BAB II

TINJAUAN PUSTAKA & LANDASAN TEORI

2.1 Tinjauan Pustaka

Analisis sentimen adalah riset mengenai opini, sentimen dan emosi yang diekspresikan secara tekstual oleh masyarakat. Menurut Shelby (2013), jika diberikan suatu set dokumen teks yang berisi opini mengenai suatu objek, maka analisi sentimen bertujuan untuk mengekstrak atribut dan komponen dari objek yang telah dikomentasi pada setiap dokumen dan untuk menentukan apakah komentar tersebut bermakna positif atau negatif. Pada bab ini akan menjelaskan beberapa studi yang sudah terlebih dahulu dilakukan sebelum Analisis Sentimen Pada Data Twitter Mengenai Bencana Gunung Merapi dengan Metode Maximum Entropy. Berikut akan dijelaskan mengenai beberapa studi yang memiliki kesamaan dengan penelitian yang dibuat penulis.

Aqsath dan Purwarianti (2011) mengembangkan sebuah sistem yang dapat mengklasifikasi sentimen pada media sosial di Indonesia yang menggunakan bahasa formal maupun non-formal dengan membandingkan algoritma Maximum Entropy dengan SVM. Dalam penelitian mengenai analisis sentimen pada umumnya hanya menggunakan kamus bahasa formal, namun terkadang pengguna media sosial menggunakan bahasa sehari-hari dalam mengekspresikan perasaan atau emosi. Pengguna media sosial yang didominasi oleh remaja dan orang muda tentu membuat status atau

kicauan menggunakan bahasa sehari-hari, dengan mengidentifikasi bahasa non-formal tentu saja membuat akurasi penelitian semakin akurat. Pada penelitian ini menggunakan kamus non-formal untuk memudahkan identifikasi dan klasifikasi. Penelitian ini membuktikan jika algoritma Maximum Entropy lebih unggul menggunakan kamus non-formal dengan akurasi 30% dibandingkan SVM yang akurasinya 26,66%.

Tegar dan Siti (2013) mengembangkan aplikasi yang mampu memantau perkembangan mahasiswa dari media sosial yang dimiliki. Masukan dalam aplikasi merupakan sebuah *trigger* yang memicu sistem untuk bekerja secara otomatis mencari informasi mahasiswa yang menjadi anak wali dan anak didik dosen yang bersangkutan didalam media sosial. Dengan menggunakan informasi tersebut, sistem melakukan pengklasifikasian topik dan analisis sentimen menggunakan algoritma Maximum Entropy. Penelitian ini membuktikan metode Maximum Entropy dapat digunakan untuk melakukan analisis sentimen dalam bahasa Indonesia dengan tingkat akurasi 70% untuk dokumen positif dan 53% untuk dokumen negatif.

Neethu dan Rajasree (2013) melakukan penelitian mengenai analisis sentimen cara orang mengekspresikan pandangan melalui media sosial seperti posting blog, diskusi online, maupun review produk. Ketika seseorang ingin memberi produk, mereka akan mencari ulasan dari produk sebelum memutuskan untuk membelinya. Penelitian sebelumnya menunjukkan bahwa pertunjukan sentimen classifiers tergantung pada topik. Karena itu kita tidak bisa

mengatakan jika suatu pengklasifikasi itu yang terbaik untuk semua topik. Analisis sentimen di twitter cukup sulit karena panjang pendeknya. Adanya emoticon, kata slang dan salah eja di tweet membuat kesulitan untuk melakukan preprocessing sebelum ekstraksi fitur. Ada metode ekstraksi fitur yang berbeda untuk mengumpulkan fitur yang relevan dari teks yang bisa diaplikasikan ke tweets. Namun ekstraksi fitur itu harus dilakukan dalam dua tahap untuk mengekstrak fitur yang relevan. Pada tahap pertama, Fitur khusus twitter diekstrak. Fitur ini akan dihapus dari tweets untuk membuat teks normal. Setelah itu, Sekali lagi ekstraksi fitur dilakukan untuk mendapatkan lebih banyak fitur. Ini adalah Ide yang digunakan dalam penelitian ini untuk menghasilkan vektor fitur yang efisien untuk menganalisis sentimen twitter. Setelah proses ekstrak fitur sudah selesai maka dilakukan analisis sentimen positif dan negatif dengan menggunakan algoritma Maximum Entropy, SVM, dan Naive Bayes. Dari penelitian ini algoritma Maximum Entropy memiliki akurasi rata-rata diatas algoritma lainnya sehingga kesimpulannya algoritma Maximum Entropy dapat digunakan pada penelitian ini.

