# BAB II

# LANDASAN TEORI

## *Text Mining*

### *Text mining* merupakan proses *mining* atau menambang suatu informasi dari data yang tersaji dalam jumlah besar, dalam hal ini adalah teks. Proses ini dilakukan dalam rangka penggalian, pengolahan, serta pengaturan pada informasi dengan menganalisa keterkaitan antara informasi satu dengan yang lainnya (Sudiantoro, dkk, 2018). Dalam definisi lain, *text mining* adalah proses penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, dan tidak diketahui sebelumnya, atau menemukan kembali informasi yang tersirat secara implisit (Vina & Wibowo, 2019).

## Analisis Sentimen

Analisis sentimen merupakan bidang penelitian yang sedang berlangsung di bidang *text mining*. Tujuan dari analisis sentimen adalah untuk menganalisis opini, sentimen, dan subjektifitas teks. Analisis sentimen juga dapat disamakan dengan *opinion mining* karena berfokus kepada pendapat, sikap, emosi yang mewakili pandangan individu terkait peristiwa atau topik tertentu (Afrizal, dkk, 2019) (Medhat, dkk, 2014). Saat ini, analisis sentimen banyak digunakan oleh peneliti sebagai salah satu cabang riset dalam ilmu komputer seiring dengan ledakan informasi di internet. Twitter merupakan salah satu media sosial yang paling populer untuk digunakan sebagai sumber data pada analisis teks (Watrianthos, 2020) (Ferdiana, Jatmiko, dkk, 2019).

## Media Sosial

Media sosial merupakan media penyampaian informasi yang banyak menjadi pilihan masyarakat, dengan adanya media sosial pengguna dapat memanfaatkan akun yang dimiliki untuk mengungkapkan perasaan baik atau buruk terhadap suatu peristiwa atau objek tertentu (Oktasari, Herry, 2016).

1. Twitter

Twitter merupakan jejaring sosial daring dan layanan *microblogging* yang memungkinkan pengguna terdaftar untuk membaca dan memposting pesan singkat yang disebut dengan kicauan (*tweet*)(Sasmito, 2018) (Andre, dkk, 2019). Twitter juga merupakan media sosial yang populer dikalangan masyarakat Indonesia, menurut penelitian dan analisis oleh statista.com tercatat negara Indonesia menempati peringkat ke-7 dengan 13.2 miliar pengguna pada Oktober 2020 (Statista Research Departement, 2020). Pada umumnya *tweet* diunggah untuk menyampaikan sebuah berita atau informasi terkait peristiwa tertentu, isi *tweet* juga dapat mengekspresikan sebuah pendapat dari penggunanya. Karena hal tersebut, Twitter banyak digunakan sebagai objek penelitian. Hal ini karena tulisan-tulisan pada media sosial Twitter (*tweet*), memiliki struktur yang sangat cocok untuk digunakan pada analisis (Ferdiana, Jatmiko, dkk, 2019).

## *Crawling*

*Crawling* merupakan proses mengumpulkan data dari sebuah laman dan menyimpannya untuk diatur dan dianalisis lebih lanjut (Nurulbaiti, Subekti, 2018). Dalam penelitian ini proses *crawling* dilakukan menggunakan *standard* *search* API Twitter dengan pustaka *tweepy*. Penggunaan pustaka *tweepy* bertujuan untuk menjembatani antara program yang dibuat dengan *standard* *search* APITwitter bedasarkan parameter seperti kata kunci dan waktu pencarian. Dalam penelitian ini, *dataset* dikumpulkan berdasarkan beberapa parameter kata kunci antara lain: X, Y dan Z dalam periode waktu 1 Desember 2020 hingga 31 Januari 2021.

