Bab 2 Landasan Teori

2.1 Kajian Penelitian Terdahulu

Orebaugh dan Allnutt (2009) melakukan penelitian dalam analisis forensik untuk mengidentifikasi pengguna *Instant Messanger* (IM) berdasarkan tingkah laku pengguna IM. Karakteristik yang digunakan yaitu *stylometric features* dengan kategori identifikasi kata menggunakan kata singkatan, struktur kalimat, karakter khusus, dan emoticon. Data yang digunakan dalam penelitian ini terdiri dari *Log* percakapan IM dengan 4 kategori *class* user yaitu user1, user2, user3 dan user4. Data tersebut diuraikan untuk menghitung struktur kalimat dan frekuensi yang telah ditetapkan pada kategori karakter khusus, emoticon, dan singkatan. Selanjutnya untuk klasifikasi digunakan metode *J48 decision tree, IBk nearest neighbor, and Naïve Bayes classifiers.* Hasilnya, akurasi pada Naif Bayes dengan atribut singkatan (97.85%) , J48 (97.86%) dan IBk classifier (97.14%), hasil ketika semua atribut dikombinasikan menggunakan Naif Bayes memberikan akurasi terbaik yaitu 99,29%.

Penelitian Dinakar (2012) untuk mendeteksi *cyberbullying* pada jejaring sosial *Youtube* dan *Formspring.me* yang merupaka situs jejaring sosial yang popular di kalangan remaja. Penelitian ini menggunakan pendekatan state-of-the-art NLP dan sense knowledge base yang mana topiknya sangat luas pada kehidupan sehari-hari. Subyek yang diteliti seperti penampilan, *knowledge, racial* dan *ethnic shrul.* Penelitian ini menghasilkan aplikasi yang disebut *Sosial Network Dashboard* untuk menganalsis dan mendeteksi Cyberbullying. Contoh penggunaanya pada *Youtubube* yang mana komentar pada *youtube* dikategorikan dalam tiga kelompok pada 1500 komentar yang telah terunduh, diantaranya yang mengandung *cyberbullying* adalah kelompok *sexuality* 627 komentar, *Race and culture* 841 komentar dan *Intelligence* 809 komentar.

Penelitian Nalini dan Sheela (2014), mengusulkan sebuah framework berbasis *contextual* dan *word level feature* untuk mendeteksi *content* yang menyinggung dan mengidentifikasi pelaku *user* didalam *IRC logs*. Metode ini dibagi menjadi dua pase yaitu pertama, untuk mendeteksi level kata yang menyinggung, kedua memperoleh level user sebagai pelaku. Pada fase pertama, peneliti menerapkan teknik NLP seperti *word level feature* dan *contextual level feature*. Untuk fase kedua, peneliti menerapkan *user level feature* dengan

menggunakan *style, sturucture* dan fitur *cyberbullying*. Hasil dari metode tersebut akan di uji dengan menggunakan algoritma *classification* yang disediakan oleh WEKA, diantaranya *Random Forest, J48* dan *Sequential Minimal Optimization.*

Setty et al (2014) melakukan penelitian pada *facebook* untuk mengklasifikasikan berita ke beberapa kategori seperti *friends posts* dan *liked pages posts*. Kemudian *Friends posts* dikategorikan lagi menjadi *life events posts* dan *entertainment posts* pada halaman facebook. Penerapannya bertujuan untuk mengevaluasi klasifikasi *news feed* dan *sentiment analys* agar halaman berita facebook lebih terorganaisir dan terlihat lebih menarik. Penelitian ini Menggunakan aplikasi WEKA dan *Learning Model* untuk membandingkan beberapa algoritma klasifikasi pada 2000 data yang diekstrak dari facebook. Hasilnya, akurasi SVM dan Bayes Net lebih baik dari pada algoritma yang lain.

Kansara (2015), penelitian ini membangun sebuah *framework* dengan menggabungkan analisa teks dan gambar untuk mendeteksi ancaman *Cyberbullying.* Metode yang digunakan untuk analisa gambar adalah kombinasi antara *bag of visual word* (BoVW), *Local binary Pattern* (LBP) dan *SVM Classifier*. Sedangkan untuk analisa text menggunakan kombinasi *bag of word* (BoW) dan *Naïve Bayes Classifier*. Hasil analisa dari gambar dan text ini di proses lagi menggunakan *Boolean System* untuk menentukan apakah conten tersebut mrengandung *cyberbullying* atau tidak.

Penelitian selanjutnya, Hafilizara dan Adisantoso (2014) melakukan klasifikasi email spam dengan teknik Naïve Bayes, Pada teknik Naïve Bayes tersebut terdapat metode *smoothing* yang telah umum digunakan yaitu *Add-One smoothing* atau *Laplace smoothing*, namun masih ada metode lainnya yaitu *Jelinek-Mercer* (JM) *smoothing*, *Dirichlet* (Dir) *smoothing*, *Absolute Discounting* (AD) *smoothing*, dan *Two-Stage* (TS) *smoothing* yang dapat digunakan sebagai fungsi klasifikasi dan diduga mampu meningkatkan akurasi melebihi Laplace *smoothing*. Penelitian ini mengimplementasikan fungsi klasifikasi metode smoothing tersebut. Pada hasil percobaan terlihat akurasi yang dihasilkan metode *Laplace smoothing* lebih rendah dari metode smoothing lainnya. Dari hasil evaluasi terhadap nilai akurasi, *false rate*, dan *miss rate* terlihat metode *smoothing* Dirichlet memiliki nilai *miss rate* yang paling rendah sebesar 1.4%, nilai *false rate* 13.4%, dan akurasi 94.82%.

Penelitian Hosseinmardi et al (2015) pada sosial media Instagram menggunakan *Snowbal Sampling Method* telah mengidentifikasi 41K Ids user. Sekitar 61% user mempunyai *profile* yang *public* jadi sekitar 25K user. Setiap *user public* dikumpulkan dengan postingan gambar beserta 150 komentar, *id user* yang meng-*follow*, *id user* yang di-*follow*, *id user* yang mengomentari atau menyukai media yang di *share* oleh user tersebut. Hasilnya, 48% dari media tersebut bukan *cyberbullying*, melainkan hanya *cyberagression*, namun menggunakan

Linear SVM classifier dapat meningkatkan akurasi hingga 87% dengan menggabungkan fitur teks, gambar dan meta data dari media tersebut.

Akaichi (2013) melakukan penelitian dengan fokus pada penggunaan *teks mining* untuk *sentiment classification*. Ilustrasi dilakukan pada pengguna *Facebook* di Tunisia. metode yang diusulkan didasarkan pada dukungan Vektor mesin (SVM) dan Naif Bayes. peneliti juga membangun sebuah lexicon sentimen, berdasarkan emoticon, kata bahasa dan akronim, dari status data yang telah diekstrak. Selanjutnya, membandingkan dua *machine learning* yaitu algoritma SVM dan Naif Bayes melalui model pelatihan untuk klasifikasi sentiment. Hasilnya, data set dari data training 60% dan sisanya yaitu data test 40%.

