TESIS – TE142599

SENTIMENT ANALYSIS MENGGUNAKAN RULE BASED METHOD PADA DATA PENGADUAN PUBLIK BERBASIS LEXICAL RESOURCES

MASFULATUL LAILIYAH
NRP 2215206714

DOSEN PEMBIMBING
Dr. Surya Sumpeno, ST, M.Sc
Dr. I Ketut Eddy Purnama, ST, MT

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA - CIO
DEPARTEMEN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI ELEKTRO
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

TESIS – TE142599

SENTIMENT ANALYSIS MENGGUNAKAN RULE BASED METHOD PADA DATA PENGADUAN PUBLIK BERBASIS LEXICAL RESOURCES

MASFULATUL LAILIYAH
NRP 2215206714

DOSEN PEMBIMBING
Dr. Surya Sumpeno, ST, M.Sc
Dr. I Ketut Eddy Purnama, ST, MT

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA - CIO
DEPARTEMEN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI ELEKTRO
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

LEMBAR PENGESAHAN

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

Masfulatul Lailiyah
NRP. 2215206714

Tanggal Ujian : 05 Juni 2017
Periode Wisuda : September 2017

Disetujui oleh:

1. Dr. Surya Sumpeno, ST, M.Sc (Pembimbing I)
   NIP: 19690613 199702 1 003
2. Dr. I Ketut Eddy Purnama (Pembimbing II)
   NIP: 19690730 199512 1 001
3. Prof.Dr.Ir. Yoyon Kusnendar S, M.Sc (Penguji)
   NIP: 19540925 197803 1 001
4. Dr. Ir. Endroyono, DEA (Penguji)
   NIP: 19650404 199102 1 001
5. Dr. Eko Mulyanto Yuniarto, ST, MT (Penguji)
   NIP: 19680601 199512 1 009
6. Dr. Istas Pratomo, ST, MT (Penguji)
   NIP: 19790325 200312 1 001

Dekan Fakultas Teknologi Elektro

Dr. Tri Arief Sardjono, S.T., M.T.
NIP. 197002121995121001

iii

PERNYATAAN KEASLIAN TESIS

Dengan ini saya menyatakan bahwa isi keseluruhan Tesis saya dengan judul "**SENTIMENT ANALYSIS MENGGUNAKAN RULE BASED METHOD PADA DATA PENGADUAN PUBLIK BERBASIS LEXICAL RESOURCES**" adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diijinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri.

Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya, Mei 2017

Masfulatul Lailiyah
NRP. 2215206714

Halaman ini sengaja dikosongkan

SENTIMENT ANALYSIS MENGGUNAKAN RULE BASED METHOD PADA DATA PENGADUAN PUBLIK BERBASIS LEXICAL RESOURCES

Nama mahasiswa : Masfulatul Lailiyah
NRP : 2215206714
Pembimbing : 1. Dr. Surya Sumpeno, ST, M.Sc
2. Dr. I Ketut Eddy Purnama, ST, M.Sc

ABSTRAK

Public complaints merupakan salah satu bentuk partisipasi masyarakat dalam mengawasi jalannya pembangunan dan pelaksanaan pelayanan publik. Sesuai undang-undang pelayanan publik no 25 tahun 2009, instansi pemerintah penyedia layanan publik wajib menyediakan wadah untuk menampung aspirasi masyarakat baik melalui media sosial maupun website resmi pemerintah. Informasi yang diperoleh dari pengaduan masyarakat baik topik maupun sentiment dari pengaduan bisa digunakan oleh pemerintah untuk meningkatkan kepuasan masyarakat.

Penelitian mengenai *sentiment analysis* sudah banyak dilakukan, baik menggunakan pendekatan statistik, semantik maupun keduanya. Pendekatan statistik sudah banyak sekali digunakan untuk menganalisa sentiment dari teks, sedangkan pendekatan semantik sedang menjadi *hot topic* saat ini. Pada pendekatan semantik, *lexical resources* merupakan komponen penting dalam menentukan sentiment dari sebuah teks. Salah satu contohnya *Sentiwordnet* dan *Indonesian sentiment lexicon*. Saat ini *lexical resources* dalam bahasa Indonesia mulai berkembang, akan tetapi *lexical resource* yang ada belum memiliki *polarity score* (bobot) yang nantinya bisa digunakan untuk menganalisa tingkat emosi yang terdapat pada teks seperti *Sentiwordnet*. *Sentiwordnet* banyak digunakan dalam *opinion mining* maupun *sentiment analysis* teks dalam bahasa inggris. Pada penelitian ini, kami mencoba memanfaatkan *Sentiwordnet* untuk menganalisa sentiment dari pengaduan masyarakat berbahasa Indonesia serta membandingkannya dengan sentimen leksikon Indonesia. Diperoleh nilai akurasi sebesar 47% untuk data pengaduan pada media *twitter* dan 56.85% untuk data pengaduan pada *media center* ketika menggunakan *Sentiwordnet*. Sedangkan pada penggunaan sentimen leksikon Indonesia diperoleh nilai akurasi sebesar 65.4% untuk data pengaduan pada media *twitter* dan 81.4% untuk data pengaduan pada *media center*.