## *Preprocessing*

*Preprocessing* atau pra-pemrosesan merupakan bagian dari *text mining* yang dilakukan untuk menghapus *noise* pada dokumen atau kalimat. Selain itu, proses ini bertujuan untuk menghindari data yang kurang sempurna; gangguan pada data; dan data yang tidak konsisten (Vina & Wibowo, 2019). Proses pengubahan data teks yang tidak terstruktur menjadi data teks yang terstruktur sangat diperlukan sehingga perlu adanya proses pra-pramrosesan data (Sudiantoro, dkk, 2018). Merujuk pada penelitian yang telah dilakukan (Watrianthos, Ronal, 2020) (Budi & Nugroho, 2020) (Nurul, dkk, 2019) (Antinasari, 2017) maka dalam penelitian ini akan dilakukan beberapa tahapan pra-pemrosesan teks antara lain: *casefolding*, *cleansing*, mengubah *slang word,* menghapus *stop word,* dan *stemming*.

### *Casefolding*

*Case folding* merupakan proses yang bertujuan untuk mengubah seluruh huruf menjadi huruf kecil (*lowercase*) (Budi & Nugroho, 2020) (Nurul, dkk, 2019).

### *Cleansing*

*Cleansing* merupakan proses yang bertujuan untuk menghapus atribut yang tidak diperlukan untuk proses analisis (Watrianthos, Ronal, 2020) (Budi & Nugroho, 2020) (Nurul, dkk, 2019). *Cleansing* yang dilakukan dalam penelitian ini terbagi menjadi beberapa tahapan antara lain: menghapus URL, *mention* (*@mention*), *hastag* (*#hastag*), angka (0-9), spesial karakter (*unicode*), tanda baca, dan spasi berlebih .

### *Mengubah slang word*

*Slang word* merupakan kata yang tidak sesuai dengan ejaan bahasa Indonesia yang baku (EYD) baik berupa kata singkatan ,kata gaul atau modern, ataupun kesalahan salah eja (Antinasari, 2017). *Slang word*  tersebut sebanyak mungkin akan ditampung ke dalam kamus *slang word*. Kamus tersebut kemudian digunakan sebagai pengetahuan untuk melakukan *replace* atau mengubah kata *slang* menjadi kata dengan bahasa Indonesia yang baku sesuai EYD.

### Menghapus *Stop word*

*Stopword* merupakan kata yang tidak berpengaruh atau kurang bermakna namun sering ditemui dalam dokumen atau kalimat, seperti kata ‘saya’, ‘dan’, ‘atau’ (Watrianthos, Ronal, 2020) (Budi & Nugroho, 2020). Dalam proses ini, kata yang tergolong ke dalam *stop word* akan ditampung ke dalam kamus *stop word.* Kamus tersebut kemudian digunakan sebagai pembanding untuk menghapus sebuah kata dalam dokumen atau kalimat yang tergolong kedalam *stop word*.

### *Stemming*

*Stemming* merupakan proses mengubah kata berimbuhan menjadi kata dasar (Watrianthos, Ronal, 2020) (Nurul, dkk, 2019). Definisi lain dari *stemming* menurut Nazief (Nazief, dkk, 2005) adalah proses pengembalian sebuah kata menjadi kata dasarnya, *stemming* bekerja dengan mengilangkan imbuhan dan akhiran kata, penghilangan imbuhan dan akhiran kata sangat berperan penting dalam *machine translation*, *document summarization* dan *text classification*. Dalam penelitian ini proses *stemming* dilakukan dengan menggunakan pustaka Sastrawi melalui dengan paket StemmerFactory.

## *Labeling*

*Labeling* atau pelabelan merupakan proses pemberian kelas berdasarkan ciri atau karakteristik yang terkandung dalam sebuah dokumen atau kalimat. Proses pelabelan terbagi menjadi tiga kategori kelas sentimen, yakni sentimen positif; sentimen negatif; dan sentimen netral (Budi & Nugroho, 2020). Dalam penelitian ini proses pelabelan dapat dilakukan dengan dua (2) cara antara lain: pelabelan manual dengan melabeli kalimat berdasarkan subjektifitas peneliti dan pelabelan dengan pendekatan kamus sentimen.

