TESIS:

KLASIFIKASI *CYBER BULLYING* PADA MEDIA SOSIAL *TWITTER* DENGAN MENGGUNAKAN ALGORITMA *NAÏVE BAYES*

ENDAH TRIHAPSARI
2214206709

DOSEN PEMBIMBING
Dr. Surya Sumpeno, S.T., M.Sc.
Dr. Adhi Dharma Wibawa, S.T., M.T.

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA-CIO
JURUSAN TEKNIK ELEKTRO
FAKULTAS TEKNIK INDUSTRI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016

THESIS

CYBER BULLYING CLASSIFICATION ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES ALGORITHM

ENDAH TRIHAPSARI
2214206709

SUPERVISOR
Dr. Surya Sumpeno, S.T., M.Sc.
Dr. Adhi Dharma Wibawa, S.T., M.T.

MASTER PROGRAM
AREAS OF EXPERTISE TELEMATIKA-CIO
ELECTRICAL ENGINEERING DEPARTMENT
FACULTY OF INDUSTRIAL TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016

LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Teknik (M.T)
di
Institut Teknologi Sepuluh Nopember
oleh:
Endah Trihapsari
NRP. 2214206709

Tanggal Ujian : 23 Juni 2016
Periode Wisuda : Wisuda ke 114

Disetujui oleh:

1. .....
Dr. Surya Sumpeno, S.T., M.Sc. (Pembimbing I)
NIP. 196906131997021003

2. .....
Dr. Adhi Dharma Wibawa, S.T., M.T. (Pembimbing II)
NIP. 197605052008121003

3. .....
Dr. Ir. Endroyono, DEA. (Penguji)
NIP. 196504041991021001

4. .....
Dr. Ir. Achmad Affandi, DEA. (Penguji)
NIP. 196510141990021001

5. .....
Eko Setijadi, S.T, M.T., Ph.D. (Penguji)
NIP. 197210012003121002

Direktur Program Pascasarjana,

Prof. Ir. Djauhar Manfaat, M.Sc, Ph.D
NIP. 196012021987011001

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH
UNTUK KEPENTINGAN AKADEMIS

Sebagai mahasiswa Institut Teknologi Sepuluh Nopember Surabaya, yang bertanda tangan di bawah ini saya :

Nama : Endah Trihapsari
NRP. : 2214206709
Jurusan / Fak. : Teknik Elektro (S2) / Bidang Keahlian Telematika - CIO
Alamat kontak
a. Email : n.trihapsari@gmail.com
b. Telp/HP : 081555647654

Menyatakan bahwa semua data yang saya *upload* di Digital Library ITS merupakan hasil final (revisi terakhir) dari karya ilmiah saya yang sudah disahkan oleh dosen penguji. Apabila di kemudian hari ditemukan ada ketidaksesuaian dengan kenyataan, maka saya bersedia menerima sanksi.

Demi perkembangan ilmu pengetahuan, saya menyetujui untuk memberikan Hak Bebas Royalti Non-Ekslusif *(Non-Exclusive Royalti-Free Right)* kepada Institut Teknologi Sepuluh Nopember Surabaya atas karya ilmiah saya yang berjudul :

KLASIFIKASI CYBER BULLYING PADA MEDIA SOSIAL TWITTER DENGAN MENGGUNAKAN ALGORITMA NAÏVE BAYES

Dengan Hak Bebas Royalti Non-Ekslusif ini, Institut Teknologi Sepuluh Nopember Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data *(database)*, mendistribusikannya, dan menampilkan / mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta. Saya bersedia menanggung secara pribadi, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya Ilmiah saya ini tanpa melibatkan pihak Institut Teknologi Sepuluh Nopember Surabaya.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Surabaya
Pada tanggal 14 Juli 2016
Yang Menyatakan,

Dosen Pembimbing 1

Dr. Surya Sumpeno, S.T., M.Sc.
NIP. 196906131997021003

Endah Trihapsari
NRP. 2214206709

KLASIFIKASI *CYBER BULLYING* PADA MEDIA SOSIAL *TWITTER* DENGAN MENGGUNAKAN ALGORITMA *NAÏVE BAYES*

Nama Mahasiswa	: Endah Trihapsari
NRP	: 2214206709
Pembimbing 1	: Dr. Surya Sumpeno, S.T., M.Sc.
Pembimbing 2	: Dr. Adhi Dharma Wibawa, S.T., M.T.

ABSTRAK

Jumlah pengguna Internet di Indonesia selalu meningkat dengan pesat setiap tahunnya, dan pada tahun 2014 mencapai 34,9% dari jumlah populasi penduduk, dengan rata-rata waktu akses 5,5 jam setiap harinya. Peningkatan penggunaan Internet ternyata juga diiringi dengan fakta lainnya, yaitu adanya peningkatan kasus *cyberbullying* di Indonesia.

Sebagian besar kasus *cyberbullying* terjadi di media sosial, salah satunya adalah melalui *Twitter*. Pengguna *Twitter* di Indonesia masuk dalam tiga besar dunia dan penduduk kota Jakarta menenempati urutan pertama di dunia dalam hal jumlah posting yang dikirim melalui *Twitter*. Melihat data tersebut, potensi terjadinya *cyberbullying* di Indonesia sangat besar sehingga perlu dilakukan upaya untuk mencegahnya. Upaya pencegahan ini perlu dilakukan karena *cyberbullying* dapat terjadi lebih cepat, berdampak lebih luas, dan berpotensi menjadi ancaman yang lebih besar jika dibandingkan dengan *bully* yang dilakukan secara langsung.

Langkah awal yang dilakukan adalah melakukan identifikasi kata dan pola kalimat yang paling berpotensi digunakan untuk melakukan *cyberbullying*, dimana data tersebut nantinya dapat digunakan oleh pengembang perangkat lunak untuk membuat sebuah sistem peringatan dini adanya *cyberbullying*. Dengan menggunakan Algoritma *Naive Baye*s diperoleh akurasi terbaik sebesar 87,67% dalam penentuan kategori "*bully*" atau "bukan *bully*"

Berdasarkan hasil penelitian, diketahui bahwa kata kunci "tolol" merupakan kata dengan bobot paling tinggi yang paling sering digunakan untuk melakukan *bully*. Selain menggunakan kata-kata kunci *bullying*, pesan dengan konten *bullying* terbentuk pula dengan tambahan subyek dalam kalimat tersebut. Subyek dengan bobot paling tinggi adalah kata "lo". Dengan demikian, jika subyek ditambahkan dengan kata-kata kunci *bullying*, akan membentuk kalimat pesan *bully*, dirumuskan dengan: "subyek + kata kunci *bully*", contohnya "lo" + "tolol".

Kata kunci: *Cyberbullying*, Klasifikasi, Naïve Bayes, Text mining

CYBER BULLYING CLASSIFICATION ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES ALGORITHM

By	: Endah Trihapsari
Student Identity Number	: 2214206709
Supervisor	: Dr. Surya Sumpeno, S.T., M.Sc.
Co-Supervisor	: Dr. Adhi Dharma Wibawa, S.T., M.T.

ABSTRACT

Number of Internet users in Indonesia is increasing rapidly every year, and in 2014 reached 34.9 % of the total population, with an access time average of 5.5 hours per day. The increase of Internet use was also accompanied by other facts, namely the increase of cyberbullying case in Indonesia.

Most of cyberbullying cases occurs in social media, one of which is through Twitter. Twitter users in Indonesia are in the top three of the world and people in Jakarta was ranked first in the world in terms of posting number sent via Twitter. Seeing these data, the potention of cyberbullying in Indonesia is very large, so it is necessary to prevent it. These prevention need to be done because cyberbullying can happen faster, wider impact, and potentially a greater threat than the traditional bullying.

The first step is identifying words and sentence patterns most potentially be used for cyberbullying, where the data can be used by software developers to create an early warning system cyberbullying. By using a Naive Bayes algorithm obtained the best accuracy of 87.67 % in the determination of the category of "bully" or "no bully".

Based on this research, it is known that the key word "tolol" is a word with the highest weighting that most often used to bully. Besides using bullying key words, message with bullying content also formed with additional subjects in the sentence. Subjects with the highest weighting is the word "lo". Thus, if the subject was added with the bullying key words will be formed bullying sentences, formulated with : "subject + bully key word", for example "lo" + "tolol".

Key words: Classifications, Cyberbullying, Naïve Bayes, Text mining

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Penetrasi pengguna internet di Indonesia meningkat cukup signifikan dari tahun ke tahun. Data Asosiasi Pengguna Jasa Internet Indonesia (APJII) menunjukkan bahwa pada tahun 2014 pengguna internet di Indonesia telah mencapai 88,1 juta jiwa atau 34,9% dari jumlah populasi. Dari jumlah tersebut 84% nya minimal mengakses internet sekali sehari. Waktu yang dihabiskan untuk berselancar di internet rata-rata per hari adalah selama 5,5 jam. Data tersebut tentu saja sangat menggembirakan ditinjau dari sudut peningkatan jumlah pengguna dan lamanya waktu yang dihabiskan masyarakat untuk menggunakan internet.

Dibalik peningkatan penetrasi pengguna internet yang cukup menggembirakan, ternyata ada fakta bahwa kasus *cyberbullying* meningkat cukup tinggi juga. Dari 40 negara yang pernah di survey antara tahun 2005-2006, Indonesia menempati urutan ketiga dalam hal kasus *cyberbullying* setelah Jepang dan Korea Selatan (Kaman, 2007). Survey lain yang pernah dilakukan oleh Ipsos terhadap 200.000 anak usia sekolah di 40 negara, Indonesia menjadi salah satu negara yang mempunyai prosentase tinggi terjadinya kasus *cyberbullying*. Hampir 91% lebih warga Indonesia menyatakan bahwa anak mereka mengalami kasus *cyberbullying* di media sosial (Gottfried, 2012).

Data lain yang cukup menarik adalah, berdasarkan penelitian yang dilakukan Kementerian Komunikasi dan Informatika bekerjasama dengan Unicef pada tahun 2011 hingga 2013, sebagian besar remaja di Indonesia telah menjadi korban *cyberbullying*. Studi melibatkan 400 anak dan remaja di 11 provinsi dengan rentang usia 10 hingga 19 tahun. Sebagian besar dari mereka menerima *bullying* melalui media sosial khususnya *Facebook* dan *Twitter*. Yang mengkhawatirkan adalah persentase yang relatif tinggi anak-anak yang menjadi korban cyberbullying, hanya 42 % responden menyadari risiko ditindas secara online, dan di antara mereka 13 % telah menjadi korban selama tiga bulan sebelumnya, yang diterjemahkan ke dalam ribuan anak-anak.

Melihat data-data yang telah ada, perlu adanya upaya nyata dan segera untuk mencegah semakin meluasnya praktik-praktik *cyberbullying* di Indonesia terutama bagi anak-anak yang paling rentan menjadi korban. Dalam tesis ini akan dibahas mengenai klasifikasi *microblogging Twitter*, apakah sebuah *tweet* mengandung unsur *bullying* atau tidak dengan menggunakan algoritma *Naïve Bayes.* Selanjutnya diharapkan akan diketahui juga kata maupun gabungan kata yang paling berpotensi untuk digunakan melakukan *cyberbullying*. Data ini nantinya dapat digunakan oleh pengembang untuk membangun sebuah perangkat lunak yang dapat memprediksi apakah pesan yang dikirim mengandung unsur bullying atau tidak sebagai sistem peringatan dini. Diharapkan dengan adanya perangkat lunak tersebut dapat menurunkan potensi terjadinya cyberbullying di masyarakat.

1.2 Masalah

Permasalahan yang akan diangkat dalam tesis ini adalah, adanya kesulitan untuk menentukan apakah suatu pesan yang dikirim melalui media sosial khususnya *Twitter*, mengandung unsur *bullying* atau tidak.

1.3 Tujuan

Tujuan yang akan dicapai dalam tesis ini adalah, menemukan kata atau gabungan kata yang paling berpotensi digunakan untuk melakukan *cyberbullying* pada media sosial khususnya Twitter, sehingga mempermudah pengembang untuk membuat sistem atau perangkat lunak yang dapat melakukan deteksi dini terhadap terjadinya *cyberbullying* di media sosial.

1.4 Manfaat

Manfaat yang akan diperoleh dalam tesis ini adalah, tersedianya data tentang kata atau gabungan kata yang paling berpotensi digunakan untuk melakukan *bullying* di media sosial khususnya *Twitter*. Selanjutnya data tersebut dapat dimanfaatkan oleh pengembang perangkat lunak untuk membuat sebuah aplikasi yang mampu melakukan deteksi dini terhadap terjadinya *cyberbullying*. Dengan tersedianya aplikasi tersebut diharapkan akan

mengurangi potensi terjadinya bullying di media sosial pada umumnya dan Twitter khususnya.

1.5 Metode

Secara garis besar metode yang akan digunakan untuk menyelesaikan tesis ini dibagi menjadi beberapa tahap seperti di bawah ini:

1. Menentukan kata-kata yang berpotensi menimbulkan *bullying*.
2. Melakukan pengumpulan data pada media sosial *Twitter* dengan menggunakan *streaming APIs* berdasarkan kata kunci yang telah ditentukan sebelumnya dalam kurun waktu tertentu dan menyimpannya dalam database.
3. Melakukan preprocessing terhadap data yang sudah tersimpan di database tersebut, meliputi cleansing, case folding dan parsing.
4. Melakukan pemilihan fitur (*feature selection*) terhadap data Twitter yang telah mengalami pra proses dengan melakukan *stopword removal* dan melakukan proses *stemming*
5. Melakukan pembelajaran dan klasifikasi *cyberbullying* terhadap data yang telah melalui tahap pra proses dengan menggunakan algoritma *naïve bayes*.
6. Melakukan validasi dan evaluasi terhadap hasil pembelajaran yang telah dilakukan.
7. Menentukan kata dan gabungan kata apa yang mempunyai potensi paling besar digunakan untuk melakukan *cyberbullying*, khususnya di media sosial *Twitter*.

(Halaman ini sengaja dikosongkan)

BAB 2
KAJIAN PUSTAKA

2.1 Bullying

Kata “bully” dikenal sejak tahun 1530-an. Pada dasarnya bullying melibatkan dua orang, pelaku bullying dan korban. Pelaku mem-bully korban secara fisik, lisan, atau cara lain untuk mendapatkan rasa kekuasaan. Tindakan ini mungkin langsung (memukul, mencela, dan lain-lain) atau secara tidak langsung (gossip, rumors, fitnah, dan lain-lain). (JurnalBullying and Cyberbullying: History, Statistics, Law, Prevention, dan Analysis).

Bullying dapat didefinisikan sebagai bentuk-bentuk perilaku kekerasan dimana terjadi pemaksaan secara psikologis ataupun fisik terhadap seseorang atau sekelompok orang yang lebih “lemah” oleh seseorang atau sekelompok orang. Pelaku bullying yang biasa disebut bully bisa seseorang, bisa juga sekelompok orang, dan ia atau mereka mempersepsikan dirinya memiliki power (kekuasaan) untuk melakukan apa saja terhadap korbannya. Korban juga mempersepsikan dirinya sebagai pihak yang lemah, tidak berdaya dan selalu merasa terancam oleh bully. Pengertian tersebut didukung oleh Coloroso (2006, 44-45) yang mengemukakan bahwa bullying akan selalu melibatkan ketiga unsur berikut :

1. Ketidakseimbangan kekuatan (imbalance power).

   Bullying bukan persaingan antara saudara kandung, bukan pula perkelahian yang melibatkan dua pihak yang setara. Pelaku bullying bisa saja orang yang lebih tua, lebih besar, lebih kuat, lebih mahir secara verbal, lebih tinggi secara status sosial, atau berasal dari ras yang berbeda
2. Keinginan untuk mencederai (desire to hurt).

   Dalam bullying tidak ada kecelakaan atau kekeliruan, tidak ada ketidaksengajaan dalam pengucilan korban. Bullying berarti menyebabkan kepedihan emosional atau luka fisik, melibatkan

tindakan yang dapat melukai, dan menimbulkan rasa senang di hati sang pelaku saat menyaksikan penderitaan korbannya.

3. Ancaman agresi lebih lanjut.

   Bullying tidak dimaksudkan sebagai peristiwa yang hanya terjadi sekali saja, tapi juga repetitifatau cenderung diulangi.

4. Teror.

   Bullying adalah kekerasan sistematik yang digunakan untuk mengintimidasi dan memelihara dominasi. Teror bukan hanya sebuah cara untuk mencapai bullying tapi juga sebagai tujuan bullying.

