BAB 2

LANDASAN KEPUSTAKAAN

Pada bab ini penulis membahas mengenai pustaka yang berkaitan dengan penelitian.

2.1 Kajian Pustaka

Terdapat beberapa metode yang diterapkan pada klasifikasi teks dokumen di dunia maya yang telah mengalami perkembangan beberapa tahun terakhir ini, dikarenakan peningkatan jumlah pengguna internet dan media sosial yang terus meningkat dari tahun-ketahun. Perkembangan teknologi melaju bersamaan dengan banyak pengumpulan dan penyimpanan data menyebabkan tumpukan data yang sangat banyak. Dengan banyaknya data yang menumpuk membuat pengguna internet juga membutuhkan informasi penting dari polapola yang terstruktur. Informasi maupun berita seolah-olah menjadi kebutuhan banyak manusia setiap harinya. Sampai ini hampir setiap manusia memiliki akun media sosial, mereka tidak hanya mengekspresikan opini, namun mereka juga seringkali menggali informasi melalui media sosial tersebut. Dari banyak media sosial yang tersedia, Twitter salah satunya yang dapat membantu penggunanya meraih segala informasi yang diinginkan. Cara yang digunakan untuk mendapatkan informasi berpola dan memiliki kualitas merupakan yang disebut dengan *data mining*. Salah satu metode klasifikasi yang cukup populer adalah K-Nearest Neighbor.

Dalam sebuah jurnal milik Sukarno (2016) dengan judul *Klasifikasi Tweets pada Twitter Menggunakan Metode Naive Bayes dan Query Expansion* berbasis Apriori*,* jurnal tersebut menyebutkan bahwa penelitian yang telah dilakukan menghasilkan akurasi terbaik berada di angka 82% dengan pada saat data latih tertinggi berjumlah 1600 data latih. Namun kekurangan dari penelitian tersebut adalah dibutuhkannya waktu yang lama saat proses *running*, karena penggunaan data dan pengujian yang dilakukan masih terlalu banyak dan kurang efisiennya skenario pengujiannya untuk memperoleh hasil akurasi maksimalnya. Selain itu penelitian lainnya yang dilakukan oleh Rungsawang (1999), dkk, Hasil yang diperoleh dengan menggunakan *query expansion* berbasis apriori sangat menjanjikan. Dengan menggunakan teknik *query expansion* oleh Rungsawang,dkk, mampu mendapatkan nilai peningkatan 19%.

Dengan demikian, peneliti mengusulkan metode klasifikasi lain, dan menggunakan *query expansion* untuk memperoleh hasil akurasi terbaik. Seperti penulisan saran pada Agung Kharisma Sukarno yaitu apakah klasifikasi teks dengan *query expansion* dengan metode yang lainnya dapat diterapkan dengan menghasilkan akurasi bernilai tinggi.

2.2 *Twitter*

Twitter adalah salah satu media sosial di dunia maya yang dapat membantu penggunanya dapat membagi dan menggali informasi dengan mem-*posting* kicauan (*tweets*) dengan maksimal 280 karakter. Pengguna Twitter akan dapat mengikuti dan memiliki pengikut pada akunnya. *Tweets* yang muncul pada beranda adalah *posting* yang muncul dari akun yang diikutinya (kwak et al., 2010). Fitur yang ada pada Twitter sebagai berikut (O'Reilly dan Milstein, 2012).

2.2.1 Halaman Utama (*Home*)

Halaman ini biasa disebut dengan beranda. Jika mengguna mulai memasuki pada akun yang pertama muncul adalah tampilan beranda. Disini akan bermunculan banyak tweets yang dikirimkan oleh pengguna yang diikutinya (*Following*).

2.2.2 Profil (*Profile*)

Pada halaman ini akan tampak postingan pemilik akun, serta informasi , gambar profil, dan jumlah pengikut bahkan yang diikutinya.

2.2.3 Favorit (*Favorite*)

Sebuah tanda dimana pengguna dapat memberikannya pada sebuah *tweets* agar tidak hilang pada halaman sebelumnya.

2.2.4 *Mention*

Menu ini adalah berguna untuk memberikan balasan percakapan antar pengguna Twitter dengan menandai @username pada teksnya.

2.2.5 Pesan Langsung (*Direct Message*)

Kegunaan pesan ini adalah seperti SMS yang dikirimkan kepada pengguna

2.2.6 Tagar (*Hashtag*/#)

Hashtag (#) digunakan sebagai tanda pada suatu topik atau teks tertentu setelah di-*posting*, agar dapat dicari keberadaanya melalui pencarian di Twitter.

2.2.7 Topik populer(*Trending Topic*)

Tweets yang sedang populer dan banyak dibicarakan oleh banyak pengguna dalam waktu tertentu.