### Perhitungan skor sentimen

Perhitungan skor sentimen merupakan proses pelabelan dengan cara pendekatan kamus sentimen. Kamus tersebut berisikan kata opini positif dan kata opini negatif. Skor suatu kata akan bernilai +1 jika kata tersebut adalah kata opini positif, dan bernilai -1 jika kata tersebut adalah kata opini negatif (Liu, dkk, 2005) (Asrofi, 2017). Maka dapat diketahui bahwa nilai skor sentimen dapat diperoleh menggunakan rumus:

Keterangan:

*x* = sebuah dokumen atau kalimat

*kata positif* = bilangan bulat positif atau nol

*kata negatif* = bilangan bulat negatif atau nol

### Pemberian kelas sentimen

Setelah melakukan proses perhitungan skor sentimen dan diketahui nilai skor dari suatu kalimat. Proses selanjutnya dalam pelabelan dengan pendekatan kamus sentimen adalah pemberian kelas pada kalimat(*tweet*) berdasarkan skor. Jika kalimat mempunyai skor > 0 akan masuk ke dalam kelas positif, jika kalimat mempunyai skor < 0 akan masuk ke dalam kelas negatif (Budi & Nugroho, 2020) (Asrofi, 2017) (Nurulbaiti, Subekti, 2018), sedangkan jika kalimat mempunyai skor = 0 maka akan diabaikan sehingga penentuan kelas sentimenya dilakukan secara manual. Adapun proses pelabelan dapat dilihat pada gambar X di bawah ini:



Gambar X. Pelabelan kelas sentimen

## *CountVectorizer*

*CountVectorizer* merupakan proses pengolahan dokumen atau teks menjadi bentuk vektor. *CountVectorizer* digunakan untuk menghitung frekuensi kata dalam dokumen atau kalimat kemudian direpresentasikan ke dalam bentuk vektor ().

## *K-Nearest Neighbors*

*K-Nearest Neighbors* (KNN) adalah algoritme klasifikasi *supervised learning* berbasis jarak. Algoritme ini bekerja dengan cara membandingkan jarak antara data tes dengan semua data latih yang ada (Novi, dkk, 2019) (Andre, dkk, 2019) (Octaviani & Adiwijaya, 2020). Untuk menghitung jarak antara data digunakan perhitungan *Euclidean Distance* dengan rumus:

Keterangan:

*distance* = nilai jarak antar data

*n* = jumlah fitur

*Li* = Fitur ke-*i* dalam data latih

*Ti* = Fitur ke-*i* dalam data tes

Proses selanjutnya setelah menghitung jarak untuk setiap data latih adalah mencari data latih dengan nilai jarak terkecil (ketetanggaan terdekat) sebanyak nilai K yang telah ditentukan. Proses akhir setelah ditemukanya data tetangga terdekat adalah pemungutan suara (*voting*). *Voting* bertujuan untuk menentukan kelas atau label dari suatu pengujian data (Octaviani & Adiwijaya, 2020).

## Pengujian

Pengujian merupakan hal penting untuk memastikan bahwa suatu algoritma yang telah dirancang dapat berjalan sesuai dengan harapan. Pengujian klasifikasi sentimen dilakukan dengan menguji aplikasi yang telah dibangun dengan membandingkan antara data prediksi dan data aktual. Data prediksi berupa hasil klasifikasi yang dihasilkan oleh aplikasi yang dibangun, sedangkan data aktual berupa yang didapatkan melalui proses pelabelan (Haryalesmana & Azhari, 2020).

Dalam penelitian ini, pengujian dilakukan pada sebuah *model* terhadap data tes yang tersedia. *Model* tersebut dibangun menggunakan data latih yang telah diproses sebelumnya dengan algoritme Naive Bayes Classifier dan pembobotan TF-IDF. Data latih yang dijadikan *model* dipilih dengan teknik sampling kuota (*quota sampling*). *Quota Sampling* merupakan teknik sampling yang menentukan jumlah sampel dari populasi yang memiliki ciri tertentu hingga jumlah kuota yang diinginkan tercapai (Vina & Wibowo, 2019). Data latih yang digunakan untuk membangun *model* adalah X data *tweet,* terdiri dengan ciri X *tweet* positif, X *tweet* netral dan X *tweet* negatif. Hasil dari pengujian tersebut akan dievaluasi menggunakan *confused matrix* untuk mengukur tingkat akurasi, presisi dan *recall*.