Kata kunci : sentiment analysis, pengaduan masyarakat, lexical resources, sentiwordnet, indonesian sentiment lexicon

Halaman ini sengaja dikosongkan

SENTIMENT ANALYSIS USING RULE BASED METHOD ON PUBLIC COMPLAINTS BASED ON LEXICAL RESOURCES

By : Masfulatul Liliyah
Student Identity Number : 2215206714
Supervisor(s) : 1. Dr. Surya Sumpeno, ST, M.Sc
2. Dr. I Ketut Eddy Purnama, ST, MT

ABSTRACT

Public complaints were one of the kinds of public participation and awareness to public service implementation. Information from public complaints can be used by the government to improve public satisfaction. In addition, the government can obtain public sentiment from public complaints either on media social or the official government site. Many researches on sentiment analysis has been done, either used statistical method approach, semantic method approach or both. Statistical method approach were widely used. While semantic method approach being hot topic recently. On semantic method approach, lexical resource was an important component to classify sentiment on text. Namely Sentiwordnet and Indonesian sentiment lexicon. Currently, Indonesian sentiment lexicon for sentiment analysis has grown. But the lexicon doesn't have polarity score that can be measure emotion on text like Sentiwordnet. Sentiwordnet has been widely used on researches in English. In this research we apply Sentiwordnet to classify sentiment on Indonesian public complaints with accuracy 47% either on media Twitter and 56.85% on the official government website's data. Furthermore, we compare it with Indonesian sentiment lexicon and get the accuracy 65.4% on media Twitter and 81.4% on the official government website.

Keywords: sentiment analysis, public complaints, lexical resources, sentiwordnet, indonesian sentiment lexicon

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Alhamdulillah, puji syukur kami haturkan kehadirat Alloh SWT atas karunianya sehingga penulis bisa menyelesaikan tesis dengan judul "Sentimen Analysis Menggunakan Rule Based Method Pada Data Pengaduan Publik Berbasis Leksikal Resources". Tesis ini disusun sebagai persyaratan dalam memperoleh gelar Magister Teknik di bidang keahlian Telematika – *Chief Information Officer* (CIO) pada jurusan Teknik Elektro Institut Teknologi Sepuluh Nopember.

Pada kesempatan ini, penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada :

1. Bapak Dr. Surya Sumpeno, ST, M.Sc dan Bapak Dr. I Ketut Eddy Purnama, ST, MT selaku dosen pembimbing atas semua waktu, bimbingan dan arahan selama proses pengerjaan penelitian ini.
2. Bapak Dr. Adhi Dharma Wibawa, ST, MT selaku dosen wali dan koordinator bidang keahlian Magister Telematika – Chief Information Officer (CIO) atas bimbingan dan saran yang telah banyak diajarkan selama proses perkuliahan.
3. Kementrian Komunikasi dan Informatika Bidang Penelitian dan Pengembangan SDM (Litbang) yang telah memberikan kesempatan bagi penulis untuk melanjutkan pendidikan di bidang Teknologi Informasi.
4. Suami atas support, tenaga dan semangat yang selalu diberikan kepada penulis selama penyelesaian studi dan penelitian ini.
5. Teman seangkatan mahasiswa S2 Telematika-CIO dan rekan Laboratorium Human Center (HCCV) atas semua semangat dan support-nya dalam penyelesaian penelitian ini.
6. Serta berbagai pihak yang telah membantu penulis dalam menyelesaikan penelitian ini dan tidak bisa kami sebutkan satu persatu.

Menyadari adanya keterbatasan ilmu dan pengalaman penulis, serta pustaka yang digunakan sebagai acuan dalam penyelesaian penelitian ini. Penulis berharap penelitian ini bisa dikembangkan lebih lanjut agar lebih bermanfaat dan bisa dimanfaatkan di bidang pemerintahan. Untuk itu, penulis sangat mengharapkan

saran dan kritik untuk kesempurnaan tesis ini serta sebagai masukan bagi penulis untuk penelitian di masa yang akan datang.

Akhir kata, penulis berharap tesis ini dapat memberikan manfaat bagi kita semua terutama untuk pengembangan ilmu pengetahuan di kepemerintahan.

Surabaya, Mei 2017
Masfulatul Lailiyah

DAFTAR ISI

Halaman ini sengaja dikosongkan

DAFTAR GAMBAR

Halaman ini sengaja dikosongkan

DAFTAR TABEL

Halaman ini sengaja dikosongkan

DAFTAR VARIABEL

pb	:	jumlah hasil klasifikasi (prediksi) yang sesuai dengan kelas sebenarnya.
num_data	:	jumlah data keseluruhan
total_kelas	:	jumlah keseluruhan data pada kelas tertentu
total_pisah	:	jumlah data yang dipisahkan *classifier* sebagai anggota kelas tertentu
TP (True Positive)	:	jumlah data dari kelas yang diprediksi benar dan hasilnya benar
TN (True Negative)	:	jumlah data dari kelas yang diprediksi salah dan hasilnya salah
FP (False Positive)	:	jumlah data dari kelas yang diprediksi benar tapi hasilnya salah
FN (False Negative)	:	jumlah data dari kelas yang diprediksi salah tapi hasilnya benar
PosScore	:	*positivity score* dari *synset* (kata)
NegScore	:	*negativity score* dari *synset* (kata)
tot_index	:	jumlah index dari *synset* yang dipisahkan oleh karakter # pada *sentiwordnet*
Score	:	bobot *score* dari kata (*term*)
$Frek_{Opini\_word}$	:	frekuensi kemunculan kata yang mengandung opini / sentimen
$Sentence_{score}$	:	*sentiment score* dari kalimat
num_words	:	jumlah kata yang mengandung opini / sentimen dalam sebuah kalimat