Margono et al (2014) melakukan penelitian untuk menganalisa kata bullying berbahasa Indonesia pada *twitter*. Dengan menggunakan text mining, telah berhasil mengidentifikasi bahwa kata "bangsat" dan "anjing" merupakan kata yang paling sering digunakan untuk pola bullying berbahasa Indonesia, sekitar 14000 data yang di proses menggunakan metode Fp-Growth dan Association Rule Hasilnya, bahwa kata *bullying* “bangsat” dan “anjing” paling banyak digunakan oleh tweeter yang tinggal di Jakarta dan Surabaya. Perbandingan penelitian terdahulu ditunjukkan pada tabel 2.1:

Tabel 2. 1 Perbandingan Penelitian Terdahulu

No	Nama Peneliti	Objek	Judul Penelitian	Metode Penelitian	Hasil Penelitian
1.	Dinakar et al (2012)	Youtube dan Formspring. me	*Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying*	NLP dan Common Sense Knowledege Base	Menghasilkan applikasi yang diberi nama *Sosial Network Dashboard* dan salah satu aplikasinya digunakan untuk Mendetekasi *Cyberbullying* dari komentar *Youtube*. Dari 15000 ribu komentar di kategorikan ke dalam tiga kelompok *cyberbullying* yaitu *Sexuality* sebanyak 627 komentar, *Race & Culture* 841 komentar dan *Inteliigence* 809 komentar.
2.	Orebaugh dan Allnutt (2009)	Instant Messanging	*Classification of Instant Messaging Communicatio ns for Forensics Analysis*	*J48, IBK classifier* dan *Naïve Bayes Classifier*	Data yang digunakan dalam penelitian ini terdiri dari *Log* percakapan IM dengan 4 kategori *class* user yaitu user1, user2, user3 dan user4. Data tersebut diuraikan untuk menghitung struktur kalimat dan frekuensi yang telah ditetapkan pada kategori karakter khusus, emoticon, dan singkatan. Selanjutnya untuk klasifikasi digunakan metode *J48 decision tree, IBk nearest neighbor, and Naïve Bayes classifiers.* Hasilnya, akurasi pada Naif Bayes dengan atribut singkatan (97.85%) , J48 (97.86%) dan IBk classifier (97.14%), hasil ketika semua atribut dikombinasikan menggunakan Naif Bayes memberikan akurasi terbaik yaitu 99,29%.
3.	Hendro Margono. et al (2014)	Twitter	*Mining Indonesian Cyber Bullying Patterns in Sosial Networks*	FP-Growth, Assosiation Rule,	Data yang digunakan sebanyak 14000 *tweet*. Dengan fokus pada *tweet* bullying yang menggunakan bahasa Indonesia. Hasilnya, bahwa kata *bullying* “bangsat” dan “anjing” paling banyak digunakan oleh tweeter yang tinggal di Jakarta dan Surabaya.
3.	Shankar Setty. et al (2014)	Facebook	*Classification of Facebook News Feeds and Sentiment Analysis*	Menggunakan Algoritma BayesNet, J48, Naïve Bayes, SVM, Logistic Regression dari WEKA.	Menggunakan aplikasi WEKA dan *Learning Model* untuk membandingkan beberapa algoritma klasifikasi pada 2000 data yang diekstrak dari facebook. Hasilnya, akurasi SVM dan Bayes Net lebih baik dari pada algoritma yang lain.

Tabel 2.1 Perbandingan Penelitian Terdahulu (Lanjutan)

No	Nama Peneliti	Objek	Judul Penelitian	Metode Penelitian	Hasil Penelitian
4.	Kansara et al (2015)	Jejaring Sosial (Twitter, Facebook, Instagram dll)	*A Framework for Cyberbullying Detection in Sosial Network*	Untuk deteksi Gambar menggunakan metode *Bag of visual word (BoVW), Local Binnary Pattern (LBP),* dan *SVM Classification.* Untuk text metode yang digunakan yaitu *bag of word* (BoW) dan *Naïve Bayes Classifier*.	Penelitian ini membangun sebuah *Framework* untuk mendeteksi *Cyberbullying* menggunakan metode *Boolean System* dengan cara mengevaluasi hasil gambar dan hasil analisa teks.
5.	Mutia hafilizara dan Julio adisantoso (2014)	Email	Metode Smoothing dalam Naïve Bayes untuk Klasifikasi Email Spam.	Naïve Bayes, *Jelinek-Mercer* (JM) *smoothing*, *Dirichlet* (Dir) *smoothing*, *Absolute Discounting* (AD) *smoothing*, dan *Two-Stage* (TS)	Data email yang digunakan adalah korpus *email* public dengan 1396 *spam* , 250 *hard ham*, dan 1400 *ham*. Dari hasil evaluasi terhadap nilai akurasi, *false rate*, dan *miss rate* terlihat metode *smoothing* Dirichlet memiliki nilai *miss rate* yang paling rendah sebesar 1.4% , nilai *false rate* 13.4%, dan akurasi 94.82%.

Tabel 2.1 Perbandingan Penelitian Terdahulu (Lanjutan)

No	Nama Peneliti	Objek	Judul Penelitian	Metode Penelitian	Hasil Penelitian
6.	Hosseinmardi et al (2015)	Instagram	*Detection of Cyberbullying Insidents on The Instagram Sosial Network*	Metode yang digunakan untuk identifikasi user pada instagram menggunakan *Snowball sampling method*. Untuk klasifikasi menggunakan *Naive Bayes* dan *Linear SVM*.	Penelitian ini menggabungkan label komentar dan gambar. Hasil yang diperoleh, sekitar 48% dianggap bukan *cyberbullying*, sebagian menunjukkan hanya cyberaggression dan bukan *cyberbullying*, tetapi ketika menggunakan Linear SVM akurasi meningkat menjadi 87% dengan memasukkan multi-modal dari text, gambar, dan meta data dari media tersebut.
7.	Jalel Akaichi (2013)	Facebook	*Sosial Networks' Facebook' Statutes Updates Mining for Sentiment Classification*	Algoritma yang digunakan adalah *Naïve Bayes multinomial event driven model, SVM, SMO (Sequential Minimal optimization).*	Data yang digunakan sebanyak 260 status, dibagi menjadi tiga *lexicon* yaitu *Interjections, acronym,* dan *emoticons.* Hasilnya, jika menggunakan bigram dalam *Naïve bayes* akurasinya tinggi, dan jika menggunakan unigram dalam SVM akurasinya lebih tinggi daripada algoritma *Naïve Bayes.*
8.	Nalini et al (2014)	*Sosial Networking sites* dan *web site*	*A Survey on Datamining in Cyberbullying*	*Naïve Bayes, SVM, Classifier term frequency* dan tf-idf.	Penelitian ini mengusulkan metode secara *contextual* dan *Level word* untuk mendeteksi *content* yang menyinggung dan *user* yang menyinggung. Sebanyak 325 transkrip yang digunakan sebagai *dataset*. Pengujiannya dilakukan dengan membandingkan beberapa algoritma klasifikasi yang di sediakan oleh WEKA. Metode ini dianggap efektif untuk mendeteksi *cyberbullying* pada *Sosial Network* dan *Web site*.
9	Hariani (2015)	Twitter	Analisis Bukti Digital *Cyberbullying Pada Jejaring Sosial* Menggunakan *Naïve Bayes Classifier* (NBC)	*Tf-Idf, 10 fold cross validation, Naïve Bayes Classifier*	Metode Tf-Idf diterapkan pada data yang telah dibersihkan, kemudian menggunakan pengujian 10 fold cross validation sebagai validasi data, selanjutnya dilakukan klasifikasi menggunakan Naïve Bayes untuk mendapatkan informasi perkembangan Cyberbullying dan jenis cyberbullying yang biasa digunakan di Indonesia.