Berdasarkan jenisnya, bullying dapat dibagi menjadi beberapa kategori sebagai berikut:

1. Fisik.

   Bullying di kategori fisik pada dasarnya melibatkan penggunaan kekuatan fisik sehingga menjadi aksi bullying yang paling mudah diidentifikasi. Mendorong, menendang, meninju, dan menampar adalah beberapa contoh aksi dari jenis bullying ini. Tujuan dari perilaku ini untuk dapat seterusnya mengontrol kehidupan korban, misalnya agar korban menuruti apa keinginan pelaku, seperti mengerjakan tugas atau perintah yang tidak masuk akal.

2. Verbal.

   Bullying verbal adalah bentuk bullying lewat lisan atau tulisan, bertujuan mengintimidasi korban melalui ejekan, hinaan, fitnah, sampai ancaman.

3. Emosional.

   Pada jenis bullying emosional, pelaku langsung menyerang korban pada tingkat emosional, pada jenis bullying ini pelaku bertujuan untuk melemahkan harga diri korban, misalnya seperti cibiran, tawa mengejek, helaan napas, pandangan yang agresif, dan bahasa tubuh yang mengejek.Bullying dalam bentuk emosional cenderung perilaku bullying yang paling sulit dideteksi dari luar dan sering kali tidak disadari oleh pelaku.

4. Cyberbullying.

   Cyberbullying adalah jenis bullying yang paling sering terjadi di era teknologi seperti saat ini.Cyberbullying adalah sebagai bentuk intimidasi yang menggunakan teknologi.Situs jejaring sosial semakin digemari belakangan ini oleh kalangan remaja, semakin banyak pula terjadi kasus cyberbullying.

2.2 Cyberbullying

Cyberbullying adalah segala bentuk kekerasan yang dialami anak atau remaja dan dilakukan teman seusia mereka melalui dunia *cybe*r atau internet. *Cyberbullying* adalah kejadian manakala seorang anak atau remaja diejek, dihina, diintimidasi, atau dipermalukan oleh anak atau remaja lain melalui media internet, teknologi digital atau telepon seluler.

Bentuk dan metode tindakan *cyberbullying* amat beragam. Bisa berupa pesan ancaman melalui e-mail, mengunggah foto yang mempermalukan korban, membuat situs web untuk menyebar fitnah dan mengolok-olok korban hingga mengakses akun jejaring sosial orang lain untuk mengancam korban dan membuat masalah. Motivasi pelakunya juga beragam. Ada yang melakukannya karena marah dan ingin balas dendam, frustrasi, ingin mencari perhatian bahkan ada pula yang menjadikannya sekedar hiburan pengisi waktu luang. Tidak jarang, motivasinya kadang-kadang hanya ingin bercanda.

Cyberbullying yang berkepanjangan bisa mematikan rasa percaya diri anak, membuat anak menjadi murung, khawatir, selalu merasa bersalah atau gagal karena tidak mampu mengatasi sendiri gangguan yang menimpanya. Bahkan ada pula korban *cyberbullying* yang berpikir untuk mengakhiri hidupnya karena tak tahan lagi diganggu. Remaja korban *cyber bullying* akan mengalami stress yang bisa memicunya melakukan tindakan-tindakan rawan masalah seperti mencontek, membolos, lari dari rumah, dan bahkan minum minuman keras atau menggunakan narkoba.

Anak-anak atau remaja pelaku cyber bullying biasanya memilih untuk menganggu anak lain yang dianggap lebih lemah, tak suka melawan

dan tak bisa membela diri. Pelakunya sendiri biasanya adalah anak-anak yang ingin berkuasa atau senang mendominasi. Anak-anak ini biasanya merasa lebih hebat, berstatus sosial lebih tinggi dan lebih populer di kalangan teman-teman sebayanya. Sedangkan korbannya biasanya anak-anak atau remaja yang sering diejek dan dipermalukan karena penampilan mereka, warna kulit, keluarga mereka, atau cara mereka bertingkah laku di sekolah. Namun bisa juga si korban *cyberbullying* justru adalah anak yang populer, pintar, dan menonjol di sekolah sehingga membuat iri teman sebayanya yang menjadi pelaku.

Cyberbullying pada umumnya dilakukan melalui media situs jejaring sosial seperti *Facebook* dan *Twitter*. Ada kalanya dilakukan juga melalui SMS maupun pesan percakapan di layanan Instant Messaging seperti Yahoo Messenger atau MSN Messenger. Anak-anak yang penguasaan komputer serta internetnya lebih bagus melakukan *cyber bullying* dengan cara lain. Mereka membuat situs atau blog untuk menjelek-jelekkan korban atau membuat masalah dengan orang lain dengan berpura-pura menjadi korban. Ada pula pelaku yang mencuri *password* akun *e-mail* atau situs jejaring sosial korban dan mengirim pesan-pesan mengancam atau tak senonoh menggunakan akun milik korban.

Cyberbullying lebih mudah dilakukan daripada kekerasan konvensional karena si pelaku tidak perlu berhadapan muka dengan orang lain yang menjadi targetnya. Mereka bisa mengatakan hal-hal yang buruk dan dengan mudah mengintimidasi korbannya karena mereka berada di belakang layar komputer atau menatap layar telepon seluler tanpa harus melihat akibat yang ditimbulkan pada diri korban. Peristiwa *cyberbullying* juga tidak mudah diidentifikasikan orang lain, seperti orang tua atau guru karena tidak jarang anak-anak remaja ini juga mempunyai kode-kode berupa singkatan kata atau emoticon internet yang tidak dapat dimengerti selain oleh mereka sendiri. Harus diwaspadai bahwa kasus *cyberbullying* ini seperti gunung es. Korban sendiri lebih sering malas mengaku. Ini karena bila mereka mengaku biasanya akses mereka akan internet (maupun HP) akan dibatasi. Korban juga terkadang malas mengaku karena sulitnya

mencari pelaku *cyberbullying* atau membuktikan bahwa si pelaku benar-benar bersalah. Ini menyebabkan munculnya kondisi gunung es tadi. Tujuannya adalah untuk mengganggu, mengancam, mempermalukan, menghina, mengucilkan secara sosial, atau merusak reputasi orang lain.

Ketika seseorang melakukan atau menjadi korban *cyberbullying*, seringkali tidak menyadari karena beragamnya tindakan yang dikategorikan sebagai *cyberbullying*. Adapun Jenis-jenis *cyberbullying* (willard, 2007) adalah sebagai berikut:

1. Flaming (kata kasar): yaitu mengirimkan pesan teks yang isinya merupakan kata-kata yang penuh amarah dan frontal. Istilah "flame" ini pun merujuk pada kata-kata di pesan yang berapi-api.
2. Harassment (gangguan): pesan-pesan yang berisi gangguan pada email, sms, maupun pesan teks di jejaring sosial dilakukan secara terus menerus.
3. Denigration (pencemaran nama baik): yaitu proses mengumbar keburukan seseorang di internet dengan maksud merusak reputasi dan nama baik orang tersebut.
4. Impersonation (peniruan): berpura-pura menjadi orang lain dan mengirimkan pesan-pesan atau status yang tidak baik
5. Outing: menyebarkan rahasia orang lain, atau foto-foto pribadi orang lain
6. Trickery (tipu daya): membujuk seseorang dengan tipu daya agar mendapatkan rahasia atau foto pribadi orang tersebut
7. Exclusion (pengeluaran) : secara sengaja dan kejam mengeluarkan seseorang dari grup online.
8. Cyberstalking: mengganggu dan mencemarkan nama baik seseorang secara intens sehingga membuat ketakutan besar pada orang tersebut.

2.3 Text Mining

Text mining (penambangan teks) adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang

tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda (Feldman & Sanger, 2007). Text mining merupakan teknik yang digunakan untuk menangani masalah klasifikasi, clustering, information extraction dan information retrival (Berry & Kogan, 2010).

Pada dasarnya proses kerja dari text mining banyak mengapdopsi dari penelitian Data Mining namun yang menjadi perbedaan adalah pola yang digunakan oleh text mining diambil dari sekumpulan bahasa alami yang tidak terstruktur sedangkan dalam Data Mining pola yang diambil dari database yang terstruktur (Han & Kamber, 2006). Tahap-tahap text mining secara umum adalah text preprocessing dan feature selection (Feldman & Sanger 2007, Berry & Kogan 2010).

Proses text mining yang khas meliputi kategorisasi teks, text clustering, ekstraksi konsep/entitas, produksi taksonomi granular, sentiment analysis, penyimpulan dokumen, dan pemodelan relasi entitas (yaitu, pembelajaran hubungan antara entitas bernama). Pendekatan manual text mining secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah memungkinkan ranah tersebut untuk berkembang selama dekade terakhir. Text mining adalah bidang interdisipliner yang mengacu pada pencarian informasi, pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistik. Dikarenakan kebanyakaninformasi (perkiraan umum mengatakan lebih dari 80%) saat ini disimpan sebagai teks, text mining diyakini memiliki potensi nilai komersial tinggi (Bridge, 2011).

Menurut Saraswati (2011), saat ini text mining telah mendapat perhatian dalam berbagai bidang diantaranya :

1. Aplikasi keamanan

   Banyak paket perangkat lunak text mining dipasarkan terhadap aplikasi keamanan, khususnya analisis plain text seperti berita internet. Hal ini juga mencakup studi enkripsi teks.

2. Aplikasi biomedis

   Berbagai aplikasi text mining dalam literatur biomedis telah disusun. Salah satu contohnya adalah PubGene yang mengkombinasikan text mining biomedis dengan visualisasi jaringan sebagai sebuah layanan Internet. Contoh lain text mining adalah GoPubMed.org. Kesamaan semantik juga telah digunakan oleh sistem text mining, yaitu, GOAnnotator.

3. Perangkat Lunak dan Aplikasi

   Departemen riset dan pengembangan perusahaan besar, termasuk IBM dan Microsoft, sedang meneliti teknik text mining dan mengembangkan program untuk lebih mengotomatisasi proses pertambangan dan analisis. Perangkat lunak text mining juga sedang diteliti oleh perusahaan yang berbeda yang bekerja di bidang pencarian dan pengindeksan secara umum sebagai cara untuk meningkatkan performansinya.

4. Aplikasi Media Online

   Text mining sedang digunakan oleh perusahaan media besar, seperti perusahaan Tribune, untuk menghilangkan ambigu informasi dan untuk memberikan pembaca dengan pengalaman pencarian yang lebih baik, yang meningkatkan loyalitas pada site dan pendapatan. Selain itu, editor diuntungkan dengan mampu berbagi, mengasosiasi dan properti paket berita, secara signifikan meningkatkan peluang untuk menguangkan konten.

5. Aplikasi Pemasaran

   Text mining juga mulai digunakan dalam pemasaran, lebih spesifik dalam analisis manajemen hubungan pelanggan. Coussement dan Poel (2008) menerapkannya untuk meningkatkan model analisis prediksi untuk churn pelanggan (pengurangan pelanggan).

6. Sentiment Analysis

   Sentiment Analysis mungkin melibatkan analisis dari review film untuk memperkirakan berapa baik review untuk sebuah film. Analisis semacam ini

mungkin memerlukan kumpulan data berlabel atau label dari efektifitas katakata. Sebuah sumber daya untuk efektivitas kata-kata telah dibuat untuk WordNet.

7. Aplikasi Akademik

   Masalah text mining penting bagi penerbit yang memiliki database besar untuk mendapatkan informasi yang memerlukan pengindeksan untuk pencarian. Hal ini terutama berlaku dalam ilmu sains, di mana informasi yang sangat spesifik sering terkandung dalam teks tertulis. Oleh karena itu, inisiatif telah diambil seperti Nature's proposal untuk Open Text Mining Interface (OTMI) dan Health's common Journal Publishing untuk Document Type Definition (DTD) yang akan memberikan isyarat semantik pada mesin untuk menjawab pertanyaan spesifik yang terkandung dalam teks tanpa menghilangkan barrier penerbit untuk akses publik.

Sebelumnya, website paling sering menggunakan pencarian berbasis teks, yang hanya menemukan dokumen yang berisi kata-kata atau frase spesifik yang ditentukan oleh pengguna. Sekarang, melalui penggunaan web semantik, text mining dapat menemukan konten berdasarkan makna dan konteks (daripada hanya dengan kata tertentu). Text mining juga digunakan dalam beberapa filter email spam sebagai cara untuk menentukan karakteristik pesan yang mungkin berupa iklan atau materi yang tidak diinginkan lainnya.

Dengan text mining tugas-tugas yang berhubungan dengan penganalisaan teks dengan jumlah yang besar, penemuan pola serta penggalian informasi yang mungkin berguna dari suatu teks dapat dilakukan. Sebagai bentuk aplikasi dari text mining, sistem klasifikasi berita menggunakan berita sebagai sumber informasi dan informasi klasifikasi sebagai informasi yang akan diekstrak dari sumber informasi. Informasi klasifikasi dapat berbentuk angkaangka probabilitas, set aturan atau bentuk lainnya. Walaupun inti dari suatu sistem klasifikasi adalah tahap penemuan pola (pattern discovery) namun secara lengkap proses text mining dibagi menjadi 3 tahap utama, yaitu proses awal terhadap teks (text preprocessing),

transformasi teks ke dalam bentuk antara (*text transformation/feature generation*), dan penemuan pola (*pattern discovery*). (Even dan Zohar, 2002). Masukan awal dari proses ini adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai hasil interpretasi.

2.4 Text Preprocessing

Tahap text preprocessing adalah tahap awal dari text mining. Tahap ini mencakup semua rutinitas, dan proses untuk mempersiapkan data yang akan digunakan pada operasi knowledge discovery sistem text mining (Feldman & Sanger, 2007). Tindakan yang dilakukan pada tahap ini adalah *toLowerCase*, yaitu mengubah semua karakter huruf menjadi huruf kecil dan *Tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter- delimiter seperti tanda titik (.), koma (,), spasi dan karakter angka yang ada pada kata tersebut (Weiss et al, 2005).

2.5 Feature Selection

Tahap seleksi fitur (*feature selection*) bertujuan untuk mengurangi dimensi dari suatu kumpulan teks, atau dengan kata lain menghapus kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dokumen sehingga proses pengklasifikasian lebih efektif dan akurat (Do et al, 2006., Feldman & Sanger, 2007., Berry & Kogan 2010). Pada tahap ini tindakan yang dilakukan adalah menghilangkan *stopword* (*stopword removal*) dan *stemming* terhadap kata yang berimbuhan (Berry & Kogan 2010., Feldman & Sanger 2007).

Stopword adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen (Dragut et al. 2009). Misalnya “di”, “oleh”, “pada”, “sebuah”, “karena” dan lain sebagainya. Sebelum proses *stopword removal* dilakukan, harus dibuat daftar stopword (stoplist). Jika termasuk di dalam *stoplist* maka kata-kata tersebut akan dihapus dari deskripsi sehingga kata-kata yang tersisa di dalam deskripsi dianggap sebagai kata-kata yang

mencirikan isi dari suatu dokumen atau keywords. Daftar kata stopword di penelitian ini bersumber dari Tala (2003).

Setelah melalui proses *stopword removal* tindakan selanjutnya adalah yaitu proses *stemming*. *Stemming* adalah proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya (stem) (Tala, 2003). Tujuan dari proses *stemming* adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata. Jika imbuhan tersebut tidak dihilangkan maka setiap satu kata dasar akan disimpan dengan berbagai macam bentuk yang berbeda sesuai dengan imbuhan yang melekatinya sehingga hal tersebut akan menambah beban database. Hal ini sangat berbeda jika menghilangkan imbuhan-imbuhan yang melekat dari setiap kata dasar, maka satu kata dasar akan disimpan sekali walaupun mungkin kata dasar tersebut pada sumber data sudah berubah dari bentuk aslinya dan mendapatkan berbagai macam imbuhan. Karena bahasa Indonesia mempunyai aturan morfologi maka proses stemming harus berdasarkan aturan morfologi bahasa Indonesia.

Berdasarkan penelitian sebelumnya, ada beberapa algoritma stemming yang bisa digunakan untuk stemming bahasa Indonesia diantaranya algoritma confix- stripping, algoritma Porter stemmer bahasa Indonesia, algoritma Arifin dan Sutiono, dan Algorima Idris (Tala 2003, Agusta 2009, Asian et al 2005, Adriani et al 2007). Dimana, Algoritma confix-stripping adalah algoritma yang akurat dalam stemming bahasa Indonesia (Tala 2003, Agusta 2009, Asian et al 2005, Adriani et al 2007).

2.6 Stop Word

Kebanyakan bahasa resmi di berbagai negara memiliki kata fungsi dan kata sambung seperti artikel dan preposisi yang hampir selalu muncul pada dokumen teks. Biasanya kata-kata ini tidak memiliki arti yang lebih di dalam memenuhi kebutuhan seorang searcher di dalam mencari informasi. Kata-kata tersebut (misalnya a, an, the on pada bahasa Inggris) disebut sebagai *stopwords*.