2.2.8 *List*

Pengelompokan akun yang di ikuti oleh pengguna Twitter dalam suatu grup agar memudahkan pengguna untuk melihat akun tertentu yang telah dipilahnya.

2.2.9 Mengikuti (*Following*)

Seorang pengguna yang telah mengikuti akun Twitter lainnya untuk saaaling berbagi informasi melalui *tweets.*

2.2.10 Pengikut (*Follower*)

Kebalikan dari *following,* Pengikut (*Follower*) yaitu seorang pengguna Twitter lainnya yang mengikuti atau menjadikan kita sebagai teman untuk saling berbagi informasi melalui *tweets.*

2.3 Text Mining

Text Mining adalah suatu proses yang dilakukan guna mengekstraksi pola berupa informasi atau pengetahuan yang bermanfaat yang berasal dari kumpulan teks. Biasanya sumber teks yang di dapat berasal dari dokumen *Word*, PDF, kutipan teks, dan lain-lain. Manfaat dari *text mining* adalah memudahkan pencarian dan membangun inovasi bagi manusia untuk mengerti dan menggunakan informasi berharga dari suatu *repository* dokumen (Hand, 2010). Adapun proses yang dapat dilakukan oleh *Text Mining a*ntara lain adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks dan lain-lain.

2.3.1 *Text Preprocessing*

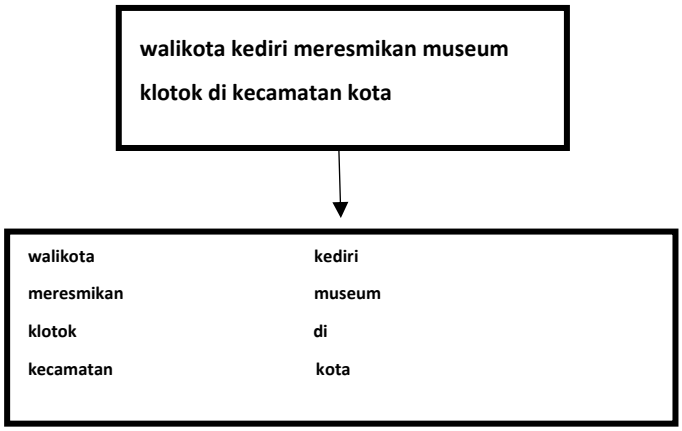
Pada proses *text mining* diperlukan adanya suatu tahapan-tahapan yang harus dilalui untuk mendapatkan informasi secara terstruktur. salah satu dari proses *text mining* adalah tahapan *preprocessing text.* Tahapan ini dilakukan untuk menyeleksi dan memfilter teks yang baik dan berguna dan siap untuk dianalisis (Hadna, et.al.,2016). Adapun tahapan *preprocessing text* yaitu meliputi *Case Folding, Tokenizing, Filtering, dan Steming.*

2.3.1.1 *Case Folding*

Case Folding merupakan tahapan proses *text preprocessing* yang dapat merubah semua huruf dari data teks menjadi huruf menjadi huruf kecil. Dengan huruf “a” hingga huruf “z” yang hanya akan diterima oleh sistem. Selain itu, karakter yang bukan huruf melainkan adalah simbol bahkan tanda baca akan di hilangkan dan dianggap oleh sistem sebagai *delimiter*. Contohnya adalah “Kantor Bumiputera merupakan kantor jenis ASURANSI yang dibuka di berbagai kota (Seluruh Indonesia)” menjadi “kantor bumiputera merupakan kantor jenis asuransi yang dibuka di berbagai kota seluruh indonesia”

2.3.1.2 *Tokenizing*

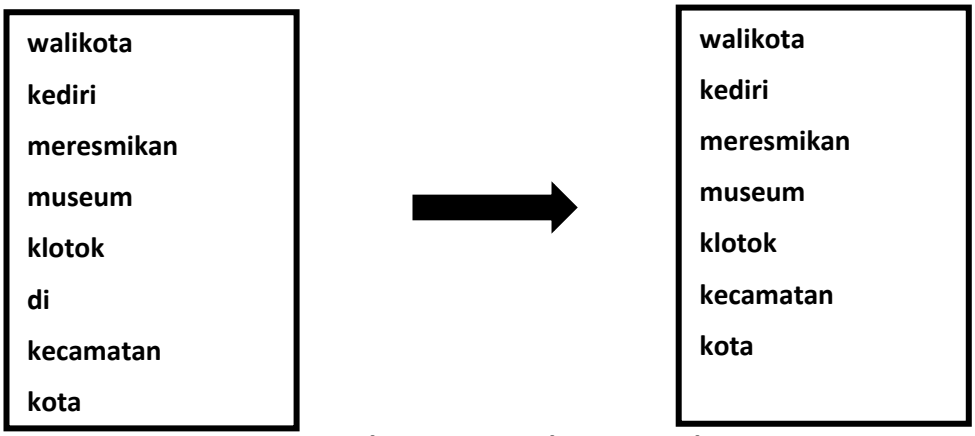
Tahapan *tokenizing* adalah proses untuk memecah semua kalimat menjadi per kata. Tahapan ini juga mempunyai tujuan untuk membuang karakter yang berbentuk tanda baca (Insan, 2013).