Halaman ini sengaja dikosongkan

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Salah satu prinsip dalam mewujudkan pemerintahan yang baik (*good governance*) adalah keterbukaan informasi publik. Hal ini diatur dalam undang-undang pelayanan publik Nomor 25 tahun 2009 pasal 36 dan 37 yang berbunyi "*penyelenggara pelayanan publik wajib menyediakan wadah untuk masyarakat dalam menyampaikan aspirasi dan memberikan masukan terhadap pelayanan yang diberikan*". Mendukung hal tersebut, baik pemerintah pusat maupun daerah berlomba-lomba dalam menyediakan wadah untuk menampung aspirasi masyarakat baik melalui media sosial seperti *twitter*, *facebook* atau website resmi layanan pengaduan terpadu. Pemerintah kota Surabaya juga menyediakan wadah untuk menampung pengaduan dan aspirasi masyarakat kota Surabaya baik melalui twitter maupun website layanan pengaduan yang diberi nama @sapawargasby dan *media center*.

Media center digagas oleh Dinas Komunikasi dan Informatika Pemkot Surabaya untuk menampung partisipasi masyarakat baik dalam bentuk keluhan, informasi maupun saran pada proses pembangunan kota (Humas Pemkot Surabaya, 2015). Selain bertujuan untuk keterbukaan informasi, *media center* juga berfungsi dalam menjembatani partisipasi publik dalam pembangunan daerah. Selain itu, *media center* juga merupakan alat komunikasi dua arah antara masyarakat dan pemerintah dalam mengawasi program pembangunan serta untuk memperbaiki kinerja pemerintah kota Surabaya (Humas Pemkot Surabaya, 2015).

Minat masyarakat dalam berpartisipasi mengawasi jalannya pembangunan daerah cukup besar. Hal ini ditunjukkan oleh data pengaduan yang masuk di *media center* pemkot Surabaya. Berdasarkan data dari Diskominfo jumlah pengaduan terus mengalami peningkatan. Ketika pertama di *launching* pada tahun 2011 sebanyak 698 keluhan. Tahun 2012 bertambah menjadi 2.717 keluhan, tahun 2013 menjadi 4.176 (Dinas Komunikasi dan Informatika, 2015). Semakin hari masyarakat semakin antusias berpartisipasi di *media center*, hingga pada tahun

2014 data keluhan masyarakat mencapai angka 4.298. Sedangkan jumlah *tweet* dan *mention* yang masuk ke twitter @sapawargaSby dari tahun 2011 sampai bulan mei 2016 ini mencapai angka 8.630 tweet. Dengan jumlah *tweet* yang masuk mencapai 200 *tweet*. Informasi yang diperoleh dari pengaduan masyarakat baik topik maupun sentimen dari pengaduan dapat dimanfaatkan oleh pemerintah untuk meningkatkan kepuasan masyarakat (Anggareska & Purwarianti, 2014).

Pengaduan yang masuk pada *twitter* dan website resmi layanan pengaduan terpadu memiliki perbedaan karakteristik dan tata bahasa. Pada website resmi, sebagian besar pengaduan yang masuk menggunakan bahasa semi-formal dan sopan serta informasi pengaduan yang disampaikan jelas. Akan tetapi, pengaduan yang melalui media *twitter* sifatnya lebih bebas baik dari segi pemakaian tata bahasa maupun permasalahan (topik) yang disampaikan. Dari segi tata bahasa sering kita jumpai penggunaan bahasa non-formal, singkatan yang sifatnya tidak baku, mengganti huruf dengan angka, serta tidak semua pengaduan mengandung informasi penting (Naradhipa & Purwarianti, 2011). Seperti mengganti kata ulang dengan angka seperti anak2 , atau menggambarkan perasaan dengan mengulang kata seperti lhooooo. Oleh karena itu, diperlukan tahapan *pre-processing noisy text* (formalization) untuk mengubah kata menjadi bentuk baku dalam bahasa Indonesia (Anggareska & Purwarianti, 2014) (Naradhipa & Purwarianti, 2011) (Lunando & Purwarianti, 2013).

Selama ini pengaduan masyarakat yang masuk direspon oleh staf *media center* dan didistribusikan ke SKPD terkait untuk dilakukan penanganan lebih lanjut. Sedangkan analisa mengenai sentimen masyarakat terhadap permasalahan (topik) yang dilaporkan belum pernah di analisa. Padahal informasi sentimen dari pengaduan yang masuk baik melalui media twitter dan *media center* bisa dimanfaatkan oleh pemerintah untuk meningkatkan kepuasan masyarakat. Analisa sentimen saat ini juga banyak dimanfaatkan oleh perusahaan besar untuk melihat opini pasar terhadap produk mereka dari review produk yang tersebar di forum maupun di media sosial. Hal ini menguntungkan perusahaan, karena tidak perlu membuat kuisioner untuk mengetahui opini *customer* (Naradhipa & Purwarianti, 2011). Oleh karena itu, *sentiment analysis* merupakan bidang yang menarik saat ini.