2.2 Cybercrime

Cybercrime merupakan bentuk-bentuk kejahatan yang timbul karena pemanfaatan teknologi internet. Menurut Hamzah (2012) dalam tulisannya mengartikan kejahatan komputer sebagai: "Kejahatan di bidang komputer secara umum dapat diartikan sebagai penggunaan komputer secara illegal". Dari pengertian di atas, secara ringkas dapat dikatakan bahwa *cybercrime* dapat didefinisikan sebagai perbuatan melawan hukum yang dilakukan dengan menggunakan internet yang berbasis pada kecanggihan teknologi komputer dan telekomunikasi. Selama ini dalam kejahatan konvensional, dikenal adanya dua jenis kejahatan sebagai berikut:

a. Kejahatan kerah biru (*blue collar crime*)

   Kejahatan ini merupakan jenis kejahatan atau tindak kriminal yang dilakukan secara konvensional seperti misalnya perampokan, pencurian, pembunuhan dan lain-lain.

b. Kejahatan kerah putih (*white collar crime*)

   Kejahatan jenis ini terbagi dalam empat kelompok kejahatan, yakni kejahatan korporasi, kejahatan birokrat, malpraktek, dan kejahatan individu.

Perkembangan penggunaan Internet telah menggeser banyak aktivitas masyarakat ke dunia maya. Kemajuan teknologi ini memberikan dampak positif dari segi kemudahan masyarakat untuk memperoleh pengetahuan dan informasi. Namun perkembangan internet tidak selamanya menghasilkan hal-hal positif. Internet dalam pengaruh negatifnya muncul dalam kasus-kasus semacam *cyberbullying, hoax, phisihing* sampai pada *cybercrime*. Beberapa kasus dampak negatif internet antara lain (Haryati, 2014):

1. Pencurian dan penggunaan account internet milik orang lain. Salah satu kesulitan dari sebuah TSP *{Internet Service Provider)* pelanggan mereka yang " dicuri " dan digunakan secara tidak sah adalah adanya *account*.
2. Membajak situs web, Salah satu kegiatan yang sering dilakukan oleh *cracker* adalah mengubah halaman *web,* yang dikenal dengan istilah *deface.*
3. Probing dan port scanning, Salah satu langkah yang dilakukan *cracker* sebelum masuk ke *server* yang ditargetkan adalah melakukan pengintaian. Cara yang dilakukan adalah dengan melakukan *"port scanning"* atau *"probing"* untuk melihat servis-servis apa saja yang tersedia di *server target.*
4. *Virus***,** seperti halnya di tempat lain, virus computer pun menyebar di Indonesia. Penyebaran umumnya dilakukan dengan menggunakan email.
5. Kejahatan yang berhubungan dengan nama *domain*. Nama domain *(domain name)* digunakan untuk mengidentifikasi perusahaan dan merek dagang.

6. *Spam*, yaitu email sampah yang berisi penawarkan produk/jasa, penipuan berkedok kerja sama bisnis, penyebaran virus, dll.
7. *Hoax*, Informasi palsu yang menyesatkan *(Hoax).*
8. *Illegal Content.* Kontenilegal, seperti perjudian dan pornografi.
9. *Phising*, Pencurian identitas untuk melakukan pencurian, penipuan, dll.
10. *Cyberbullying.* Kekerasan menggunakan media internet, seperti pelecehan, ancaman, dan tuduhan.
11. *Copyright Plagiarism.* hak cipta. *infringement*, Pelanggaran.

2.3 Data Mining dan Digital Forensik

Data mining merupakan aplikasi yang terdiri dari beberapa algoritma untuk mengekstrak pola data pengetahuan yang berguna dalam pengambilan keputusan. Data mining memiliki beberapa aplikasi digital forensic diantaranya adalah korelasi dalam forensic data (*association*), menemukan dan menyortir data ke dalam kelompok tertentu (*classification*), mengidentifikasi kelompok tertentu (clustering) dan menemukan pola data untuk memprediksikan sesuatu (*forecasting*). Data Mining tidak terbatas dalam digital forensic, *tools* dan model yang dikembangkan dapat membantu penyidik untuk menemukan bukti digital lebih efisien dan lebih cepat (Kayarkar 2014). Klasifikasi Tools digital forensic dalam data mining dapat dilihat pada table 2.2 dibawah ini:

Tabel 2. 2 Teknik Digital Forensik dan Data Mining

Digital Forensic Techniques	Data Mining Techniques	Tools
Data Recovery, Data generation and Pre-Processing	Statistical test Analysis Bartlett's test of sphericity Kaiser-Meyer-Olkin (KMO)	Recuva FTK Encase Sleuth kit/Autopsy ProDiscover
Data Analysis	Clustering –K-means, EM, Hierarchical Clustering	Weka
	Classification Supervised learning - Decision Tree, Neural Networks, SVM, Naïve Baiyesian	Weka
	Unsupervised learning PCA, Karnohuen Map	-

Tabel 2.2 Teknik Digital Forensics dan Data Mining (Lanjutan)

Digital Forensic Techniques	Data Mining Techniques	Tools
Data Analysis	Frequent Pattern Mining/Association rule Mining - Apriori, Eclat	Weka
	Named Entity recognition	Lingepipe
	Visualization	Cyber Forensics Time Lab
	Statistical Analysis and Anamoly Detection	EMT/MET
	Recursive data Mining	-
	Phishing	Invisible Witness
	Regression	-

Digital Forensik merupakan aplikasi bidang ilmu pengetahuan dan teknologi komputer untuk kepentingan pembuktian hukum (*pro justice*), yang dalam hal ini adalah untuk membuktikan kejahatan berteknologi tinggi atau *computer crime* secara ilmiah *scientific* hingga bisa mendapatkan bukti-bukti digital untuk menjerat pelaku kejahatan tersebut. Pencarian bukti-bukti digital untuk menjerat pelaku sering sering kali merupakan pekerjaan yang sangat kompleks dimana seorang *digital forensics analyst/investigator* harus mengikuti proesdur-prosedur yang diakui secara hukum baik nasional maupun Internasional, termasuk juga mereka harus memahami secara teoritis hal-hal yang berkaitan dengan bukti digital yang ditemukan, disamping juga memahami bagaimana penggunaan *software-software* forensic untuk mencari bukti-bukti digital tersebut dengan benar (Nuh, 2012).