Gambar 2.1 Contoh Proses *Tokenizing*

2.3.1.3 *Filtering*

Filtering berfungsi untuk pengambilan kata-kata yang dianggap penting dari hasil *tokenizing*. Bisa menggunakan algoritme *stoplist* (menghilangkan kata kurang penting) atau juga bisa menggunakan algoritme *wordlist* (pengambilan kata penting dari *string(Insan, 2013).* Penghapusan *stopword* ini juga memiliki efek sebagai berikut:



Gambar 2.2 Contoh Proses *FIltering*

2.3.1.4 *Stemming*

Metode ini digunakan sebagai pencarian kata dasar pada kata. Agar memperoleh hasil yang mendekati kata sempurna dalam suatu pemrosesan teks maka harus dilakukan *stemming* dikarenakan pada saat kata yang baru melalui proses *filtering* pasti adalah kata yang masih asli dan pada setiap kata mempunyai imbuhan-imbuhan yang berbeda-beda dari kata satu dengan kata

lainnya walau kata dasarnya sama. Pada proses *stemming* ini bergantung pada domain kata atau bahasa yang digunakan domain bahasa Indonesia berbeda dengan domain yang menggunakan bahasa Inggris. Pada bahasa Inggris proses yang digunakan hanya penghilangan sufiks. Sedangkan pada bahasa Indonesia terdapat banyak yang perlu diperhatikan selain sufiks ada juga prefiks dan konfiks yang perlu dihapus. Sebagai contoh dari proses *stemming*, Kata bersama, kebersamaan, menyamai, akan diubah menjadi kata dasar "sama" (Insan, 2013).

2.4 *Query Expansion*

Metode ini merupakan salah satu teknik dasar pada *relevance feedback* di mana sistem akan menambahkan *query* tambahan pada pencarian pertama (Fachrudin, 2011). Pada klasifikasi *short text* seringkali terjadi kendala pada kata yang muncul dalam dokumen *short text* sering tidak terdapatnya di dokumen latih. Selain itu terkadang muncul permasalahan yang baru seperti proses klasifikasi tidak berjalan dengan baik karena akan ada banyak kata yang tidak mampu terdeteksi masuk pada kategori yang mana. Proses pengelompokkan dokumen pada kategori-kategori yang harus mempunyai kesamaan kata dengan dokumen latih.

Solusi dari permasalahan tersebut adalah mengunakan teknik *query expansion.* Teknik tersebut merupakan perluasan *query* dengan memformulasikan kembali *query* awal dengan melakukan penambahan kata untuk meningkatkan kinerja. Dengan kata lain, perbendaharaan kata yang ada di *dataset* semakin banyak sehingga dapat membantu kinerja pada klasifikasi untuk lebih baik lagi.

2.5 *Distributional Semantic*

Distributional semantic merupakan teori yang mempelajari tentang metode untuk mengukur dan mengkategorikan teks yang mempunyai makna yang sama berdasarkan sifat distribusinya dari data informasi yang besar dan berasal dari eksternal. Dalam hal ini, dapat dimaksud dengan kombinasi dua kata berbeda namun memiliki arti maupun level yang sama (Nikolaos, 2013). Terdapat salah satu parameter yang digunakan pada penelitian ini, yaitu *similarity measure* dengan menggunakan teknik yang dipakai adalah *Euclidean Distance*. Teknik ini berguna untuk menghitung jarak antar dokumen untuk mengetahui tingkat kemiripan berdasarkan nilai jarak.

2.5.1 *Euclidean Distance*

Metode ini adalah salah satu pendekatan yang bisa digunakan dalam membantu proses perhitungan jarak antar dokumen maupun teks yang terdapat

pada *dataset* yang telah tersedia. *Euclidean Distance* adalah metrika yang paling sering digunakan untuk menghitung kesamaan dua vektor (titik). Rumus *Euclidean Distance* adalah akar dari kuadrat perbedaan dua titik data (Sendhy, 2014). Persamaan 2.1 merupakan bentuk perhitungan *Euclidean Distance.*

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})2} \quad (2.1)$$

Keterangan:

D_{ij} = Tingkat perbedaan (*dissimilarity degree*)