## Studi Pustaka

Berdasarkan landasan teori yang telah dijelaskan, terdapat penelitian yang sudah ada sebelumnya, di rangkum dalam Tabel X berikut :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Penulis | Judul | Terbitan | Deskripsi |
| 1 | Ronal Watrianthos | Analisis Pembelajaran Daring di Era Pandemic Covid-19 | Green Press, Hal 55-64, 2018, P-ISBN: 978-623-93614-2-6, *e*-ISBN: 978-623-93614-3-3 | Melakukan analisis terhadap pembelajaran daring melalui sosial media Twitter, berdasarkan kata kunci pada tanggal 1 Juli - 31 Juli 2020. Menggunakan metode analisis sentimen dengan algoritma Naive Bayes. Hasil analisis menunjukkan sentimen negatif sangat tinggi mencapai 83%; 16% sentimen positif; 1% sentimen netral dan pada periode Juli 2020. |
| 2 | Eko Budi Santoso, Aryo Nugroho | Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di Facebook | Jurnal Eksplora Informatika, Vol. 9, No. 1, Hal 60-69, September 2019, P-ISSN: 2089-1814, *e*-ISSN: 2460-3694 | Melakukan analisis komentar masyarakat pada media sosial Facebook terhadap popularitas dari seorang calon presiden. Metode klasifikasi yang digunakan adalah Naive Bayes disertai dengan proses asosiasi teks, juga menggunakan fitur kamus *(lexicon)* pada proses pelabelan kelas sentimen. Berdasarkan data yang telah dikumpulkan sejak tanggal 17 April 2019 sampai 22 Mei 2019, sebanyak 10.000 dataset komentar berhasil dikumpulkan melalui halaman *fanpage* Joko Widodo dan Prabowo Subianto. Penelitian ini menghasilkan persentase setimen (positif dan negatif) tiap pasangan calon presiden dan serta pengujian akurasi untuk metode Naïve Bayes Classifier yaitu sebesar 86,4%. |
| 3 | Novelty Octaviani Faomasi Daeli, Adiwijaya | *Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor* | J. Data SCI APPL, Vol. 3, No. 1, Hal. 001-007, Januari 2020, *e*-ISSN 2614-7408 | Melakukan pengujian untuk mencari nilai K yang optimal untuk K-Nearest Neighbor (KNN) dengan perhitungan jarak euclidean distance. Dataset yang digunakan adalah dataset review film Cornell Polarity v2.0 dengan total data 1000 dokumen negatif dan 1000 dokumen positif. Dengan melibatkan Information Gain, nilai K optimal yang diperoleh untuk KNN adalah 3 (K=3) dengan memberikan akurasi sebesar 96.8%. |
| 4 | Nova Tri Romadloni, Imam Santoso, Sularso Budilaksono | Perbandingan Metode Naive Bayes, KNN Dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line | Jurnal IKRA-ITH Informatika, Vol. 3, No. 2, Juli 2019, ISSN: 2580-4316 | Melakukan perbandingan metode Naive Bayes, K-Nearest Neighbor (KNN) dan Decision Tree untuk melakukan analisis sentimen pada data media sosial Twitter. Tahapan yang dilakukan terhadap dataset tweet (Twitter) sebelum melakukan perbandingan diantaranya: *Convert Emoticon*, *Cleansing*, *Case Folding*, *Tokenizing*, *Stemming.* Pengujian dilakukan terhadap 127 data yang telah diberikan label positif atau negatif, menghasilkan akurasi 80% menggunakan algoritma Naive Bayes; 80% menggunakan algorima KNN; 100% menggunakan Decision Tree. |
