# BAB II

# TINJAUAN PUSTAKA

## 2.1 Analisis Sentimen

Analisis Sentimen merupakan *Natural Language Proccessing* yang bertujuan untuk membangun sebuah metode yang dapat diimplementasikan menjadi sebuah tools yang dapat digunakan untuk mengekstraksi informasi subjektif berupa sentimen atau opini dalam sebuah data text. kecenderungan penelitian tentang analisis sentimen berfokus pada pendapat yang menyatakan suatu sentimen memiliki nilai positif atau negatif. (Liu, 2015).

Liu mencontohkan permasalahan kalimat sentimen pada kasus pemilihan berikut :

(1) Saya membeli IPhone 7 enam bulan yang lalu.

(2) Saya menyukai kameranya yang memiliki kualitas gambar menakjubkan.

(3) Ketahanan pemakaian baterai juga sangat bagus.

(4) Namun kurang baik karena tidak memiliki fitur fast charging“

maka dapat dijelaskan sebagai berikut :

maka dapat dijelaskan sebagai berikut :

Tulisan ini memiliki pendapat tentang suatu entitas/domain yaitu IPhone 7. Pada kalimat (2) dan kalimat (3) kecenderungan pendapat positif terhadap IPhone 7 tentang kamera yang menghasilkan kualitas gambar yang bagus dan ketahanan daya baterai yang bagus. Sedangkan pada kalimat (4) mengekspresikan pendapat

negatif tentang fitur charging yang tidak ada

## 2.2 Twitter

Twitter merupakan salah satu situs jejaring sosial yang masih popular hingga saat ini. *Twitter* didirikan oleh Jack Dorsey pada bulan Maret 2006. *Twitter* membatasi kata yang akan di post sebanyak 140 karakter. Namun, tidak hanya tulisan yang dapat diunggah pada jejaring sosial tersebut, *twitter* juga bisa mengunggah foto, video, url, dan lain-lain. *Twitter* banyak digunakan untuk berbagai informasi, menjalin relasi bisnis, menuangkan isi hati dan pikiran dalam bentuk tulisan.

Pada aplikasi *twitter* disediakan sebuah *search engine* yang dapat digunakan oleh pengguna *twitter*. Dengan *search engine* tersebut kita bisa mendapatkan informasi terkait *tweets* yang ada pada aplikasi *twitter*. Informasi tersebut banyak digunakan untuk mencari wawasan terkait permasalahan yang sedang dihadapi. Informasi yang diberikan dapat berupa *tweets*, pengguna yang menuliskan *tweets* tersebut, lokasi *tweets* tersebut, dan lain-lain. (Adiyana dan Hakim 2015 dalam Ghifari 2018).

### 1.3.1 Twitter API

*Application Programming Interface* (API) merupakan fungsi- fungsi/perintah-perintah untuk menggantikan bahasa yang digunakan dalam system calls dengan bahasa yang lebih terstruktur dan mudah dimengerti oleh programmer. Fungsi yang

dibuat dengan menggunakan API tersebut kemudian akan memanggil system calls sesuai dengan sistem operasinya. Tidak tertutup kemungkinan nama dari system calls sama dengan nama di API. Twitter menyediakan API yang diperuntukkan untuk developer yang ada pada website https://developer.twitter.com.

Twitter API terdiri dari 3 bagian yaitu :

a. *Search API*

Search *API* dirancang untuk memudahkan *user* dalam mengelola query *search* di konten twitter. *User* dapat menggunakannya untuk mencari tweet berdasarkan keyword khusus atau mencari tweet lebih spesifik berdasarkan *username* twitter. *Search API* juga menyediakan akses pada data *trending topic*.

b. *REST API*

*REST API* memperbolehkan developer untuk mengakses inti dari Twitter seperti timeline, status update dan informasi *user*. *REST API* digunakan dalam membangun sebuah aplikasi Twitter yang kompleks yang memerlukan inti dari Twitter

c. *Streaming API*

*Streaming API* digunakan developer untuk kebutuhan yang lebih intensif seperti melakukan penelitian dan analisis data. Streaming API dapat menghasilkan aplikasi yang dapat mengetahui statistik status update, follower dan lain sebagainya

Pada penelitian ini, bagian Twitter *API* yang digunakan adalah *REST API*, dengan menggunakan *REST API* proses pengumpulan data komentar pengguna twitter akan lebih mudah didapat.