N = Jumlah vektor/titik

X_{ik}= Titik *input*

Xj_k= Titik pembanding

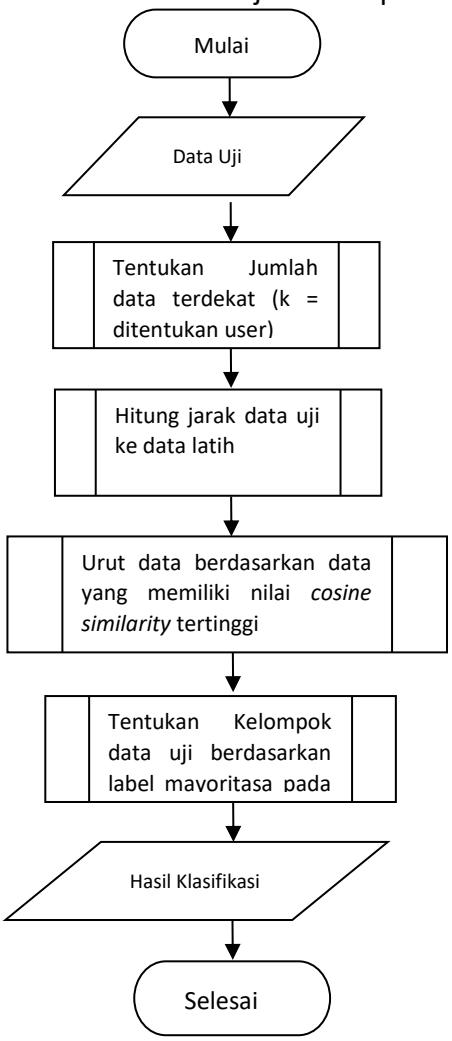
2.6 Klasifikasi Teks

Klasifikasi merupakan teknik bagaimana untuk mengkategorikan teks yang sesuai dengan karakteristiknya. Dengan terdapatnya teknik tersebut,dapat memberikan pandangan konseptual mengenai cara bagaimana untuk mengelompokkan dokumen yang mempunyai peran penting pada dunia nyata (Sriram eal, 2010). Dalam hal ini, penulis menggunakan metode K-Nearest Neighbor untuk teknik klasifikasi teksnya.

2.6.1 *K-Nearest Neighbor*

K-Nearest Neighbor adalah teknik pengelompokan data baru dengan menghitung jarak dari data baru dengan beberapa data terdekat (Santosa, 2007). Dalam arti kata lain yaitu suatu pendekatan guna mencari permasalahan dengan menghitung jarak terdekat antara kasus baru dengan beberapa kasus yang lama, dan mencocokkan berdasarkan bobot dari sejumlah data yang ada. Algoritme ini sangat mudah jika digunakan untuk teknik klasifikasi, yang sangat membantu unuk mengklasifikasikan data baru berdasarkan atribut dan *training sample* yang selanjutnya dihasilkan titik *training* paling dekat dengan titik *query.*

Adapun algoritme dari KNN ditunjukkan pada *flowchart* berikut:



Gambar 2.3 Diagram Alir *K-Nearest Neighbor*

2.6.1.1 *Cosine Similarity*

Cosine Similarity merupakan metode yang digunakan sebagai perhitungan tingkat kemiripan atau kesamaan antar kedua buah teks maupun objek. Umumnya perhitungan dengan metode tersebut sebelumnya berdasarkan *vector space similarity measure*. Metode *cosine similarity* ini menghitung dua buah obyek atau dinyatakan pada dua buah *vector* dengan kata kunci (*keywords*) dari dokumen sebagai tolak ukur.

$$Cosine\ (d_i, q_i) = \frac{q_i.d_i}{|q_i||d_i|} = \frac{\sum_{j=1}^{t}(q_{ij}.d_{ij})}{\sqrt{\sum_{j=1}^{t}(q_{ij})^2}.\sum_{j=1}^{t}(d_{ij})^2} \quad (2.2)$$

Keterangan:

q_{ij} = bobot istilah *j* pada dokumen *i*=$tf_{ij}\ .idf_j$

d_{ij} = bobot istilah *j* pada dokumen *i*=$tf_{ij}\ .idf_j$

2.7 Evaluasi

Proses evaluasi digunakan sebagai membandingkan antara hasil implementasi dengan kriteria standar yang telah ditetapkan untuk memperoleh seberapa nilai keberhasilannya. Sehingga dari hasil evaluasi yang didapatkan akan tersedia informasi mengenai sejauh mana nilai yang diperoleh selisih dari standar yang ditetapkan dengan hasil yang bisa dicapai. Untuk menghitung nilai akurasi dengan menggunakan rumus persamaan berikut.

$$Akurasi = \frac{Jumlah\ Data\ Uji\ Benar}{Jumlah\ Seluruh\ data\ uji} * 100\% \quad (2.3)$$