Bukti digital diekstrak dari barang bukti elektronik. Beberapa contoh bukti digital yang dapat dianalisis adalah:

1. *Logical File*, yaitu file-file yang masih ada dan tercatat di file system di suatu partisi. Contoh: file aplikasi, library, office logs, multimedia dan lain-lain.
2. *Deleted File*, disebut juga *unallocated cluster* yang merujuk pada tempat penyimpanan yang telah terhapus.
3. *Lost File*, yaitu file yang sudah tidak tercatat lagi di file system yang sedang berjalan disuatu partisi.
4. *File slack*, yaitu sector penyimpanan yang berada diantara *end of file* (EoF) dengan *End of Cluster* (EoC).

5. *Log File*, yaitu file-file yang merekam aktivitas dari suatu keadaan tertentu misalnya log dari system operasi, internet browser, aplikasi, internet traffic, jejaring sosial dan lain-lain.
6. *Encrypted File*, yaitu file yang isinya sudah dilakukan enkripsi dengan menggunakan algoritma kriptografi sehingga tidak bias dilihat atau dibaca secara normal.

Dalam melakukan tindakan penegak hukum mempunyai acuan yang benar dan prosedural di dalam melakukan investigasi *computer crime* dan *computer-related crime*, serta memeriksa dan menganalisis barang bukti mempunyai *guidelines* yang telah diakui secara Internasional diantaranya adalah (Nuh, 2012):

1. *Good Practice Guide For Computer-Based Electronic Evidence* yang dikeluarkan *Association Of Chief Police Officers* (ACPO) yang merupakan asosiasi para pimpinan kepolisian di Inggris yang bekerja sama dengan 7safe.
2. *Forensics Examination of Digital Evidence: Guide for Law Enforcement,* yang dikeluarkan oleh *National Institute Of Justice* yang berada di bawah *U.S. Department of Justice.*
3. *Electronic Crime Scene Investigation: A Guide for First Responders*, yang juga dikeluarkan oleh *National Institute of Justice* yang berada di bawah *U.S. Departement of Justice.*

2.4 Jejaring Sosial

Seperti halnya di dunia nyata, menjalin hubungan persahabatan bisa juga dilakukan di dunia maya (internet). Bedanya, melalui internet ini tidak bisa bertatap muka secara langsung untuk berjabat tangan dan menanyakan siapa namanya. Di internet, seseorang bisa berkenalan dengan siapa pun, kapan pun, latar belakang, suku bangsa yang berbeda, bahkan antar negara dengan bahasa yang berbeda pula. Tentu saja, seperti halnya di dunia nyata, etika ketika mengajak berkenalan harus tetap dijaga. Meskipun yang diajak adalah teman lama, namun jangan sampai memaksa orang lain untuk menjadi teman kita. Proses untuk menjalin hubungan di dunia maya (internet) seperti itu, sering disebut *sosial networking* (jejaring sosial) Sudarma dalam (Akbar 2013).

Menurut modul CHFI (2011), jejaring sosial adalah semua orang berkomunikasi dengan berbagai perangkat dan media komunikasi yang digunakan. Seseorang dapat melakukan hubungan dengan orang yang lain, organisasi dan organisasi yang lain hanya bermodalkan sebuah perangkat. Berkomunikasi dengan kesamaan hobi, ide, teman dan sebagainya.

Pendapat lainnya, Jejaring sosial atau terjemahan Inggrisnya adalah *sosial networking* yang merupakan hubungan antar individu atau organisasi yang di bentuk karena adanya kesamaan, misalnya kesamaan visi, misi, pertemanan, keturunan, suku, dan sebagainya Utomo dalam (Akbar 2013).

Dengan demikian sebuah situs jejaring sosial adalah situs yang digunakan sebagai media untuk berinteraksi sosial. Selain itu situs jejaring sosial berfungsi sebagai media komunikasi antar anggotanya. Saat ini banyak situs jejaring sosial yang digunakan dan dikenal oleh para pengguna internet di Indonesia, antara lain *Friendster, Facebook, Twitter, Myspace.* Selain itu, masih banyak situs jejaring sosial yang lainnya.

Sampai saat ini, banyak sekali situs yang menyediakan khusus untuk menjalin hubungan di dunia maya. Meskipun inti tujuannya sama, masing-masing situs memiliki fitur yang berbeda. Ada yang khusus untuk menjaring pertemanan saja, menjaring pertemanan dengan lebih interaktif dan menguak memori dengan teman lama, atau lebih menonjolkan komunikasi dan interaksi dengan teman lewat *blog.*

2.5 Twitter

Twitter sebagai salah satu *micro-blogging system* merupakan media jejaring sosial yang memberikan fasilitas bagi para penggunanya untuk mem-*posting* segala hal terkait aktivitas, opini, dan segala hal yang terjadi kepada public atau suatu grup melalui pesan yang dikenal dengan sebutan *tweet*. Karakteristik yang dimiliki oleh *tweet* diantaranya Go dalam (Hidayatullah 2014):

1. *Length*

   Hal yang membedakan Twitter dengan jejaring sosial lainnya adalah adanya batasan dalam mem-posting *tweet* yaitu maksimum sebanyak 140 karakter sehingga pengguna dituntut untuk dapat mengekspresikan pandangan mereka menggunakan satu atau dua kata kunci.

2. *Data Availability*

   Keberadaan Twitter API memberikan kemudahan dalam mengumpulkan jutaan *tweet* atau lebih untuk *training*.

3. *Languange Model*

Model bahasa yang disampaikan dalam *tweet* berasal dari berbagai macam media yang berbeda sehingga dimungkinkan banyak terdapat ejaan yang salah ataupun penggunaan bahasa '*slang'* di dalamnya.

4. *Domain*

Pengguna twitter mengirimkan pesan singkat tentang bermcam-macam topik. Berbeda dengan situs *review* yang memberika opini yang difokus kan pada topik tertentu.

2.6 Cyberbullying

Cyberbullying merupakan istilah yang merujuk kepada penggunaan teknologi Informasi untuk menggertak orang dengan mengirim atau posting teks yang bersifat mengintimidasi atau mengancam. Istilah ini telah ditambahkan ke dalam kamus OED (Oxford English Dictionary). OED menunjukkan penggunaan istilah ini pertama kali di Canberra pada tahun 1998, namun sebelumnya istilah telah di gunakan pada artikel New York Times 1995 dimana banyak sarjana dan penulis Besley seorang Kanada yang meluncurkan website *cyberbullying* tahun 2013 dengan istilah coining (Machsun 2016).