| 5 | Fransiska Vina Sari, Arief Wibowo | Analisis Sentimen Pelanggan Toko Online JD.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi | Jurnal SIMETRIS, Vol. 10, No. 2 November 2019, P-ISSN: 2252-4983, *e*-ISSN: 2549-3108 | Melakukan analisis terhadap opini pelanggan atau konsumen terkait toko online JD.id. Menggunakan data yang bersumber pada media sosial Twitter dengan metode klasifikasi Naive Bayes dan pembobotan TF-IDF disertai fitur konversi ikon emosi *(emoticon).* Hasil penelitian menunjukkan bahwa metode Naïve Bayes tanpa penambahan fitur mampu mengklasifikasi sentimen dengan nilai akurasi sebesar 96,44%, sementara jika ditambahkan fitur pembobotan TF-IDF disertai konversi ikon emosi mampu meningkatkan nilai akurasi menjadi 98%. |
| 6 | Ghulam Asrofi Buntoro | Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter | Integer Journal, Vol 2, No 1, Maret 2017, Hal. 32-41, P-ISSN: 2477-5274, *e*-ISSN: 2579-566X | Melakukan analisis terkait opini masyarakat terhadap pemilihan gubernur DKI Jakarta tahun 2017 pada media sosial Twitter. Proses penentuan sentimen menggunakan metode *Lexicon-Based* dan proses klasifikasinya menggunakan metode Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). Akurasi tertinggi didapat saat menggunakan metode klasifikasi Naïve Bayes Classifier (NBC), dengan nilai rata-rata akurasi mencapai 95%, nilai presisi 95%, nilai recall 95% nilai TP rate 96,8% dan nilai TN rate 84,6%. |
| 7 | Jeremy Andre Septian, Tresna Maulana Fahrudin, Aryo Nugroho | Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor | Journal of Intelligent Systems And Computation, Vol. 1 No. 1, Oktober 2019, P-ISSN: 2621-9220, *e-*ISSN: 2722-1962 | Melakukan analisis sentimen pada setiap kalimat dari pengguna twitter terhadap persepakbolaan Indonesia apakah memiliki sentimen negatif atau positif menggunakan K-Nearest Neighbor (KNN) dengan pembobotan kata TF-IDF. Data yang digunakan dalam didapatkan dari hasil crawling dari media sosial twitter terkait persepakbolaan di Indonesia yang diambil dari akun twitter resmi PSSI. Dari 2000 data tweet berbahasa indonesia didapatkan hasil akurasi optimal pada nilai k=23 sejumlah 79.99%. |
| 8 | Walaa Medhat, Ahmed Hassan, Hoda Korashy | *Sentiment Analysis Algorithms and Applications: A Survey* | Ain Shams Engineering Journal, Vol 5, No. 4, Hal. 1093–1113, Desember 2014, https://doi.org/10.1016/j.asej.2014.04.011 | Melakukan survey terkait analisis sentimen. Meliputi proses melakukan analisis sentimen menggunakan pendekatan *machine learning* dan *lexicon based.* Penelitian ini juga membahas macam-macam teknik klasifikasi sentimen dan cara pengaplikasianya secara singkat untuk mengolah data teks. |
| 9 | Sitti Nurul Jannah Fitriyyah, Novi Safriadi, Enda Esyudha Pratama | Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes | JEPIN (Jurnal Edukasi dan Penelitian Informatika) Vol. 5, No. 3, Desember 2019, P-ISSN: 2460-0741, e-ISSN:2548-9364 | Melakukan analisis sentimen terhadap pasangan calon(paslon) presiden melalui media sosial Twitter. Penelitian ini juga melakukan penerapan metode Naive Bayes untuk klasifikasi sentimen pengguna twiter dengan dua kelas sentimen (negatif, positif) dan tiga kelas sentimen (negatif, positif, netral). Hasil dari penelitian ini menunjukkan metode Naive Bayes memiliki performa lebih baik dalam mengklasifikasikan 2 kelas sentimen (negatif, positif) dibandingkan pengujian dengan 3 kelas sentimen. |
