BAB II
LANDASAN TEORI

2.1 Twitter

Menurut Jansen dkk (2009), Twitter adalah situs web yang dimiliki oleh Twitter Inc yang merupakan jejaring sosial berupa *microblog* yang memungkinkan penggunanya mengirim dan membaca pesan yang disebut *tweet*. *Tweet* adalah teks tulisan dengan panjang 140 karakter yang ditampilkan pada halaman *profile* penggunanya, karena hanya 140 karakter. Pengaturan standar pada *tweet* adalah publik yang memungkinkan pengguna lain mengikuti dan membaca *tweet* tanpa perlunya izin. Media sosial Twitter saat ini yang semakin berkembang digunakan masyarakat untuk mendapatkan informasi, Twitter yang menyajikan suatu informasi secara *realtime* dapat menyebarkan informasi secara cepat. Twitter tidak hanya digunakan untuk mencari informasi, media sosial Twitter menjadi keuntungan untuk memasarkan suatu produk atau jasa. Dengan melakukan promosi pada media sosial, menjadikan target pemasaran produk lebih luas dan biaya promosi yang murah.

2.2 Iklan

Iklan adalah memberitahukan dan memasarkan suatu produk melalui media seperti televisi, radio, koran, majalah dan media sosial. semakin berkembangnya teknologi media sosial saat ini tidak hanya digunakan untuk menjalin komunikasi, tetapi digunakan untuk memasarkan produk, Twitter yang selalu menjadi sorotan di masyarakat menjadikan Twitter cocok untuk memasarkan suatu produk barang dan jasa.Untuk membuat iklan di media sosial, pemasang iklan harus mengetahui faktor-faktor yang dapat mempengaruhi keputusan konsumen untuk membeli suatu produk, seperti memberitahukan identitas lengkap kepada konsumen agar lebih mudah melakukan transaksi (Junia dan Rosyad, 2013).

Menurut Bittner dalam (Rendra, 2005) ada 2 jenis iklan sebagai berikut:

1. Iklan Standar

   Iklan yang ditata secara khusus untuk keperluan memperkenalkan barang, jasa, pelayanan untuk konsumen melalui media periklanan. Iklan ini memiliki keuntungan ekonomis

2. Iklan Layanan Masyarakat

   Iklan yang bersifat *non profit*, disebut *non profit* karena iklan yang tidak mencari keuntungan secara langsung. Namun keuntungan iklan ini bertujuan untuk keuntungan sosial

Menurut Robert V. Zacher dalam (Sumantoro, 2002) beberapa tujuan iklan diantaranya adalah:

1. Memberikan informasi mengenai suatu produk berupa barang, jasa dan ide
2. Menimbulkan rasa suka kepada atas produk yang diiklankan tersebut dengan memberikan prefensi-prefensi.
3. Meyakinkan produk tersebut sehingga mereka berusaha untuk memiliki atau menggunakan produk tersebut.
4. Dari sudut pandang konsumen, konsumen mengetahui informasi produk tersebut, baik harga, spesifikasi, fungsi, dan lain-lain.

2.3 *Information Extraction*

Information Extraction (IE) merupakan proses untuk menemukan dan mencari kata dan entitas untuk mengambil kesimpulan isi yang tidak terstruktur. Secara singkat ekstraksi informasi adalah sebuah proses membuat data menjadi relevan dan terstruktur, seperti relasi pada database dan juga seperti basis pengetahuan. Oleh karena itu, kegiatan utama untuk melakukan proses ekstraksi informasi adalah pengenalan entitas (*Named Entity Recognition*) dan ekstraksi relasinya(Jiang, 2012)(Jiang, 2012)(Jiang, 2012)(Jiang, 2012)(Jiang, 2012)(Jiang, 2012). Pengenalan entitas memanfaatkan pola kemuncuan entitas pada teks. Oleh karena itu, terdapat dua pendekatan dalam pengenalan entitas, yaitu pendekatan berbasis aturan dan pendekatan berbasis pembelajaran (Jiang, 2012).