Sentiment Analysis sendiri merupakan cabang dari *text classification* yang bertujuan untuk mengklasifikasikan sentimen (opini) dari sebuah teks secara otomatis. Apakah teks mengandung opini negatif (*negatif sentiment*), opini positif (*positif sentiment*) atau netral (*netral sentiment*) (Liu, 2010). Fitur dan algoritma merupakan dua hal utama dalam *text classification* (Atmadja & Purwarianti, 2015). Secara garis besar ada dua algoritma di *sentiment analysis*, yakni *rule based method* dan *statistical based method* (Atmadja & Purwarianti, 2015). Pada pendekatan statistik proses pemisahan data sesuai kelas masing-masing menggunakan perhitungan matematis atau dikenal sebagai *machine learning*. Metode yang sering digunakan pada *sentiment analysis* diantaranya *naïve bayes*, *support vector machine* (SVM), dan *maximum enthropy*.

Sedangkan pada pendekatan *rule based method* menggunakan bantuan *human expert* berupa aturan (*rule*) yang digunakan untuk memisahkan data sesuai kelas masing-masing. Pendekatan statistik sudah banyak digunakan untuk menyelesaikan problem *sentiment analysis* saat ini (Manurung & Manurung, 2008) (Akbarisanto, et al., 2016) (Susilawati, 2016) (Fiarni, et al., 2016) (Anastasia & Budi, 2016).

Sama halnya dengan algoritma yang digunakan dalam proses klasifikasi, proses ekstraksi fitur juga merupakan poin utama dalam *sentiment analysis*. Ada dua pendekatan yang bisa digunakan dalam proses ekstraksi fitur, yakni *statistical approach* dan *semantic approach*. Pendekatan statistik memanfaatkan perhitungan statistik untuk ekstraksi fitur, seperti frekuensi kemunculan kata dalam dokumen (TF) atau frekuensi kemunculan kata dalam dokumen terhadap keseluruhan dokumen (TF-IDF). Kedua teknik tersebut digunakan untuk menghitung bobot dari kata (*term*) sebagai fitur dalam *sentiment analysis*. Sama seperti *statistical based method*, pendekatan statistik untuk ekstraksi fitur juga sudah banyak digunakan dan dikembangkan.

Pendekatan lainnya untuk proses ekstraksi fitur adalah pendekatan semantik (makna). Saat ini pendekatan semantik menjadi topik yang sedang hangat dikembangkan (Naradhipa & Purwarianti, 2011) (Lunando & Purwarianti, 2013) (Agarwal, et al., 2016) (Cernian & Sgarciu, 2015). Pada pendekatan semantik lexical resource atau biasa disebut dengan sentimen leksikon memegang peranan

penting untuk proses ekstraksi fitur. Salah satu contohnya adalah *sentiwordnet* dan sentimen leksikon Indonesia.

Sentiwordnet merupakan *public lexical resources* yang dibangun dengan tujuan untuk mendukung *opinion mining* dan *sentiment analysis*. Banyak penelitian mengenai *sentiment analysis* dalam bahasa Inggris menggunakan *sentiwordnet* (Agarwal, et al., 2016) (Cernian & Sgarciu, 2015). Sedangkan untuk teks bahasa Indonesia *sentiwordnet* mulai dikembangkan sebagai penunjang dalam *text classification*. Selain *sentiwordnet*, kita juga bisa menggunakan sentimen leksikon Indonesia untuk menganalisa sentimen pada teks bahasa Indonesia.

Saat ini sentiment leksikon dalam bahasa Indonesia sudah mulai dikembangkan (Naradhipa & Purwarianti, 2011) (Vania, et al., 2014) (Wicaksono, et al., 2014). Akan tetapi, leksikon tersebut belum memiliki *polarity score* yang nantinya bisa digunakan untuk mendeteksi emosi pada teks seperti Sentiwordnet. Selain itu, sentimen leksikon Indonesia masih sedikit yang bersifat publik. Banyak penelitian di bidang *sentiment analysis* dan *emotion classification* menggunakan *Sentiwordnet* dalam bahasa inggris. Beberapa penelitian dalam bahasa Indonesia juga memanfaatkan *Sentiwordnet* sebagai penunjang (Lunando & Purwarianti, 2013) (Vania, et al., 2014).

Pada penelitian ini, dilakukan *sentiment analysis* pada pengaduan publik yang masuk melalui media *twitter* dan website resmi layanan pengaduan terpadu menggunakan pendekatan semantik, dimana *sentiwordnet* dan sentimen leksikon Indonesia digunakan untuk ekstraksi fitur. Metode yang digunakan adalah *rule based method*. Fitur yang digunakan dalam penelitian ini adalah *unigram* (kata) yang sesuai dengan *sentiwordnet* dan sentimen lexicon Indonesia. Karena data yang digunakan dalam bahasa Indonesia, sedangkan *sentiwordnet* menggunakan bahasa Inggris. Maka diperlukan proses *translate* untuk mengubah data ke bahasa inggris menggunakan *API Google Translate* (Lunando & Purwarianti, 2013). Penggunaan pendekatan semantik untuk *sentiment analysis* pada data pengaduan masyarakat diharapkan bisa meningkatkan akurasi hasil klasifikasi.