Sebuah sistem *text retrieval* biasanya disertai dengan sebuah *stoplist. Stoplist* berisi sekumpulan kata yang 'tidak relevan', namun sering sekali muncul dalam sebuah dokumen. Dengan kata lain Stoplist berisi sekumpulan Stopwords.

Stopwords removal adalah sebuah proses untuk menghilangkan kata yang 'tidak relevan' pada hasil parsing sebuah dokumen teks dengan cara membandingkannya dengan *stoplist* yang ada.

Word	Root	Part of Speech	Word	Root	Part of Speech
ada	ada	verb	lah	lah	particle
adanya	ada	noun	lain	lain	adjective
adalah	adalah	verb	lainnya	lain	adjective
adapun	adapun	particle	melainkan	lain	verb
agak	agak	adverb	selaku	laku	particle
agaknya	agak	adverb	lalu	lalu	verb
agar	agar	particle	melalui	lalu	verb
akan	akan	particle	terlalu	lalu	adverb
akankah	akan	particle	lama	lama	adjective
akhirnya	akhir	noun	lamanya	lama	noun
aku	aku	pronomia	selama	lama	noun
akulah	aku	pronomia	selama-lamanya	lama	adjective
amat	amat	adverb	selamanya	lama	adjective
amatlah	amat	adverb	lebih	lebih	adjective
anda	anda	noun	terlebih	lebih	adverb
andalah	anda	noun	bermacam	macam	adjective
antar	antar	particle	bermacam-macam	macam	adjective
diantaranya	antar	verb	macam	macam	noun
antara	antara	noun	semacam	macam	adverb
antaranya	antara	particle	maka	maka	particle
diantara	antara	verb	makanya	maka	particle
apa	apa	pronomia	makin	makin	adverb
apaan	apa	pronomia	malah	malah	adverb
mengapa	apa	pronomia	malahan	malah	adverb
apabila	apabila	particle	mampu	mampu	adjective
apakah	apakah	pronomia	mampukah	mampu	adjective
apalagi	apalagi	pronomia	mana	mana	pronoun
apatah	apatah	pronomia	manakala	manakala	particle
atau	atau	particle	manalagi	manalagi	particle
ataukah	atau	particle	masih	masih	adverb
ataupun	atau	particle	masihkah	masih	adverb
bagai	bagai	noun	semasih	masih	adverb
bagaikan	bagai	particle	masing	masing	pronomia

continue to next page

Gambar 2.1 Contoh *Stoplis*t dalam Bahasa Indonesia (Tala, 2003)

2.7 Kata dalam Bahasa Indonesia

Kamus Besar Bahasa Indonesia (KBBI) (1997) memberikan beberapa definisi mengenai kata sebagai berikut:

1. Elemen terkecil dalam sebuah bahasa yang diucapkan atau dituliskan dan merupakan realisasi kesatuan perasaan dan pikiran yang dapat digunakan dalam berbahasa
2. Konversasi, bahasa
3. Morfem atau kombinasi beberapa morfem yang dapat diujarkan sebagai bentuk yang bebas
4. Unit bahasa yang dapat berdiri sendiri dan terdiri dari satu morfem (contoh kata) atau beberapa morfem gabungan (contoh perkataan)

Berdasarkan beberapa definisi tersebut, dapat dijelaskan bahwa kata adalah suatu unit dari suatu bahasa yang mengandung arti dan terdiri dari satu atau lebih morfem. Kata adalah merupakan bahasa terkecil yang dapat berdiri sendiri. Umumnya kata terdiri dari satu akar kata tanpa atau dengan beberapa afiks. Gabungan kata-kata dapat membentuk frasa, klausa, atau kalimat.

Berdasarkan bentuknya, kata dapat digolongkan menjadi empat yaitu:

1. Kata dasar

   Kata dasar adalah kata yang belum diberi imbuhan. Dengan kata lain, kata dasar adalah kata yang menjadi dasar awal pembentukan kata yang lebih besar. Contoh kata dasar antara lain: bangun, datang, pergi, tinggal, pulang.
2. Kata turunan

   Kata turunan atau disebut dengan kata berimbuhan adalah kata – kata yang telah berubah bentuk dan makna. Perubahan ini dikarenakan kata-kata tersebut telah diberi imbuhan yang berupa awalan (prefiks), akhiran (sufiks), sisipan (infiks), dan awalan-akhiran (konfiks). Contoh yang termasuk kata turunan antara lain: berlari, catatan, gemetar, membersihkan.

3. Kata ulang

   Kata ulang adalah bentuk kata yang merupakan pengulangan kata dasar. Pengulangan ini dapat memiliki atau menciptakan arti baru. Kata ulang terdiri atas dua macam bentuk yaitu pengulangan seluruh dan pengulangan sebagian. Contoh kata ulang antara lain: anak-anak, buku-buku, tetangga.

4. Kata majemuk / kata gabung

   Kata majemuk adalah bentuk kata yang terdiri dari dua kata yang berhubungan secara padu dan membentuk arti atau makna baru. Kata majemuk tidak bisa dipisahkan karena akan kehilangan maknanya. Cara-cara penulisan kata majemuk bisa dilakukan sebagai berikut:

   a. Unsur-unsurnya ditulis secara terpisah untuk menyebutkan suatu istilah khusus, contoh: rumah sakit, kereta api, duta besar.
   b. Unsur-unsurnya ditulis menggunakan tanda hubung untuk mengaskan pertalian unsur yang bersangkutan, contoh: anak-istri, ibu-bapak, simpan-pinjam.
   c. Unsur-unsurnya ditulis secara serangkai, contoh: beasiswa, dukacita, saputangan.

2.8 Kelas Kata

Dalam tata bahasa baku bahasa Indonesia, kelas kata terbagi menjadi tujuh kategori, yaitu:

1. Nomina (kata benda); nama dari seseorang, tempat, atau semua benda dan segala yang dibendakan, misalnya buku, kuda.
2. Verba (kata kerja); kata yang menyatakan suatu tindakan atau pengertian dinamis, misalnya baca, lari.
   a. Verba transitif (membunuh),
   b. Verba kerja intransitif (meninggal),
   c. Pelengkap (berumah)
3. Adjektiva (kata sifat); kata yang menjelaskan kata benda, misalnya keras, cepat.

4. Adverbia (kata keterangan); kata yang memberikan keterangan pada kata yang bukan kata benda, misalnya sekarang, agak.
5. Pronomina (kata ganti); kata pengganti kata benda, misalnya ia, itu.
    a. Orang pertama (kami),
    b. Orang kedua (engkau),
    c. Orang ketiga (mereka),
    d. Kata ganti kepunyaan (-nya),
    e. Kata ganti penunjuk (ini, itu)
6. Numeralia (kata bilangan); kata yang menyatakan jumlah benda atau hal atau menunjukkan urutannya dalam suatu deretan, misalnya satu, kedua.
    a. Angka kardinal (duabelas),
    b. Angka ordinal (keduabelas)
7. Kata tugas adalah jenis kata di luar kata-kata di atas yang berdasarkan peranannya dapat dibagi menjadi lima subkelompok:
    a. Preposisi (kata depan) (contoh: dari),
    b. Konjungsi (kata sambung) - Konjungsi berkoordinasi (dan), Konjungsi subordinat (karena),
    c. Artikula (kata sandang) (contoh: sang, si) - Umum dalam bahasa Eropa (misalnya the),
    d. Interjeksi (kata seru) (contoh: wow, wah), dan
    e. Partikel.

2.9 Struktur Kalimat Bahasa Indonesia

Yang dimaksud dengan kalimat ialah bagian terkecil ujaran atau teks (wacana) yang mengungkapkan pikiran yang utuh, merupakan satuan gramatikal yang dapat berdiri sendiri sebagai satu kesatuan, terdiri atas satu atau lebih klausa yang ditata menurut sistem bahasa yang bersangkutan, dan mempunyai pola intonasi final.

Dalam suatu kalimat terdiri dari beberapa unsur antara lain subyek, predikat, obyek, pelengkap, dan keterangan. Kalimat dikatakan sempurna jika minimal memliki unsur Subyek dan Predikat.

Dipandang sari segi jumlah dan jenis klausa yang terdapat pada dasar, kalimat dapat dibedakan sebagai berikut:

1. Kalimat tunggal

   Kalimat tunggal adalah kalimat yang terdiri atas satu klausa bebas, tanpa klausa terikat.

   Contoh:

   a. Ali tidur.
   b. Budi makan.

2. Kalimat bersusun

   Kalimat bersusun adalah kalimat yang terdiri atas satu klausa bebas, dan sekurang-kurangnya satu klausa terikat.

   Contoh:

   a. Caca datang sesudah matahari terbenam.
   b. Kami akan bertanding kalau wasitnya bukan dia.

   Penjelasan:

   Kalimat tersebut merupakan contoh kalimat bersusun, “Caca datang” dan “Kami akan bertanding” merupakan klausa bebas, sedangkan “sebelum matahari terbit” dan “kalau wasitnya bukan dia” merupakan klausa terikat.

3. Kalimat majemuk

   Kalimat mejemuk adalah kalimat yang terdiri atas beberapa klausa bebas.

   Contoh:

   a. Saya menyuruhnya datang, tetapi dia tidak berkenan.
   b. Dudi tidak akan bekerja, kecuali gaji bulan lalu telah dibayar.

2.10 Stemming

Stemming merupakan suatu proses yang terdapat dalam sistem IR yang mentransformasi kata- kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (root word) dengan menggunakan aturan-aturan tertentu. Sebagai contoh, kata bersama, kebersamaan, menyamai, akan distem ke root word-nya yaitu “sama”. Proses stemming pada teks berbahasa Indonesia

berbeda dengan stemming pada teks berbahasa Inggris. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia, selain sufiks, prefiks, dan konfiks juga dihilangkan.

Metode stemming memerlukan input berupa term yang terdapat dalam dokumen. Sedangkan outputnya berupa stem. Ada tiga jenis metode stemming, antara lain :

1. ***Successor Variety*** (SV) : lebih mengutamakan penyusunan huruf dalam kata dibandingkan dengan pertimbangan atas fonem. Contoh untuk kata-kata : *corpus*, *able*, *axle*, *accident*, *ape*, *about* menghasilkan SV untuk kata *apple*:
   a. Karena huruf pertama dari kata " *apple*" adalah "a", maka kumpulan kata yang ada *substring* "a" diikuti "b", "x", "c", "p" disebut SV dari "a" sehingga "a" memiliki 4 SV.
   b. Karena dua huruf pertama dari kata "apple" adalah "ap", maka kumpulan kata yang ada substring "ap" hanya diikuti "e" disebut SV dari "ap" sehingga "ap" memiliki 1 SV.
2. ***N-Gram Conflation*** : ide dasarnya adalah pengelompokan kata-kata secara bersama berdasarkan karakter-karakter (substring) yang teridentifikasi sepanjang N karakter.
3. ***Affix Removal*** (penghilangan imbuhan) : membuang prefix (awalan) dan suffix (akhiran) dari term menjadi suatu stem. Yang paling sering digunakan adalah algoritma Porter Stemmer karena modelnya sederhana dan efisien.

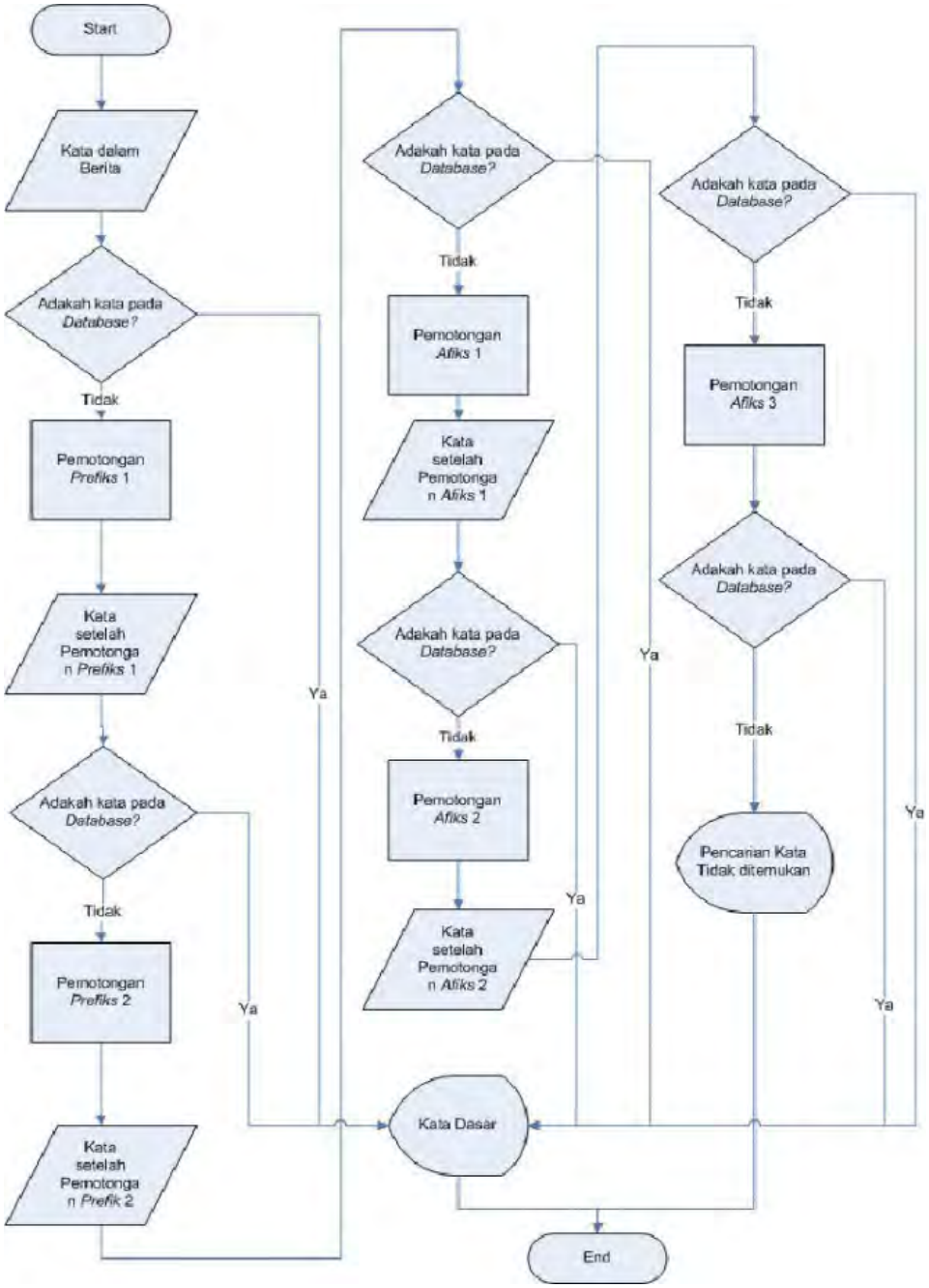
   a. Jika suatu kata diakhiri dengan "ies" tetapi bukan "eies" atau "aies", maka "ies" di-replace dengan "y"
   b. Jika suatu kata diakhiri dengan "es" tetapi bukan "aes" atau "ees" atau "oes", maka "es" di-replace dengan "e"
   c. Jika suatu kata diakhiri dengan "s" tetapi bukan "us" atau "ss", maka "s" di-replace dengan "NULL"

2.11 Stemming Bahasa Indonesia dengan Algoritma Nazief dan Andriani

Stemming adalah salah satu cara yang digunakan untuk meningkatkan performa IR (*information retrieval*) dengan cara mentransformasi kata-kata dalam sebuah dokumen teks ke bentuk kata dasarnya. Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma *stemming* untuk bahasa lainnya. Sebagai contoh bahasa Inggris memiliki morfologi yang berbeda dengan bahasa Indonesia sehingga algoritma *stemming* untuk kedua bahasa tersebut juga berbeda. Proses stemming pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan *root word* (kata dasar) dari sebuah kata. Pada umumnya kata dasar pada bahasa Indonesia terdiri dari kombinasi:

Prefiks 1 + Prefiks 2 + Kata dasar + Sufiks 3 + Sufiks 2 + Sufiks 1

Dari kombinasi tersebut dapat dibuat sebuah flowchart seperti tampak pada gambar 2.2.



Gambar 2.2 Algoritma Nazief dan Andriani

Algoritma Nazief & Andriani yang dibuat oleh Bobby Nazief dan Mirna Andriani ini memiliki tahap-tahap sebagai berikut:

1. Pertama cari kata yang akan distem dalam kamus kata dasar. Jika ditemukan maka diasumsikan kata adalah root word. Maka algoritma berhenti.