Pang dkk (2002) melakukan penelitian menyangkut masalah penggolongan dokumen review film. Pada umumnya penggolongan dokumen hanya dilakukan berdasarkan topik sehingga analisis sentimen yang dilakukan kurang akurat dan menyeluruh. Untuk mengatasi masalah itu maka dilakukan analisa terhadap sentimen secara keseluruhan, contohnya

menentukan apakah suatu ulasan positif atau negatif dengan menggunakan metode pembelajaran Maximum Entropy, SVM, dan Naive Bayes sebagai metode pengklasifikasian sentimen. Pengklasifikasian berdasarkan topik tradisional berbasis kategorisasi dianggap kurang menyeluruh dan kurang efektif oleh karena itu penulis mencoba untuk mengurutkan dokumen menurut untuk materi pelajaran mereka (misalnya, olahraga vs politik). Selain itu review produk di situs jual beli dengan label positif atau negatif akan memberikan ringkasan singkat kepada pembaca. Dengan adanya nilai sentimen maka dapat diterapkan skema rating yang berbeda yang masing-masing pengulas, dimana skema ini didapatkan dengan menggunakan klasifikasi sentimen. Berdasarkan penelitian ini diketahui jika akurasi algoritma Maximum Entropy lebih menggungguli akurasi dari algoritma Naive Bayes dan SVM terutama di bagian fitur bigram dengan akurasi 77,4% dan fitur *adjectives* dengan akurasi 77,7 %. Maka dari itu dapat disimpulkan algoritma Maximum Entropy efektif digunakan dalam penelitian ini.

Yan dan Huang (2015) melakukan penelitian tentang kalimat Tibet analisis sentimen dilakukan dengan belajar dari bahasa China dan Inggris analisis sentimen berdasarkan metode statistik. Penulis membangun sistem analisis sentimen kalimat berbahasa Tibet menggunakan model Maximum Entropy. Hasil menunjukkan bahwa parameter evaluasi yang baik dan dapat dicapai kepraktisan sampai batas tertentu. Dengan pesatnya perkembangan Web2.0, semakin banyak

pengguna terlibat dalam pembuatan konten situs, dan ada sejumlah besar pesan komentar berharga yang mengandung karakter, acara, produk di internet. Pengguna dapat menganalisis informasi ini, dan memanfaatkan pandangan orang dan pendapat dari satu hal, maka buatlah bisnis yang efektif keputusan, keputusan politik dan sebagainya. Bagaimana menggunakan komputer untuk membantu pengguna menganalisa dan mengolahnya teks web cepat dan otomatis, lalu ekstrak informasi emosi telah menjadi fokus penulis. Analisis sentimen teks adalah proses analisis, pengolahan, rangkum dan pembuangan kata-kata, kalimat dan teks dengan warna emosional. Karena itu, sentimen tingkat kalimat klasifikasi memiliki nilai penelitian yang penting, dan ini adalah fokus penelitian ini. Dalam penelitian ini, penulis mengusulkan dan menerapkan kalimat bahasa Tibet untuk melihat kecenderungan sistem penilaian sentimen menggunakan Maximum Entropy. Lalu penulis menguji hasil penelitian di korpus yang berisi 10000 kalimat sentimen Tibet. F-nilai hasil mencapai 82,8%, yang pada dasarnya membuktikan bahwa algoritma Maximum Entropy efektif digunakan dalam klasifikasi kalimat bahasa Tibet.