Cyberbullying adalah tindakan seorang atau remaja secara sengaja mengintimidasi, mengancam, atau mempermalukan seseorang, atau sekelompok anak lain melalui teknologi informasi, seperti media sosial atau *mobile device*. Hasil penelitian Ipsos Global menunjukkan sebanyak 60 persen responden mengatakan *cyberbullying* terjadi di sejumlah laman media sosial terkemuka seperti *facebook* (Firman dan Ngazis, 2012). *Cyberbullying* tidak hanya berlaku pada remaja dan anak-anak namun juga hal yang sama berlaku pada orang dewasa. Perbedaan dalam kelompok umur ini disebut *cyberstalking* atau *cyberharassment. Cyberstalking* adalah penggunaan komunikasi internet, *e-mail*, atau elektronik lainnya dan umumnya mengacu pada pola perilaku mengancam atau berbahaya. Sementara *Cyberharassment* berbeda dari *Cyberstalking*, *Cyberharassment* biasanya berkaitan dengan mengancam atau melecehkan pesan *e-mail*, pesan instan, atau *blog* dan *website* yang didekasikan sepenuhnya untuk menyiksa individu (Haryati 2014).

Professor Dan Olweus pada tahun 1993 dalam (Rudi 2010) telah mendefinisikan bullying yang mengandung tiga unsur mendasar perilaku bullying, yaitu:

1. Bersifat menyerang (agresif) dan negatif.
2. Dilakukan secara berulang kali.
3. Adanya ketidakseimbangan kekuatan antara pihak yang terlibat.

Olweus kemudian mengidentifikasikan dua subtipe bullying, yaitu perilaku secara langsung (*Direct bullying*), misalnya penyerangan secara fisik dan perilaku secara tidak langsung (*Indirect bullying*), misalnya pengucilan secara sosial. Underwood, Galen, dan Paquette di tahun 2001, mengusulkan istilah "Sosial Aggression" untuk perilaku menyakiti secara tidak langsung. Riset menunjukkan bahwa bentuk bullying tidak langsung, seperti pengucilan atau penolakan secara sosial, lebih sering digunakan oleh perempuan daripada laki-laki. Sementara anak laki-laki menggunakan atau menjadi korban tipe bullying secara langsung, misalnya penyerangan secara fisik, dalam Nansel et al. 2001; Olweus 1997 (Rudi 2010).

Menurut survey global yang diadakan oleh *Latitude News*, Indonesia merupakan Negara dengan kasus bullying tertinggi kedua dunia setelah Jepang, dan mengalahkan kasus bullying di Amerika Serikat yang menempati urutan ketiga. Bullying paling banyak dilakukan melalui jejaring sosial. Sebagai Negara dengan jumlah populasi terbanyak keempat di sunia, Indonesia memiliki jumlah pengguna *Facebook* terbesar ketiga di dunia dan penyumbang 15 persen *tweet* setiap hari untuk *Twitter*. Berdasarkan penelitian 91% responden asal Indonesia mengaku telah melihat kasus *Cyberbullying*, dan paling sering terjadi melalui media sosial. Di Indonesia, 74% responden menunjuk *Facebook* sebagai biangnya *Cyberbullying*, dan 44% menyebut media website yang lain (Satalina 2014).

2.7 Tinjauan Undang-Undang

Cyberbullying di Indonesia diatur dalam Pasal 27 UU ITE, namun terdapat beberapa tindakan yang termasuk *cyberbullying* yakni *Flaming*, *Harassment* (gangguan), *Impersonation* (peniruan), *Outing* (menyebarkan rahasia orang lain), *Trickery* (tipu daya), Exclusion(pengeluaran), *Cyberstalking*. UU ITE hanya memuat unsur penghinaan dan pengancaman, padahal tindakan *cyberbullying* lainnya juga kerap kali terjadi dan menjadi langkah awal tindak pidana lain. Dengan berkembangnya situs jejaring sosial maka hal tersebut akan memudahkan pelaku *cyberbullying* melakukan tindakannya (satyawati 2014).

Didalam Pasal 27 ayat (3) UU ITE yang menyatakan bahwa Setiap Orang dengan sengaja dan tanpa hak mendistribusikan dan/atau mentransmisikan dan/atau membuat dapat diaksesnya Informasi Elektronik dan/atau Dokumen Elektronik yang memiliki muatan penghinaan dan/atau pencemaran nama baik. Kemudian dalam Pasal 27 ayat (4) UU ITE yang menyatakan bahwa Setiap Orang dengan sengaja dan tanpa hak mendistribusikan dan/atau mentransmisikan dan/atau membuat dapat diaksesnya Informasi Elektronik dan/atau Dokumen Elektronik Cyber bullying dalam UU ITE tidak terdapat unsur yang jelas. Hanya

terdapat unsur penghinaan, pencemaran nama baik, pengancaman dan pemerasan. Sedangkan jenis *cyberbullying* tidak hanya mengandung unsur penghinaan, pencemaran nama baik, pengancaman dan pemerasan saja. Pasal 27 ayat (3) dan (4) UU ITE belum menyangkut unsur dari *Flaming*, *Harassment* (gangguan), Impersonation (peniruan), *Outing* (menyebarkan rahasia orang lain), *Trickery* (tipu daya), *Exclusion* (pengeluaran), *Cyberstalking*. yang memiliki muatan pemerasan dan pengancaman (satyawati 2014).

2.8 Data Mining

Istilah data mining memiliki beberapa padanan, seperti *knowledge discovery* ataupun *pattern recognition*. Kedua istilah tersebut sebenarnya memiliki ketepatannya masing-masing. Istilah *knowledge discovery* atau penemuan pengetahuan digunakan karena tujuan utama dari data mining memang untuk mendapatkan pengetahuan yang masih tersembunyi di dalam bongkahan data. Istilah *pattern recognition* atau pengenalan pola pun tepat untuk digunakan karena pengetahuan yang hendak digali memang berbentuk pola-pola yang mungkin juga masih perlu digali dari dalam bongkahan data yang tengah dihadapai. Jadi apakah sebenarnya data mining itu? Banyak definisi bagi istilah ini dan belum ada yang dibakukan atau disepakati semua pihak. Namun demikian, istilah ini memiliki hakikat (notion) sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki (Susanto dan Suryadi, 2010).

