Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., dan Schmidhuber, J. 2009. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31(5), hal. 855-868. <https://doi.org/10.1109/TPAMI.2008.137>.
- Güngör, T. 2010. Part-of-speech tagging. *Handbook of Natural Language Processing. Second Edition*. hal. 205–235. CRC Press. <https://doi.org/10.1016/b0-08-044854-2/00952-4>.
- Guttag, J. V. 2016. *Introduction to Computation and Programming using Python with application to understanding data. Journal of Chemical Information and Modeling*. Vol. 53, hal. 1689–1699. <https://doi.org/10.1017/CBO9781107415324.004>.
- Holton, A. E., Baek, K., Coddington, M., dan Yaschur, C. 2014. Seeking and Sharing: Motivations for Linking on Twitter. *Communication Research Reports*. 31(1), hal. 33–40. <https://doi.org/10.1080/08824096.2013.843165>.
- Hotho, A., Nürnberger, A. dan Paaß, G. 2005. A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*. 20(1), hal. 19–62. <https://doi.org/10.1111/j.1365-2621.1978.tb09773.x>.
- Hu, M., dan Liu, B. 2004. Mining opinion features in customer reviews. *Proceedings of the National Conference on Artificial Intelligence*. hal. 755–760.
- Karpathy, A. 2015. The Unreasonable Effectiveness of Recurrent Neural Networks. *Web Page*. hal 1–28. Tersedia pada <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>
- Kelly, R. 2009. Twitter Study. *New York*. 2010(August), hal. 1–17. [https://doi.org/10.1016/S1361-3723\(09\)70038-7](https://doi.org/10.1016/S1361-3723(09)70038-7).

- Koppel, M., dan Schler, J. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*. 22(2), hal. 100–109.  
<https://doi.org/10.1111/j.1467-8640.2006.00276.x>.
- Kuhlman, D. 2012. A Python Book: Beginning Python, Advanced Python, and Python Exercises. *Book*. hal. 12–13. Tersedia pada:  
[https://www.davekuhlman.org/python\\_book\\_01.pdf](https://www.davekuhlman.org/python_book_01.pdf).
- Le, Q. dan Mikolov, T. 2014. Distributed representations of sentences and documents, *31st International Conference on Machine Learning, ICML 2014*, 4, hal. 2931–2939.
- Lecun, Y., Bengio, Y., dan Hinton, G. 2015. Deep learning. *Nature*. Nature Publishing Group. 521(7553), hal. 436–444.  
<https://doi.org/10.1038/nature14539>.
- Liu, B. 2010. Sentiment Analysis and Subjectivity in: Handbook of Natural Language Processing, Second Edition. *Handbook of Natural Language Processing, Second Edition*, 2, 568.
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., dan Khudanpur, S. 2010. Recurrent neural network based language model. *Proceedings of the 11th Annual Conference of the International Speech Communication Association. INTERSPEECH 2010*. (September), hal. 1045–1048.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., dan Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. Neural information processing systems foundation. hal. 1–9.
- Mikolov, T., Chen, K., Corrado, G., dan Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR. hal. 1–12.
- Novantirani, A., Sabariah, M. K., dan Effendy, V. 2015. Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. *E-Proceeding of Engineering*.

- Ortony, A., L. Clore, G. dan Collins, A. 2011. Introduction. In *The Cognitive Structure of Emotions*. hal 1-14. Cambridge University Press.  
<https://doi.org/10.1017/cbo9780511571299.002>
- Putra, R. S. 2017. Analisis sentimen twitter dengan klasifikasi naïve bayes menggunakan seleksi fitur mutual information dan inverse document frequency. (Skripsi). Institut Pertanian Bogor. Bogor. 40 hlm.
- Řehůřek, R. 2011. *Scalability Of Semantic Analysis In Natural Language Processing*. Masaryk University. Tersedia pada:  
[https://radimrehurek.com/phd\\_rehurek.pdf](https://radimrehurek.com/phd_rehurek.pdf).
- Rehurek, R. dan Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. hal. 45–50.
- Saranya, K. dan Jayanthi, S. 2018. Onto-based sentiment classification using machine learning techniques. *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIECS 2017*, Vol 2018-Januari, hal. 1–5. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ICIECS.2017.8276047>
- Sarkar, D. 2019. *Text Analytics with Python. Text Analytics with Python*. Apress.  
<https://doi.org/10.1007/978-1-4842-4354-1>
- Singh, G., Kumar, B., Gaur, L., dan Tyagi, A. 2019. Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. In *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*. hal. 593–596. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ICACTM.2019.8776800>
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., dan Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. hal. 1631-1642. Association for Computational Linguistics (ACL).
- Soong, H., Jalil, N. B. A., Ayyasamy, R. K., dan Akbar, R. 2019. The essential of sentiment analysis and opinion mining in social media. *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*.



hal. 272–277. <https://doi.org/10.1109/ISCAIE.2019.8743799>.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., dan Stede, M. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*. 37(2), hal. 267-302. [https://doi.org/10.1162/COLI\\_a\\_00049](https://doi.org/10.1162/COLI_a_00049).

UNDESA 2019. *Accelerating SDG 7 Achievement: SDG 7 Policy Briefs in Support of the High-Level Political Forum 2019*. United Nations. hal. 9-207. Tersedia pada:  
[https://sustainabledevelopment.un.org/contact/%0Ahttps://sustainabledevelopment.un.org/content/documents/22877un\\_final\\_online\\_webview.pdf](https://sustainabledevelopment.un.org/contact/%0Ahttps://sustainabledevelopment.un.org/content/documents/22877un_final_online_webview.pdf).

Vargas, R. dan Ruiz, L. 2018. Deep Learning : Previous and Present. *Journal of Awareness*, 2(3), hal. 11–20. Tersedia pada:  
<https://ratingacademy.com.tr/ojs/index.php/joa/article/view/306>.

Vidya, N. A. 2015. Twitter Sentiment Analysis Terhadap Brand Reputation: Studi Kasus Pt Xl Axiata Tbk. *Metrologia*. 53(5). hal. 1–116.  
<https://doi.org/10.1590/s1809-98232013000400007>

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., dan Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories Technical Report*, (89).

Zulfa, I. dan Winarko, E. 2017. Sentimen Analisis Tweet Berbahasa Indonesia dengan Deep Belief Network. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 11(2), 187 hlm. <https://doi.org/10.22146/ijccs.24716>.