1.2 Rumusan Masalah

Permasalahan yang muncul dari uraian latar belakang diatas adalah masih minimnya sentimen leksikon dalam bahasa Indonesia. Baik sentimen leksikon yang memiliki *polarity score* seperti *Sentiwordnet* maupun yang bersifat publik (*free*).

1.3 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah menganalisa sentimen dari pengaduan publik menggunakan pendekatan semantik (makna kata), dengan *rule based* sebagai metode klasifikasi dan *lexical resources* untuk ekstraksi fitur. Serta untuk membandingkan tingkat kesesuaian penggunaan sentimen leksikon Inggris yakni *Sentiwordnet* dan sentimen leksikon Indonesia untuk proses ekstraksi fitur dalam *sentiment analysis* pengaduan publik berbahasa Indonesia. Dengan mengetahui tingkat kesesuaian penerapan *Sentiwordnet* pada bahasa Indonesia, diharapkan bisa digunakan untuk menganalisa emosi yang terdapat pada teks kedepannya.

1.4 Batasan Masalah

Data yang digunakan dalam penelitian ini hanyalah data pengaduan masyarakat yang masuk melalui akun *twitter* (@sapawarga) dan *media center* Surabaya dalam bentuk teks dan menggunakan bahasa Indonesia.

1.5 Kontribusi

Penelitian ini diharapkan dapat berkontribusi dalam menambah *lexical resource* yang dapat digunakan untuk *sentiment analysis* dan *emotion classification* pada teks bahasa indonesia.

1.6 Sistematika Penulisan

Bab I : Bab ini merupakan bagian yang akan menguraikan latar belakang permasalahan yang diangkat dan pengambilan judul penelitian, rumusan masalah yang diangkat dalam penelitian ini, tujuan

penelitian, batasan masalah, kontribusi dan metodologi skema penulisan penelitian.

Bab II : Bab ini berisi tentang tinjauan pustaka yang digunakan sebagai acuan dalam membangun sistem untuk menyelesaikan permasalahan yang diuraikan dalam bab pertama.

Bab III : Bab ini berisi tentang uraian alur metodologi dan metode yang digunakan untuk membangun sistem klasifikasi secara rinci dan metode pengujian dan validasi yang digunakan dalam penelitian ini.

Bab IV : Bab ini menyajikan pembahasan hasil yang diperoleh dan model klasifikasi yang dibangun dalam penelitian ini. Serta evaluasi dan pengujian sistem yang telah dihasilkan untuk melihat kesesuaian dengan tujuan yang diharapkan.

Bab V : Bab ini berisi kesimpulan dari hasil penelitian yang sudah dilakukan dan saran untuk perbaikan penelitian ini kedepannya.

BAB 2
KAJIAN PUSTAKA

2.1 Pengaduan Publik

Menurut PERMENPAN Nomor 3 Tahun 2015, definisi pengaduan adalah sebuah tindakan untuk menyampaikan keluhan terhadap ketidak sesuaian pelayanan yang diberikan oleh penyelenggara pelayanan publik dengan standar pelayanan. Disebutkan juga pada Undang-undang Pelayanan Publik Nomor 25 tahun 2009 pasal 36 dan 37 bahwa penyelenggara pelayanan publik wajib menyediakan wadah untuk masyarakat dalam menyampaikan aspirasi dan memberikan masukan terhadap pelayanan yang diberikan.

Dengan adanya wadah untuk menuangkan aspirasi, dapat mendorong partisipasi dan peran aktif masyarakat dalam proses pengambilan kebijakan publik. Serta penyelenggaraan Negara akan lebih transparan, sesuai dengan tujuan undang-undang keterbukaan informasi publik nomor 14 tahun 2008. Pengelolaan pengaduan publik diatur dalam permenpan nomor 3 tahun 2015, dimana pengaduan harus dikelola dengan baik dengan cara menyediakan sarana pengaduan berupa aplikasi pengelolaann pengaduan pelayanan publik, serta menugaskan pelaksana yang berkompeten untuk menindaklanjuti pengaduan tersebut.

Jenis Pengaduan atau komplain sangat beragam. Komplian dilihat dari sisi pelaksanaan program pembangunan, bisa dibedakan menjadi :

a. Komplain yang didasarkan oleh kebutuhan informasi.
b. Komplain yang didasarkan pada keinginan melakukan upaya perbaikan atau penyampaian ide, usulan/gagasan.
c. Komplain dikarenakan pelanggaran aturan main.

2.2 Media Social (*Twitter*)

Twitter merupakan salah satu jenis *microblogging* yang banyak digunakan untuk berkomunikasi, mengungkapkan isi hati, *sharing* informasi baik berupa kata maupun gambar dan video. *Twitter* berisi pesan singkat dengan jumlah karakter

maksimal 140 karakter biasa disebut dengan *tweet* (kicauan). *Tweet* bisa berisi pesan, web *link*, *mention* (@) atau *hashtag* (#). *Mention* berarti menyebutkan pengguna lain dalan *tweet* kita, sedangkan *hashtag* berarti *tweet* kita terhubung dengan suatu topik tertentu di *twitter*:

Twitter merupakan jenis *social media* satu arah, dimana pengguna yang mem-*follow* twitter kita bisa melihat semua kicauan kita di *twitter*. Tetapi tidak berlaku sebaliknya. kita tidak bisa melihat isi *tweet* dia. Untuk mendapatkan data dari *twitter*, kita bisa menggunakan *tools* #TAGS berdasarkan kata kunci yang diinginkan.