2. Inflection Suffixes (“-lah”, “-kah”, “-ku”, “-mu”, atau “-nya”) dibuang. Jika berupa particles (“-lah”, “-kah”, “-tah” atau “-pun”) maka langkah ini diulangi lagi untuk menghapus Possesive Pronouns (“-ku”, “-mu”, atau “-nya”), jika ada.
3. Hapus Derivation Suffixes (“-i”, “-an” atau “-kan”). Jika kata ditemukan di kamus, maka algoritma berhenti. Jika tidak maka ke langkah 3a
    a) Jika “-an” telah dihapus dan huruf terakhir dari kata tersebut adalah “-k”, maka “-k” juga ikut dihapus. Jika kata tersebut ditemukan dalam kamus maka algoritma berhenti. Jika tidak ditemukan maka lakukan langkah 3b.
    b) Akhiran yang dihapus (“-i”, “-an” atau “-kan”) dikembalikan, lanjut ke langkah 4.
4. Hapus Derivation Prefix. Jika pada langkah 3 ada sufiks yang dihapus maka pergi ke langkah 4a, jika tidak pergi ke langkah 4b.
    a) Periksa tabel kombinasi awalan-akhiran yang tidak diijinkan. Jika ditemukan maka algoritma berhenti, jika tidak
    b) Pergi ke langkah 4b.
    c) For i = 1 to 3, tentukan tipe awalan kemudian hapus awalan. Jika root word belum juga ditemukan lakukan langkah 5, jika sudah maka algoritma berhenti. Catatan: jika awalan kedua sama dengan awalan pertama algoritma berhenti.
5. Melakukan Recording.
6. Jika semua langkah telah selesai tetapi tidak juga berhasil maka kata awal diasumsikan sebagai root word. Proses selesai.

Tipe awalan ditentukan melalui langkah-langkah berikut:

1. Jika awalannya adalah: “di-”, “ke-”, atau “se-” maka tipe awalannya secara berturut-turut adalah “di-”, “ke-”, atau “se-”.
2. Jika awalannya adalah “te-”, “me-”, “be-”, atau “pe-” maka dibutuhkan sebuah proses tambahan untuk menentukan tipe awalannya.

3. Jika dua karakter pertama bukan “di-”, “ke-”, “se-”, “te-”, “be-”, “me-”, atau “pe-” maka berhenti.
4. Jika tipe awalan adalah “none” maka berhenti. Jika tipe awalan adalah bukan “none” maka awalan dapat dilihat pada Tabel 2. Hapus awalan jika ditemukan.

Tabel 2.1 Kombinasi awalan akhiran yang tidak diijinkan

Awalan	Akhiran yang tidak diijinkan
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se-	-i, -kan

Tabel 2.2 Cara Menentukan Tipe Awalan Untuk awalan “te-”

Following Characters				Tipe Awalan
Set 1	Set 2	Set 3	Set 4	
“-r-“	“-r-“	-	-	none
“-r-“	“-r-“	-	-	Ter-luluh
“-r-“	not (vowel or “-r-”)	“-er-“	vowel	ter
“-r-“	not (vowel or “-r-”)	“-er-“	not vowel	ter-
“-r-“	not (vowel or “-r-”)	not “-er-“	–	ter
not (vowel or “-r-”)	“-er-“	vowel	–	none
not (vowel or “-r-”)	“-er-“	not vowel	–	te

Tabel 2.3. Jenis Awalan Berdasarkan Tipe Awalannya

Tipe Awalan	Awalan yang harus dihapus
di-	di-
ke-	ke-
se-	se-
te-	te-
ter-	ter-
Ter-luluh	Ter

Untuk mengatasi keterbatasan pada algoritma di atas, maka ditambahkan aturan-aturan dibawah ini:

1. Aturan untuk reduplikasi.
    a) Jika kedua kata yang dihubungkan oleh kata penghubung adalah kata yang sama maka root word adalah bentuk tunggalnya, contoh : “buku-buku” root word-nya adalah “buku”.
    b) Kata lain, misalnya “bolak-balik”, “berbalas-balasan, dan ”seolah-olah”. Untuk mendapatkan root word-nya, kedua kata diartikan secara terpisah. Jika keduanya memiliki root word yang sama maka diubah menjadi bentuk tunggal, contoh: kata “berbalas-balasan”, “berbalas” dan “balasan” memiliki root word yang sama yaitu “balas”, maka root word “berbalas-balasan” adalah “balas”. Sebaliknya, pada kata “bolak-balik”, “bolak” dan “balik” memiliki root word yang berbeda, maka root word-nya adalah “bolak-balik”.
2. Tambahan bentuk awalan dan akhiran serta aturannya.
    a) Untuk tipe awalan “mem-“, kata yang diawali dengan awalan “memp-” memiliki tipe awalan “mem-”.
    b) Tipe awalan “meng-“, kata yang diawali dengan awalan “mengk-” memiliki tipe awalan “meng-”.

Berikut contoh-contoh aturan yang terdapat pada awalan sebagai pembentuk kata dasar.

1. Awalan SE-

   Se + semua konsonan dan vokal tetap tidak berubah

   Contoh :

   Se + bungkus = sebungkus

   Se + nasib = senasib

   Se + arah = searah

   Se + ekor = seekor

2. Awalan ME-

   Me + vokal (a,i,u,e,o) menjadi sengau “meng”

   Contoh :

   Me + inap = menginap

   Me + asuh = mengasuh

   Me + ubah = mengubah

   Me + ekor = mengekor

   Me + oplos = mengoplos

   Me + konsonan b menjadi “mem”

   Contoh :

   Me + beri = memberi

   Me + besuk = membesuk

   Me + konsonan c menjadi “men”

   Contoh :

   Me + cinta = mencinta

   Me + cuci = mencuci

   Me + konsonan d menjadi “men”

   Contoh :

   Me + didik = mendidik

   Me + dengkur = mendengkur

   Me + konsonan g dan h menjadi “meng”

Contoh :

Me + gosok = menggosok

Me + hukum = menghukum

Me + konsonan j menjadi “men”

Contoh :

Me + jepit = menjepit

Me + jemput = menjemput

Me + konsonan k menjadi “meng” (luluh)

Contoh :

Me + kukus = mengukus

Me + kupas = mengupas

Me + konsonan p menjadi “mem” (luluh)

Contoh :

Me + pesona = mempesona

Me + pukul = memukul

Me + konsonan s menjadi “meny” (luluh)

Contoh :

Me + sapu = menyapu

Me + satu = menyatu

Me + konsonan t menjadi “men” (luluh)

Contoh :

Me + tanama = menanam

Me + tukar = menukar

Me + konsonan (l,m,n,r,w) menjadi tetap “me”

Contoh :

Me + lempar = melempar

Me + masak = memasak

Me + naik = menaik

Me + rawat = merawat

Me + warna = mewarna

3. Awalan KE-

   Ke + semua konsonan dan vokal tetap tidak berubah

   Contoh :

   Ke + bawa = kebawa

   Ke + atas = keatas

4. Awalan PE-

   Pe + konsonan (h,g,k) dan vokal menjadi “per”

   Contoh :

   Pe + hitung + an = perhitungan

   Pe + gelar + an = pergelaran

   Pe + kantor + = perkantoran

   Pe + konsonan “t” menjadi “pen” (luluh)

   Contoh :

   Pe + tukar = penukar

   Pe + tikam = penikam

   Pe + konsonan (j,d,c,z) menjadi “pen”

   Contoh :

   Pe + jahit = penjahit

   Pe + didik = pendidik

   Pe + cuci = pencuci

   Pe + zina = penzina

   Pe + konsonan (b,f,v) menjadi “pem”

Contoh :

Pe + beri = pemberi

Pe + bunuh = pembunuh

Pe + konsonan “p” menjadi “pem” (luluh)

Contoh :

Pe + pikir = pemikir

Pe + potong = pemotong

Pe + konsonan “s” menjadi “peny” (luluh)

Contoh :

Pe + siram = penyiram

Pe + sabar = penyabar

Pe + konsonan (l,m,n,r,w,y) tetap tidak berubah

Contoh :

Pe + lamar = pelamar

Pe + makan = pemakan

Pe + nanti = penanti

Pe + wangi = pewangi

2.12 N-Gram

N-gram adalah potongan n karakter dalam suatu string tertentu atau potongan n kata dalam suatu kalimat tertentu (Cavnar & Trenkle, 1994). Misalnya dalam kata “Teknik” akan didapatkan n-gram sebagai berikut.

Tabel 2.4. Contoh pemotongan N-gram berbasis karakter

Nama	n-gram karakter
Uni-gram	T, E, K, N, I, K
Bi-gram	_T, TE, EK, KN, NI, IK, K_
Tri-gram	_TE, TEK, EKN, KNI, NIK, IK_, K_ _
Quad-gram	_TEK, TEKN, EKNI, KNIK, NIK_, IK_ _, K

Karakter blank "_" digunakan untuk merepresentasikan spasi di depan dan di akhir kata. Dan untuk word-based n-gram contohnya adalah sebagai berikut.

Kalimat : "N-gram adalah potongan n karakter dalam suatu string tertentu"

Tabel 2.5. Contoh pemotongan N-gram berbasis kata

Nama	n-gram kata
Uni-gram	n-gram, adalah, potongan, n, karakter, dalam, suatu, string, tertentu
Bi-gram	n-gram adalah, adalah potongan, potongan n, n karakter, karakter dalam, dalam suatu, suatu string, string tertentu
Tri-gram	n-gram adalah potongan, adalah potongan n, potongan n karakter, n karakter dalam, karakter dalam suatu, dalam suatu string, suatu string tertentu
Dst..	

2.13 Pembobotan Kata (*term weighting*)

Sistem Temu Kembali Informasi berhadapan dengan pencarian informasi yang sesuai dengan query pengguna dari koleksi dokumen. Koleksi dokumen tersebut terdiri dari dokumen-dokumen yang beragam panjangnya dengan kandungan term yang berbeda pula. 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 tingkat kepentingan yang berbeda dalam dokumen, maka untuk setiap kata tersebut diberikan sebuah indikator, yaitu *term weight*.

Term weighting atau pembobotan term sangat dipengaruhi oleh hal-hal berikut ini (Mandala, 2004):

1. *Term Frequency (tf) factor,* yaitu 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 kesesuian yang semakin besar.
2. *Inverse Document Frequency (idf) factor,* yaitu pengurangan dominansi term yang sering muncul di berbagai dokumen. Hal ini diperlukan karena term yang banyak muncul di berbagai dokumen, dapat dianggap sebagai term umum (*common term*) sehingga tidak penting nilainya. Sebaliknya faktor kejarangmunculan kata (*term scarcity*) dalam koleksi dokumen harus diperhatikan dalam pemberian bobot. Menurut Mandala (dalam Witten, 1999) 'Kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting (uncommon tems) daripada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata (*inverse document frequency*). Hal ini merupakan usulan dari George Zipf. Zipf mengamati bahwa frekuensi dari sesuatu cenderung kebalikan secara proposional dengan urutannya.

Metode TF-IDF merupakan metode pembobotan term yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Pada metode ini, perhitungan bobot term t dalam sebuah dokumen dilakukan dengan mengalikan nilai Term Frequency dengan Inverse Document Frequency.

Pada Term Frequency (tf), terdapat beberapa jenis formula yang dapat digunakan yaitu (Mandala, 2004):

1. tf biner (*binery tf*), hanya memperhatikan apakah suatu kata ada atau tidak dalam dokumen, jika ada diberi nilai satu, jika tidak diberi nilai nol

2. tf murni (*raw tf*), nilai tf diberikan berdasarkan jumlah kemunculan suatu kata di dokumen. Contohnya, jika muncul lima kali maka kata tersebut akan bernilai lima.
3. tf logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit kata dalam query, namun mempunyai frekuensi yang tinggi.

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

4. tf normalisasi, menggunakan perbandingan antara frekuensi sebuah kata dengan jumlah keseluruhan kata pada dokumen.

$$tf = 0.5 + 0.5 \text{ x } [tf / max\ tf] \quad (2.2)$$

Inverse Document Frequency (idf) dihitung dengan menggunakan formula

$$Idfj = log\ (D / dfj\ ) \quad (2.3)$$

dimana

D adalah jumlah semua dokumen dalam koleksi

df j adalah jumlah dokumen yang mengandung term tj

Menurut Defeng (dalam Robertson, 2004) 'Jenis formula yang akan digunakan untuk perhitungan term frequency (tf) yaitu tf murni (raw tf). Dengan demikian rumus umum untuk TF-IDF adalah penggabungan dari formula perhitungan raw tf dengan formula idf (rumus b.3) dengan cara mengalikan nilai term frequency (tf) dengan nilai inverse document frequency (idf) :

$$wij = tfij \times idfj$$

$$wij = tfij \times log\ (D/dfj) \quad (2.4)$$

Keterangan :

wij adalah bobot term tj terhadap dokumen di

tfij adalah jumlah kemunculan term tj di dalam dokumen di

D adalah jumlah semua dokumen yang ada dalam database

df j adalah jumlah dokumen yang mengandung term tj (minimal ada satu kata yaitu term t)

Berdasarkan rumus b.4, berapapun besarnya nilai tfij , apabila *D = df j* maka akan didapatkan hasil 0 (nol) untuk perhitungan idf. Untuk itu, dapat ditambahkan nilai 1 pada sisi *idf*, sehingga perhitungan bobotnya menjadi sebagai berikut:

$$wij=tfij \times (\log (D/dfj) + 1) \quad (2.5)$$

2.14 Microblogging Twitter

Twitter adalah sebuah situs web yang dimiliki dan dioperasikan oleh Twitter Inc, yang menawarkan jaringan sosial berupa mikroblog sehingga memungkinkan penggunanya untuk mengirim dan membaca pesan Tweets. Mikroblog adalah salah satu jenis alat komunikasi online dimana pengguna dapat memperbarui status tentang mereka yang sedang memikirkan dan melakukan sesuatu, apa pendapat mereka tentang suatu objek atau fenomena tertentu. Tweets adalah teks tulisan hingga 140 karakter yang ditampilkan pada halaman profil pengguna. Tweets bisa dilihat secara publik, namun pengirim dapat membatasi pengiriman pesan ke daftar teman-teman mereka saja. Pengguna dapat melihat Tweets pengguna lain yang dikenal dengan sebutan pengikut (*follower*).

Tidak seperti Facebook, LinkedIn, dan MySpace, Twitter merupakan sebuah jejaring sosial yang dapat digambarkan sebagai sebuah graph berarah (Wang, 2010), yang berarti bahwa pengguna dapat mengikuti pengguna lain, namun pengguna kedua tidak diperlukan untuk mengikutinya kembali. Kebanyakan akun berstatus publik dan dapat diikuti tanpa memerlukan persetujuan pemilik..

Semua pengguna dapat mengirim dan menerima Tweets melalui situs Twitter, aplikasi eksternal yang kompatibel (telepon seluler), atau dengan pesan singkat (SMS) yang tersedia di negara-negara tertentu. Pengguna dapat menulis pesan berdasarkan topik dengan menggunakan tanda # (hashtag). Sedangkan untuk menyebutkan atau membalas pesan dari pengguna lain bisa menggunakan tanda @.

Pesan pada awalnya diatur hanya mempunyai batasan sampai 140 karakter disesuaikan dengan kompatibilitas dengan pesan SMS, memperkenalkan singkatan notasi dan slang yang biasa digunakan dalam pesan SMS. Batas karakter 140 juga meningkatkan penggunaan layanan memperpendek URL seperti bit.ly, goo.gl, dan tr.im, dan jasa hosting konten, seperti Twitpic, Tweephoto, memozu.com dan NotePub untuk mengakomodasi multimedia isi dan teks yang lebih panjang daripada 140 karakter. Twitter menggunakan bit.ly untuk memperpendek otomatis semua URL yang dikirim-tampil. Fitur yang terdapat dalam Twitter, antara lain:

1. Laman Utama (*Home*)

   Pada halaman utama kita bisa melihat Tweets yang dikirimkan oleh orang-orang yang menjadi teman kita atau yang kita ikuti (*following*).
2. Profil (*Profile*)

   Pada halaman ini yang akan dilihat oleh seluruh orang mengenai profil atau data diri serta Tweets yang sudah pernah kita buat.
3. Followers

   Pengikut adalah pengguna lain yang ingin menjadikan kita sebagai teman. Bila pengguna lain menjadi pengikut akun seseorang, maka Tweets seseorang yang ia ikuti tersebut akan masuk ke dalam halaman utama.
4. Following

   Kebalikan dari pengikut, following adalah akun seseorang yang mengikuti akun pengguna lain agar Tweets yang dikirim oleh orang yang diikuti tersebut masuk ke dalam halaman utama.

5. Mentions

   Biasanya konten ini merupakan balasan dari percakapan agar sesama pengguna bisa langsung menandai orang yang akan diajak bicara.

6. Favorite

   Tweets ditandai sebagai favorit agar tidak hilang oleh halaman sebelumnya.

7. Pesan Langsung (Direct Message)

   Fungsi pesan langsung lebih bisa disebut SMS karena pengiriman pesan langsung di antara pengguna.