Data mining bukan hanya pelengkap saja dalam suatu database, melainkan mempunyai fungsi yang yang penting untuk membantu penggunanya mendapatkan informasi yang berguna serta meningkatkan pengetahuan bagi sang pengguna itu sendiri dan dapat nantinya berguna untuk orang banyak. Pada dasarnya, data mining mempunyai empat fungsi dasar yaitu (Berson 2000):

1. Fungsi Klasifikasi (classification)

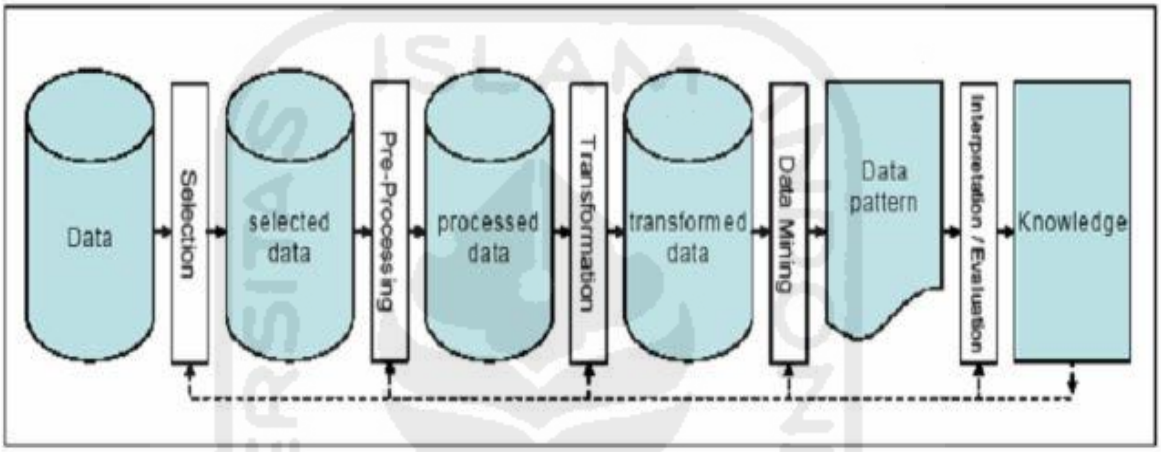
   Data mining dapat digunakan untuk mengelompokkan data-data yang jumlahnya besar menjadi data-data yang lebih kecil
2. Fungsi Segmentasi (*Segmentation*)

   Disini data mining juga digunakan untuk melakukan segementasi (pembagian) terhadap data berdasarkan karakteristik tertentu.
3. Fungsi Asosiasi (*Association*)

   Disini data mining juga digunakan untuk mencari hubungan antara karakteristik tertentu .
4. Fungsi pengurutan (*Sequencing*)

Pada Fungsi ini, data mining digunakan untuk mengidentifikasikan perubahan pola yang telah terjadi dalam jangka waktu yang tertentu.

Istilah data mining dan knowledge discovery in databases (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda akan tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining (Fayyad, 1996). Proses KDD secara garis besar dapat dijelaskan pada Gambar 2.1:



Gambar 2. 1 Fase-fase Dalam Data Mining (Fayyad, 1996)

1. Data Selection

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Pre-processing/ Cleaning

   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses Enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformation

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation/ Evaluation

   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

2.9 Text Mining

Seiring perkembangan teknologi komunikasi, manusia memungkinkan untuk saling berkomunikasi satu sama lain. Salah satu cara untuk berkomunikasi adalah melalui teks. Bidang data mining yang bertujuan untuk menambang informasi dari data teks adalah text mining.

Ungkapan text mining umumnya digunakan untuk menunjukan system yang menganalisis data yang besar yang berupa teks bahasa alami dan mendeteksi pola penggunaan leksial atau linguistik dalam upaya mengekstrak informasi yang berguna, sebastiani dalam (Sandi 2012). Sementara itu, witten (2005) menyatakan bahwa text mining adalah cabang ilmu baru yang sedang berkembang yang mencoba untuk mengumpulkan informasi yang bermanfaat dari data teks yang natural (*natural language text*). Hal ini biasanya dilakukan dengan mengekstrak informasi dari data teks yang berguna untuk tujuan tertentu. Dengan demikian text mining adalah penggalian informasi yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang diekstrak secara otomatis dari sumber data yang berupa teks yang berbeda-beda.

Proses text mining memerlukan beberapa tahapan, mengingat data teks memiliki karakteristik yang lebih komplek dari data biasa. Berdasarkan presentasi dari (Auvil 2003) yang menyatakan bahwa secara garis besar sebuah dokumen memiliki karakteristik sebagai berikut:

1. Database teks memiliki ukuran yang besar (*Large textual database*)
2. Memiliki dimensi tinggi, satu kata satu dimensi (*high dimensionaly*)

3. Mengandung frase dan antara frase satu dengan frase lainnya dapat memiliki arti yang berbeda dan saling bergantung satu sama lain (*dependency*)
4. Banyak mengandung kata/kalimat yang ambigu (*Ambiguity*)
5. Mengandung data noise, seperti singkatan. Istilah dan spelling mistake.
6. Mengandung struktur yang tidak baku misalkan singkatan-singkatan pada kata seperti "km di mn?"

Dengan demikian, proses penggalian informasi dari sekumpulan dokumen teks seperti halaman web, *twitter*, dokumen dan lain-lain membutuhkan beberapa proses yang saling terkait. Mengolah dokumen dari yang tidak terstruktur menjadi lebih terstruktur dengan menerapkan beberapa teknik ekstraksi dan penyaringan kata-kata dalam dokumen sekaligus dengan pembobotan tingkat kepenntingan kata-kata dengan metode pembobotan. Data hasil pembobotan tersebut kemudian diolah dengan menggunakan teknik data mining sesuai dengan tujuan pengolahan data tersebut. Pada bagian ini akan menjelaskan dasar-dasar teori yang digunakan dalam proses ekstraksi dan pembobotan dokumen dengan metode pembobotan tf-idf.

Bentuk pengubahan dokumen teks yang masih tidak terstruktur menjadi lebih terstruktur adalah dengan mengubahnya menjadi bentuk *spreadsheet*. Kolom menunjuk kepada dokumen dan baris menunjuk kepada kata. Sedangkan sel menunjukan frekuensi kemunculan suatu kata dalam suatu dokumen.

Agar sebuah dokumen dapat dianalisis dengan menggunakan teknik-teknik data mining, maka dilakukan proses *text transformation* dilakukan dengan mengubah data kata-kata menjadi kata-kata yang lebih baik dan sesuai kemudian dilakukan penghapusan kata-kata yang kurang penting dalam proses selanjutnya dengan proses *stopword list*. Setelah mendapatkan kata-kata unik dan bersih. Maka tahapan feature selection dilakukan dengan melakukan perhitungan jumlah kata yang muncul ataupun dengan teknik statistic lainnya, dimana dimungkinkan dilakukan proses pembobotan kata atau *term weight*. Data *term weight* inilah yang kemudian diolah dengan teknik data mining untuk menghasilkan informasi yang lebih berguna.

1. *Casefolding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Dimana hanya huruf 'a' sampai 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap sebagai delimiter (manning, 2008). Dengan demikian data teks yang akan diproses adalah data teks yang semua hurufnya adalah huruf kecil dan juga menghilangkan karakter angka dan symbol lain.
2. Tokenizing

Proses tokenisasi pada data teks adalah melakukan pemecahan sekumpulan karakter (kalimat) menjadi kata-kata (token). Setiap token adalah objek dari suatu tipe, sehingga jumlah token akan lebih banyak daripada tipenya. Contoh, jika di dalam sebuah teks terdapat kata “saya” sebanyak sepuluh buah, maka kesepuluhnya bertipe “saya”.