Tata bahasa yang digunakan dalam menuliskan pesan di *twitter* bersifat bebas, tidak mengikat [6]. Oleh karena itu, sering kita jumpai penggunaan kata yang tidak standar di isi *twitter*. Seperti singkatan kata yang tidak standar (misal : ank berarti anak), mengganti kata berulang dengan angka (misal : ga2l berarti gagal), mengulang penggunaan huruf tertentu untuk mengekspresikan kalimat (misal : Surabaya kota ramah anak lhoooo).

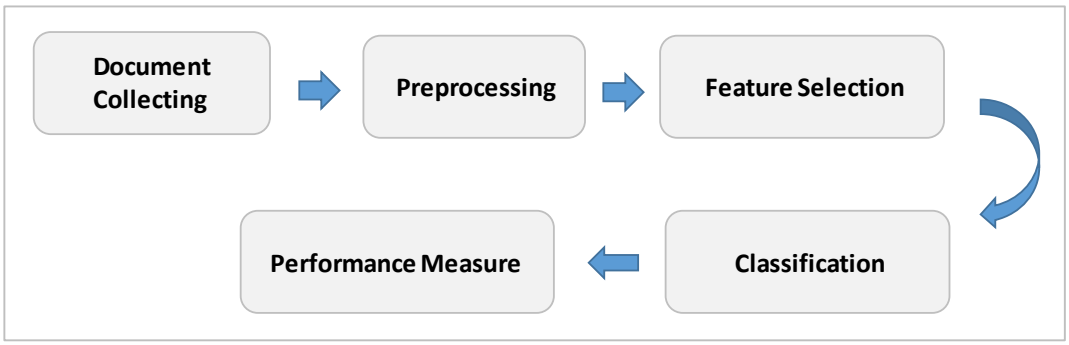
Sebelum pengolahan lebih lanjut, data *twitter* terlebih dahulu harus melalui tahapan *pre-processing noisy text* untuk mengubah kata menjadi bentuk standar sesuai kamus bahasa indonesia seperti yang sudah dilakukan oleh Aqsath (Naradhipa & Purwarianti, 2011), Edwin (Lunando & Purwarianti, 2013), dan Dekha (Anggareska & Purwarianti, 2014).

2.3 Text Classification

Dengan semakin banyaknya dokumen elektronik dan informasi berupa teks dari berbagai sumber, text mining menjadi bidang yang potensial untuk dipelajari. Text mining adalah ilmu yang mempelajari cara meng-ekstrak informasi dan mencari pola dari sebuah dokumen secara otomatis (Korde & Mahender, 2012). Prinsip kerja text mining sama dengan data mining, yakni melakukan penambangan informasi dari sekumpulan data berukuran besar. Hanya saja data yang digunakan pada text mining berupa teks. Ada banyak penerapan text mining seperti klasifikasi text (*text classification*), *information retrieval*, *topic clustering*, ekstraksi topik (*topic extraction*), *document summary* dan *sentiment analysis*.

Klasifikasi teks adalah proses untuk mengelompokkan sebuah dokumen ke dalam kategori atau kelas tertentu yang sudah ditentukan. Misalkan d_i adalah sebuah dokumen dari kumpulan dokumen (D), dan {C_1, C_2, C_3, C_4 dan C_5} adalah kelas/kategori yang ditentukan. Maka klasifikasi teks bertugas untuk menentukan kategori (C_i) dari masing-masing dokumen (d_i) berdasarkan karakteristik yang dimiliki masing-masing kelas. Jika setiap dokumen hanya dimasukkan ke salah satu kelas, maka klasifikasi tersebut dinamakan *single label*. Akan tetapi jika satu dokumen bisa dimasukkan lebih dari satu kelas, maka dinamakan *multi label*.

Sama halnya dengan data mining, tahapan proses pada klasifikasi teks meliputi *document collecting*, *pre-processing*, ekstraksi fitur, klasifikasi, dan validasi seperti ditunjukkan oleh gambar 2.1. *Document collecting* merupakan tahapan pengumpulan data yang akan digunakan, selanjutnya tahapan *preprocessing* yang bertujuan untuk menormalisasikan data dan membuang karakter yang tidak digunakan pada proses klasifikasi selanjutnya. Sedangkan tahapan *feature selection* bertujuan untuk mengekstraksi fitur yang digunakan sebagai ciri dari data. Fitur yang diperoleh selanjutnya digunakan pada tahapan klasifikasi. Tahapan terakhir adalah *performance measure* yang bertujuan untuk mengukur tingkat keakuratan *classifier* (pemisah) dan model klasifikasi yang dihasilkan dalam memisahkan data masing-masing kelas. Selain itu, pada tahapan *performance measure* kita bisa melihat adanya penyimpangan data dalam klasifikasi dengan melihat perolehan nilai *precision* dan *recall*.