8. Hashtag

   Hashtag "#" yang ditulis di depan topik tertentu agar pengguna lain bisa mencari topik yang sejenis yang ditulis oleh orang lain juga

9. List

   Pengguna Twitter dapat mengelompokkan ikutan mereka ke dalam satu grup sehingga memudahkan untuk dapat melihat secara keseluruhan para nama pengguna (*username*) yang mereka ikuti (*follow*).

10. Topik Terkini (*Trending Topic*)

    Topik yang sedang banyak dibicarakan banyak pengguna dalam suatu waktu yang bersamaan.

2.15 Algoritma Naïve Bayes

Algoritma naive bayes classifier merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat (Feldman & Sanger 2007). Dalam penelitian ini yang menjadi data uji adalah dokumen Tweets. Ada dua tahap pada klasifikaasi dokumen. Tahap pertama adalah pelatihan terhadap dokumen yang sudah diketahui kategorinya. Sedangkan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya.

Dalam algoritma naïve bayes classifier setiap dokumen direpresentasikan dengan pasangan atribut "x1, x2, x3,...xn" dimana x1 adalah kata pertama, x2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori Tweet.

Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (VMAP), dimana persamaannya adalah sebagai berikut :

$$V_{MAP} = \underset{VjeV}{\arg max} \frac{P(x_1, x_2, x_3, \dots x_n|V_j)P(V_j)}{P(x_1,x_2,x_3,\dots x_n)} \quad (2.6)$$

Untuk P(x1, x2, x3,...xn) nilainya konstan untuk semua kategori (Vj) sehingga persamaan dapat ditulis sebagai berikut :

$$V_{MAP} = \underset{VjeV}{\arg max}\, P(x_1, x_2, x_3, \dots x_n|V_j)P(V_j) \quad (2.7)$$

Persamaan diatas dapat disederhanakan menjadi sebagai berikut :

$$V_{MAP} = \underset{VejV}{arg\,max} \prod_{i=1}^{n} P(x_i\,|V_j)P(V_j) \quad (2.8)$$

Keterangan :

V_j = Kategori *tweet* j =1, 2, 3,…n. Dan dalam penelitian ini
- j_1 = kategori *tweet* sentimen negatif,
- j_2 = kategori *tweet* sentimen positif
- j_3 = kategori *tweet* sentiment netral, dan
- j_4 = kategori *tweet* sentiment tanya

P(xi|Vj) = Probabilitas xi pada kategori Vj

P(Vj) = Probabilitas dari Vj

Untuk P(Vj) dan P(xi|Vj) dihitung pada saat pelatihan dimana persamaannya adalah sebagai berikut :

$$P(V_j) = \frac{|docs\ j|}{|contoh|} \quad (2.9)$$

$$P(x_i|V_j) = \frac{n_k+1}{n + |kosakata|} \quad (2.10)$$

Keterangan :

|docs j| = jumlah dokumen setiap kategori j

|contoh| = jumlah dokumen dari semua kategori

nk = jumlah frekuensi kemunculan setiap kata

n = jumlah frekuensi kemunculan kata dari setiap kategori

|kosakata| = jumlah semua kata dari semua kategori

2.16 Perangkat Lunak Weka

Weka (*Waikato Environment for Knowledge Analysis*) adalah sebuah paket *tools machine learning* praktis, yang dibuat di Universitas Waikato, New Zealand untuk penelitian, pendidikan dan berbagai aplikasi. Weka mampu menyelesaikan masalah-masalah *data mining* di dunia nyata, khususnya klasifikasi yang mendasari pendekatan-pendekatan *machine learning*. Perangkat lunak ini ditulis dalam hirarki *class Java* dengan metode berorientasi obyek dan dapat berjalan hampir di semua *platform*.

Weka mudah digunakan dan diterapkan pada beberapa tingkatan yang berbeda. Tersedia pilihan algoritma pembelajaran terbaru yang dapat kita terapkan pada dataset yang kita miliki. Weka memiliki perangkat untuk melakukan pra-proses data, klasifikasi, regresi, klastering, aturan asosiasi, dan visualisasi. Pengguna dapat melakukan tahap pra proses pada data, memasukkannya dalam sebuah skema pembelajaran, untuk selanjutnya menganalisa klasifikasi dan performansi yang dihasilkan tanpa melakukan pemrograman.

Tools yang dapat digunakan untuk pre-processing dataset membuat user dapat berfokus pada algoritma yang digunakan tanpa terlalu memperhatikan detail seperti pembacaan data dari file-file, implementasi algoritma filtering, dan penyediaan kode untuk evaluasi hasil.

Gambar 2.3 Weka GUI Chooser

2.17 Penelitian Sebelumnya

Penelitian mengenai analisis sentimen untuk mendeteksi kasus bullying telah dilakukan oleh Sanchez (2011). Di dalam papernya, Sanchez melakukan klasifikasi sentimen terhadap data tweet yang berisi istilah umum yang mengandung kekerasan, yang akan digunakan untuk mendeteksi intimidasi. Teknik pembelajaran mesin yang digunakan yaitu Naïve Bayes.

Penelitian lain tentang bullying yang dilakukan melaui media sosial Twitter dilakukan oleh Margono (2014). Margono membuat daftar kumpulan kata kunci yang berpotensi mengandung bullying, dan selanjutnya menggunakan metode Association Rule and FP-Growth untuk menemukan trend kata bullying di Indonesia melalui Twitter. Berdasarkan penelitian Margono, diketahui bahwa kata “bangsat” dan “anjing” merupakan kata-kata yang paling banyak digunakan untuk menyampaikan pesan dengan kategori bullying.

Tabel 2.6 Penelitian Sebelumnya

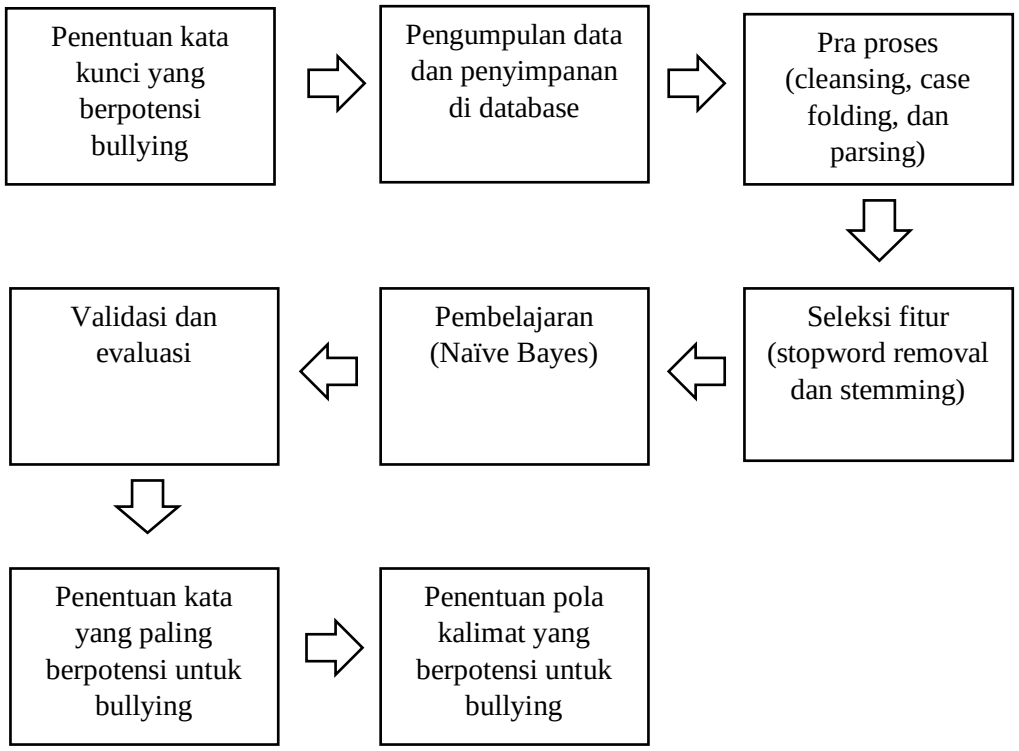
No	Peneliti/Tahun	Judul
1.	Sanchez, 2011	Twitter Bullying Detection
2.	Margono, 2014	Mining Indonesian Cyber Bullying Paterns in Social Networks

(Halaman ini sengaja dikosongkan)

BAB 3
METODOLOGI

3.1 Diagram Alir Metodologi Penelitian

Berikut ini adalah diagram alir yang akan digunakan untuk menyelesaikan atau mencapai tujuan penelitian:



Gambar 3.1 Metodologi Penelitian

3.2 Penentuan Kata Kunci

Pada tahap ini akan dilakukan pemilihan kata yang akan menjadi kata kunci untuk melakukan pencarian *tweet* dari pengguna *Twitter* yang berpotensi mengandung unsur *cyberbullying*. Diharapkan hasil pencariannya sesuai dengan ekspektasi yang diinginkan, dalam hal ini adalah memperoleh informasi tentang adanya *cyberbullying* melalui sosial media *Twitter*. Berikut ini adalah daftar kata kunci yang berpotensi digunakan untuk melakukan *cyberbullying*:

Tabel 3.1 Kata kunci

No	Kategori	Kata Kunci
1.	Binatang	Bangsat
		Anjing
		Babi
		Monyet
		Kunyuk
2.	Kebodohan dan Psikologi	Goblok
		Idiot
		Geblek
		Gila
		Tolol
		Sarap
		Udik
		Kampungan
3.	Kecacatan	Buta
		Budek
		Jelek
4.	Umum	Setan
		Iblis
		Keparat
		Gembel
		Brengsek
		Sompret
		Bajingan
5.	Sikap	Bejad

3.3 Pengumpulan Data

Proses pengumpulan data (*harvesting*) tweet dilakukan dengan memanfaatkan *Twitter Streaming APIs*. Pencarian dan pengumpulan opini masyarakat di Twitter didasarkan pada kata kunci yang telah ditetapkan sebelumnya dalam kurun waktu dua bulan. Data yang didapat dari hasil *harvesting* disimpan ke dalam database.

3.4 Pra Proses

Sebelum dilakukan proses seleksi fitur terhadap tweet yang telah didapatkan dan untuk mendapatkan hasil yang lebih akurat untuk analisa sentimen tweet, dilakukan pra proses (*pre processing*) terhadap data tweet yang ada yang meliputi:

1. *Cleansing*

   Pada proses *cleansing* dilakukan penghapusan URL, @mention , #hashtag dan delimiter(karakter angka & simbol)
2. *Case Folding*

   Pada tahap ini dilakukan perubahan semua karakter huruf menjadi huruf kecil.
3. *Parsing*

   Pemisahan sebuah tweet menjadi kata atau penguraian kalimat menjadi per kata dilakukan pada tahap ini

3.5 Seleksi Fitur

Seleksi fitur dilakukan sebelum proses pembelajaran dan klasifikasi. Pada tahap ini ada dua proses yang dilakukan yaitu:

1. *Stop Word Removal*

   Penghapusan kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen (mis: “di”, “oleh”, “pada”, “sebuah”, “karena”)
2. *Stemming*

   Proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya (stem), dengan cara menghilangkan

menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata

3.6 Pembelajaran dan Klasifikasi

Dari hasil seleksi fitur yang telah dilakukan, selanjutnya dilakukan proses pembelajaran dan klasifikasi dengan menggunakan algoritma naïve bayes yang dibagi menjadi dua tahap yaitu:

1. *Tahap pertama*

   Pelatihan terhadap dokumen tweet yang sudah diketahui kategorinya (*tweet bully*, bukan *tweet bully*).

2. *Tahap kedua*

   Proses klasifikasi dokumen yang belum diketahui kategorinya (*tweet bully*, bukan *tweet bully*).

3.7 Validasi dan Evaluasi

Tahap ini diperlukan untuk menvalidasi dan melakukan evaluasi sejauh mana tingkat keakuratan proses pembelajaran dan klasifikasi dengan menggunakan algoritma naïve bayes yang telah dilakukan. Perubahan pada beberapa parameter seperti *term weighting, min term freq, stop word, tokenizer, percentage split* diperlukan untuk meningkatkan akurasi. Selain itu, langkah validasi dapat dilakukan dengan cara membandingkan penggunaan algoritma yang berbeda untuk mengetahui perbandingan nilai akurasi yang diperoleh dari dua algoritma yang berbeda tersebut.

3.8 Penentuan Kata yang Berpotensi Digunakan untuk Cyberbullying

Dari hasil prediksi terhadap dokumen yang belum diketahui kategorinya apakah masuk dalam kelas *tweet bullying* atau kelas bukan *tweet bullying* dengan menggunakan *mechine learning* yang telah dilakukan sebelumnya, akan didapatkan kata yang prosentasenya paling rendah sampai paling tinggi yang digunakan untuk melakukan *cyberbullying*. Dengan dasar tersebut, dapat ditentukan kata yang berpotensi paling kecil sampai paling besar untuk digunakan melakukan *cyberbullying*.

3.9 Penentuan Pola Kalimat yang Berpotensi untuk Bullying

Pada tahapan ini dilakukan analisa gabungan kata atau pola kalimat yang mempunyai prosentase tertinggi terjadinya *cyberbullying* dengan menganalisa fitur yang mempunyai nilai bobot tertinggi.

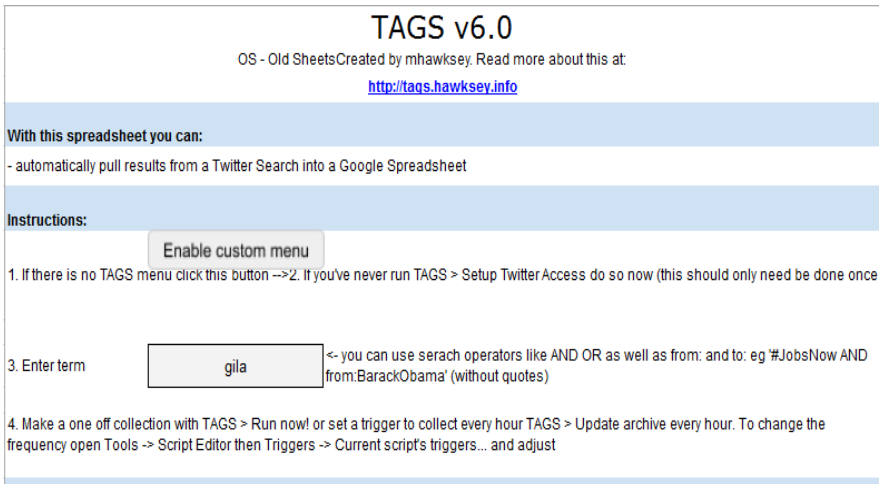
(Halaman ini sengaja dikosongkan)

BAB 4
HASIL DAN PEMBAHASAN

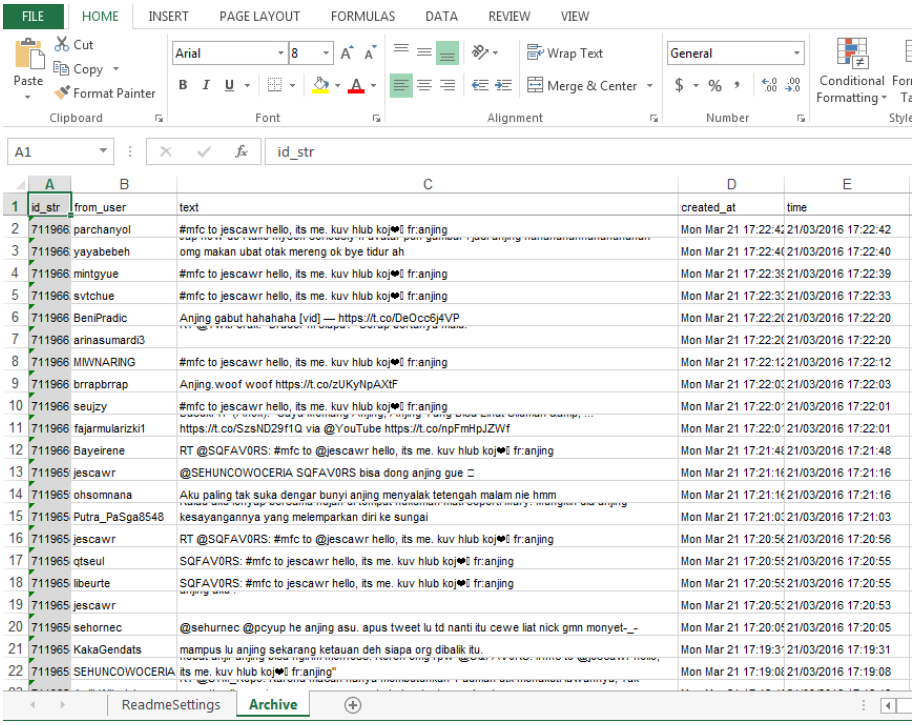
4.1 Verifikasi Metode

4.1.1 Pengumpulan Data

Diawali dengan penentuan kata kunci, proses pengumpulan data tweet dilakukan pada rentang waktu bulan Januari sampai dengan Maret 2016. Berdasarkan kata kunci yang telah ditentukan sebelumnya yaitu dalam bahasa Indonesia, maka tweet yang diambil juga dipilih yang berbahasa Indonesia. Proses pengumpulan data tweet (tweet harvesting) dilakukan dengan menggunakan perangkat lunak Tags V 6.0 yang terhubung dengan Twitter Streaming APIs, dengan demikian dapat diperoleh data tweet secara realtime.