Yan dan Yi (2010) melakukan penelitian tentang klasifikasi teks dengan Maximum Entropy. Dalam penelitian ini, kerangka kerja Maximum Entropy (ME) digunakan untuk mengklasifikasikan dokumen teks. Maximum Entropy memiliki banyak keunggulan bila dibandingkan dengan yang lain algoritma pembelajaran laiinya seperti Naive Bayes Classifier. Misalnya,

tidak ada kondisi yang asumsi independen antar istilah. Dengan empat set data berlabel, eksperimen dilakukan untuk membandingkan keakuratan algoritma ME dengan yang dari Naive Bayes dan Support Vector Machine (SVM), yang merupakan dua algoritma populer untuk klasifikasi teks. Sejak Algoritma Maximum Entropy diusulkan, beberapa penelitian telah dilakukan dilakukan untuk melakukan percobaan untuk membandingkan kinerja Maximum Entropy dengan metode lainnya. Bila jumlah *data training* lebih besar dari pada fitur kata maka distribusi dapat diestimasi secara tepat. Sayangnya, hal ini tidak terjadi pada banyak aplikasi. Keadaan seperti ini memaksa kita untuk menggunakan fitur pilihan. Pilihan fitur memiliki banyak manfaat: Pertama, itu dapat menghindari overfitting dan dengan demikian meningkatkan akurasi; Kedua, juga bisa menjaga akurasi saat membuang sebanyak mungkin fitur. Selanjutnya, percobaan ekstensif dilakukan dengan menguji metode pemilihan fitur. Hasil akhirnya adalah metode ME memiliki performa yang konsisten terus naik dibandingkan Naive Bayes dan algoritma SVM dalam hal akurasi. Untuk data set WebKB dan Industri Vector, keakuratan algoritma Maximum Entropy meningkat dari 81,38% menjadi 85,52% dan dari 85,73% masing menjadi 89,78%. Pada data set 20 Newsgroups keakuratan algoritma Maximum Entropy meningkat dari 94,76% menjadi 96,16%. Dari hasil penelitian ini dapat disimpulkan bahwa algoritma Maximum Entropy lebih efektif dibandingkan metode Naive Bayes dan SVM dalam klasifikasi teks.

Tsatsoulis & Hofmann (2014) melakukan penelitian tentang klasifikasi lirik dari lagi musisi rock Tom Waits. Lirik lagu yang diteliti adalah bagian dari album Swordfishtrombones yang memulai fase baru dalam 40 tahun karir Waits. Penulis mengklasifikasikan lirik menjadi dua kelas yaitu visual(A) dan komputasi(B) berdasarkan analisis teks dengan menggunakan model Maximum Entropy. Penelitian ini mengunakan data dari 10 pelajar yang ditugaskan untuk memisahkan lirik menjadi ruang dimensis tinggi dan vektor kata. Untuk validasi, penulis membandingkan metode *Maximum Entropy* dengan metode lainnya yaitu *Support Vector Machine*(SVM), *Random Forest*, *Bagging*, *Boosting*, *GLMnet*, dan *Scaled Linear Discriminant Analysis*(SLDA). Pada penelitian ini, penulis melakukan pengujian data dengan membedakan level *prunning*, berdasarkan pengujian ini terbukti metode *Maimum Entropy* menunjukan konsistensi performa yang baik sementara metode *Random Forest* menunjukan hasil yang konsistens meskipun performanya tidak begitu baik, SVM menunjukan kinerja yang stabil, namun dengan akurasi yang buruk yang hanya mampu mengklasifikasikan kelas B dengan baik. Sementara metode *Bagging*, *Boosting*, *GLMnet*, dan SLDA menunjukan hasil yang tidak konsisten dengan naik turunnya akurasi metode. Berdasarkan hasil penelitian ini menunjukan metode *Maximum Entropy* terbukti dapat mengklasifikasikan lirik menjadi dua kelas dnegan akurasi mencapai 95% dengan konsistensi terbaik.