### 1.3.2 Analisis Sentimen Pada Twitter

Definisi analisis sentimen twitter pada dasarnya merujuk pada pendapat komentar yang ada pada media twitter. Pesan twitter lebih mudah untuk dilakukan analisis karena penulisan yang dibatasi. Kalimat seringkali memuat pendapat tunggal, meskipun tidak bersifat mutlak bahwa setiap kalimat berisi pendapat tunggal. Dalam kasus lain terdapat kalimat dengan pendapat lebih dari satu pada suatu kalimat namun ini hanya sebagian kecil (Liu, 2015). Pada dasarnya analisis sentimen merupakan tahapan klasifikasi. Namun tahapan klasifikasi sentimen pada twitter yang tidak terstruktur menyebabkan sedikit lebih sulit dibanding dengan klasifikasi dokumen terstruktur. Langkah pertama adalah untuk mengklasifikasikan apakah kalimat mengungkapkan pendapat atau tidak. Langkah kedua adalah mengklasifikasikan kalimat-kalimat pendapat menjadi positif dan kelas negatif.

### 1.3.3 Struktur Data Twitter

Untuk mendalami permasalahan analisis sentimen twitter diperlukan pemahaman terhadap struktur data twitter itu sendiri. Twitter menjadi sumber yang hampir tak terbatas yang digunakan pada *text classification*. Menurut Go (2009), terdapat banyak karakteristik pada *tweets* twitter. Pesan pada twitter memiliki banyak *attribute* yang unik, yang membedakan dari media sosial lainnya :

1. Twitter memiliki maksimal panjang karakter yaitu 140 karakter.

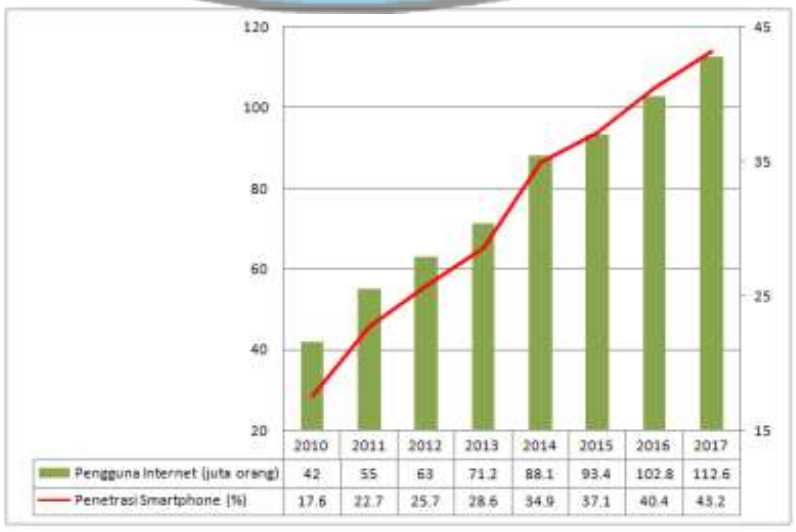
2. Twitter menyediakan data yang bisa diakses secara bebas dengan menggunakan Twitter *API*, mempermudah saat proses pengumpulan *tweets* dalam jumlah yang sangat banyak.
3. Pengguna twitter mem-*posting* pesan melalui banyak media berbeda untuk mengungkapkan pendapat tentang suatu topik atau kejadian tertentu, sehingga merupakan sumber yang bagus dalam menemukan pendapat orang lain.
4. Terdapat ragam topik didalamnya. Setiap pengguna dapat menuliskan topik apapun pada pesan twitter.

