BAB II
LANDASAN TEORI

2.1 Hate Speech

Hate speech adalah suatu ujaran kebencian yang dilakukan di berbagai media, yang membuat semakin populer karena perbedaan yang sampai mewakili berbagai kelompok seperti suku, ras, etnis dan agama (Rohman, 2016). *Hate speech* ini biasanya semakin meningkat intensitasnya di media sosial menjelang pemilihan umum kepala daerah. Dasar yang paling banyak menyebabkan perselisihan atau perbedaan adalah masalah sara(suku, ras, agama diantara golongan). Kejahatan ini memilki potensi mengancam ke stabilitas negara dan keamanan. Terkait dengan permasalahan di atas pemerintah mengeluarkan aturan terkait penanganan ujaran kebencian *(Hate Speech).*

2.2 Twitter

Twitter adalah media jejaring sosial unik yang memfasilitasi penggunanya untuk dapat mengirim dan menerima terkait segala aktivitas, opini, serta segala sesuatu hal terhadap pengguna lainnya secara publik yang biasa disebut *tweet* atau juga dapat mengirim pesan secara pribadi dalam komunitas *Twitter*. Komunitas Twitter itu adalah:

a. *Following*

   Following adalah komunitas ini diartikan dengan mengikuti pengguna media jejaring sosial Twitter lainnya. pengguna juga dapat melihat *tweets* yang ditampilkan oleh semua pengguna yang diikuti. Dengan mengikuti pengguna lain di *Twitter* dapat diartikan pengguna berlangganan dengan tampilan *tweets* mereka.

b. *Followers*

   Followers adalah pengguna lain yang membaca tampilan *tweets* pengguna dan mengikuti pengguna di media jejaring sosial *Twitter*. *Followers* atau pengikut dapat melihat *tweets* yang pengguna kirim ke jejaring sosial *Twitter*.

Tweets adalah kiriman pesan singkat yang memiliki panjang yang terdiri dari 140 karakter, sehingga gampang untuk difilter (Crow Communications, 2011).

2.3 Sentimen *Analysis*

Sentiment Analysis (SA) atau biasa di sebut juga sebagai *opinion mining* adalah suatu riset komputasi nal dari emosi yang diungkapkan atau diekspresikan berupa tulisan (*tekstual*) dan *opini sentiment* (Zulfa & Winarko, 2017). *Sentiment Analysis* (SA) merupakan suatu proses untuk memahami data, mengolah data dan mengekstrak data tekstual secara otomatis dengan tujuan mendapatkan informasi sentimen atau intisari dari data yang terdapat di dalam suatu kalimat opini. *Sentiment Analysis* (SA) ini sendiri untuk melihat pendapat atau kecenderungan opini terhadap suatu masalah atau objek oleh seseorang, apakah kecenderungan tersebut mengarah ke hal positif atau negatif (Rozi, Pramono, & Dahlan, 2012).

Sentiment Analysis (SA) dibedakan berdasarkan sumber dari datanya, Adapun beberapa level yang paling banyak digunakan dalam penelitian adalah *sentiment analysis* (SA) berdasarkan level elemen dan sentimen *analysis* (SA) berdasarkan level kalimat (Falahah & Nur, 2015). Berdasarkan sumber datanya *sentiment analysis* dibagi menjadi 2 kelompok besar yaitu :

a. *Coarse-grained Sentiment Analysis*

   Sentiment analysis ini dilakukan pada level dokumen. Secara garis besar *sentiment analysis* jenis ini fokus utama dengan seluruh isi dokumen yang akan di analisis sebagai sentimen negatif dan sentimen negatif (Falahah & Nur, 2015).

b. *Fined-grained Sentiment Analysis*

   Sentiment analysis ini dilakukan pada level kalimat. Dalam *sentiment analysis* ini fokus untuk menganalisis data dari setiap kalimat (Falahah & Nur, 2015).

2.4 *Text Mining*

Text mining merupakan konsep terapan dalam teknik *data mining* untuk mencari pola inti suatu teks, dengan tujuan mendapatkan informasi yang terkandung dalam suatu teks yang dapat di manfaatkan dengan tujuan tertentu. Dari ketidak ter aturan suatu data teks dan banyaknya kandungan kata-kata imbuhan serta kiasan dalam suatu data teks, dalam proses text *mining* memerlukan tahapan-tahapan untuk mendapatkan data teks yang lebih terstruktur.

