BAB II
TINJAUAN PUSTAKA

B ab 2 menjelaskan terkait tinjauan pustaka dan studi literatur yang digunakan pada penelitian mengenai *sentiment analysis* media sosial Twitter pada kasus Jaminan Hari Tua menggunakan metode *naive bayes*. Tujuannya untuk mempelajari literatur yang berkaitan dengan penelitian. Teori yang dibahas antara lain media sosial, Jaminan Hari Tua, *Sentiment Analysis*, *Text Mining*, *Naïve Bayes* dan *Natural Language Processing*

2.1 Media Sosial

Media sosial merupakan media untuk sosialisasi dengan yang lain dan dilakukan secara *online* yang memungkinkan manusia untuk berinteraksi satu sama lain tanpa dibatasi oleh ruang dan waktu. Seorang individu dapat terpengaruh dalam lingkungan jaringan sosialnya oleh peristiwa-peristiwa yang terjadi di sekitar lingkungan mereka. Media sosial pertama dimulai dengan peluncuran SixDegrees.com di 1997 pengguna atau *user* dapat membuat profil dan mendaftarkan teman-temannya untuk menjajaki pertemanan pertama kali dimulai pada tahun 1998. Pada tahun 2000-an, banyak jejaring sosial mulai bermunculan mulai dari Friendster, Match. Com, MySpace, Twitter, untuk Facebook yang memiliki fitur untuk pengembang di luar Facebook untuk membangun aplikasi yang memungkinkan pengguna untuk profil mereka. Jejaring sosial memberikan peluang untuk berinteraksi lebih mudah (M.boyd, 2010).

2.2 Twitter

Twitter adalah layanan Microblogging yang resmi dirilis pada tahun 13 Juli 2006 (M.Mostafa, 2013) . Aktivitas dari Twitter adalah untuk mengunggah sesuatu yang singkat melalui web atau seluler. Batasan maksimum sebuah *tweet* adalah 280 karakter. Twitter adalah sebuah sumber yang tidak terbatas digunakan dalam klasifikasi teks. Ada banyak karakteristik dari *tweet* Twitter (Go, 2009). Pesan di

Twitter memiliki banyak atribut unik, yang membedakannya dari media sosial lainnya:

- Twitter mempunyai batas karakter maksimal 280 karakter.
- Twitter mengizinkan untuk data dapat diakses secara bebas dengan menggunakan Twitter API, sehingga lebih mudah untuk mengumpulkan dalam jumlah besar dari *tweet*.
- Pengguna Twitter membahas berbagai topik secara singkat sesuai dengan topik tertentu dan berlaku juga secara global.

2.3 Jaminan Hari Tua

Jaminan Hari Tua adalah program yang berbentuk jaminan yang memiliki tujuan terjaminnya keamanan. Penjaminan yang berguna untuk para tenaga kerja serta keluarganya dari konsekuensi yang dapat di jangkau oleh pengusaha. Jaminan dikenal dari beberapa pendekatan yang saling melengkapi (Muhtar dan Habibullah, 2009). Pendekatan pertamanya yaitu asuransi (*compulsory social insurance*) yang dibiayai dari kontribusi yang dibayarkan oleh tenaga kerja atau pihak perusahaan yang di mana kontribusi harus dikaitkan dengan tingkat penghasilan yang dibayar oleh perusahaan tersebut. Cara yang kedua biasanya berbentuk bantuan, biasanya berupa bantuan uang tunai dan juga jasa yang dibiayai oleh negara dan juga bantuan masyarakat lainnya.

Menurut Sutrisno (2016), yang menyebabkan para tenaga kerja merasa puas dengan pekerjaannya ialah, umur, status, jaminan keuangan, jabatan serta kualitas pengawasan. Dalam hal ini, biasanya keuangan dan jaminan kebanyakan memiliki pengaruh yang besar terhadap kepuasan kerja karena dia merasa mendapatkan balasan yang setimpal yang sudah ia kerjakan bagi perusahaan tempat ia bekerja.