Berikut adalah tabel perbandingan tinjauan pustaka dari penelitian yang sudah dilakukan sebelumnya mengenai Analisis Sentimen dengan Maximum Entropy.

Tabel 2.1 Perbandingan Tinjauan Pustaka

No	Pembanding / Penelitian	Algoritma	Pokok Penelitian	Klasifikasi pada	Hasil
1	Aqsath dan Purwarianti (2011)	Maximum Entropy, SVM	Nilai akurasi dari algoritma	Data media sosial	Jenis kamus lexicon berpengaruh terhadap nilai akurasi. Algoritma *Maximum Entropy* menghasilkan akurasi yang baik
2	Tegar dan Siti (2013)	Maximum Entropy	Nilai akurasi dari algoritma	Data media sosial mahasiswa	Algoritma *Maximum Entropy* efektif digunakan
3	Neethu dan Rajasree (2013)	Maximum Entropy, Naive Bayes, SVM, Esemble	Nilai Akurasi, precision, recall, dan F-measure dari 3 algoritma	Data Twitter	Algoritma *Maximum Entropy* mempunyai nilai akurasi terbaik.
4	Pang dkk (2002)	Maximum Entropy, Naive Bayes, SVM	Nilai Akurasi, precision, recall, dan F-measure dari 3 algoritma	Data Review film	Setiap Algoritma memiliki akurasi yang tinggi di bidang-bidang tertentu.

5	Yan dan Huang (2015)	Maximum Entropy	Nilai Akurasi, precision, recall, dan F-measure algoritma	Data review bahasa Tibet	Algoritma *Maximum Entropy* efektif digunakan
6	Yan dan Yi (2010)	Maximum Entropy, Naive Bayes, SVM	Nilai Akurasi, precision, recall, dan F-measure dari 3 algoritma	data set WebKB dan Industri Vector	Algoritma *Maximum Entropy* mempunyai nilai akurasi terbaik.
7	Tsatsoulis & Hofmann (2014)	*Maximum Entropy*	Nilai akurasi dari algoritma	Lirik lagu Tom Waits	Algoritma *Maximum Entropy* dapat mengklasifikasikan lirik dengan akurasi terbaik

2.2 Landasan Teori

2.2.1 Penanggulangan Bencana

Penanggulangan bencana adalah kegiatan atau proses yang berkaitan dengan langkah-langkah penanganan, berupa rangkaian kegiatan yang meliputi pencegahan, mitigasi, kesiapsiagaan, tanggap darurat, rehabiltasi dan pembangunan kembali pasca bencana. Menurut Undang-Undang Republik Indonesia nomor 24 tahun 2007 tentang Penanggulangan Bencana Bab I Pasal 1 ayat 6, penyelenggaraan penanggulangan bencana adalah serangkaian upaya yang meliputi penetapan kebijakan pembangunan yang berisiko timbulnya bencana, kegiatan pencegahan bencana, tanggap darurat, dan rehabilitasi. Lalu pada pasal 4 menjelaskan tujuan penanggulangan bencana yaitu :

a. Memberikan perlindungan kepada masyarakat dari ancaman bencana;
b. Menyelaraskan peraturan perundang-undangan yang sudah ada;
c. Menjamin terselenggaranya penanggulangan bencana secara terencana, terpadu, terkoordinasi, dan menyeluruh;
d. Menghargai budaya lokal;
e. Membangun partisipasi dan kemitraan publik serta swasta;
f. Mendorong semangat gotong royong, kesetiakawanan, dan kedermawanan; dan
g. Menciptakan perdamaian dalam kehidupan bermasyarakat, berbangsa, dan bernegara.