Tahapan proses yang harus di lewati *text mining* di bagi menjadi 5 bagian untuk memperoleh hasil yang diinginkan. Adapun 5 proses tersebut yang harus dijalankan dalam *text mining*, adalah (Hidayatullah, 2014):

a. *Text preprocessing*

Tahapan awal dalam *text mining* adalah *text prepocessing* dengan tujuan mempersiapkan data teks yang nantinya akan mengalami pengolahan data teks berikutnya. Selain itu biasanya dalam proses *text preprocessing* ini juga menggunakan *case folding*, yaitu pengubahan data teks pada karakter huruf besar dalam data teks menjadi huruf kecil.

b. *Text transformation*

Dalam tahap ini hasil yang di dapatkan dari proses *text preprocessing* akan dilakukan proses transformasi. Proses transformasi ini dilakukan dengan mengurangi jumlah dari setiap kata dalam data teks *stop word removal* dan mengubah kata-kata menjadi kata dasar dalam data teks *stemming.*

Stop word removal adalah suatu kata yang memiliki keunikan kata dari data teks seperti kata sambung, serta kata kepunyaan yang nantinya pada proses transformasi kata-kata tersebut tidak akan dihitung. Selain itu proses *stop word removal* dapat mengurangi beban kinerja sistem, karena kata yang akan di ambil adalah kata-kata yang dianggap penting.

Stemming adalah suatu proses dalam teks transformasi yang digunakan sebagai memproses kata-kata di dalam data teks agar menjadi kata dasar.

c. *Feature selection*

Dalam tahapan *feature selection* adalah tahapan penting dalam *text mining.* Karena dalam tahap ini dilakukan proses pembuangan beberapa *term atau kata* yang tidak terkait sehingga memperoleh *term* atau kata penting sebagai wakil kumpulan dokumen yang di analisis. Dalam *feature selection* terdapat beberapa metode yang digunakan, diantaranya adalah sebagai berikut:

1. *Document Frequency*

   Document Frequency adalah seberapa banyak kemunculan suatu *term* atau kata dalam data dokumen yang akan dianalisis.

2. *Term Frequency*

   Term frequency ($tf_{t,d}$) adalah menghitung banyaknya kemunculan *term* atau kata dalam suatu *corpus* terhadap suatu bobot *term t* atau kata pada dokumen *d.*

3. *Term Frequency-Inverse Document Frequency* (TF-IDF)

   TF-IFD itu sendiri terdiri dari *Term Frequency* dan *Inverse Document Frequency.*

d. *Pattern discovery*

Tahap *pattern discovery* berguna untuk menemukan suatu *knowledge* atau pola dengan menggunakan beberapa teknik data *mining* sebagai contoh *classification* dan *clustering.*

e. *Interpretation*

Tahapan terakhir ini adalah melakukan proses interpretasi ke sebuah bentuk kemudian di evaluasi.

2.5 *Naive Bayes Classifier*

Naïve Bayes Classifier adalah algoritme yang terdapat dalam teknik data *mining* yang menerapkan teori *Naïve Bayes* dalam klasifikasi, semua itu mendasarkan pada nilai suatu atribut secara kondisional saling bebas jika diberikan suatu nilai *output* (Ridwan, Suryono, & Sarosa, 2013). *Naïve Bayes Classifier* yaitu suatu metode pengklasifikasian berakar pada teorema *bayes*. Teorema *bayes* adalah pendekatan statistik yang fundamental dalam *pattern recoginition* (pengenalan pola).

Keuntungan menggunakan metode *Naïve Bayes Classifier* adalah metode ini hanya memerlukan nilai atau jumlah data pelatihan (*Training Data*) yang kecil sebagai penentu estimasi parameter yang nantinya diperlukan dalam proses klasifikasi data (Manalu, Sianturi, & Manalu, 2017). Berikut adalah persamaan 2.1 *Teorema Bayes* (Hidayatullah, 2014):

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \qquad (2.1)$$

Keterangan :

E = Data yang belum diketahui *class*nya

H = Suatu *class* spesifikasi hipotesis data E

P(H|E) = *probabilitas posterior*, *probabilitas* maka akan muncul H jika diketahui E

P(E|H) = *probabilitas posterior*, *probabilitas* maka akan muncul E jika diketahui H

P(H) = *probabilitas prior*, *probabilitas* kejadian H

P(E) = *probabilitas prior*, *probabilitas* kejadian E

Peraturan dari Naïve Bayes Classifier :

Jika $P(h1|e) < P(h2|e)$, maka e dapat diklasifikasikan h2. $P(h1|e)$ mengindentifikasi probabilitas h1 berdasarkan terjadi pada kondisi e, begitu pula sebaliknya dengan h1. Klasifikasikan dari e sesuai dengan probabilitas terbesar antara probabilitas e dengan semua kelas.