Gambar 4.1 Perangkat Lunak Tags V.6.0



	id_str	from_user	text	created_at	time
2	711966	parchanyol	#mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:22:42	21/03/2016 17:22:42
3	711966	yayabebeh	omg makan ubat otak mereng ok bye tidur ah	Mon Mar 21 17:22:40	21/03/2016 17:22:40
4	711966	mintgyue	#mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:22:39	21/03/2016 17:22:39
5	711966	svtchue	#mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:22:33	21/03/2016 17:22:33
6	711966	BeniPradic	Anjing gabut hahahaha [vid] — https://t.co/DeOcc6j4VP	Mon Mar 21 17:22:20	21/03/2016 17:22:20
7	711966	arinasumardi3		Mon Mar 21 17:22:20	21/03/2016 17:22:20
8	711966	MIWNARING	#mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:22:12	21/03/2016 17:22:12
9	711966	brrapbrrap	Anjing.woof woof https://t.co/zUKyNpAXtF	Mon Mar 21 17:22:03	21/03/2016 17:22:03
10	711966	seujzy	#mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:22:01	21/03/2016 17:22:01
11	711966	fajarmularizki1	https://t.co/SzsND29f1Q via @YouTube https://t.co/npFmHpJZWf	Mon Mar 21 17:22:01	21/03/2016 17:22:01
12	711966	Bayeirene	RT @SQFAV0RS: #mfc to @jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:21:48	21/03/2016 17:21:48
13	711965	jescawr	@SEHUNCOWOCERIA SQFAV0RS bisa dong anjing gue ☐	Mon Mar 21 17:21:16	21/03/2016 17:21:16
14	711965	ohsomnana	Aku paling tak suka dengar bunyi anjing menyalak tetengah malam nie hmm	Mon Mar 21 17:21:16	21/03/2016 17:21:16
15	711965	Putra_PaSga8548	kesayangannya yang melemparkan diri ke sungai	Mon Mar 21 17:21:03	21/03/2016 17:21:03
16	711965	jescawr	RT @SQFAV0RS: #mfc to @jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:20:56	21/03/2016 17:20:56
17	711965	qtseul	SQFAV0RS: #mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:20:55	21/03/2016 17:20:55
18	711965	libeurte	SQFAV0RS: #mfc to jescawr hello, its me. kuv hlub koj❤ fr:anjing	Mon Mar 21 17:20:55	21/03/2016 17:20:55
19	711965	jescawr		Mon Mar 21 17:20:53	21/03/2016 17:20:53
20	711965	sehornec	@sehurnec @pcyup he anjing asu. apus tweet lu td nanti itu cewe liat nick gmn monyet-_-	Mon Mar 21 17:20:05	21/03/2016 17:20:05
21	711965	KakaGendats	mampus lu anjing sekarang ketauan deh siapa org dibalik itu.	Mon Mar 21 17:19:31	21/03/2016 17:19:31
22	711965	SEHUNCOWOCERIA	its me. kuv hlub koj❤ fr:anjing"	Mon Mar 21 17:19:08	21/03/2016 17:19:08

Gambar 4.2 Contoh Hasil *Tweet Harvesting* dengan Tags V.6.0

Berdasarkan kata kunci yang telah ditentukan, berikut ini merupakan hasil pengumpulan data *tweet* yang ditunjukkan pada Tabel 4.1:

Tabel 4.1 Perolehan Hasil *Tweet Harvesting*

No	Kategori	Kata Kunci	Jumlah Tweet
1.	Binatang	Bangsat	4043
		Anjing	32560
		Babi	5075
		Monyet	16698
		Kunyuk	1261
2.		Goblok	4607

No	Kategori	Kata Kunci	Jumlah Tweet
	Kebodohan dan Psikologi	Idiot	442
		Geblek	3359
		Gila	1321
		Tolol	3832
		Sarap	198
		Udik	982
		Kampungan	3063
3.	Kecacatan	Buta	1406
		Budek	1403
		Jelek	39764
4.	Umum	Setan	22267
		Iblis	3589
		Keparat	660
		Gembel	10564
		Brengsek	5713
		Sompret	237
		Bajingan	5588
5.	Sikap	Bejad	1217
		Total Tweet	**169849**

4.1.2 Data Training

Berdasarkan perolehan data hasil tweet harvesting, selanjutnya diambil data tweet yang akan digunakan sebagai data pelatihan dan data uji sebanyak 3000 tweet. Pengambilan 3000 data dari seluruh data yang

diperoleh dilakukan dengan membuat prosentase pada masing-masing kata kunci agar semua kata kunci dapat terwakili dalam data pelatihan dan data uji tersebut.

Langkah berikutnya, 3000 data tweet tersebut dilabeli secara manual untuk diklasifikasikan dalam kelompok tweet yang mengandung unsur bullying atau bukan, dengan format seperti ditunjukkan pada Tabel 4.2.

Tabel 4.2 Contoh data *tweet* yang dilabeli secara manual

No	Tweet ID	User ID	Tweet	Time	Klasifikasi
1.	711940631640330240	theresia_zsa	GAGARA LO CEWE RUSAK! LO AJ YG MATI ANJING! INGET GOBLOG LU BAKAL DAPET PEMBALASAN DR GW!!!! GW G PEDU… — so scared https://t.co/wOedNjrW9l	21/03/2016 15:40:48	Bully
2.	704272189688655873	solidculli	Perempuan Ini Dilarang Pelihara Hewan Satu Dekade Karena Telantarkan Anjing https://t.co/5VBs8esylb	29/02/2016 11:49:09	Bukan

No	Tweet ID	User ID	Tweet	Time	Klasifikasi
			https://t.co/utJGnYmTq7		
3.	704296105979224064	Klub Cinta Buku	“Hendaklah kamu tetap berbuat baik kepada orang yang berbuat jelek kepadamu” (Lukman Hakim)	29/02/2016 13:24:11	Bukan
4.	707899072267485184	Jawaaa97	Dasar tai dsar anjing gak berguna lo monyet	10/03/2016 12:01:05	Bully
5.	708009654933008384	novaarinarin	jgn pd kyk anjing loe semua kelurag ma loe resss okee selamnya ,, anjing babi bangsat loe smua ,tau diri loe bikin malu aja loe pd .,goblok	10/03/2016 19:20:30	Bully
6.	709047729209520128	yanuar_rizqi	Besok pagi buta masih ada perjalanan ke Bogor . Semangat	13/03/2016 16:05:26	Bukan

Setelah pelabelan manual, selanjutnya dilakukan tahap praproses pada data tweet tersebut dengan cara diimpor ke dalam database MySQL terlebih dahulu.

4.1.3 Pra Proses

Tahapan pra proses dilakukan sebelum proses seleksi fitur terhadap 3000 tweet data training dijalankan, dimana tahap pra proses ini meliputi:

1. Case Folding

   Pada tahap case folding semua karakter huruf pada data pelatihan diubah menjadi huruf kecil.

2. Cleansing

   Pada tahap cleansing dilakukan penghapusan URL, @mention, #hashtag, RT @ dan delimiter (karakter angka & simbol).

3. Parsing

   Pada tahap parsing sebuah data tweet yang semula berbentuk kalimat dipisahkan menjadi kata.

Contoh hasil dari tahap pra proses ditunjukkan pada Tabel 4.3 berikut:

Tabel 4.3 Hasil Tahap Pra Proses *Data Training*

Tweet	Case Folding	Cleansing	Parsing
RT @sejonc ANJING LO. SINI LO MACEM APAAN LO MARAH AMA GUE CMN NGEFAV TWEET GUE DOANG. BAGUS LO HAH	rt @sejonc anjing lo. sini lo macem apaan lo marah ama gue cmn ngefav tweet gue doang. bagus lo hah https//t.co/c8 xlpke8rf	anjing lo. sini lo macem apaan lo marah ama gue cmn ngefav tweet gue doang. bagus lo hah	word [1] = anjing word [2] = lo word [3] = sini word [4] = lo word [5] = macem word [6] = apaan word [7] = lo word [8] = marah word [9] = ama word [10] = gue word [11] = cmn word [12] = ngefav word [13] = tweet word [14] = gue

Tweet	Case Folding	Cleansing	Parsing
https//t.co/C8xlpKe8rF			word [15] = doang word [16] = bagus word [17] = lo word [18] = hah
@suparman152017 @asharsyah @Yusrilihza_Mhd menurutmu orang muslim memilih pemimpin mereka dr golongan sendiri itu dunggu? Dsr otak babi lo□□	@suparman152017 @asharsyah @yusrilihza_mhd menurutmu orang muslim memilih pemimpin mereka dr golongan sendiri itu dunggu? dsr otak babi lo□□	menurutmu orang muslim memilih pemimpin mereka dr golongan sendiri itu dunggu dsr otak babi lo	word [1] = menurutmu word [2] = orang word [3] = muslim word [4] = memilih word [5] = pemimpin word [6] = mereka word [7] = dr word [8] = golongan word [9] = sendiri word [10] = itu word [11] = dunggu word [12] = dsr word [13] = otak word [14] = babi word [15] = lo

4.1.4 Seleksi Fitur

Setelah tahap pra proses, tahap berikutnya yang dilakukan pada data training adalah seleksi fitur, dengan menjalankan langkah-langkah sebagai berikut:

1. Stop Word Removal

   Berdasarkan penelitian Tala (Tala, F. Z. (2003)) diperoleh daftar kata yang dianggap tidak mempunyai makna atau dapat diabaikan, sehingga bisa dihilangkan dari suatu dokumen tweet data training melalui tahap stop word removal.

2. Stemming

   Pada tahap stemming ini dilakukan proses pengembalian berbagai bentukan kata ke dalam bentuk kata dasar, menggunakan PHP library Sastrawi yang dibuat berdasarkan algoritma stemming Nazief dan Andriani.

Tabel 4.4 Hasil Stemming Menggunakan PHP Library Sastrawi

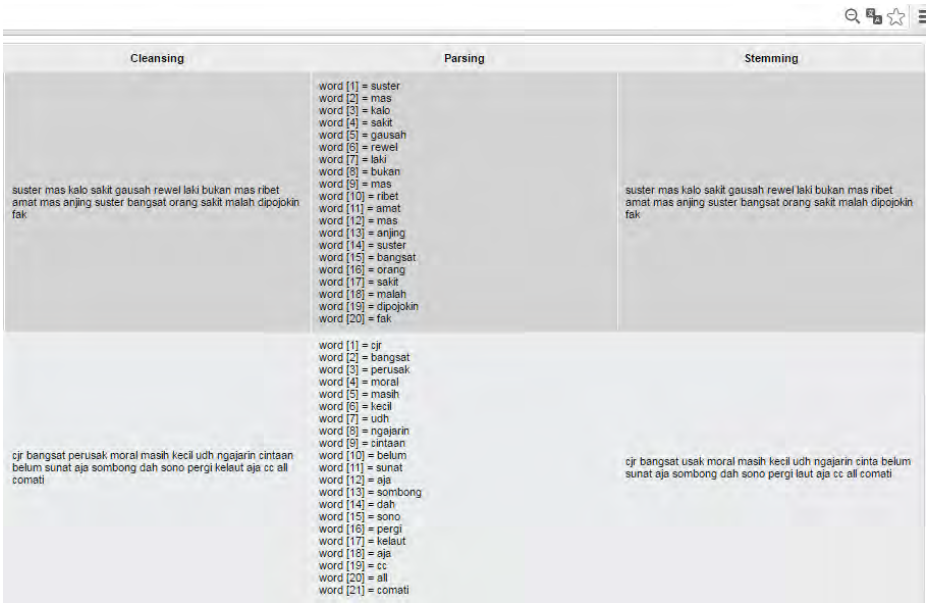
Hasil Parsing Tweet	Stemming
word [1] = anjing word [2] = lo word [3] = sini word [4] = lo word [5] = macem word [6] = apaan word [7] = lo word [8] = marah word [9] = ama word [10] = gue word [11] = cmn word [12] = ngefav word [13] = tweet word [14] = gue word [15] = doang word [16] = bagus word [17] = lo word [18] = hah	word [1] = anjing word [2] = lo word [3] = sini word [4] = lo word [5] = macem word [6] = apaan word [7] = lo word [8] = marah word [9] = ama word [10] = gue word [11] = cmn word [12] = ngefav word [13] = tweet word [14] = gue word [15] = doang word [16] = bagus word [17] = lo word [18] = hah
word [1] = menurutmu word [2] = orang word [3] = muslim word [4] = memilih	word [1] = turut word [2] = orang word [3] = muslim word [4] = pilih

Hasil Parsing Tweet	Stemming
word [5] = pemimpin	word [5] = pimpin
word [6] = mereka	word [6] = mereka
word [7] = dr	word [7] = dr
word [8] = golongan	word [8] = golongan
word [9] = sendiri	word [9] = sendiri
word [10] = itu	word [10] = itu
word [11] = dunggu	word [11] = dunggu
word [12] = dsr	word [12] = dsr
word [13] = otak	word [13] = otak
word [14] = babi	word [14] = babi
word [15] = lo	word [15] = lo

Script PHP yang dibuat untuk tahap pra proses dan seleksi fitur serta hasil akhir masing-masing ditunjukkan pada Gambar 4.3, Gambar 4.4, dan Gambar 4.5.

Gambar 4.3 Script PHP untuk Melakukan Pra Proses dan Seleksi Fitur

Gambar 4.4 Tweet Asli, Proses *Case Folding*, dan *Cleansing*



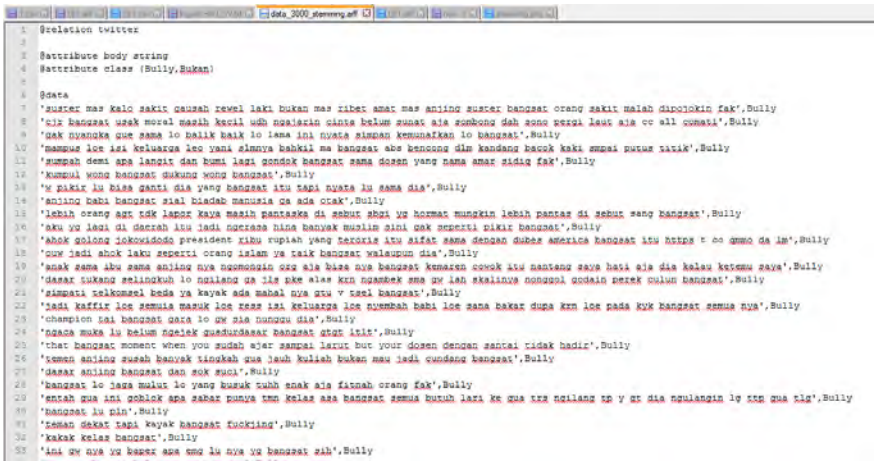
Gambar 4.5 Proses Parsing dan Stemming

4.1.5 Pembelajaran dan Klasifikasi

Setelah tahap seleksi selesai dilakukan, tahap berikutnya yang dilakukan adalah proses pembelajaran dan klasifikasi. Proses ini dilakukan menggunakan algoritma Naïve Bayes melalui dua tahap sebagai berikut:

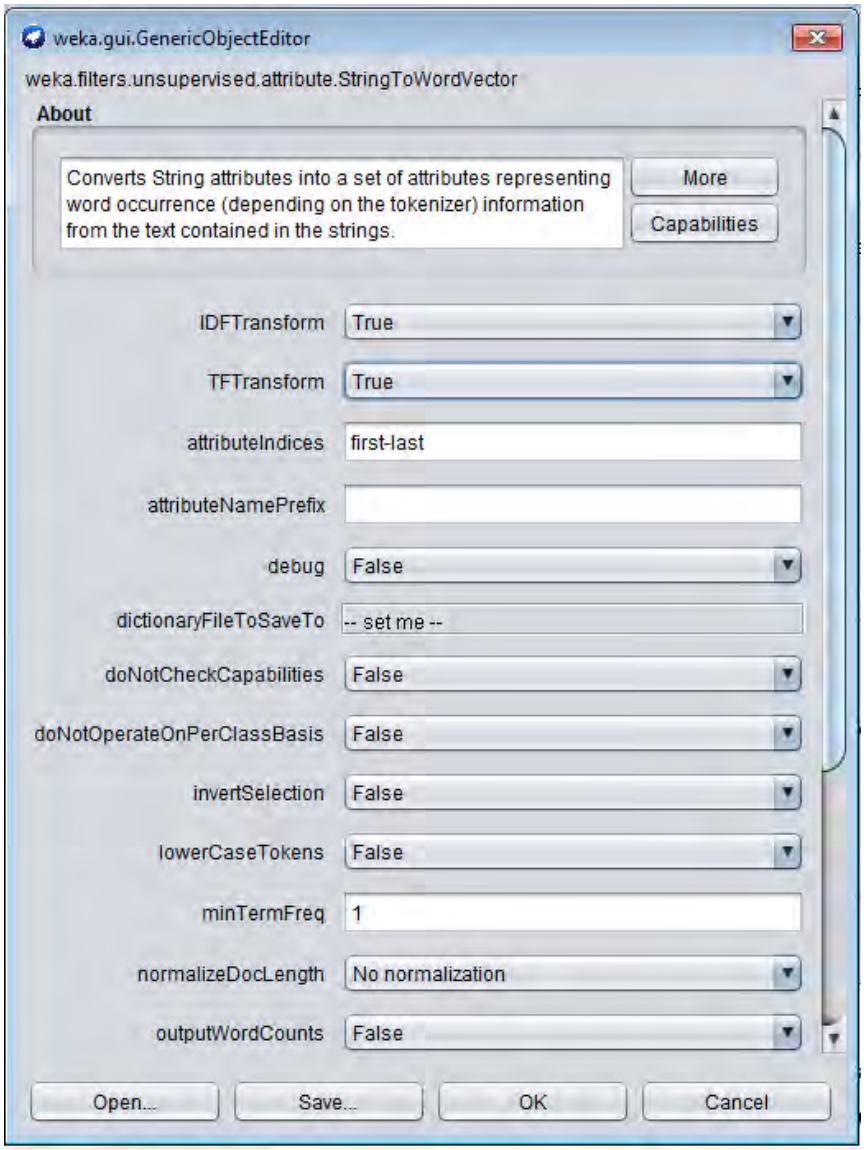
1. Tahap pertama

Setelah diketahui klasifikasi data training berdasarkan hasil seleksi fitur (*bully* atau bukan), berikutnya dilakukan proses pembelajaran terhadap data training tersebut. Proses pembelajaran dilakukan dengan menggunakan perangkat lunak Weka 3.8.0. Data yang dapat diolah menggunakan perangkat lunak Weka adalah data dengan format arff, sehingga perlu dilakukan perubahan format data dari csv ke arff. Selanjutnya, data dengan format arff dapat dianalisa menggunakan algoritma Naïve Bayes. Format data arff ditunjukkan pada Gambar 4.6.



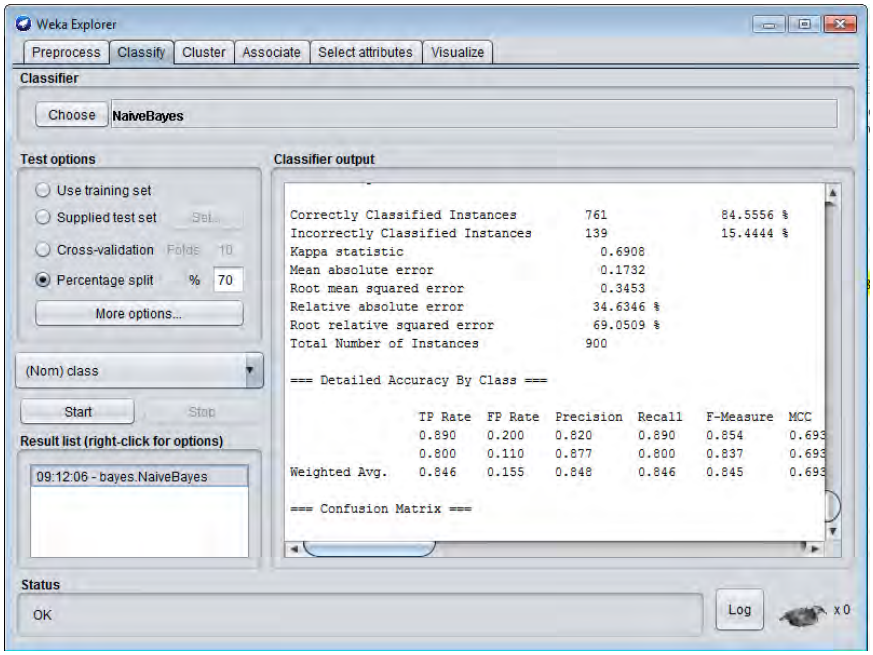
Gambar 4.6 Data dengan Format arff

Proses analisa data di Weka dengan algoritma Naïve Bayes dapat dilakukan dengan menggunakan beberapa kombinasi filtering yang terdiri atas *term weighting, min term freq, stop word dan tokenizer* seperti terlihat pada Gambar 4.7, dengan tujuan memperoleh hasil perhitungan yang optimum.



Gambar 4.7 Kombinasi pada Proses *Filtering String to Word Vector*

Tahap selanjutnya yang dilakukan setelah proses konversi string to word vector dan filtering adalah melakukan pelatihan terhadap data training, dengan memilih algoritma Naïve Bayes untuk pengklasifikasian. Prosentase jumlah data yang digunakan sebagai data training dan data uji dapat ditentukan pada tab percentage split di dalam menu test options. Setelah dilakukan klasifikasi, diperoleh nilai akurasi seperti ditunjukkan pada Gambar 4.8.



Gambar 4.8 Contoh Hasil Klasifikasi pada Weka

Kombinasi filtering dalam proses pengklasifikasian data yang bertujuan untuk memperoleh hasil perhitungan nilai akurasi yang optimum disajikan pada Tabel 4.5, 4.6, 4.7, dan 4.8 berikut ini.

Tabel 4.5 Hasil pembelajaran tanpa pembobotan TF atau IDF

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
	1	Tala	Word	50%	80,80%
				60%	81,92%
				70%	82,89%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,53%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
-				60%	85,08%
				70%	85,33%
	3	Tala	Word	50%	80,73%
				60%	81,92%
				70%	77,44%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,47%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%

Tabel 4.6 Hasil pembelajaran dengan pembobotan TF

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
TF	1	Tala	Word	50%	80,80%
				60%	81,92%
				70%	82,89%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,53%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%
	3	Tala	Word	50%	80,73%
				60%	81,92%
				70%	83,00%

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,47%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%

Tabel 4.7 Hasil pembelajaran dengan pembobotan IDF

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
IDF	1	Tala	Word	50%	80,80%
				60%	81,92%
				70%	82,89%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
			Word	50%	84,53%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%
	3	Tala	Word	50%	80,73%
				60%	81,92%
				70%	83,00%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,47%
				60%	85,25%

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%

Tabel 4.8 Hasil pembelajaran dengan pembobotan TF-IDF

term weighting	min term freq	stop word	tokenizer	percentage split	accuracy
TF-IDF	1	Tala	Word	50%	80,80%
				60%	81,92%
				70%	82,89%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,53%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%
	3	Tala	Word	50%	80,73%
				60%	81,92%
				70%	83,00%
			Ngram	50%	80,67%
				60%	82,25%
				70%	82,22%
		-	Word	50%	84,47%
				60%	85,25%
				70%	86,00%
			Ngram	50%	84,73%
				60%	85,08%
				70%	85,33%

Hasil yang ditampilkan pada Tabel 4.5, 4.6, 4.7, dan 4.8 merupakan hasil klasifikasi yang dilakukan tanpa melakukan stemming terhadap tweet data training. Berdasarkan tabel tersebut, terlihat bahwa nilai akurasi tertinggi adalah sebesar 86%. Nilai akurasi ini dapat diperoleh pada kondisi:

a. Pembelajaran dilakukan tanpa pembobotan TF-IDF dengan *minimal term frequency* senilai 1, tanpa menjalankan proses *stopword*, menggunakan *word tokenizer*, dan *percentage split sebesar* 70%.
b. Pembelajaran dilakukan tanpa pembobotan TF-IDF dengan *minimal term frequency* senilai 3, tanpa menjalankan proses *stopword*, menggunakan *word tokenizer*, dan *percentage split sebesar* 70%.

Nilai akurasi yang berbeda pada hasil pembelajaran akan diperoleh jika dilakukan proses stemming pada tweet data training. Hal ini ditunjukkan pada pada Tabel 4.9, 4.10, dan 4.11.

Tabel 4.9 Hasil Pembelajaran dengan Lovins Stemmer

term weighting	stemmer	min term freq	stop word	tokenizer	percentage split	Accuracy
		1	Tala	Word	50%	82,80%
					60%	83,58%
					70%	84,22%
				Ngram	50%	82,53%
					60%	83,50%
					70%	84,11%
			-	Word	50%	85,00%
					60%	86,17%
					70%	86,56%
				Ngram	50%	85,20%
					60%	85,83%
					70%	86,11%

term weighting	stemmer	min term freq	stop word	tokenizer	percentage split	Accuracy
TF-IDF	Lovins Stemmer	3	Tala	Word	50%	82,80%
					60%	83,58%
					70%	84,22%
				Ngram	50%	82,53%
					60%	83,50%
					70%	84,11%
			-	Word	50%	85,00%
					60%	86,17%
					70%	86,56%
				Ngram	50%	85,20%
					60%	85,83%
					70%	86,11%

Tabel 4.10 Hasil Pembelajaran dengan Iterated Lovins Stemmer

term weighting	stemmer	min term freq	stop word	tokenizer	percentage split	Accuracy
TF-IDF		1	Tala	Word	50%	83,00%
					60%	84,08%
					70%	85,00%
				Ngram	50%	83,00%
					60%	84,08%
					70%	85,00%
			-	Word	50%	85,33%
					60%	86,58%
					70%	87,67%
				Ngram	50%	85,67%
					60%	86,67%
					70%	87,22%
			Tala	Word	50%	83,00%
					60%	84,08%
					70%	85,00%

term weighting	stemmer	min term freq	stop word	tokenizer	percentage split	Accuracy
	Iterated Lovins Stemmer	3		Ngram	50%	83,00%
					60%	84,08%
					70%	85,00%
			-	Word	50%	85,33%
					60%	86,58%
					70%	87,67%
				Ngram	50%	85,67%
					60%	86,67%
					70%	87,22%

Tabel 4.11 Hasil Pembelajaran dengan Algoritma Stemming Nazief-Andriani

term weighting	stemmer	min term freq	stop word	tokenizer	percentage split	Accuracy
TF-IDF		1	Tala	Word	50%	80,20%
					60%	81,67%
					70%	82,67%
				Ngram	50%	72,00%
					60%	72,33%
					70%	71,44%
			-	Word	50%	83,93%
					60%	84,00%
					70%	84,56%
				Ngram	50%	81,53%
					60%	82,17%
					70%	81,78%
			Tala	Word	50%	80,13%
					60%	81,75%
					70%	82,67%

term weighting	stemmer	min term freq	stop word	tokenizer	percentage split	Accuracy
	Nazief - Andriani Stemmer	3		Ngram	50%	72,00%
					60%	72,33%
					70%	71,44%
			-	Word	50%	83,93%
					60%	84,00%
					70%	84,56%
				Ngram	50%	81,53%
					60%	82,17%
					70%	81,78%

Pada Tabel 4.9 ditunjukkan data hasil pembelajaran dengan metode stemming menggunakan Lovins Stemmer, dan diperoleh nilai akurasi paling optimum yaitu sebesar 86,56%, yang dicapai pada kondisi:

a. Pembelajaran dilakukan dengan pembobotan TF-IDF dengan minimal term frequency senilai 1, tanpa menjalankan proses stopword, menggunakan word tokenizer, dan percentage split sebesar 70%.
b. Pembelajaran dilakukan dengan pembobotan TF-IDF dengan minimal term frequency senilai 3, tanpa menjalankan proses stopword, menggunakan word tokenizer, dan percentage split sebesar 70%.

Data pada Tabel 4.10 merupakan data hasil pembelajaran dengan metode stemming menggunakan Iterated Lovins Stemmer, dan dengan metode ini didapatkan nilai akurasi optimum sebesar 87,67%, dimana nilai ini dicapai pada kondisi:

a. Pembelajaran dilakukan dengan pembobotan TF-IDF dengan minimal term frequency senilai 1, tanpa menjalankan proses

stopword, menggunakan word tokenizer, dan percentage split sebesar 70%.

b. Pembelajaran dilakukan dengan pembobotan TF-IDF dengan minimal term frequency senilai 3, tanpa menjalankan proses stopword, menggunakan word tokenizer, dan percentage split sebesar 70%.

Tabel 4.11 merupakan data hasil pembelajaran dengan metode stemming PHP Sastrawi yang dibuat berdasarkan algoritma Nazief – Andrianie. Di antara ketiga proses pembelajaran yang dilakukan dengan menggunakan metode stemming, proses pembelajaran dengan stemming berdasarkan algoritma Nazief – Andrianie menghasilkan nilai akurasi optimum yang paling rendah yaitu sebesar 84,56%, yang dicapai pada kondisi:

a. Pembelajaran dilakukan dengan pembobotan TF-IDF dengan minimal term frequency senilai 1, tanpa menjalankan proses stopword, menggunakan word tokenizer, dan percentage split sebesar 70%.
b. Pembelajaran dilakukan dengan pembobotan TF-IDF dengan minimal term frequency senilai 3, tanpa menjalankan proses stopword, menggunakan word tokenizer, dan percentage split sebesar 70%.

Berdasarkan semua data hasil pembelajaran yang ditampilkan pada tabel, baik menggunakan proses stemming atau tanpa stemming, diperoleh nilai akurasi tertinggi pada proses pembelajaran dengan metode stemming Iterated Lovins Stemmer.

2. Tahap kedua

Setelah melakukan proses pembelajaran pada data training menggunakan algoritma Naïve Bayes dan diperoleh nilai akurasi yang paling optimum, berikutnya dilakukan prediksi terhadap data

tweet yang belum diketahui klasifikasinya, apakah tergolong kategori bullying atau bukan bullying. Model pembelajaran dengan nilai akurasi yang paling optimum akan digunakan dalam proses prediksi ini.

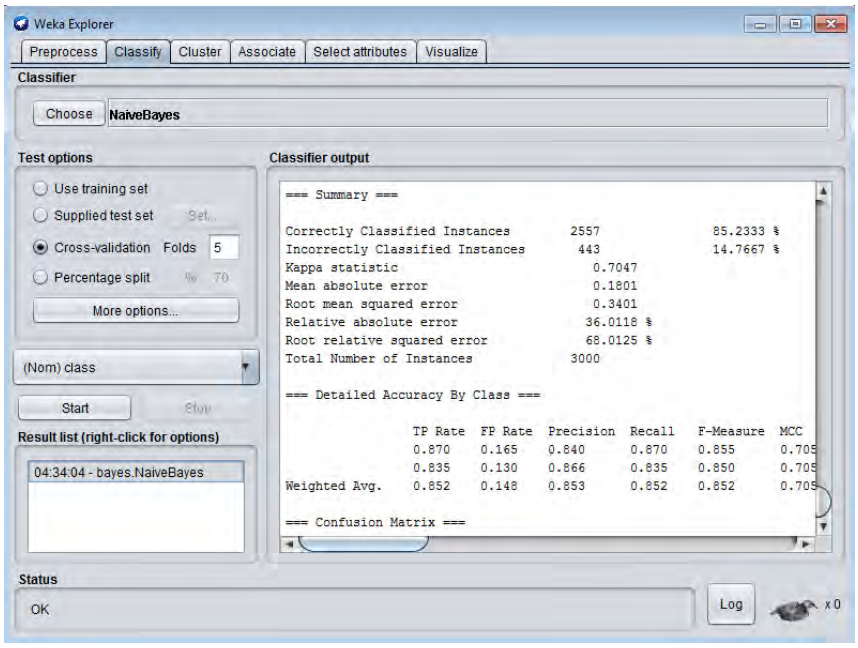
Langkah yang harus dilakukan terlebih dahulu adalah melakukan tahap praproses dan stemming pada semua data diluar data training hasil *tweet harvesting* dengan tahapan yang sama dengan data training. Data tersebut dibuat dengan format ARFF atau CSV dengan klasifikasi berupa tanda tanya. Dengan menggunakan menu *supplied test set*, file ARFF yang telah disiapkan sebelumnya dapat diambil dan digunakan untuk melakukan prediksi terhadap data di luar data training.

4.1.6 Validasi dan Evaluasi

Validasi diperlukan untuk membuktikan bahwa metode yang dipilih sudah sesuai sehingga dapat menghasilkan data yang valid. Dalam penelitian ini, validasi yang dilakukan adalah dengan dua cara, yang pertama yaitu dengan menggunakan mode test options yang berbeda pada algoritma yang sama, dan cara yang kedua yaitu dengan menggunakan metode pengujian yang berbeda pada data training yang telah disediakan. Berdasarkan kedua proses tersebut akan diperoleh nilai akurasi, kemudian besarnya nilai akurasi yang diperoleh akan dibandingkan dengan hasil akurasi optimum yang disajikan pada Tabel 4.10.