Penanggulangan bencana tentunya memiliki proses sampai suatu bencana benar-benar tertanggulangi

terutama dari segi dampak yang dirasakan. Oleh karena itu menurut Sena dan Michael (2006) dalam buku berjudul *Disaster Prevention and Preparedness* membagi tahap penanggulangan bencana menjadi beberapa tahap yaitu :

1. **Mitigation**

Mitigasi adalah tindakan berkelanjutan untuk mengurangi atau menghilangkan risiko terhadap nyawa dan harta benda dari bahaya bencana dan dampaknya. Fungsi mitigasi berbeda dengan disiplin manajemen bencana lainnya yang melihat solusi jangka panjang untuk mengurangi risiko dibandingkan dengan kesiapan untuk bahaya, sementara mitigasi merupakan respon yang langsung dilakukan terhadap bencana, atau pemulihan jangka pendek dari dampak bencana. Upaya mitigasi mencakup pembangunan pemukiman darurat, penyediaan air minum yang cukup, perbaikan prasarana sanitasi, dan sistem peringatan dini untuk suatu penyakit.

2. **Preparedness**

Preparednes atau kesiapsiagaan adalah bagaimana rencana untuk merespon saat terjadi bencana. Selama fase kesiapsiagaan, pemerintah, organisasi, dan individu mengembangkan rencana untuk menyelamatkan nyawa, meminimalkan kerusakan akibat bencana, dan meningkatkan operasi penanggulangan bencana. Langkah-langkah persiapan meliputi rencana kesiapsiagaan meliputi latihan/pelatihan darurat, sistem peringatan, sistem komunikasi darurat, rencana evakuasi dan pelatihan, persediaan sumber

daya, kesiagaan petugas darurat, penyebaran kontak kesiagaan bencana. Seperti upaya mitigasi, tindakan kesiapsiagaan bergantung pada penggabungan langkah-langkah yang tepat dalam rencana pembangunan nasional dan regional. Selain itu, keefektifannya bergantung pada ketersediaan informasi tentang bahaya, risiko darurat dan tindakan penanggulangan yang harus dilakukan.

3. **Response**

Response atau tanggap darutat adalah kegiatan yang bertujuan untuk memberikan bantuan segera untuk mempertahankan kehidupan, memperbaiki kesehatan dan mendukung moral masyarakat yang terkena dampak. Bantuan semacam itu berkisar dari pemberian bantuan khusus namun terbatas, seperti membantu pengungsi dengan transportasi, tempat penampungan sementara, dan makanan, untuk mendirikan pemukiman semi permanen di kamp-kamp dan lokasi lainnya. Ini juga mungkin memerlukan perbaikan awal terhadap infrastruktur yang rusak. Fokus dalam fase tanggap darurat adalah pada memenuhi kebutuhan dasar masyarakat sampai solusi yang lebih permanen dan berkelanjutan dapat ditemukan.

4. **Recovery**

Recovery atau pemulihan adalah kegiatan penanggulangan bencana yang bertujuan untuk memulihkan keadaan seperti keadaan normal. Karena keadaan darurat sudah terkendali, penduduk yang

terkena dampak mampu melakukan sejumlah aktivitas yang terus berlanjut untuk memulihkan kehidupan dan infrastruktur mereka yang mendukungnya. Tidak ada titik yang jelas dimana keadaan pemberian bantuan segera berubah menjadi pemulihan dan kemudian berlanjut ke pembangunan berkelanjutan jangka panjang. Banyak kesempatan selama masa pemulihan untuk meningkatkan pencegahan dan meningkatkan kesiapan, sehingga mengurangi kerentanan. Idealnya, harus ada transisi yang mulus dari pemulihan menuju pembangunan yang sedang berjalan. Aktivitas pemulihan berlanjut sampai semua sistem kembali normal atau lebih baik. Langkah-langkah pemulihan, baik jangka pendek maupun jangka panjang, mencakup mengembalikan sistem pendukung kehidupan vital ke standar operasi minimum seperti perumahan sementara, informasi Publik, pendidikan kesehatan dan keselamatan, rekonstruksi, program konseling, dan studi dampak ekonomi. Sumber dan layanan informasi mencakup pengumpulan data yang berkaitan dengan pembangunan kembali, dan dokumentasi pelajaran yang dapat dipetik.