Ekstraksi informasi merupakan bagian dari *Natural Language Processing* (NLP). Fitur-fitur NLP adalah sebagai berikut:

1. *Orthography* atau *Case*, adalah penggunaan huruf besar dan huruf kecil oleh token.
2. *Tokenkind*, adalah jenis token: kata, angka, simbol, atau tanda baca.
3. *Lemma*, adalah bentuk dasar dari token, merupakan hasil dari analisis morfologikal.
4. *Part of Speech* (POS), yaitu tata bahasa dari token, apakah merupakan kata benda, kata kerja, dan sebagainya.
5. *Lookup* atau *gazetteer*, adalah daftar kata dan istilah berbagai kategori, misalnya kategori negara berisi daftar seluruh negara yang ada di dunia.
6. *Entity*, adalah salah satu fitur *Named Entity Recognition* yang dimiliki ANNIE, dan bekerja berdasarkan aturan ekstraksi yang sudah terdefinisi.

2.4 *Text Mining*

Text mining dapat diartikan sebagai penemuan informasi yang baru dan tidak diketahui sebelumnya oleh komputer, secara otomatis mengekstrak informasi dari sumber-sumber yang berbeda. Pada dasarnya proses kerja dari *text mining* banyak mengapdopsi dari penelitian *data mining*, dibandingkan dengan jenis data yang disimpan dalam database, *text mining* memanfaatkan teks yang tidak terstruktur dan sulit untuk diselesaikan dengan algoritma (Witten, 1999). *Text mining* merupakan teknik yang mampu menangani permasalahan seperti klasifikasi, *information retrieval, information extraction* dan *clustering.* (Berry dan Kogan, 2010)

2.4.1 *Preprocessing*

Tahapan awal dari *text mining* adalah *preprocessing* yang bertujuan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahapan berikutnya. *Preprocessing* dilakukan dengan cara mengubah data ke bentuk yang mudah diproses oleh sistem. Hasil *preprocessing* digunakan pada tahap selanjutnya. Tahap pada *preprocessing* mencakup semua langkah yang untuk mempersiapkan data yang digunakan pada operasi *knowledge discovery* pada sistem *text mining.*

Pada penelitian ini menggunakan data Twitter yang sebagian besar berisi kata-kata atau kalimat tidak baku yang memiliki banyak *noise.* (Mujilahwati, 2016)

2.4.2 *Cleaning*

Cleaning adalah proses membersihkan dokumen dari komponen-komponen yang tidak diperlukan. Kata yang dihilangkan adalah karakter HTML, link url (http://situs.com), *hashtag* (#) dan RT (*retweet*).

2.4.3 *Case Folding*

Case folding adalah proses mengubah semua huruf kapital dalam dokumen menjadi huruf kecil. Proses *case folding* dilakukan untuk meratakan semua kata agar dapat mengatasi kata atau *term* ganda karena penulisan kata yang tidak sama.

2.4.4 *Tokenizing*

Tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Proses ini berfungsi untuk membagi karakter dalam dokumen teks menjadi berupa token yang digunakan untuk proses selanjutnya. Tahapan ini juga berfungsi memisahkan karakter yang dianggap sebagai tanda dibaca yang tidak diperlukan untuk pemrosesan selanjutnya selesai.

2.5 POS *Tagging*

Part of Speech Tagging (POS-Tagging) adalah suatu proses untuk memberi label pada setiap kata dalam kalimat dengan tag yang sesuai dengan kata tersebut. POS *Tagging* merupakan bagian dari NLP dan digunakan untuk menyelesaikan proses NLP seperti *word sense disambiguation, parsing, question answering, machine translation, speech recognition* dan *named entity recognition.* Ada beberapa jenis tagset untuk bahasa Indonesia, yaitu 35 jenis tag, 21 jenis tag, 37 jenis tag, dan 29 jenis tag. Penggunaan *tagset* untuk Bahasa Indonesia dan bahasa Inggris berbeda dikarenakan tidak semua *tagset* dapat diimplementasikan dari bahasa Inggris ke bahasa Indonesia.