2.4 *Sentiment Analysis*

Sentiment analysis adalah metode yang berfungsi untuk identifikasi ekspresi berupa teks dan maka dari hal tersebut dapat dihasilkan kategori sebagai kalimat positif atau negatif (T. Nasukawa, 2003). Menurut Indurkhya dan Damerau (2010), penggalian mempunyai bidang yang luas seperti, komputasi, pemrosesan dan penambangan pada teks yang bertujuan untuk analisis pendapat, perasaan,

penilaian, sikap, penilaian dan perasaan pembicara atau penulis tentang topik, produk, layanan atau aktivitas tertentu lainnya. Analisis sentimen digunakan untuk mempelajari komentar yang dibuat oleh pengguna yang menjelaskan bagaimana suatu produk atau merek diterima oleh pengguna (I. P. Cvijikj, 2011). Pengguna internet biasanya menulis opini, pengalaman dan semua yang berhubungan dengannya berdasarkan emosinya. Ini adalah emosi positif yang dapat diekspresikan dengan cara yang sangat kompleks (C. Troussas, 2013).

Analisis sentimen biasanya berguna untuk menganalisis pendapat pelanggan secara otomatis tentang produk dan layanan. Informasi digunakan menjadi data yang diolah sesuai kebutuhan agar bermanfaat bagi pengguna yang perlu membantu membuat keputusan. Informasi sendiri dapat dibagi menjadi dua hal, yaitu opini dan fakta. Fakta adalah objektif pernyataan tentang sesuatu yang telah terjadi dan biasanya disertai dengan bukti, sedangkan pendapat lebih banyak subjektif dalam cara seseorang mengekspresikan diri terhadap segala sesuatu yang terjadi sesuai dengan keadaannya masing-masing persepsi dan asumsi (Liu, 2012). Analisis sentimen didefinisikan sebagai tugas menemukan pendapat penulis tentang sebuah entitas tertentu (Ghiassi, 2018).

2.5 *Text Mining*

Penambangan teks bertujuan untuk menemukan yang sebelumnya tidak diketahui tetapi berpotensi berguna pengetahuan dari data teks tidak terstruktur atau semi terstruktur (Ding, 2018). Penambangan teks juga menghadapi masalah seperti jumlah besar data, dimensi sebuah data dan struktur akan berubah-ubah dan "*noise*". Tidak seperti penambangan data, terutama saat proses agar data menjadi terstruktur, data yang dipakai oleh penambangan teks biasanya tidak terstruktur, setidaknya teks semi terstruktur.

Penambangan teks adalah analisis teks di mana sumber data yang didapatkan dari sebuah dokumen, memiliki tujuan yaitu untuk menemukan beberapa kata yang bisa mewakili isi dari dokumen tersebut agar bisa dilakukannya analisis keterlibatan, keterkaitan dan kelasnya (Lesmeister, 2015). Penambangan teks melibatkan pengolahan dokumen menjadi pembagian teks, ekstraksi kata dan

informasi. Metode Ini biasa dipakai untuk mendapat informasi yang bersumber dari data lewat identifikasi dan pengamatan data yang diinginkan (N. Indurkhya, 2010). Penambangan teks biasanya digunakan untuk penanganan klasifikasi, pengelompokan, informasi ekstraksi dan masalah pencarian informasi (M. W. Berry, 2010). Penambangan teks mampu mengidentifikasi secara emosional tentang suatu pernyataan (L. Zhang, 2011). Proses pengerjaan penggalian data diadopsi secara luas dari riset penambangan data. Namun yang menjadi pembedanya ialah pola pada penggalian data yang didapat dari kumpulan yang tidak teratur atau acak, sedangkan pola penambangan data diambil dari basis data yang teratur. Proses penambangan teks yang umum termasuk ke dalam pembagian kelompok teks, ekstraksi entitas, analisis sentimen, produksi taksonomi granula, inferensi dan *entity-relationship*, yaitu untuk mempelajari hubungan antar entitas (P. Bhargavi, 2009).