2.6 *Cross Validation*

Cross Validation adalah salah satu teknik sebagai penilaian memvalidasi kan keakuratan dari suatu model yang dibuat berdasarkan dataset tertentu. Pembuatan model ini biasanya

bertujuan sebagai penentu prediksi maupun pengklasifikasian terhadap suatu data baru yang dapat dikatakan belum pernah muncul di dalam dataset. Data yang dipergunakan sebagai proses pembuatan model dapat disebut juga sebagai data latih atau data *training*, sedangkan data yang akan sebagai validasi model disebut sebagai data *test*. Salah satu metode *Cross-Validation* yang paling banyak digunakan adalah *K-Fold Cross Validation. K-fold* bekerja melipat data sebanyak K dan melakukan proses mengulang sebanyak K juga.

2.7 *Performance Evaluation Measure*

Performance Evaluation Measure (PEM) atau juga bisa disebut sebagai pengukuran evaluasi performa. Pengukuran evaluasi performa adalah suatu proses tahapan yang berguna sebagai pengukur performa suatu sistem. *Performance Evaluation Measure* ini banyak di pergunakan dalam kasus training data. Dibuatnya proses ini bertujuan untuk mengevaluasi model yang sudah dibuat. Beberapa perhitungan yang terdapat dalam *Performance Evaluation Measure* untuk menemukan nilai *Performance Evaluation Measure,* biasanya diterapkan secara parsial ataupun sebagai kombinasi. Beberapa perhitungan yang terdapat dalam *Performance Evaluation Measure* seperti (Amin , 2012):

a. *Precision*.

   Precision adalah tingkat ketepatan atau ketelitian dari hasil antara pengujian request pengguna dengan jawaban sistem.

b. *Recall*.

   Recall adalah ukuran ketepatan atau ketelitian antara informasi yang sama dengan informasi yang sudah pernah ada sebelumnya.

c. *Accuration.*

   Accuration adalah sebagai pembanding antara informasi yang dijawab oleh sistem dengan benar dengan keseluruhan informasi.

Rumus *precision* (pre) :

$$pre = \frac{TP}{TP + FP} \tag{2.2}$$

Rumus *recall (rec) :*

$$rec = \frac{TP}{TP + FN} \tag{2.3}$$

Rumus *accuration (acc) :*

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4)$$

Performance Evaluation Measure biasanya digambarkan dalam bentuk tabel atau *confusion matrix*, tabel ini berisi dari hasil pengujian model yang telah melalui proses perbandingan dengan *dataset*, tabel ini terdiri dari kelas *true* dan *false*, seperti pada Tabel 2.1.

Tabel 2.1 *Confusion Matrix*

	Class	
True Class	***Positive***	***Negative***
Positive	TP	FN
Negative	FP	TN

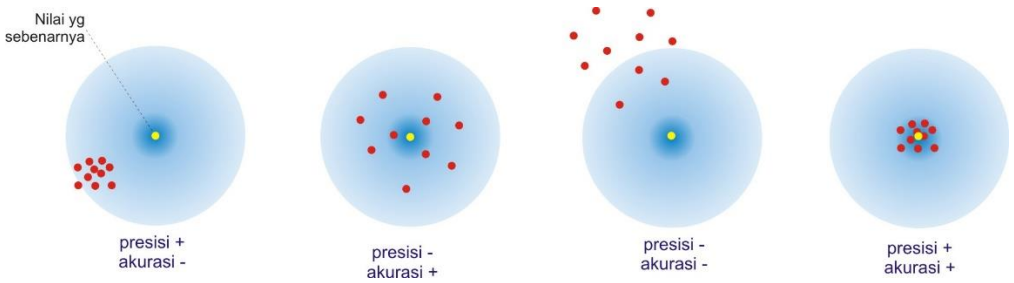
Keterangan:

TP (*true positive*) : contoh data bernilai positif yang diprediksi benar sebagai positif