2.2.2 Logistik

Menurut Peraturan Kepala Badan Nasional Penanggulangan Bencana Nomor 18 tahun 2010 tentang Pedoman Distribusi Bantuan Logistik dan Peralatan Penanggulangan Bencana, logistik adalah segala sesuatu yang berujud yang dapat digunakan untuk memenuhi suatu kebutuhan dasar manusia yang habis

pakai terdiri atas pangan, sandang dan papan atau turunannya. Termasuk dalam kategori logistik adalah barang yang habis pakai atau dikonsumsi, misalnya: sembako (sembilan bahan pokok), obat-obatan, pakaian dan kelengkapannya, air, kantong tidur (sleeping bag), perlengkapan bayi, perlengkapan keluarga (pembalut wanita, odol, sabun mandi, shampo, detergen, handuk).

Menurut Peraturan Kepala Badan Nasional Penanggulangan Bencana Nomor 18 tahun 2009 bab II tentang Pedoman Standarisasi Logistik, kategori bantuan logistik dalam penanggulangan bencana dapat dibedakan menjadi beberapa kategori yaitu :

1. Pangan, yang termasuk dalam kategori ini adalah makanan pokok (beras/sagu/jagung/ubi,dll), lauk-pauk, air bersih, bahan makanan pokok tambahan seperti mi, susu, kopi, teh, perlengkapan makan (food ware) dan sebagainya.
2. Sandang, yang termasuk dalam kategori ini adalah perlengkapan pribadi berupa baju, kaos dan celana anak-anak sampai dewasa laki-laki dan perempuan, sarung, kain batik panjang, handuk, selimut, daster, perangkat lengkap pakaian dalam, seragam sekolah laki-laki dan perempuan (SD dan SMP), sepatu/alas kaki sekolah dan turunannya.
3. Logistik lainnya, termasuk dalam kategori ini adalah, obat dan alat kesehatan habis pakai, tenda gulung, tikar, matras, alat dapur keluarga, kantong tidur (sleeping bag) dan sebagainya.

4. Paket kematian, termasuk dalam kategori ini adalah, kantong mayat, kain kafan dan sebagainya.

2.2.3 Text Mining

Text Mining merupakan penerapan konsep *data mining* untuk penambangan data bentuk teks untuk mendapatkan pola-pola yang nantinya menghasilkan informasi-informasi. *Text mining* sudah banyak didefinisikan oleh ahli riset dan praktisi. *Text mining* memiliki definisi menambang data yang berupa teks di mana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata - kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen. *Text mining* juga merupakan proses penemuan akan informasi atau trend baru yang sebelumnya tidak terungkap dengan memproses dan menganalisis data dalam jumlah besar.

Sistem *text mining* terdiri dari komponen *text preprocessing*, *feature selection*, dan komponen *data mining*. Komponen *text preprocessing* berfungsi untuk mengubah data tekstual yang tidak terstruktur seperti dokumen, kedalam data terstruktur dan disimpan ke dalam basis data. *Feature selection* akan memilih kata yang tepat dan berpengaruh pada proses klasifikasi. Komponen terakhir akan menjalankan teknik *data mining* pada output dari komponen sebelumnya.

Menurut Rian (2012), permasalahan yang dihadapi pada *text mining* sama dengan permasalahan yang terdapat pada *data mining*, yaitu jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data *noise*. Perbedaan di antara keduanya adalah pada data yang digunakan. Pada data mining, data yang digunakan adalah *structured data*, sedangkan pada text mining, data yang digunakan text mining pada umumnya adalah *unstructured data*, atau minimal *semistructured*. Hal ini menyebabkan adanya tantangan tambahan pada text mining yaitu struktur text yang kompleks dan tidak lengkap, arti yang tidak jelas dan tidak standar, dan bahasa yang berbeda ditambah translasi yang tidak akurat.