**2.4. Shopee**

Shopee merupakan e-commerce international yang pertama kali muncul pada awal tahun 2015 di singapore sebagai mobile marketplace pertama di Asia Tenggara. Dengan kemajuan zaman yang begitu cepat, shopee membuka store di Thailand, Vietnam dan Malaysia untuk memenuhi kebutuhan gaya hidup pria dan wanita Asia Tenggara. Setelah itu, pada awal tahun 2016 Shopee memasuki wilayah Indonesia dengan membuka store di Indonesia untuk memenuhi gaya hidup pria dan wanita di Indonesia yang beraneka ragam. Shopee juga hadir sebagai wadah bagi para penjual yang menjual seluruh kebutuhan pria dan wanita serta memudahkan pria dan wanita di Indonesia untuk mengikuti gaya hidup dengan menggunakan produk yang ada di Shopee. Shopee hingga saat ini masih menjadi mobile marketplace yang pertama di Indonesia yang menawarkan kemudahan untuk melakukan jual beli langsung pada forum jual beli online shopee di Indonesia

Saat ini, masyarakat Indonesia sudah banyak menggunakan teknologi informasi yang berbasis internet untuk memenuhi kebutuhannya. Perkembangan teknologi internet ini telah memberikan dampak positif bagi Indonesia dimana layanan internet sudah banyak dilakukan oleh individu, perusahaan, instansi pemerintahan maupun swasta (miung, 2015). Menggunakan internet bagi masyarakan Indonesia sudah menjadi hal yang menyatu dalam kehidupan sehari-hari. Terlebih lagi biaya untuk mengakses internet relatif murah, didukung pula dengan semakin murahnya harga ponsel pintar di Indonesia yang dijadikan sebagai penunjang kegiatan tersebut, hal tersebut, berdampak pula pada bertambahnya pengguna internet di Indonesia (Sumber: https://www.emarketer.com/,2015).

Peningkatan penggunaan internet dapat dilihat melalui data jumlah pengguna internet di Indonesia (Sumber:https://www.apjii.or.id/ diakses 10 oktober 2016) Proyeksi pengguna internet di Indonesia dapat dilihat pada Gambar 1.2 sebagai berikut ini :



**Gambar 2.1 Proyeksi Pengguna Internet di Indonesia**

Berdasarkan Gambar 2.1 jumlah pengguna internet di Indonesia mengalami peningkatan setiap tahunnya. Berdasarkan data asosiasi penyelenggara jasa internet di Indonesia(APJII) pada tahun 2011 pengguna internet mengalami kenaikkan drastis sebanyak 55 juta pengguna, dibandingkan tahun 2010 yang hanya mencapai 42 juta pengguna. Pada tahun-tahun berikutnya jumlah pengguna internet terus mengalami peningkatan yang cukup signifikan hingga pada tahun 2017. Pada tahun 2017 APJII telah memperkirakan bahwa pengguna internet di Indonesia akan mencapai 112,6 juta pengguna. Pengguna internet yang semakin meningkat berpengaruh pada perkembangan e-commerce di Indonesia, khususnya forum jual-beli. Banyak peluang bisnis yang muncul pada beberapa forum jual-beli di Indonesia. Tidak bisa dipungkiri bahwa belanja online menjadi pilihan oleh banyak konsumen untuk memperoleh barang yang diinginkan tanpa menggunakan banyak waktu dan tenaga. Alasan banyak konsumen menggunakan belanja online adalah kemudahan untuk melakukan transaksi, harga yang cukup bersaing dan kualitas barang yang bagus sesuai dengan keinginan konsumen. Penelitian yang dilakukan oleh Brand Marketing Institute (BMI) pada tahun 2014 mengenai tren belanja online di dunia menunjukkan 26% pengguna internet di indonesia atau 1.231 orang menunjukan tendensi untuk melakukan belanja online. BMI memprediksi pasar belanja online akan tumbuh hingga 57% di tahun 2015. Nilai total belanja online per orang selama satu tahun mencapai Rp 825 ribu, atau jika diakumulasikan dengan jumlah pengguna internet di