2.6 *Naives Bayes*

Algoritma *Naïve Bayes* biasanya dipakai untuk mencari suatu nilai probabilitas tertinggi untuk mengklasifikasikan data uji ke pengelompokan yang cocok. Algoritma ini pertama kali ditemukan oleh ilmuwan Inggris Thomas Bayes, menggunakan metode probabilitas yang memprediksi kemungkinan yang akan terjadi berdasarkan pengalaman dan disebut dengan Teorema Bayes. *Naïve Bayes Classifier* merupakan algoritma yang sering digunakan untuk tujuan penambangan data dalam menggunakan algoritma (M. Hall, 2007) untuk kecepatan dalam prosesnya bisa terbilang cukup cepat, pengimplementasiannya cukup mudah dengan struktur yang sederhana dan tingkat efektivitas yang cukup akurat (S. Taheri, 2013). *Naïve Bayes* menentukan kelas yang mempunyai kemungkinan tertinggi dan menghitung probabilitas kelas berdasarkan dari atributnya. Klasifikasi menggunakan *Naïve Bayes* biasanya pengelompokan kelas berdasarkan kemungkinan yang cukup sederhana dengan asumsi bahwa setiap atribut dalam data memiliki nilai eksklusif masing-masing. Dalam model probabilitas, pada semua kelas *k* memiliki banyak atribut dapat ditulis persamaannya seperti pada Persamaan berikut:

$$P(Y|X) = \frac{P(X|Y)\ P(Y)}{P(X)} \quad (2.1)$$

Perhitungan *Naïve Bayes* adalah peluang munculnya dokumen X di dalam kelas, probabilitas tersebut dikalikan dengan kategori kelas P. Maka menghasilkan, distribusi dokumen P(x) akan terjadi. Jadi, kita mendapatkan rumus perhitungan *Naïve Bayes* ditulis dalam persamaan:

$$Y_{MAP} = \underset{V}{^{argmax}} \frac{P\left(x_1, x_2, x_3, naïve, x_n \middle| Y_j\right) P(Y_j)}{P(x_1, x_2, x_3, \dots, x_n)} \quad (2.2)$$

Jika $P(X_1, \dots, X_n)$ adalah nilai untuk semua kategori dari Y_j maka persamaannya adalah:

$$Y_{MAP} = \underset{V}{^{argmax}} P\left(x_1, x_2, x_3, naïve, x_n \middle| Y_j\right) P(Y_j) \quad (2.3)$$

Dan persamaan bisa ditulis menjadi:

$$Y_{MAP} = \underset{V}{^{argmax}} \prod_{i=1}^{n} P(x_i|Y_j)\ P(Y_j) \quad (2.4)$$

Keterangan:

Y_{MAP}: semua kategori yang diuji

Y_j: merupakan kategori komentar

V : merupakan kategori komentar

j_1: merupakan komentar positif

j_2: merupakan komentar negatif

j_3: merupakan komentar netral

$P(X_i|Y_j)$: probabilitas X_i pada kategori Y_j

$P(Y_j)$: probabilitas dari Y_j

Untuk meningkatkan hasil prediksi maka dapat dilakukan pembobotan atribut pada kelas, dengan menghitung berat kelas pada atribut. Akurasi klasifikasi akan menjadi dasar yang berarti bukan hanya probabilitasnya saja, tetapi juga bobot pada semua yang menjadi dasar atribut kelas (I. Rish, 2001). Perhitungan $P(Y_j)$ dan $P(X_i|Y_j)$ pada saat melatih data dengan menggunakan persamaan:

$$P(Y_j) = \frac{docs_j}{contoh} \quad (2.5)$$

$$P(Y_j) = \frac{n_k + 1}{n + kosakata} \quad (2.6)$$

Keterangan:

$docs_j$: merupakan jumlah dokumen pada setiap kategori j

$contoh$: merupakan jumlah dokumen dari semua kategori

n_k: merupakan jumlah frekuensi kemunculan dari setiap kata

n : merupakan jumlah frekuensi kemunculan dari setiap kategori

$kosakata$: merupakan jumlah semua kata dari semua kategori

2.7 *Natural Language Processing*

Natural language processing adalah serangkaian motivasi teoritis untuk menganalisis dan mengekspresikan teks yang muncul secara alami pada banyak tingkatan analisis. Tujuan utama dari NLP adalah "agar bisa memproses teks layaknya manusia". Untuk kata 'pengolahan' akan ditekankan dan tidak diperbolehkan menggunakan kata 'memahami'. Karena NLP sendiri memiliki nama lain yaitu *Natural Language Understanding* (NLU) di antaranya sebagai berikut (Chowdhary, 2020):

1 Parafrase teks masukan
2 Menjawab konteks dari teks tersebut yang ditanyakan
3 Mendapatkan simpulan dari teks tersebut

Sementara NLP adalah bidang penelitian dan aplikasi yang baru, dibandingkan dengan yang lain pendekatan teknologi informasi, ada cukup keberhasilan sampai saat ini bahwa menyarankan bahwa teknologi akses informasi berbasis NLP akan terus menjadi area utama penelitian dan pengembangan informasi sekarang dan jauh ke depan. Selain itu, mencakup area spesifik seperti penguraian probabilitas, ambiguitas dan resolusinya, ekstraksi informasi, analisis wacana, antarmuka, pemikiran, penalaran dan keragaman kausal NLP (Chowdhary, 2020).

2.8 *Bag of Word*

Bag-of-Word (BOW) adalah sebuah model yang ada di dalam *natural language processing* (NLP), untuk mengetahui nilai dari sebuah dokumen atau kalimat diwakilkan oleh tas (*bag*) *multiset* yang berisi semua kata dari semua dokumen yang digunakan. *Bag-of-Word* menggunakan kalimat diubah menjadi *vector* untuk mewakili dokumen. Walaupun model mengabaikan urutan katanya akan tetapi terbukti bisa menangkap setiap kata dan informasi pada topik tersebut. *Bag-of-Word* akan memecah kalimat per kata, lalu setiap kata selain *stopwords* akan dibandingkan dengan setiap kalimat, *Bag-of-Word* mencari frekuensi dari kata-kata yang memiliki arti sama. Maka dari itu *Bag-of-Word* (BOW) juga disebut *sentence to vector,* model ini umum dan biasa digunakan dalam NLP (Basuki, 2020).

2.9 Evaluasi Model

Evaluasi model merupakan proses untuk mengukur keberhasilan performa dari suatu model dengan membandingkan hasil yang didapat. Pengukuran evaluasi model bertujuan untuk data latih agar bisa mengevaluasi model yang sudah dibuat. Pada tahap evaluasi model ini memiliki beberapa perhitungan *confusion matrix,* yaang berisi *accuracy, precision, recall* dan *f-measure.* Hasil tersebut berisi *true positive, true negative, false positive* dan *false negative.* Berikut merupakan *confusion matrix* seperti pada Tabel 2.1

Tabel 2.1 Tabel *Confusion Matrix*

	Prediksi Positif	Prediksi Negatif
Aktual Positif	*True Positive* (TP)	*False Negative* (FN)
Aktual Negatif	*False Positive* (FP)	*True Negative* (TN)

Tabel 2.1 dijelaskan bahwa *true positive* merupakan jumlah pada prediksi kelas yang diprediksi positif dan juga kelas aktualnya positif. Untuk *true negative* merupakan jumlah pada prediksi kelas yang diprediksi negatif dan juga kelas aktualnya. Untuk *false positive* merupakan jumlah pada prediksi kelas yang diprediksi positif sedangkan kelas aktualnya negatif. Untuk *false negative*

merupakan jumlah pada prediksi kelas, sedangkan kelas aktualnya positif. Berikut merupakan perhitungan pada evaluasi model:

$$Accuracy: \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

$$Precision: \frac{TP}{TP+FP} \quad (2.8)$$

$$Recall: \frac{TP}{TP+FN} \quad (2.9)$$

$$F1\text{-}Score: \frac{2*recall*precision}{recall+precision} \quad (2.10)$$

Accuracy merupakan evaluasi dari perhitungan prediksi yang memiliki nilai benar pada semua data yang diprediksi. *Precision* merupakan evaluasi dari hasil yang diperoleh dengan cara menghitung prediksi benar dan kelas aktualnya positif dari semua data yang memiliki nilai positif. *Recall* merupakan cara menghitung untuk kondisi prediksi benar dan kelas aktualnya positif dari semua data yang memiliki nilai positif. *F-Measure* merupakan perhitungan akhir dari semua hasil perhitungan evaluasi yang berguna untuk mencari nilai tengah (Basuki, 2020).

2.10 Penelitian Terdahulu

Penelitian terdahulu menjelaskan beberapa literatur penelitian yang telah dipelajari sebagai referensi untuk penelitian mengenai sentimen analisis media sosial Twitter pada kasus jaminan hari tua menggunakan metode *naïve bayes*. Pada Tabel 2.2 berikut merupakan penelitian terdahulu yang digunakan:

Tabel 2.2 Penelitian Terdahulu

No	Nama	Hasil penelitian
1	(Angelina, Ardiansyah, Tuti, Laela & Windu, 2020)	Penelitian ini berhasil mendapatkan algoritma yang efektif dan terbaik dalam mengklasifikasikan komentar positif dan komentar negatif terkait dengan aplikasi Ruang Guru menggunakan algoritma *Naive Bayes* (NB), *Support Vector Machine* (SVM), *K-Nearest Neighbour* (K-NN) dan *feature selection* dengan algoritma *Particle Swarm Optimization* (PSO)
2	(Akhmad & Muhammad, 2018)	Penelitian ini dibuat untuk mengimplementasikan algoritma KNN (*K – Nearest Neighbor*) dalam analisis pengguna Twitter tentang topik Pilkada DKI 2017
3	(Krisdiyanto, 2021)	Pada penelitian ini akan dilakukan proses opini masyarakat mengenai kebijakan PPKM dengan mengklasifikasikan opini ke dalam 2 sentimen yaitu positif atau negatif. Klasifikasi dilakukan dengan menggunakan metode *Naïve Bayes Clasifiers*
4	(Salim & Mayary, 2020)	Pada penelitian ini mengadopsi analisis sentimen dengan metode *Lexicon Based* dan *K-Nearest Neighbor*
5	(Sari, Insan & Rini, 2018)	Pada penelitian ini untuk mendapatkan informasi apakah sebuah *tweet* adalah opini positif, opini negatif atau opini netral menggunakan metode *Naive Bayes*
6	(S.Basuki, S.Maghfiroh,	Pada penelitian ini mendapat informasi *tweet* tindak kejahatan menggunakan ekstraksi fitur *Bag of Word*

No	Nama	Hasil penelitian
	Y.Azhar 2020)	
7	(Jyoti, Mihir, Nitima, Abhishek, Rabindra & Ankush, 2021)	Pada penelitian ini dilakukan analisis *tweet* yang di *posting* oleh pengguna layanan microblogging. Twitter, menggunakan pemrosesan data alami menggunakan metode *Naïve Bayes Classifier*
8	(Dedi, Eka & A Ferico, 2020)	Pada penelitian ini dilakukan analisis media sosial Twitter mengenai KPK RI menggunakan metode *Support Vector Machine* (SVM)
9	(Yonathan & Eri, 2018)	Pada penelitian ini dilakukan analisis sentimen merupakan proses klasifikasi dokumen tekstual, yaitu kelas negatif dan positif menggunakan algoritma *naïve bayes classifier*
10	(Rian, Agung & Ira, 2020)	Pada penelitian ini dilakukan analisis sentimen analisis untuk memahami apakah data tekstual tersebut termasuk opini negatif atau opini positif keluhan atau kepuasan terhadap layanan Indihome