2.2.4 Lexicon Based

Pendekatan *Lexicon Based* adalah metode untuk menentukan sentimen melalui perhitungan orientasi dari suatu dokumen. Perhitungan orientasi ini berdasarkan kata atau frasa dalam dokumen (Turney, 2002). Pendekatan ini menggunakan kamus sebagai acuan dalam menentukan kelas suatu dokumen. Kamus dibentuk secara manual atau secara otomatis menggunakan daftar kata yang telah dibentuk sebelumnya.

Sebagian besar penelitian *Lexicon Based* berfokus pada penggunaan kata sifat sebagai indikator dari orientasi teks semantik (Hatzivassiloglou and McKeown1997; Wiebe 2000; Hu and Liu 2004; Taboada, Anthony, and Voll 2006). Proses pendekatan dimulai dengan menentukan daftar

kamus kata-kata yang sesuai dengan penelitian. Kemudian dalam proses analisis kalimat, masing-masing kata sifat dalam kalimat akan diberi nilai yang nantinya akan digabung menjadi skor tunggal akhir yang menentukan kelas suatu dokumen.

Berikut adalah formulasi dari perhitungan skor suatu kalimat :

$$S_{positive} = \sum_{i \in t}^{n} positive\ score_i$$
$$S_{negative} = \sum_{i \in t}^{n} negative\ score_i \quad (2.1)$$

Kedua persamaan ini digunakan utnuk menghitung sentimen dari sebuah kalimat. Selanjutnya untuk menghitung sentimen suatu kalimat secara keseluruhan dilakukan perhitungan dengan menjumlahkan skor positif ($S_{positive}$) dan skor negatif ($S_{negative}$) dengan persamaan sebagai berikut :

$$Sentence_{sentiment} \begin{cases} positive\ if\ S_{positive} > S_{negative} \\ neutral\ if\ S_{positive} = S_{negative} \\ negative\ if\ S_{positive} < S_{negative} \end{cases} \quad (2.2)$$

Dari persamaan diatas diketahui proses pemberian kelas pada pendekatan *Lexicon Based*, jika jumlah skor positif lebih besar dari negatif maka disimpulkan kalimat memiliki sentimen positif, jika jumlah skor positif sama dengan negatif maka disimpulkan kalimat memiliki sentimen netral, sementara jika skor positif lebih kecil dari negatif maka disimpulkan kalimat memiliki sentimen negatif

2.2.5 TF.IDF

Term Frequency (TF) merupakan frekuensi kemunculan *term* pada dokumen. TF suatu dokumen dengan dokumen yang lain akan berbeda, bergantung pada tingkat kepentingan sebuah *term* dalam dokumen. *Inverse Document Frequency* (IDF) merupakan sebuah perhitungan dari bagaimana *term* didistribusikan secara luas pada koleksi dokumen yang bersangkutan. Semakin sedikit dokumen yang mengandung *term* yang dimaksud, maka nilai idf semakin besar. Jika setiap dokumen dalam koleksi mengandung *term* yang bersangkutan, maka nilai dari idf dari term tersebut adalah nol. Hal ini menunjukkan bahwa sebuah term yang muncul pada setiap dokumen dalam koleksi tidak berguna untuk membedakan dokumen berdasarkan topik tertentu. Nilai IDF sebuah *term* t dirumuskan dalam persamaan berikut:

$$IDF(t) = \log (N/df(t)) \quad (2.3)$$

N adalah jumlah dokumen dan df(t) adalah jumlah dokumen yang mengandung *term* yang bersangkutan.