Gambar 2.1 Tahapan proses pada klasifikasi teks (sumber dari jurnal Vandana and C Namrata)

Pada *text* classification, data yang digunakan adalah data teks yang memiliki sifat tidak terstruktur dan lebih komplek, maka problem utama dari

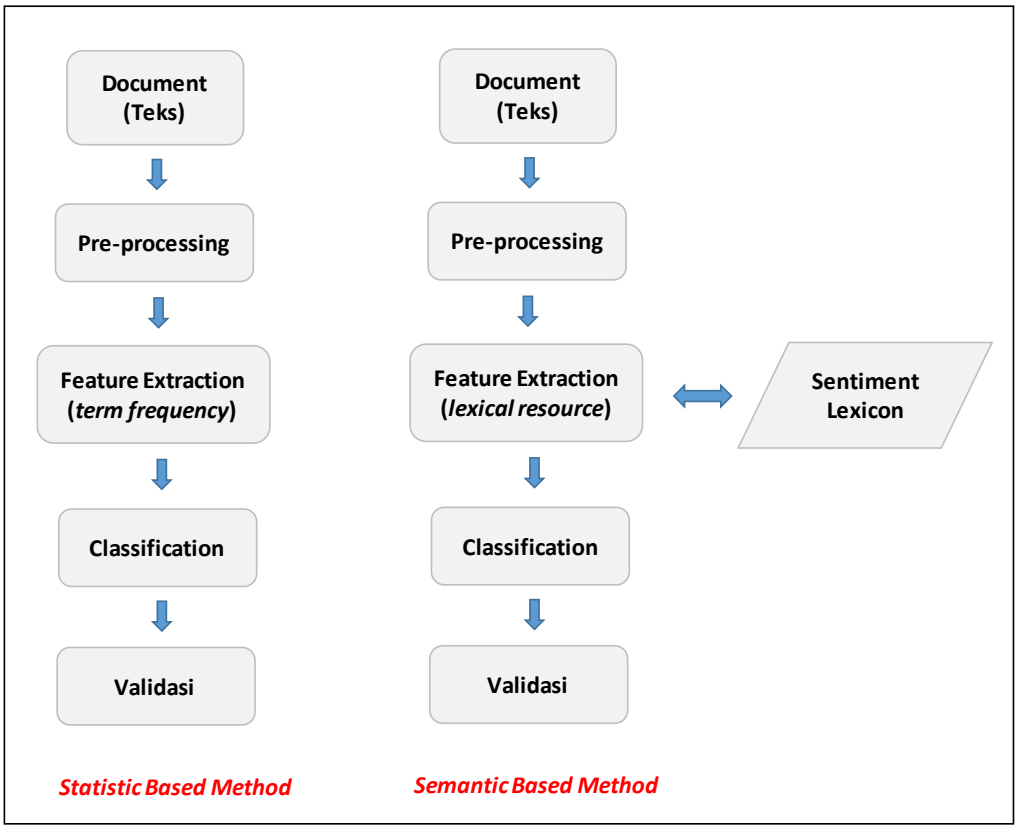
klasifikasi text diantaranya menghasilkan data berdimensi tinggi, karena satu kata mewakili satu dimensi. Problem kedua adanya frase yang berbeda dengan kata dasarnya, seperti kata "terima kasih" bermakna ucapan syukur atas bantuan seseorang. Sedangkan kata penyusunnya memiliki makna yang berbeda, yakni "terima" bermakna mendapatkan sesuatu dan "kasih" bermakna perasaan sayang. Problem terakhir adalah adanya makna ambigu untuk kata yang sama.

Pada tahapan klasifikasi ada 3 teknik dasar yang bisa digunakan, yakni *unsupervised learning*, *supervised learning* dan *semi supervised learning*. Ciri khas dari *supervised learning* adalah *output* kelas yang diharapkan sudah ditentukan dari awal. Dan data yang digunakan dibagi menjadi data *training* dan data *testing*. Sedangkan pada *semi supervised* output kelas yang diharapkan tidak ditentukan diawal, akan tetapi data yang digunakan untuk proses *training* adalah data berlabel. Teknik terakhir yakni *unsupervised learning* memiliki ciri output kelas tidak diketahui dan data yang digunakan tidak berlabel. Jadi teknik ini mencoba menemukan fitur yang tersembunyi pada data tanpa label.

2.4 Sentiment Analysis

Sentiment analysis atau dikenal juga dengan sebutan *opinion mining* merupakan cabang dari klasifikasi teks. Tujuan dari *sentiment analysis* adalah mengelompokkan teks (dokumen) yang mengandung opini sebagai *positive sentiment*, *negative sentiment*, atau netral (Liu, 2010). Dalam *sentiment analysis* fitur yang dihasilkan dan algoritma klasifikasi menjadi poin utama (Atmadja & Purwarianti, 2015). Ada dua pendekatan (algoritma) yang sering digunakan yakni *rule based method* dan *statistical based method*. Pada umumnya, *rule based method* berbasis semantik (*semantic based*). Yakni memanfaatkan *lexical resource* untuk proses ekstraksi fitur. Sedangkan pada *statistic based method* menggunakan perhitungan statistik otomatis dalam ektraksi fitur. Pendekatan statistik sudah banyak digunakan, sedangkan pendekatan semantik sedang hangat dikembangkan dengan harapan akurasi yang diperoleh lebih tinggi dari pendekatan statistik. Karena pendekatan semantik menggunakan makna kata untuk ektraksi fitur. Pada pendekatan ini, *lexical resource* (sentimen lexicon) memegang peranan penting (Vania, et al., 2014). Semakin lengkap sentiment lexicon yang digunakan, maka