Indonesia mencapai Rp 21 triliun (Sumber: www.Swa.co.id, 2015). Hasil penelitian BMI mengenai trend belanja online dapat dilihat pada Gambar 2.2 berikut ini :


**Gambar 2.2 Hasil Penelitian BMI tentang Trend Belanja Online**

Gambar 2.2 menunjukkan bahwa Indonesia memang menjadi salah satu negara yang aktif dalam kegiatan pembelian online, terbukti Indonesia masuk dalam peringkat 26 dunia dari 28 negara. Pihak-pihak di Indonesia yang terlibat di dalam kegiatan jual beli secara online telah melakukan banyak cara untuk mendorong masyarakat agar beradaptasi dengan tren penjualan online tersebut. Mulai dari provider telekomunikasi yang memberikan akses mudah bagi masyarakat untuk dapat mengakses situs-situs penjualan online, Bank dengan berbagai macam produk yang memudahkan dalam bertansaksi secara virtual, pemerintah dengan undang-undang ITE (Informasi dan Transaksi Elektronik) sampai dengan pengelola situs jual beli online yang semakin aktif beriklan (Sumber:www.wearesocial.net,2016). Berbagai hal dapat meningkatkan peluang pasar bagi pelaku pasar untuk menggunakan internet

sebagai alat untuk memasarkan produknya. Pelaku pasar dapat menggunakan berbagai cara yang dianggap cocok untuk memasarkan produknya oleh teknologi internet. Alat yang dapat dimanfaatkan dengan adanya internet adalah penggunaan media sosial, toko online, forum jual beli online serta dengan aplikasi pada smartphone. Pelaku pasar yang memiliki modal besar memiliki kesempatan untuk memaksimalkan internet dalam memasarkan produknya, kegiatan perusahaan tersebut melakukan pengembangan pembuatan online di Indonesia yang menarik dan terintegrasi dengan sistem perusahaan, kemudian pengembangan pada aplikasi mobile untuk penjualan barang dan pengembangan sosial media marketing (sumber:www.swa.co.id, 2016)

### 1.5 *Text Mining*

*Text Mining* merupakan suatu proses penggalian informasi berdasarkan suatu sumber data dokumen yang berupa teks dalam suatu proses yang dilakukan dengan komputer (Feldman dan Sanger 2007 dalam Fatimah 2018). *Text mining* juga dikenal sebagai *data mining text* atau penemuan pengetahuan dari *database* tekstual. Menurut buku *The Text Mining Handbook*, *text mining* didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen *data mining.*

Perbedaan antara *text mining* dengan *data mining* terletak pada sumber data yang digunakan. Dalam *text mining* pola-pola yang diekstrak dari data tekstual yang tidak terstruktur bukan berasal dari suatu database. Sumber data yang digunakan dalam *text mining* adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal *semi* terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorian teks dan pengelompokkan teks. (Nurhuda et al. 2013). Sedangkan dalam *data mining* data yang diolah adalah data yang terstruktur dari proses *warehousing* sehingga lebih mudah diproses oleh mesin/komputer. Persamaan dari *text mining* dan *data mining* adalah data yang digunakan merupakan data besar dan data berdimensi tinggi dengan struktur yang terus berubah.

Adapun tahapan-tahapan dalam *text mining* secara umum adalah *text preprocessing* dan *feature selection* (Feldman dan Sanger 2007 dalam Fatimah 2018). Dimana penjelasan dari tahapan-tahapan diatas sebagai berikut:

### 1.5.1 *Text Preprocessing*

Tahap pertama dalam melakukan *text mining* yaitu *text preprocessing*, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu dapat digunakan untuk proses utama. *Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur. Secara umum proses yang dilakukan dalam tahapan *preprocessing* adalah sebagai berikut:

a.*Spelling Normalization*

*Spelling Normalization* merupakan proses perbaikan atau subtitusi kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Subtitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya mempunyai maksud dan arti yang sama tetapi akan dianggap sebagai entitas yang berbeda pada saat proses penyusunan matriks

b. *Case Folding*

*Case Folding* adalah proses penyamaan *case* dalam sebuah dokumen. Hal ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu peran *case holding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar.