| 10 | Agus Sasmito Aribowo | Analisis Sentimen Publik pada Program Kesehatan Masyarakat menggunakan Twitter Opinion Mining | Seminar Nasional Informatika Medis, Hal. 17-23, 2018, ISSN: 9-772301-936005 | Melakukan penelitian untuk mengembangkan model untuk mengetahui sentimen publik terhadap enam macam program kebijakan pemerintah yaitu imunisasi, asuransi kesehatan, stunting, gizi buruk, pelayanan kesehatan, dan jaminan kesehatan masyarakat. Metodenya adalah dengan melakukan ekstraksi pengetahuan dari opini di media sosial menggunakan analisis sentimen berbasis leksikon. Dataset yang diperoleh dalam kurun waktu 3 - 9 Agustus 2018 sebanyak total 3311 data. Hasil penelitian berupa sentimen yang dituangkan ke dalam bentu grafik. |
| 11 | Fahmi Nurulbaiti, Retno Subekti | Analisis Sentimen Terhadap Data Tweet Untuk Badan Penyelenggara Jaminan Sosial (BPJS) Menggunakan Program R | Jurnal Matematika-S1, Vol. 7, No. 1, Hal. 83-107, 2018, https://eprints.uny.ac.id/56367 | Melakukan penelitian dengan tujuan untuk mengetahui langkah-langkah dalam melakukan analisis sentimen masyarakat Indonesia dengan menggunakan Program R, mengetahui topik-topik yang sering diperbincangkan masyarakat Indonesia terkait dengan BPJS serta mengetahui tanggapan masyarakat terhadap keberadaan BPJS.Data yang digunakan yaitu sebanyak 1000 tweet dan dibatasi hanya tweet berbahasa Indonesia saja. Teknik analisis data pada penelitian ini dilakukan dengan text mining dan scoring sentiment.Pengambilan data, pengolahan data, serta visualisasi informasi dilakukan dengan bantuan Program R. Hasil analisis sentimen tweets masyarakat pada jejaring sosial Twitter terhadap BPJS cenderung bersentimen negatif. |
| 12 | Bing Liu, Minqing Hu, Junsheng Cheng | *Opinion Observer: Analyzing and Comparing Opinions on the Web* | Proceedings of the 14th International World Wide Web Conference (WWW-2005), May 10-14, Chiba, Japan | Melakukan penelitian untuk menganalisa pendapat konsumen terhadap suatu produk. Mengelompokkan data pendapat berdasarkan ulasan komsumen kedalam bentuk ulasan positif atau negatif, kemudian dijadikan sebuah pengetahuan untuk dibandingkan dengan ulasan lainya. Penelitian ini juga membuahkan daftar kata positif dan negatif yang dapat digunakan kembali untuk proses klasifikasi pendapat. |
| 13 | Adhi Viky Sudiantoro, Eri Zuliarso | Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naïve Bayes Classifier | Dinamika Informatika Vol.10, No.2, Oktober 2018, Hal. 69-73, P-ISSN: 2085-3343, e-ISSN : 2714-8769 | Melakukan analisis dengan tujuan untuk mengklasifikasi data tweet menjadi dua sentimen yaitu positif dan negatif. Dataset bersumber dati tweet teks berbahasa Indonesia yang terdapat di sosial media Twitter, kemudian digunakan sebagai bahan analisis sentimen untuk mengetahui sentimen masyarkat terhadap pilkada Jawa Barat. Hasil pengujian akurasi terhadap 100 data uji, algoritma Naïve Bayes Classifier memberikan nilai akurasi sebesar 84%. |
| 14 | Prananda Antinasari, Rizal Setya Perdana, M. Ali Fauzi | Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 1, No. 12, Desember 2017, Hal. 1733-1741, *e*-ISSN: 2548-964X | Melakukan analisis komentar atau tweet mengenai film yang ada pada Twitter, dengan tujuan dapat sebagai evaluasi dalam menonton film dan meningkatkan produksi film. Pada penelitian ini, proses analisis melibatkan tahap perbaikan kata tidak baku dan normalisasi Levenshtein Distance. Sementara metode pengklasifikasian yang digunakan adalah Naive Bayes Classifier. Berdasarkan hasil pengujian yang telah dilakukan didapatkan akurasi tertinggi dengan nilai *accuracy,* *precision,* *recall,* dan *f-measure* sebesar 98.33%, 96.77%, 100%, dan 98.36%. |