Hal lain yang berhubungan dengan tokenisasi adalah bagaimana cara memisahkan kata-katanya. Sebagai contoh, karakter *whitespace*. Seperti ada juga karakter-karakter lain seperti *semicolon*, titik, dan symbol-simbol lain yang juga dapat digunakan sebagai pemisah antar kata. Contohnya pada kalimat “siang, sore,malam,” dapat dipisahkan menjadi kata siang, kata sore dan kata malam.

3. Menghapus *Stopword*

   Terkadang muncul beberapa kata yang sangat umum pada semua dokumen. Kata-kata tersebut dikenal dengan istilah stop word. *Stopword* adalah kata umum (*common words*) yang biasanya muncul dalam jumlah besar di setiap dokumen dan dianggap tidak memiliki makna. Proses penghapusan *stopword* ini dilakukan untuk setiap dokumen, apabila di dalam dokumen ditemukan kata yang termasuk kedalam daftar *stopword* maka kata tersebut dihapus, sehingga dimensi dokumen menjadi berkurang.

4. *Stemming*

   Tahap *stemming* adalah tahap mencari *root* kata dari tiap kata hasil *filtering.* Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa inggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki rumus bentuk baku yang permanen.

2.10 N-Gram

Pada dasarnya, model n-gram adalah model probabilistik yang awalnya dirancang oleh ahli matematika dari Rusia pada awal abad ke-20 dan kemudian dikembangkan untuk memprediksi item berikutnya dalam urutan item. Item bisa berupa huruf / karakter, kata, atau yang lain sesuai dengan aplikasi. Salah satunya, model n-gram yang berbasis kata digunakan untuk memprediksi kata berikutnya dalam urutan kata tertentu. Dalam arti bahwa sebuah n-gram hanyalah sebuah wadah kumpulan kata dengan masing-masing memiliki panjang n kata.Sebagai contoh, sebuah n-gram ukuran 1 disebut sebagai unigram; ukuran 2 sebagai “bigram”; ukuran 3 sebagai "trigram", dan seterusnya.

Pada pembangkitan karakter, *N-gram* terdiri dari substring sepanjang n karakter dari sebuah string, dalam definisi lain n-gram adalah potongan sejumlah n karakter dari sebuah

string. Metode *N-gram* ini digunakan untuk mengambil potongan-potongan karakter huruf sejumlah n dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen. Sebagai contoh: kata ”TEXT” dapat diuraikan ke dalam beberapa n-gram berikut:

uni-gram : T, E,X, T
bi-gram : TE, EX, XT
tri-gram : TEX, EXT
quad-gram : TEXT, EXT dan seterusnya.

Sedangkan pada pembangkitan kata, metode n-gram ini digunakan untuk mengambil potongan kata sejumlah n dari sebuah rangkaian kata (kalimat, paragraf, bacaan) yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen. Sebagai contoh : kalimat ”saya dapat melihat cahaya itu. ” dapat diuraikan ke dalam beberapa n-gram berikut:

uni-gram : saya, dapat, melihat, cahaya, itu
bi-gram : saya dapat, dapat melihat, itu ada
tri-gram : saya dapat melihat, dapat melihat dia
dan seterusnya.

Salah-satu keunggulan menggunakan n-gram dan bukan suatu kata utuh secara keseluruhan adalah bahwa n-gram tidak terlalu sensitive terhadap kesalahan penulisan yang terdapat pada suatu dokumen (Hanafi 2009).

2.11 Naïve Bayes Classifier

Naïve Bayes adalah sebuah metode yang mendasarkan pada suatu asumsi penyederhanaan dimana dimana nilai atribut secara kondisional saling bebas apabila diberikan nilai output (Santosa, 2007). *Naïve Bayes Classifier* merupakan sebuah metode klasifikasi yang berakar pada teorema Bayes. Ciri utama *Naïve Bayes Classifier* adalah asumsi yang sangat kuat (naif) dari independensi masing-masing kondisi atau kejadian (Natalius, 2010). Secara sederahana NBC mengasumsikan bahwa ada atau tidaknya fitur tertentu dari suatu kelas tidak memiliki keterkaitan dengan keberadaan fitur lainnya.

Keuntungan penggunaan NBC adalah hanya diperlukan sejumlah kecil data pelatihan untuk mengestimasi parameter yaitu rata-rata dan varian dari variable yang diperlukan untuk klasifikasi (Saraswati, 2011). Naïve bayes adalah sebuah algoritma analisa statistic, yang melakukan pengolahan data terhadapat data numeric menggunakan probabilitas Bayesian. Klasifikasi-klasifikasi bayes adalah klasifikasi statistic yang dapat memprediksi kelas suati anggota probabilitas. Untuk klasifikasi bayes sederhana yang lebih dikenal sebagai Naïve Bayesian classifier dapat diasumsikan bahwa efek dari suatu nilai atribur sebuah kelas tidak

dipengaruhi nilai dari atribut lainnya. Asumsi ini disebut *class conditional independence* yang diciptakan untuk memudahkan perhitungan, pengertian ini dianggap "naïve", dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kata-kata yang lain, sehingga dalam analisis sentiment kata yang muncul memiliki bobot seluruhnya apakah kalimat tersebut positif ataupun negatif (saputra 2015).

Secara garis besar, cara kerja metode ini dapat direpresentasekan sebagai berikut:

1. Ambil probabilitas positif, dan negatif tiap kata.
2. Hitung rata-rata probabilitas ketiganya
3. Tentukan klasifikasi berdasarkan nilai probabilitas diatas.

Untuk mendapatkan probabilitas dari tiap kata, terlebih dahulu melakukan pembelajaran terhadap setiap kata dan probabilitasnya. Dalam proses pembelajaran ini, diperlukan sebuah training set, yang merupakan sekumpulan kalimat positif, negatif dan netral yang telah diklasifikasikan. Naïve bayes merupakan teknik klasifikasi yang sederhana dan cepat. Teknik ini bekerja dengan baik dengan representase statistic. Berbeda dengan metode *rule-based*, naïve Bayesian dapat belajar secara incremental. Namun kekurangan dari naïve Bayesian adalah ukuran dari vektof fitur yang dihasilkan cukup besar dan butuh teknik untuk memperkecil ukuran vector tersebut (Darujati 2012).

Untuk kemungkinan kategori bagi dokumen yang diberikan, berikut ini adalah penjelasan dari naïve bayes (Nuraini 2011):

1. Setiap data direpresentasekan sebagai vector berdimensi –n yaitu $X = (x_1, x_2, x_3 \dots x_{n)}$ adalah gambaran dari ukuran yang dibuat di test dari n atribut yaitu $A_1, A_2, A_3 \dots A_m$ dimana m adalah kumpulan kategori yaitu $C_1, C_2, C_3 \dots C_m$. Diberikan data test X yang tidak diketahui kategorinya, maka classifier akan memprediksi bahwa X adalah milik kategori dengan posterior probabilitas tertinggi berdasarkan kondisi X. oleh karena itu, Naïve Bayes Classifier menandai bahwa test X yang tidak diketahui tadi ke kategori C_1 jika dan hanya jika dilihat pada persamaan $P(C_1|X) > P(C_j|X)$ untuk $l \leq j \leq m, j \neq i$
   Kemudian kita perlu memaksimalkan P (C_i | X) berdasarkan persamaan $P(C_1|X) = \frac{P(X|C_1)\,.\,P(C_1)}{P\,(X)}$
2. Px adalah konstan untuk semua kategori, hanya $P(X\,|C_1).P(C_1)$ yang perlu dimaksimalkan. Jika prior probabilitas mungkin diperkirakan dengan perhitungan $P(C_i) = \frac{Si}{S}$ dimana Si adalah jumlah dari data training dari kategori C_i, dan S adalah jumlah total data training.