c. *Tokenizing*

*Tokenizing* adalah proses memecah kalimat menjadi kata-kata yang dilakukan untuk menjadikan sebuah kalimat menjadi lebih bermakna. Tahap pertama yang dilakukan adalah normalisasi kata dengan mengubah semua karakter huruf menjadi huruf kecil atau *to LowerCase.* Proses tokenisasi diawali dengan menghilangkan delimiter-delimiter yaitu *symbol* dan tanda baca yang ada pada teks tersebut seperti @, $, &, tanda titik (.), koma (,), tanda Tanya (?), tanda seru

(!). tahap tokenisasi selanjutnya yaitu proses penguraian teks yang semula berupa kalimat-kalimat yang berisi kata-kata. Proses pemotongan *string* berdasarkan tiap kata yang menyusunnya, umumnya setiap kata akan terpisahkan dengan karakter spasi, proses tokenisasi mengandalkan karakter spasi pada dokumen teks untuk melakukan pemisahan. Hasil dari proses ini adalah kumpulan kata saja (Putri,2016).

d. *Filtering*

Tahap filtrasi adalah tahap mengambil kata-kata penting dari hasil token. Algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata yang penting) dapat digunakan pada tahap ini. *Stopword* adalah kata-kata yang tidak deskriptif dan bukan merupakan kata penting dari suatu dokumen sehingga dapat dibuang. Contoh *stopword* adalah "yang", "dan", "di","dari", dan seterusnya (Putri, 2016). Dalam filtrasi ini menggunakan *stoplist/stopword* agar kata-kata yang kurang penting dan sering muncul dalam suatu dokumen dibuang sehingga hanya menyisakan kata-kata yang penting dan mempunyai arti yang diproses ke tahap selanjutnya.

e. *Stemming*

*Stemming* bertujuan untuk mengurangi jumlah kata dan mendapatkan kata dasar yang benar-benar sesuai.


**Gambar 2.3 Sistem Arsitektur *Text Mining***

Penelitian dibidang *text mining* menangani masalah yang berkaitan dengan representasi teks, klasifikasi, *clustering*, ekstraksi informasi atau pencarian dan pemodelan pola. Dalam hal ini pemilihan karakteristik, juga domain penelitan dan prosedur penelitian menjadi peran penting. Oleh karena itu, adaptasi dari algoritma *data mining* dari teks yang diketahui sangat diperlukan. Maka dari itu untuk mencapai hal ini seringkali berdasarkan penelitian sebelumnya *text mining* bergantung pada *information retrieval, natural language processing* dan *information extraction*. Selain itu juga penerapan metode *data mining* dan statistic juga diterapkan untuk menangani masalah ini (Hotho, 2005). *Information Retrieval* (IR) adalah menemukan bahan (biasanya dokumen) dari suatu keadaan yang tidak terstruktur (biasanya teks) yang memenuhi kebutuhan informasi dari dalam kumpulan data yang besar (biasanya disimpan didalam komputer) (Manning, dkk. 2009). *Natural Language Processing* (NLP) bertujuan untuk mencapai hasil yang lebih baik dalam pemahaman bahasa alami dengan menggunakan komputer. Sedangkan Ekstraksi Informasi (IE). Bertujuan untuk menemukan informasi tertentu dari dokumen teks yang kemudian Ini disimpan dalam basis data seperti pola sehingga dapat digunakan dan dimanfaatkan (Hotho, 2005). Hotho (2005). juga mengatakan bahwa pada

penelitian *text mining* diperlukan tahapan *text preprocessing* pada koleksi dokumen dan menyimpan informasi tersebut dalam struktur data. Pendekatan *text mining* didasarkan pada pemikiran bahwa dokumen teks dapat diwakili oleh satu *set* kata-kata, yaitu dokumen teks digambarkan berdasarkan pada set kata-kata yang terkandung di dalamnya.