Dengan menggunakan tf.idf maka dapat diketahui deskripsi terbaik dari dokumen adalah *term* yang banyak muncul dalam dokumen tersebut dan sangat sedikt kemunculannya pada dokumen yang lain. Bobot terendah akan diberikan pada term yang muncul sangat jarang pada beberapa dokumen *(low-frequency documents)* dan term yang muncul pada hampir atau seluruh dokumen *(high-frequency documents.*

Penelitian belakangan ini telah mengkombinasikan TF dan IDF untuk menghitung bobot *term* dan menunjukkan bahwa gabungan keduanya menghasilkan performansi yang lebih baik. Kombinasi bobot dari sebuah *term* t pada text didefinisikan sebagai berikut :

$$TFIDF(d.t)=TF(d.t).IDF(t) \quad (2.4)$$

2.2.6 Maximum Entropy

Maximum Entropy adalah algoritma klasifikasi probabilistic yang termasuk dalam kelas model eksponensial. Maximum Entropy didasarkan pada prinsip Entropi Maksimum. Maximum Entropy dapat digunakan untuk memecahkan masalah klasifikasi teks seperti seperti deteksi Bahasa, klasifikasi topik, dan analisis sentimen. Menurut Nigam, Lafferty dan McCallum (1999, hal.61), Maximum Entropy pada umumnya, adalah teknik yang membantu kita untuk memperkirakan distribusi probabilitas dari data. Selain itu, menurut mereka, prinsip MaxEnt adalah distribusi harus seragam semaksimal mungkin, bila tidak ada yang diketahui.

Di dalam bidang Information Theory, kita sering menggunakan entropy sebagai suatu parameter untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan sampel data. Jika sampel data semakin heterogen, maka nilai entropi-nya semakin besar. Secara matematis entropy dirumuskan sebagai berikut:

$$Entropy(S) = \sum_{i}^{c} -p_i log_2 p_i \quad (2.5)$$

Di mana c adalah jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi). Sedangkan p_i menyatakan jumlah sampel untuk kelas i.

Menurut F.Kosta (2005), secara khusus dalam klasifikasi teks tugas klasifikasi teks adalah menjadikan proses acak Y sebagai masukan dokumen d dan menghasilkan output label kelas c. Output dari Y acak dapat dipengaruhi oleh beberapa informasi kontekstual X, yang domainnya adalah semua informasi tekstual yang mungkin terkandung dalam dokumen d. Tujuannya adalah untuk menentukan model p(y|x) yang menunjukkan probabilitas model untuk y ∈ Y dengan informasi kontekstual adalah x ∈ X.

Untuk menentukan model p(y|x), kita perlu mengetahui persebaran informasi kontekstual (term) dari suatu dokumen dengan menggunakan rumus berikut:

$$f(x,y) = \begin{cases} 1 \text{ If y=some particular value\_1 and x=some particular value2} \\ 0 \text{ otherwise} \end{cases} \quad (2.6)$$

Dimana f(x,y) bernilai 1 jika y masuk ke dalam label terpilih dan x mengandung kata yang dimaksud dan bernilai 0 selain itu.

Maximum Entropy dapat memanfaatkan data kontinu dan kategoris, dan dapat menggabungkan interaksi antara variabel yang berbeda. Selain itu efisien, karena sudah dikembangkan dan dijamin dapat

bertemu dengan distribusi probabilitas optimal. Namun Maximum Entropy memiliki kekurangan yaitu performa rendah dengan fitur independen dan memakan memori karena komputasi yang rumit dan banyak.

2.2.7 *K-Fold Cross Validation*

Metode *k-fold Cross Validation* merupakan metode pembagian data ke dalam *k* bagian secara acak. Tujuannya untuk melakukan klasifikasi sentimen pada suatu percobaan. Dengan menggunakan metode *k-fold cross validation* maka percobaan akan dilakukan sebanyak *k* buah. Masing-masing percobaan menggunakan 1-*k* bagian yang akan menjadi data *training*.

Data *training* adalah data yang digunakan untuk melakukan pembelajaran untuk klasifikasi sentimen. Sementara data *testing* adalah data yang belum digunakan sebagai data pembelajaran dan memiliki fungsi sebagai data penguji untuk memperoleh suatu nilai akurasi. Ada banyak faktor yang mempengaruhi nilai akurasi dalam metode *k-fold Cross Validation* seperti banyaknya *k*, metode, jumlah data, serta bagaimana persentase pembagian data *training* dan data *testing*.