Pada POS *Tagging* kelas kata dapat dibedakan menjadi kelas Kata Benda (*Noun*), Kata Kerja (*Verb*), Kata Sifat (*Adjective*), Kata Keterangan (*Adverb*) dan tagset bahasa Indonesia yang telah dikembangkan oleh UI dan ITB. Untuk tagset

dan kelas kata bahasa Indonesia menggunakan kamus dari kateglo lebih jelasnya pada Tabel 2.1 berikut:

Tabel 2.1 Kelas Kata

Kelas Kata	Pos Tag	Deskripsi	Contoh
Nomina	N	Kata benda	Buku, meja
Numerelia	NUM	Kata bilangan	Satu, dua
Preposisi	PRE	Kata depan	Dari, ke
Pronomina	PRO	Kata ganti	Aku, anda
Verba	V	Kata kerja	Bacok, baca
Adverbial	ADV	Kata keterangan	Terkadang
Konjungsi	K	Kata sambung	Atau, bahwa
Interjeksi	I	Kata seru	Aduh, ah
Adjektiva	ADJ	Kata sifat	Baik, abadi
Kata asing	FW	Kata asing	*Online, handphone*

Pada proses POS *Tagging* untuk pemberian kelas kata secara manual akan memakan waktu yang lama. Maka diperlukan proses *tagging* secara otomatis. Berbagai pendekatan untuk proses *tagging* telah banyak dikembangkan. Beberapa diantaranya menggunakan perhitungan probabilistik, statistika, dan berbasis *rule*.

2.6 *Chunking*

Chunking adalah metode yang efisien untuk mengidentifikasi frase pendek dalam teks. *Chunk tagging* adalah sebuah task di bidang NLP yang bertugas untuk memberi batas setiap *chunk* pada kalimat (Akhmad dkk, 2013). Frasa adalah gabungan dua kata atau lebih yang memiliki sebuah makna, menurut KBBI Bahasa Indonesia, frasa merupakan kumpulan kata nonpredikatif. Berikut Tabel 2.2 jenis frasa yang telah disusun mengacu pada pendeteksian frasa berdasarkan Putranto dkk (2016).

Tabel 2.2 Jenis-jenis frasa

No	Jenis Frasa	Frasa tag	Kelas kata	Contoh
1	Frasa Verbal	FV	V + ADJ	Kerja keras
			V + N	Banting tulang
			V + ADV	Hidup abadi
2	Frasa Nomina	FN	NUM + N	Sebuah apel
			N + N	Dunia malam
			N + ADV	Buku Harian
			N + ADJ	Baju bagus
3	Frasa Adverbial	FADV	ADV + ADJ	Begitu indah
			ADV + N	Keras kepala
			PRE + ADV	Dengan keras
4	Frasa Pronominal	FPRO	PRO + PRO	Saya ini
			PRE + PRO	Dengan saya
			PRO + N	Dia orang
5	Frasa Adjektival	FAJD	ADJ + ADV	Bagus sekali
			PRE + ADJ	Dengan manis
			ADJ + ADJ	Panas dingin

Proses POS *tagging* belum memberikan informasi mengenai struktur kalimat, proses *chunking* dilakukan untuk mendapatkan informasi mengenai struktur kalimat. Menurut Steven Bird dkk (2006), *chunking* adalah untuk mengidentifikasi token dengan melakukan penandaan tag yang benar seperti BIO tag dan *Chunk parsing* adalah untuk mengidentifikasi string token dengan mengelompokan tipe chunk yang benar seperti *parsing tree.*

2.6.1 IO tags

IO (*Inside, Outside*) *Encoding* yang paling sederhana adalah pengkodean, yang menandai setiap token sebagai salah satu tipe. *Encoding* ini tidak dapat mewakili dua entitas di samping satu sama lain, karena tidak ada tag batas.

2.6.2 BIO tags

BIO (*Beginning, Inside, Outside*) adalah format pemberian tag yang umum untuk menandai token, pada format BIO dimana B-*begin* dan I-*inside* menunjukkan token milik Entitas Bernama dan "O-*outside*" digunakan untuk semua token.