### 1.5.2 *Feature Selection*

*Feature Selection* merupakan tahap lanjutan dari pengurangan dimensi. Walaupun di tahap sebelumnya sudah melakukan penghapusan kata-kata yang tidak deskriptif (*stopword*), tidak semua kata-kata didalam dokumen memiliki arti penting. Sehingga untuk mengurangi dimensi, pemilihan hanya dilakukan pada kata-kata yang relevan dan yang benar-benar mempresentasikan isi dari suatu dokumen. Kata-kata yang dinilai penting dilihat dari intensitas kemunculan dan yang paling informatif dari keseluruhan.

a. Pembobotan Kata (*Term Weighting*)

Hal yang perlu diperhatikan dalam pencarian informasi dari koleksi dokumen yang heterogen adalah pembobotan *term*. *Term* dapat berupa kata, *frase* atau unit hasil *indexing* lainnya dalam suatu dokumen yang dapat digunakan untuk mengetahui konteks dari dokumen tersebut. Karena setiap kata memiliki tingkatan kepentingan yang berbeda dalam dokumen , maka untuk setiap kata tersebut diberikan sebuah indikator, yaitu *term weight* (Zafikri, 2008). Zafikri (2008)

menyatakan *term weighting* atau pembobotan *term* sangat dipengaruhi oleh hal-hal sebagai berikut:

1. *Term Frequency* (TF)

*Term Frequency* (TF) adalah faktor yang menentukan bobot *term* pada suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Nilai jumlah kemunculan suatu kata (*term frequency*) diperhitungkan dalam pemberian bobot terhadap suatu kata. Semakin besar jumlah kemunculan suatu term (tf tinggi) dalam dokumen, semakin besar pula bobotnya dalam dokumen atau akan memberikan nilai kesesuaian yang semakin besar.

2. *Inverse Document Frequency* (IDF)

*Inverse Document Frequency* (IDF) adalah pengurangan dominansi *term* yang sering muncul di berbagai dokumen. Hal ini diperlukan karena *term- term* yang banyak muncul di berbagai dokumen, dapat dianggap sebagai *term* umum (*common term*) sehingga tidak penting nilainya. Sebaliknya faktor jarang munculnya kata (*term scarcity*) dalam koleksi dokumen harus diperhatikan dalam pemberian bobot. Kata yang muncul pada sedikit dokumen harus dilihat sebagai kata yang lebih penting (*uncommon terms*) daripada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen

yang mengandung suatu kata (*Inverse Document Frequency* ) (Putranti dan Winarko, 2014).

Metode TF-IDF merupakan metode pembobotan *term* yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru.

Setiap *term* yang telah di-*index* diberikan bobot sesuai dengan struktur pembobotan yang dipilih, apakah pembobotan lokal, global atau kombinasi keduanya. Jika menggunakan pembobotan lokal maka, pembobotan *term* diekspresikan sebagai *tf* (*term frequency*). Namun, jika pembobotan global yang digunakan maka, pembobotan *term* didapatkan melalui nilai *idf* (*inverse document frequency*). Beberapa aplikasi juga ada yang menerapkan pembobotan kombinasi keduanya yaitu, dengan mengalikan bobot lokal dan global (*tf* . *idf*) (Bintana, 2012).

1. *Term Frequency*

   Empat cara yang dapat digunakan untuk memperoleh nilai *term frequency* (*tf*), yaitu:

*a.* *Raw term frequency*. Nilai *tf* sebuah *term* diperoleh berdasarkan jumlah kemunculan *term* tersebut dalam dokumen. Contoh kasus dimana *term* muncul sebanyak dua kali dalam suatu dokumen maka, nilai *tf term* tersebut adalah 2.

b. *Logarithm term frequency*. Hal ini untuk menghindari dominasi dokumen yang mengandung sedikit *term* dalam *query*, namun mempunyai frekuensi yang tinggi.