TN (*true negative*) : contoh data bernilai negatif yang diprediksi benar sebagai negatif

FP (*false positive*) : contoh data bernilai negatif yang diprediksi salah sebagai positif

FN (*false negative*) : contoh data bernilai positif yang diprediksi salah sebagai negative



Gambar 2.1 Ilustrasi Gambar *Precision* dan *Accuracy*

Dari ilustrasi Gambar 2.1 di atas dapat dijelaskan gambaran persebaran data dengan *precision* dan *accuracy*. Dapat diilustrasikan dengan permisalan di bawah ini:

Misalkan peneliti ingin mengukur kinerja terhadap mesin pemisah ikan yang memiliki tugas sebagai pemisah antara ikan arwana dari semua ikan yang telah dikumpulkan oleh

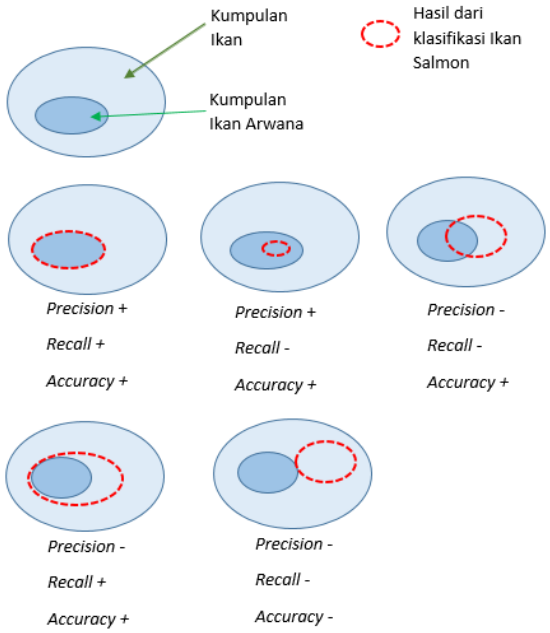
peneliti. Proses pengujiannya akan memasukkan 100 ikan arwana dan 900 adalah ikan-ikan lainnya (bukan ikan arwana). Dari proses memasukkan tadi hasil dari mesin memisahkan 110 yang terdeteksi bahwa itu adalah ikan dan hanya 90 ikan yang terdeteksi sebagai ikan arwana, sedangkan 20 lainnya adalah ikan lainnya (bukan ikan arwana), dapat diperjelas dengan melihat Tabel 2.2.

Tabel 2.2 Contoh Hasil *Confusion Matrix*

		Nilai Sebenarnya	
		True	***False***
Nilai Prediksi	***True***	90	20
	False	10	880

Dapat dilihat dari Tabel 2.2 di atas bisa dihitung dengan menggunakan persamaan (2.2), persamaan (2.3), dan persamaan (2.4) di atas. Dari kasus Tabel 2.2 di atas dapat disimpulkan bahwa mesin tersebut memiliki nilai *precision* sebesar 82%, nilai *recall* 90%, dan nilai *accuracy* sebesar 97%

Dari hasil kasus di atas bisa dapat disimpulkan gambaran seperti Gambar 2.3 di atas, apabila membandingkan dari nilai *precision,* nilai *recall* dan nilai *accuracy* :



Gambar 2.2 Perbandingan Precision, nilai Recall dan nilai Accuracy

2.8 Penelitian Serupa

Dalam pembuatan penelitian peneliti, ada beberapa penelitian sebelumnya yang sudah pernah ada dilakukan oleh orang lain yang mirip dan bahkan dijadikan sebagai acuan dari penelitian. Beberapa penelitian yang serupa dapat dilihat sebagai berikut:

a. Terdapat penelitian yang menganalisis terkait kepribadian seseorang (Sarwani & Mahmudy, 2015). kepribadian seseorang adalah hal penting untuk mengambil suatu kesimpulan atau keputusan yang berdampak baik atau buruk. Sistem ini mengambil salah satu layanan sosial yang masih bisa populer hingga saat ini yaitu *Twitter*. *T*witter hingga saat ini masih aktif menghasilkan 110 juta *tweet* per hari dan masih mempunyai lebih dari 200 juta pengguna. Dalam memproses data *Twitter* untuk menganalisis kepribadian seseorang sangat dibutuhkan metodologi yang tepat sebagai menentukan ke akuratan dari hasil. *Tweet* pada *T*witter adalah kumpulan kata yang tidak baku yang nantinya perlu diolah agar menjadi data kata yang dapat diproses. Oleh karena itu sebagai pengolahan data diperlukannya suatu proses *pre-processing* sebagai awal pengolahan kata yang kemudian akan diteruskan ke proses klasifikasi. Metode yang digunakan sebagai klasifikasi adalah metode *Naïve Bayes Classifier*. Metode tersebut dipilih karena memberikan kemudahan dan sederhana dalam proses pengolahan data serta memberikan tingkat ke akurasi yang baik. Hasil dari penelitian ini adalah menyatakan bahwa kepribadian karakter seseorang dapat diketahui dari postingan *tweet Twitter* mereka.

b. Terdapat penelitian berkaitan dengan kenaikan popularitas media jejaring sosial terus meningkat dalam beberapa tahun terakhir seperti *Twitter, Facebook,* dan *Youtube*. Salah satu dari beberapa media jejaring sosial tersebut dapat dimanfaatkan dalam bidang pemilihan umum adalah *Twitter* (Hidayatullah, 2014). Data statistik menunjukkan sejak kemunculan *Twitter* tahun 2006 terus mengalami peningkatan, *Twitter* sendiri mempunyai seratus juta lebih pengguna aktif 50 persen dari pengguna melakukan posting dan *sign in* setiap hari dengan 250 *tweets* lebih di-*posting.* Kebiasaan mem*posting tweet* pengguna mejadi salah satu sebagai acuan menentukan sentimen pengguna terhadap tokoh publik. Adapun metode yang dipergunakan sebagai mengklasifikasi data kata adalah *Naïve Bayes Classifier* dengan fitur tambahan fitur negasi berguna mengetahui negasi pada postingan *tweet*. Dengan adanya penelitian ini dapat membantu berbagai pihak yang ingin mengerti dan mengetahui tanggapan publik terkait tokoh publik yang layak untuk dapat maju sebagai pilpres dengan melalui media postingan tweet pada *Twitter*. Selain dari pada itu,

peneliti ini dapat dijadikan sebagai referensi penelitian fitur negation dalam penelitian sentimen analisis.

c. Penelitian menganalisis peran Twitter yang memiliki pengaruh yang sangat besar sebagai kesuksesan atau kehancuran citra seseorang. (Buntoro, 2016). Banyaknya gerakkan-gerakan yang dikerjakan di Twitter dapat mempengaruhi dari perspektif positif hingga perspektif negatif. Penelitian ini menganalisis *hastag* atau tagar pada Twitter dengan menggunakan dua sentimen yaitu *HateSpeech* dan *GoodSpeech*. Proses yang digunakan untuk menganalisis data di penelitian ini yaitu *Naïve Bayes Classifier (NBC)* dan *Support Vector Machine (SVM)* dengan mungumpulkan 522 *tweet.* Hasil akurasi tertinggi didapatkan saat menggunakan metode klasifikasi Support Vector Machine (SVM) dengan *tokenisasi unigram*, *stopword list* Bahasa Indonesia dan emoticons, dengan nilai rata-rata akurasi mencapai 66,6%, nilai presisi 67,1%, nilai recall 66,7% nilai TP rate 66,7% dan nilai TN rate 75,8%.

d. Dalam penelitian analisis sentimen terhadap media sosial, analisis sentimen adalah salah satu proses untuk menentukan emosi, opini dan sikap yang dicerminkan seseorang dari teks biasanya analisis ini untuk mengklasifikasi menjadi opini negatif dan opini positif (Cindo, Rini, & Ernitita, 2019). Selain itu analisis ini juga dapat digunakan sebagai menganalisis opini terkait produk atau layanan dan bisa juga topik tertentu di berbagai media, dalam penelitian ini menggunakan 3 objek penelitian data Twitter, Facebook, dan Web Scraping. Peneliti menggunakan metode *Naïve Bayes Classifier* dan *Support Vector Machine* pada saham perusahaan. Selain itu peneliti juga membandingkan beberapa metode seperti *logistic regression* dan *lexical-based*. Hasil akhir yang diperoleh *logistic regression* lebih unggul 93% dibandingkan dengan *Naïve Bayes Classifier* 88.20%, *SVM* 85.20%, dan *lexical-based* 92% pada tahun 2014 hingga 2018 terkait saham.