Penjelasan Ramshaw dan Marcus (1995) dengan judul "*Text Chunking using Transformation-Based Learning*" adalah tag *B* menunjukkan awal dari entitas dan tag *I* menunjukkan bahwa tag masih berhubungan dengan entitas dan sebagai pembatas dari entitas, dan pada tag *O* menunjukkan bahwa sebuah token tidak termasuk entitas bernama. IOB dapat mewakili informasi yang sama persis seperti notasi pada tanda kurung. Dengan tidak adanya tag *B*, tag *IO* tidak dapat membedakan antara dua entitas dengan tipe yang sama.

2.7 Named Entity Recognition (NER)

Named Entity Recognition (NER) adalah suatu proses mengidentifikasi pada suatu kata sehingga dapat diketahui entitas dari kata tersebut. Entitas adalah suatu kelas kata yang menunjukkan arti atau keterangan dari kata yang memiliki entitas tersebut, mengidentifikasi nama entitas yang didalam teks yang merupakan salah satu pekerjaan yang ada di *information extraction*, seperti nama orang, organisasi, lokasi, pernyataan waktu, nilai uang, dan sebagainya (Nadeau dan Sakine, 2007). Sebagai contoh pengenalan entitas adalah sebagai berikut:

Person	budi, ridwan kamil
Organization	Twitter inc
Location	Jakarta, bandung
Date	desember, 2015-12-17
Time	Jam 12:00
Money	Rp. 12000

2.8 Pembobotan TF-IDF

Term Frequency Inverse Document Frequency atau TF-IDF merupakan suatu metode yang digunakan dalam melakukan pembobotan terhadap kemunculan kata dalam suatu dokumen. TF menyatakan jumlah kata yang muncul dalam suatu

dokumen. Sedangkan IDF menunjukkan tingkat kepentingan suatu kata yang terdapat dalam kumpulan dokumen. Metode ini menghitung bobot setiap token didalam dokumen dan dirumuskan dengan:

$$Wdt = tfdt * IDFt \quad (2.1)$$

Keterangan:

W : bobot dokumen ke-d terhadap kata ke-t
d : dokumen ke-d
t : kata ke-t dari kata kunci
tf : banyaknya kata yang dicari pada sebuah dokumen
IDF : banyaknya kata yang sering muncul di semua dokumen

Nilai IDF didapatkan dari

$$IDF = \log2 (D/df) \quad (2.2)$$

Keterangan:

D : total dokumen
df : banyak dokumen yang mengandung kata yang dicari.

2.9 K-Nearest Neighbor (k-NN)

Algoritma *k-Nearest Neighbor* (k-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data (Insanudin, 2013). k-NN termasuk algoritma *supervised learning*, hasil query instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada k-NN. Kelas yang paling banyak muncul yang akan menjadi kelas hasil klasifikasi.

Adapun langkah-langkah dari algoritma k-NN sebagai berikut:

1. Menentukan parameter *k* (jumlah tetangga paling dekat).
2. Menghitung kuadrat jarak query euclid masing-masing objek terhadap data sample yang diberikan.
3. Urutkan objek-objek tersebut berdasarkan urutan jarak euclid terkecil.
4. Hitung jumlah mayoritas kelas berdasarkan *k* (jumlah terdekat).

Proses perhitungan kuadrat jarak query menggunakan *Cosine Similarity*, dengan rumus sebagai berikut:

$$\boldsymbol{D(i,k)} = \frac{\sum_k (\boldsymbol{d_1 d_k})}{\sqrt{\sum_k \boldsymbol{d_{ik}^2}}\sqrt{\sum_k \boldsymbol{d_{jk}^2}}} \quad (2.3)$$

$\sum_k (\boldsymbol{d_i d_k})$: *vector* dot produk dari i, dan k (2.4)

$\sqrt{\sum_k \boldsymbol{d_{ik}^2}}$: Panjang *vector* I (2.5)