Cara ini menggunakan fungsi logaritmik matematika untuk memperoleh nilai *tf*.

$$tf = 1 + \log(tf) \quad (2.1)$$

c. *Binary term frequency*. Hanya memperhatikan apakah suatu *term* ada atau tidak dalam dokumen. Jika ada, maka *tf* diberi nilai 1, jika tidak ada diberi nilai 0. Pada cara ini jumlah kemunculan *term* dalam dokumen tidak berpengaruh.

d. *Augmented Term Frequency*. Nilai *tf* adalah jumlah kemunculan suatu *term* pada sebuah dokumen, sedangkan nilai *max(tf)* adalah jumlah kemunculan terbanyak sebuah *term* pada dokumen yang sama.

$$idf_{j=} \log \frac{D}{df} \quad (2.2)$$

## 2.6 Klasifikasi

Klasifikasi merupakan metode yang digunakan untuk mengelompokkan sebuah objek ke dalam kelompok atau kelas tertentu. Algoritma klasifikasi yang banyak digunakan secara luas, yaitu *Decision* atau *Classification Trees*, *Bayesian Classifiers* atau *Naïve Bayes classifiers*, *Neural Networks*, Analisa Statistik, Algoritma Genetika, *Rough Sets*, *K-Nearest Neighbor*, Metode *Rule Based*, *Memory Based Reasoning*, dan *Support Vector Machines (SVM)*. Proses ini dilakukan agar data atau citra dapat dikategorikan dalam suatu kelas tertentu yang telah ditentukan(Suyanto, 2017).

## 2.7 Metode Klasifikasi Sentimen

Ada dua pendekatan utama dalam menentukan orientasi sentimen yaitu, pendekatan *supervised learning* dan *unsupervised learning.*

### 2.7.1 *Surpervised Learning*

*Supervised learning* adalah sebuah pendekatan dimana sudah terdapat data yang dilatih, dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan (klasifikasi) suatu data ke data yang sudah ada. Algoritma ini mengamati sekumpulan pasangan masukan- keluaran dalam jumlah yang cukup besar dan kemudian menghasilkan sebuah model yang mampu memetakan masukan yang baru menjadi keluaran yang tepat. terdapat empat isu yang harus dipertimbangkan dalam menggunakan teknik *supervised learning* (Feldman, dkk. 2007) yaitu perlunya memutuskan kategori yang akan digunakan untuk mengklasifikasikan kasus. Kedua, dibutuhkan satu set pelatihan untuk masing-masing kategori. Ketiga, perlu menentukan fitur dari setiap kategori. Biasanya, lebih baik untuk menghasilkan fitur sebanyak mungkin karena sebagian besar algoritma akan dapat fokus hanya pada fitur yang relevan. Terakhir, perlu memutuskan algoritma yang akan digunakan untuk kategorisasi tersebut. Beberapa algoritma yang biasa digunakan terhadap pendekatan *supervised learning*, diantaranya *naïve bayes*, dan *support vector machines* (SVM). *Supervised learning* bergantung pada data pelatihan. Model klasifikasi berdasarkan data latih yang telah diberi label dalam satu domain,

sering berkinerja buruk dengan domain yang berbeda. Meskipun adaptasi domain telah dipelajari oleh para peneliti, namun teknologi ini masih jauh dari sempurna (Liu, 2015).

### 2.7.2 *Unsupervised Learning*

Dalam pendekatan *unsupervised learning*, metode diterapkan tanpa adanya latihan (*training*) dan tanpa adanya guru (*teacher*). Guru yang dimaksud adalah label dari data. Misalkan ada sekelompok pengamatan atau data tanpa ada label (*output*) tertentu, makan dalam *unsupervised learning* harus mengelompokkan data tersebut ke dalam beberapa kelas yang kita kehedaki. Ini terutama dilakukan karena data yang ada tidak memiliki label. Label menandai kemana data akan dikelompokkan. Untuk melakukan tugas (*task*) ini bias kita menerapkan metode *unsupervised learning.*