Berdasarkan data yang disajikan pada Tabel 4.10 yang menunjukkan nilai akurasi tertinggi pengujian data training menggunakan algoritma Naïve Bayes, dimana nilai tertinggi tersebut dicapai pada kondisi term frequency 1 atau 3, tanpa menjalankan stopword, menggunakan word tokenizer, dan percentage split sebesar 70%, maka dapat dilakukan pilihan validasi yang pertama yaitu dengan cara mengubah mode pada test options. Jika sebelumnya digunakan mode test options percentage split, maka mode test options yang digunakan pada proses validasi ini

adalah cross-validation. Mekanisme pada cross-validation adalah membagi data menjadi beberapa subset, atau yang di dalam perangkat lunak Weka disebut dengan fold. Jumlah fold ini dapat dipilih pada beberapa macam nilai, kemudian dihitung nilai akurasi yang diperoleh pada masing-masing nilai fold yang sudah dipilih. Selanjutnya beberapa nilai akurasi yang diperoleh tersebut akan dirata-rata untuk memperoleh nilai akurasi optimum. Proses penghitungan dengan mode test options cross-validation pada perangkat lunak Weka 3.8.0 ditunjukkan pada Gambar 4.9.



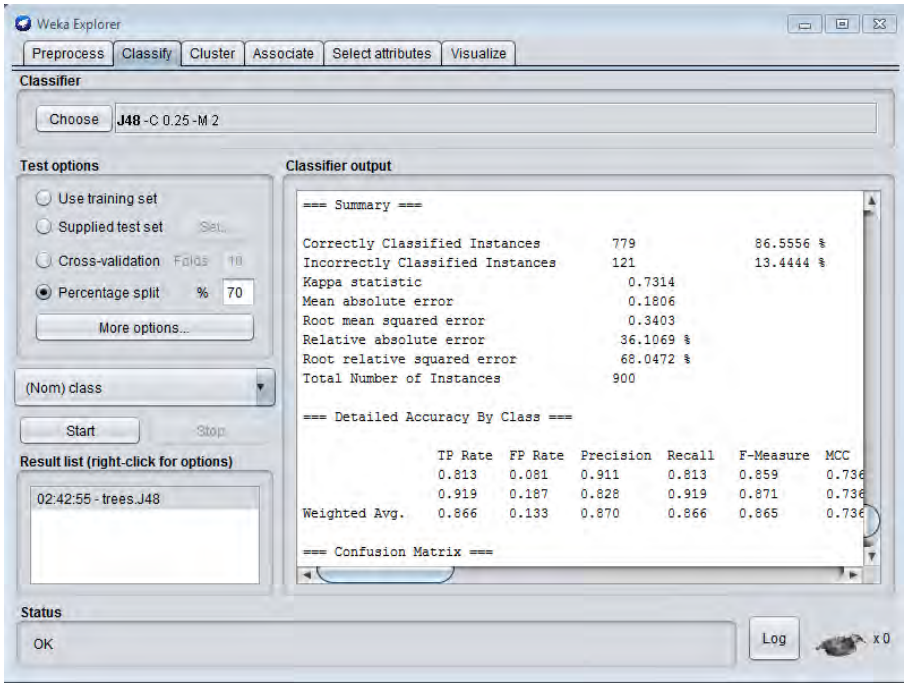
Gambar 4.9 Pengujian Data Traning dengan Algoritma Naïve Bayes pada Mode Test Options Cross-validation

Pengujian dengan mode test options cross-validations dilakukan dengan menggunakan nilai fold 5, 8, 10. Hasil pengujian ditunjukkan pada Tabel 4.12.

Tabel 4.12 Hasil Validasi Pengujian dengan Cross-validations

term weighting	stemmer	min term freq	stop word	tokenizer	folds	Accuracy
TF-IDF	Iterated Lovins Stemmer	3	-	Word	5	85,23%
					8	85,07%
					10	84,97%
					Rata-rata	85,09%

Selain algoritma Naive Bayes, algoritma Decision Tree J48 yang terdapat pada perangkat lunak Weka 3.8.0 merupakan metode yang juga dipilih untuk melakukan pengujian pada data training. Berdasakan data pada Tabel 4.10 yang menunjukkan nilai akurasi tertinggi pengujian data menggunakan algoritma Naïve Bayes, dimana nilai akurasi tertinggi tersebut diperoleh pada kondisi yang telah disebutkan sebelumnya, maka kondisi yang digunakan pada saat diperoleh nilai akurasi tertinggi tersebut dapat diterapkan pada proses pengujian data training dengan menggunakan algoritma Decision Tree J48. Hasil klasifikasi yang diperoleh ditunjukkan pada Gambar 4.10 sebagai berikut:



Gambar 4.10 Hasil Klasifikasi dengan Algoritma Decision Tree J48

Berdasarkan hasil pengujian data training yang dilakukan dengan algoritma Decision Tree J4.8, diperoleh hasil klasifikasi yang menunjukkan nilai akurasi sebesar 85, 56%.

Mengacu pada nilai akurasi yang diperoleh pada ketiga metode pengujian yang berbeda, diperoleh hasil bahwa pengujian dengan menggunakan algoritma Naïve Bayes memberikan nilai akurasi yang paling tinggi pada proses klasifikasi dengan kategori bullying atau bukan bullying.

4.2 Hasil Klasifikasi

Dengan proses pembelajaran yang dilakukan pada 3000 tweet data training, juga dapat ditentukan kata kunci dengan bobot paling tinggi yang paling berpotensi digunakan untuk melakukan bullying. Dari 24 kata kunci yang ditentukan, kata dengan bobot tertinggi yang paling berpotensi untuk melakukan bullying adalah kata “tolol”.

Selain diperoleh kata dengan bobot tertinggi tersebut, diketahui juga bahwa kalimat pesan yang disampaikan melalui Twitter yang berpotensi digunakan untuk melakukan bullying adalah berupa kalimat yang di dalamnya kata kunci bullying dan ditambahkan subyek sebelum kata kunci tersebut. Subyek "lo" merupakan subyek dengan bobot tertinggi menyampaikan pesan bully. Dengan demikian akan terbentuk kalimat yang berpotensi paling besar untuk melakukan *cyberbullying* yaitu kalimat dengan pola "Subyek + Kata Bully". Pada Tabel 4.13 ditunjukkan contoh data tweet dengan pola "Subyek + Kata Bully" yang menghasilkan kalimat bullying:

Tabel 4.13. Contoh data tweet dengan pola kalimat "Subyek + Kata Bully"

No	User ID	Data tweet dengan pola "Subyek + Kata Bully"
1	ddskendall	GOBLOK LU ANJING @SarahMOnline
2	thcmblr	anjing so soan lo monyet https://t.co/DdMgh26MH2
3	EPWX10	Woy njing!! ibumu pelacur ya pantesan kelakuan lo bejad! Dasar anak haram!! @MekelSungg
4	edododen	Lo goblok yg kaga ngertiin ke adaan gue! Untung lo blom jadi pacar gue, 11-12 deh lo ama mantan gue yg GTM!
5	purdijantoro	@sirjhonnasri1 sok tau lo tolol,kafirun darimananya,

4.3 Prediksi Kategorisasi

Berdasarkan hasil pembelajaran dan pengujian pada data training, selanjutnya yang harus dilakukan adalah menguji data tweet yang belum diketahui klasifikasinya, dengan tujuan memprediksi potensi bullying pada data tweet tersebut. Dari proses ini akan didapatkan hasil klasifikasi seperti ditampilkan pada tabel 4.14 sebagai berikut:

Tabel 4.14 Hasil Klasifikasi dan Prediksi

No	Kategori	Kata Kunci	Jumlah Tweet	Kategori		Prosentase Bully
				Bully	Bukan Bully	
1	Binatang	Bangsat	4043	749	3294	18,53%
		Anjing	32560	3655	28905	11,23%
		Babi	5075	591	4484	11,65%
		Monyet	16698	2132	14566	12,77%
		Kunyuk	1261	175	1086	13,88%
2	Kebodohan dan Psikologi	Goblok	4607	1667	2940	36,18%
		Idiot	442	25	417	5,66%
		Geblek	3359	434	2925	12,92%
		Gila	1321	117	1204	8,86%
		Tolol	3832	2744	1088	71,61%
		Sarap	198	23	175	11,62%
		Udik	982	110	872	11,20%
		Kampungan	3063	585	2478	19,10%
3	Kecacatan	Buta	1406	246	1160	17,50%
		Budek	1403	309	1094	22,02%
		Jelek	39764	5509	34255	13,85%
4	Umum	Setan	22267	2231	20036	10,02%
		Iblis	3589	411	3178	11,45%
		Keparat	660	111	549	16,82%

No	Kategori	Kata Kunci	Jumlah Tweet	Kategori		Prosentase Bully
				Bully	Bukan Bully	
		Gembel	10564	1385	9179	13,11%
		Brengsek	5713	1089	4624	19,06%
		Sompret	237	53	184	22,36%
		Bajingan	5588	589	4999	10,54%
5	Sikap	Bejad	1217	392	825	32,21%
Jumlah			169849	25332	144517	

Berdasarkan data pada tabel 4.14, dapat diperoleh informasi sebagai berikut:

1. Jika pada data training kata “tolol” merupakan kata yang memiliki bobot tertinggi sebagai kata yang paling berpotensi untuk melakukan bullying, maka hasil yang sama diperoleh juga pada proses prediksi untuk data yang belum diketahui klasifikasinya. Dari 24 kata kunci yang ditentukan, kata “tolol” memiliki prosentase tertinggi untuk digunakan sebagai kata bullying, yaitu sebesar 71,61%.
2. Kata “jelek” yang memiliki jumlah paling banyak untuk digunakan dalam menuliskan pesan melalui Twitter, ternyata tidak banyak digunakan dalam konteks penyampaian pesan bullying, dengan nilai prosentase 13,85%.
3. Kata “sompret” dan kata “sarap” merupakan jenis kata yang paling jarang digunakan dalam menyampaikan pesan melalui Twitter, ditunjukkan dengan jumlah pesan Tweet yang rendah menggunakan kedua kata tersebut.
4. Kata “idiot” merupakan kata kunci yang memiliki prosentase paling rendah yaitu sebesar 5,66%, sehingga secara umum kata “idiot” tidak banyak digunakan dalam kalimat yang mengandung bullying.

Mengacu pada data yang disajikan pada Tabel 4.14, dapat dianalisa bahwa data tweet dengan jumlah banyak yang diperoleh pada saat melakukan tweet harvesting belum tentu terdeteksi sebagai kata yang digunakan untuk melakukan bullying. Hal ini ditunjukkan pada kata “anjing” yang memperoleh

jumlah tweet tertinggi yaitu sebesar 32.560 tweet, tetapi yang digunakan untuk menyampaikan pesan bullying hanya sebesar 11,23%, sehingga dapat diartikan bahwa kata “anjing” masih banyak digunakan untuk menyampaikan pesan tweet yang tidak mengandung unsur bullying.

Berkebalikan dengan kondisi tesebut, kata “tolol” yang diperoleh sebanyak 3832 tweet, justru menghasilkan prosentase bullying paling besar di antara semua kata kunci, yaitu sebesar 71,61%. Hal ini menunjukkan bahwa ketika orang memilih menggunakan kata “tolol” dalam sebuah pesan tweet yang disampaikannya, orang tersebut mempunyai kecenderungan untuk melakukan bullying.

Dalam keseluruhan data tweet yang diperoleh, muncul juga kata-kata selain 24 kata kunci yang telah ditentukan sebelumnya yang berpotensi untuk digunakan sebagai kata-kata bullying, antara lain: “bangkai”, “tai”, dan “sampah”. Penelitian lebih lanjut diperlukan untuk mengetahui apakah kata-kata tersebut banyak digunakan untuk menyampaikan pesan bullying melalui media sosial Twitter.

(Halaman ini sengaja dikosongkan)

BAB 5
KESIMPULAN

Dari hasil penelitian yang telah dilakukan dalam tesis ini, dapat diambil kesimpulan sebagai berikut:

1. Penelitian yang telah dilakukan berhasil membuktikan bahwa terdapat kata – kata yang mempunyai potensi besar digunakan untuk melakukan aktivitas *cyberbullying*, dan diketahui bahwa kata "tolol" merupakan kata yang mempunyai potensi paling besar tersebut.
2. Pola kalimat yang berpotensi paling besar merupakan kalimat *cyberbullying* adalah "Subyek + Kata Bully", dengan subyek "lo" berpotensi paling besar digunakan untuk melakukan aktifitas cyberbullying.
3. Kata "idiot" merupakan kata kunci bullying dengan prosentase paling rendah untuk digunakan sebagai kata buulying, yaitu sebesar 5,66%.
4. Akurasi tertinggi klasifikasi tweet ke dalam dua kelas (*tweet bullying*, bukan *tweet bullying*) yang bisa dicapai dalam penelitian ini adalah sebesar 87,67 % dengan menggunakan algoritma Naïve Bayes, pembobotan yang digunakan adalah TF-IDF dengan minimal term frequency=3, tanpa melakukan proses stopword, tokenizer yang dipakai adalah= word tokenizer dengan percentage split sebesar 70%.
5. Klasifikasi cyberbullying menggunakan algoritma Naïve Bayes dengan mode test options percentage split dapat menghasilkan nilai akurasi yang paling tinggi jika dibandingkan dengan klasifikasi menggunakan mode test options cross-validations ataupun dengan metode Decision Tree J48.

DAFTAR PUSTAKA

Aliandu, P. 2013. Twitter Used by Indonesian President: An Sentiment Analysis of Timeline. Dalam Information Systems International Conference (ISICO), 2 – 4 December 2013.al. 713-717. Bali: Indonesia.

Berry, M.W. & Kogan, J. 2010. Text Mining Aplication and theory. WILEY : United Kingdom.

Feldman, R & Sanger, J. 2007. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press : New York.

Han, J & Kamber, M. 2006 Data Mining: Concepts and Techniques Second Edition. Morgan Kaufmann publisher : San Francisco.

Margono, Hendro. 2014. Mining Indonesian Cyber Bullying Patterns in Social Networks. Proceedings of the Thirty-Seventh Australasian Computer Science Conference (ACSC 2014), Auckland, New Zealand

Nazief dan Andriani. 1996. Confix Stripping : Approach to Stemming Algorithm for Bahasa Indonesia.Technical report, Faculty of Computer Science, University of Indonesia, Depok, 1996

Pang, B., Lee, L., & Vithyanathan, S. (2002). Sentiment Classification Using Machine Learning Techniques. Dalam Proceedings of The ACL-02 conference on Empirical methods in natural language processing, pp. 79-86. Stroudsburg: Association for computational Linguistic.

Prasad, S. 2011. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods.

Sanchez, Huascar. 2011. Twitter Bullying Detection. Dept of Computer Science UC Santa Cruz Santa Cruz CA.

Sunni, I. & Widyantoro, D. H. 2012. Analisis Sentimen dan Ekstraksi Topik PenentuSentimen pada Opini Terhadap Tokoh Publik

Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands

Wang, A. H. 20100. Don't Follow Me: Twitter Spam Detection. Proceedings of 5th International Conference on Security and Cryptography (SECRYPT) Athens 2010: pp. 1-10. California:IEEE.

BIODATA PENULIS

Penulis yang oleh kedua orang tuanya diberi nama Endah Trihapsari ini lahir pada tanggal 07 Juli 1982 di wilayah kabupaten Malang, propinsi Jawa Timur, sebagai anak bungsu dari tiga bersaudara. Menghabiskan masa kecil di kabupaten Blitar sejak bersekolah TK sampai menamatkan Sekolah Dasar (SD), penulis kemudian melanjutkan Sekolah Lanjutan Tingkat Pertama (SLTP) di wilayah kabupaten Sidoarjo dikarenakan mengikuti kepindahan orang tua. Masih dikarenakan perpindahan tempat tugas orang tua, penulis menamatkan pendidikan pada jenjang Sekolah Menengah Umum (SMU) di kota Malang, dan dilanjutkan sampai ke jenjang pendidikan Diploma 3. Berikutnya penulis melanjutkan pendidikan S1 di sebuah universitas di kabupaten Jombang. Saat ini penulis bekerja sebagai seorang guru di SMKN 1 Dlanggu Mojokerto, dan kemudian penulis mendapatkan kesempatan untuk melanjutkan studi S2 pada bidang keahlian Telematika CIO Institur Teknologi Sepuluh Nopember, Surabaya. Penulis dapat dihubungi pada alamat e-mail: n.trihapsari@gmail.com.