$\sqrt{\sum_k \boldsymbol{d_{jk}^2}}$: Panjang *vector* k (2.6)

2.10 Confusion Matrix

Confusion matrix adalah metode yang digunakan untuk melakukan proses perhitungan akurasi pada konsep *Data Mining* (Kohavi dan Provost, 1998). Rumus ini melakukan perhitungan dengan 4 keluaran, yaitu: *recall, precision, accuracy.*

1. *Recall* adalah proporsi kasus positif yang diidentifikasi dengan benar.
2. *Precision* adalah proporsi kasus dengan hasil positif yang benar.
3. *Accuracy* adalah perbandingan yang diidentifikasi benar dengan jumlah semua kasus.

Tabel 2.3 Pengujian Confusion Matrix

		Nilai sebenarnya	
		TRUE	FALSE
Nilai prediksi	TRUE	TP (*True Positive*) *Corect result*	TN (*True Negative*) *Unexpected result*
	FALSE	FP (*False Positif*) *Missing result*	FN (False *Negative*) *Corect obsence of result*

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$\boldsymbol{accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (2.7)$$

Keterangan:

TP = jika hasil prediksi positif dan nilai sebenarnya positif.

TN = jika hasil prediksi negatif sedangkan nilai sebenarnya positif.

FP = jika hasil prediksi positif sedangkan nilai sebenarnya negatif.

FN = jika hasil prediksi negatif dan nilai sebenarnya positif.

2.11 Penelitian Terkait

Berikut ini Tabel 2.4 adalah beberapa penelitian terkait yang pernah dilakukan tentang sistem yang akan dibangun oleh penulis.

Tabel 2.4 Penelitian Terkait

No	Judul	Peneliti	Tahun	Keterangan
1	Ekstraksi Informasi Transaksi *Online* pada Twitter	Masayu Leylia Khodra dan Ayu Purwarianti	2013	Penelitian tentang ekstraksi informasi transaksi *online* pada Twitter dengan menggunakan pendekatan ekstraksi informasi berbasis klasifikasi, dilakukan klasifikasi tweet dan klasifikasi token. Model klasifikasi token untuk tahapan ekstraksi menunjukkan bahwa model terbaik dengan akurasi 81.49% didapatkan dengan menggunakan algoritma pembelajaran IBk (*Instance-based learning*).

2	*Short Text Clasification Using kNN Based on Distance Function*	Khushbu Khamar	2013	Penelitian tersebut mencoba membandingkan tingkat akurasi dari 3 algoritma yaitu *k-Nearest Neighbor* (k-NN), *Naive Bayes* dan *Support Vector Machine* (SVM) didapat bahwa metode k-NN merupakan teknik yang baik untuk melakukan klasifikasi teks pendek
3	*Named Entity Recognition on Indonesian Microblog Messages*	Natanael Taufik, Alfan F Wicaksono dan Mirna Adriani	2016	penelitian tentang pengenalan entitas pada pesan *microblog* yang biasanya sangat singkat dan ditulis dengan cara yang tidak semestinya. Penelitian tersebut menggunakan pendekatan *Conditional Random Fields* (CRF) untuk pelabelan, namun masih terdapat masalah seperti masalah dalam urutan pelabelan.
4	*Semi-Supervised Learning Approach for Indonesian Named Entity Recognition* (NER)	Bayu Aryoyu danta, Teguh Bharata Adji dan Indriana Hidayah	2016	Penelitian tersebut menggunakan data dari DBPedia Indonesia dan artikel berita bahasa Indonesia dari situs berita

	Using Co-Training Algorithm			Indonesia yaitu kompas.com, cnnindonesia.com, tempo.co, merdeka.com dan viva.co.id. Langkah-langkah *pre-processing* yang diterapkan untuk menganalisis teks tidak terstruktur adalah segmentasi kalimat, tokenisasi, stemming dan POS *Tagging*. Hasil dari uji coba menggunakan pendekatan *semi-supervised learning* yaitu nilai *precission* 73,6%, *recall* 80,1% dan *F1 mean* 76,5%.