*Unsupervised* (Pang, dkk. 2008) adalah teknik yang terlebih dahulu menciptakan sebuah sentimen tanpa data latih, dan kemudian menentukan orientasi sentimen dari unit teks melalui beberapa fungsi berdasarkan positif dan negatif. Menentukan sentimen dengan pendekatan *unsupervised* adalah melalui kata-kata atau frase dengan polaritas sentimen, juga disebut sebagai orientasi semantik. Masuk kedalam kelompok ini adalah Metode *Lexicon.*

## 2.8 *K-Nearest Neighbor*

*K- Nearest Neighbor (K-NN)* adalah algoritma untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples* (data latih). Dimana hasil dari sampel uji

yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *K-NN*. Algoritma *K-NN* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru (Krisandi, dkk. 2013). Data latih akan dibangun dengan memperhatikan keseimbangan dokumen satu sama lain. Adapun algoritma *K-NN* dapat dijelaskan (Oktinas, 2017) dengan keterangan berikut :

a. Hitung jarak antara data sampel (data uji) dengan data latih yang telah dibangun. Salah satu persamaan dalam menghitung jarak kedekatan dapat menggunakan persamaan *Cosine Similarity*.

b. Menentukan parameter nilai *k* = jumlah tetanggaan terdekat.

c. Mengurutkan jarak terkecil dari data sample

d. Pasangkan kategori sesuai dengan kesesuaian

e. Cari jumlah terbanyak dari tetanggaan terdekat. Kemudian tetapkan kategori. Jarak yang digunakan dalam penelitian ini adalah *Cosine Similarity.*

Jarak yang digunakan dalam penelitian ini adalah *Cosine Similarity.*

$$\text{Cos(i,k)}\frac{\sum k(di,dk)}{\sum\sqrt{k\ d^2\ ik}\sqrt{\sum k d^2\ jk}} \qquad (2.3)$$

Keterangan:

$\sum k(di, dk)$ = vector dot produk dari i ke k

$\sum kd^2\ jk$ = panjang vektor i

Algoritma *K-NN* (Krisandi, dkk. 2013) adalah algoritma yang menentukan nilai jarak pada pengujian data *testing* dengan data *training* berdasarkan nilai terkecil dari nilai ketetanggaan terdekat didefinisikan sebagai berikut:

$$Dnn\ (c1,c2) = min\ 1 \leq I \leq r\ ,1 \leq j \leq s\ d\ (yi,zi) \qquad (2.4)$$

## 2.9 *Confussion Matrix*

Dalam mengukur tingkat akurasi sistem orientasi sentimen, maka digunakan tabel *confussion matrix*. *Confusion matrix* adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan. Contoh *confusion matrix* untuk klasifikasi biner ditunjukkan pada tabel 2.1 berikut:

**Tabel 2.1 *Confution Matrix***

| Kelas Sebenarnya | Kelas hasil prediksi | | |
|---|---|---|---|
| | Positif=1 | Negatif=-1 | Netral=0 |
| Positif= 1 | TP | FN | TN |
| Negatif= -1 | FP | TN | FN |
| Netral= 0 | TN | FN | TN |

Keterangan untuk tabel diatas dinyatakan sebagai berikut :

1. *True Posstive* (TP), yaitu jumlah dokumen dari kelas 1 yang benar dan diklasifikan sebagai kelas 1.

2. *True Negative* (TN), yaitu jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0.
3. *False Positive* (FP), yaitu jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1.
4. *False Negative* (FN) yaitu jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0.
5. *Netral Positive* (NP) yaitu jumlah dokumen dari kelas 0 yang bersufat netral
6. *Netral Negative* (NF) yaitu jumlah dokumen dari kelas 0 yang bersifat Netral

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN+NF+FN} \quad (2.5)$$

Keterangan :

TP = jumlah dokumen dari kelas 1 yang benar dan diklasifikan sebagai kelas 1.

TN = jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0.

FP = jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1.

FN = jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0.

NP = jumlah dokumen dari kelas 0 yang bersufat netral

NF = jumlah dokumen dari kelas 0 yang bersifat Netral