3. Diberikan data dengan banyak atribut, ini akan menjadi komputasi yang kompleks untuk mengkomputasi $P(X|C_i)$. Untuk mengurangi komputasi pada saat mengevaluasi $P(X|C_i)$, maka dapat dihitung menggunakan persamaan:

   $$P\ (X\ |C_i) = \sum_{k=1}^{n} P\ (x^k|C_i)$$

   dimana X adalah nilai-nilai atribut dalam sampel X dan probabilitas $P(X_1|C_i)$, $P(X_2|C_i)$,…….,$P(X_n\ |\ C_i)$ dapat diperkirakan dari data training.

2.12 Bag-of-words Model

Bag-of-words adalah model yang mengambil setiap kata dalam sebuah kalimat sebagai fitur, dengan asumsi bahwa setiap fitur memiliki sifat *conditional independence*. Teks direpresentasikan sebagai koleksi dari kata-kata yang tidak berurutan. Setiap fitur mewakili keberadaan satu kata. Semua kata (fitur) dalam merupakan *dictionary*. Tantangan dengan pendekatan ini adalah pilihan kata yang tepat untuk menjadi fitur.

Model *bag-of-words* tidak mampu untuk menangkap relasi antar kata. Sebagai contoh, kalimat '*this movie is excellent*' dan '*this film is great*' akan dianggap sebagai 2 kalimat yang memiliki arti yang berbeda. Bagi kita, sangat jelas kesamaan antara kata '*excellent*' dan '*great*' serta '*movie*' dan '*film*'. Hal ini dapat diatasi dengan menggunakan *semantic handling*, yakni masing-masing pasangan kata tersebut dapat diwakili oleh sebuah fitur tunggal dengan mencari kata sinonimnya, Yessenov dan Misailovic, 2009 dalam (Hilmawan 2014). Model *bag-of-words* dimodifikasi sehingga fitur-fitur yang memiliki makna yang sama hanya memiliki satu kata yang diwakili oleh *cluster* sinonimnya.

2.13 Teknik Pencarian Data

Dalam setiap waktu yang mengakses media sosial seperti Facebook dan Twitter dapat mencapai ratusan ribu bahkan jutaan pengguna, sehingga *database* media sosial mempunyai kapasitas yang sangat besar. Jika melakukan pencarian tanpa suatu teknik yang baik, akan banyak memakan waktu dan tenaga. Untuk itu, diperlukan suatu teknik pencarian yang efektif dan efisien dalam menemukan data sesuai kebutuhan (Saputra 2015).

Search Techniques pada penelitian ini menggunakan salah satu metode pencarian yang dikemukakan oleh Jackie Skinner dari University of Reading yang merupakan Universitas top 1% dunia. Search techniques tersebut memberikan petunjuk tentang opinion mining pada twitter untuk bahasa Indonesia dengan metode Naïve Bayes Classifier:

1. Menggunakan symbol untuk mencari alternative akhiran dan ejaan
2. Menggabungkan konsep dalam pencarian

3. Mencari frase
4. Melakukakan pencarian secara lebih spesifik

Metode dalam *search techniques* yang dipakai dalam penelitian ini adalah menggunakan operator pencarian. Metode ini menggabungkan kata-kata pencarian mencakup sinonim dan dikenal juga sebagai Boolean Searching. Pada metode ini memungkinkan untuk memasukkan banyak kata ataupun konsep dalam pencarian. Sudah banyak peneliti menggunakan Boolean *Searching* dalam teknik pencarian datanya, diantaranya pencarian katalog universitas, OPACs (Online Public Access Catalogs), Katalog perpustakaan online dan terdapat juga strategi pencarian positif dan negatif untuk web.

Pada operator "AND", misalkan untuk kata kunci pencarian data "bangsat dan brengsek" akan menghasilkan data yang terdapat kata baik itu kata bangsat maupun brengsek. Untuk operator "OR", misalkan untuk kata kunci pencarian data "bangsat or brengsek" akan menghasilkan data baik itu data yang berisikan bangsat saja, atau brengsek saja maupun data yang berisikan bangsat dan brengsek. Untuk operator "NOT", misalkan untuk kata kunci pencarian data "bangsat not brengsek" akan menghasilkan data yang berisikan bangsat saja.

2.14 API Twitter

API adalah singkatan dari Application Programming Interface, merupakan applikasi pendukung dari Jejaring Sosial, salah satu fungsinya yaitu untuk meng-ekstrak data dalam bentuk JSON. API adalah cara yang ditetapkan untuk sebuah program dalam menyelesaikan tugas, biasanya dalam mengambil dan memodifikasi data. Twitter menyediakan API pada hampir setiap fitur yang dapat dimanfaatkan untuk membuat aplikasi, *website, widget,* dan proyek lain yang berinteraksi dengan Twitter. Komunikasi antara aplikasi yang dibuat dengan Twitter API dilakukan melalui *Hypertext Transfer Protocol (*HTTP). Twitter API memiliki beberapa komponen yang semuanya bersifat *free* atau gratis.

Namun demikian, Twitter memberikan batasan jumlah *tweet* yang bisa didapat dalam satu jam. Beberapa bagian dari Twitter API tidak memerlukan otentikasi dasar penggunaan berupa *uesename* dan *password*, dan beberapa otentikasi menggunakan *OAuth.* Sebagian besar bagian dari API pada Twitter menggunakan model REST, kecuali untuk API streaming, yang selalu membutuhkan koneksi dengan Internet. Data yang diperoleh dari Twitter dikembalikan kedalam format XML dan JSON. Namun seiring berjalannya waktu, pengembalian data dalam bentuk XML mulai dihapus (Aliandu 2012).

Twitter Search API adalah API yang digunakan untuk menjalankan pencarian terhadap indeks *real time* dari *tweet* terakhir. Terdapat beberapa batasan dengan Twitter Search API, antara lain (Aliandu 2012):

a. Search API tidak lengkap mengindeks semua tweet tetapi indeks 1500 *tweets* terakhir.
b. Search API tidak dapat digunakan untuk menemukan *tweet* yang umurnya lebih dari seminggu.
c. Search API mengembalikan maksimal 100 *tweet* yang umurnya lebih dari seminggu.
d. Query yang terlalu kompleks mungkin tidak akan berhasil
e. Search tidak mendukung otentifikasi yang berarti semua *Query* adalah anonym.