hasil yang diperoleh akan semakin tepat. Sentimen lexicon berisi daftar kata-kata dengan *polarity* positif seperti "baik", "aman" atau *polarity* negatif seperti "lambat", "buruk". Gambar 2.2 menyajikan tahapan yang ada pada pendekatan *statistical based method* dan *semantic based method*. Secara keseluruhan tahapan pada *statistic based method* dan *semantic based method* hampir sama, yakni meliputi *pre-processing*, *feature extraction*, *classification* dan validasi. Perbedaan antara *statistic based method* dan *semantic based method* terletak pada tahapan *feature extraction*. Pada pendekatan statistik, proses *feature extraction*-nya memanfaatkan perhitungan matematis / statistik seperti frekuensi kemunculan kata (*term*) pada dokumen yang biasa disebut dengan istilah *term frequency* (TF) atau kemunculan kata (*term*) terhadap keseluruhan dokumen yang biasa disebut dengan istilah *term frequency – inverse document frequency* (TF-IDF). Sedangkan pada pendekatan semantik, proses *feature extraction*-nya memanfaatkan *lexical resource* / *sentiment lexicon* seperti terlihat pada Gambar 2.2.



Gambar 2.2 Perbedaan tahapan *statistik based method* dan *semantik based method*

2.4.1 *Preprocessing Data*

Tahapan ini merupakan proses ekstraksi teks, dari data tidak terstruktur diubah menjadi data yang terstruktur agar bisa diolah lebih lanjut untuk proses klasifikasi. Tujuan dilakukannya *pre-processing* untuk mempersiapkan dokumen teks menjadi data yang siap diolah pada tahapan selanjutnya dan untuk mengurangi *noise*. Tahapan yang ada pada preprocessing adalah *tokenizing* (memotong kata), *formalization* (mengubah ke bentuk standar sesuai kamus KBBI), *translate*, *pos tagging*, *filtering* (membuat *stopword*), *stemming* (mengubah ke bentuk dasar).

2.4.1.1 Cleansing Data

Tahapan *cleansing data* bertujuan untuk membersihkan data dari karakter-karakter yang tidak berpengaruh dalam proses klasifikasi dan membuang adanya duplikasi data. Pada tahapan ini dilakukan pemenggalan setiap kata yang menyusun sebuah dokumen / kalimat, atau biasa disebut dengan istilah *tokenizing*. Hasil dari proses *tokenizing* adalah daftar kata yang berdiri sendiri yang menyusun kalimat dalam sebuah dokumen.

Karena data twitter banyak mengandung *noise* seperti singkatan (*abbreviation*), penggunaan angka yang menggantikan huruf, *mention* (@), *hashtag* (#), kode html (http://www) dan memposting ulang *tweet* orang lain dengan tujuan membagikan kembali informasi tersebut (*retweet*). Maka sebelum melakukan pemisahan setiap kata dalam kalimat, terlebih dahulu perlu dilakukan tahapan *retweet removal* (RT) untuk menghindari adanya duplikasi data, sehingga diperoleh data yang lebih akurat.

Tahapan selanjutnya untuk menghilangkan *noise* adalah membuang *mention*, *hashtag*, dan *html* yang terdapat pada data *tweet*. Karena ketiga karakter tersebut tidak berpengaruh dalam proses klasifikasi. Dan terakhir menghapus angka yang terdapat pada data, hanya karakter huruf saja yang akan diproses lebih lanjut. Setelah semua tahapan diatas dilakukan, selanjutnya semua huruf diubah menjadi huruf kecil. Dan dilakukan pengecekan apakah ada kalimat yang sama di database, jika ada maka data ditolak untuk menghindari duplikasi data. Pada umumnya setiap kata dipisahkan oleh spasi atau *delimimter* seperti titik, atau tanda koma.

2.4.1.2 Formalization

Formalization merupakan tahapan untuk mengubah kata yang tidak standar ke bentuk standar sesuai dengan struktur KBBI (Kamus Besar Bahasa Indonesia). Dalam isi pesan *twitter* sering kita jumpai penulisan yang tidak sesuai dengan standar KBBI seperti penggunaan singkatan yang tidak baku, menggunakan bahasa lokal, mengganti huruf dengan angka (misal : ber2, anak2), menambahkan karakter huruf untuk mengekspresikan isi pesan (misal : Surabaya kora ramah anak lhooo). Oleh karena itu, diperlukan adanya kamus tersendiri untuk mengubah struktur kata menjadi bentuk standar sesuai KBBI. Kami membuat kamus lokal untuk mengubah kata yang tidak sesuai standar KBBI ke bentuk baku. Selain menggunakan kamus lokal, proses *formalization* juga menggunakan algoritma *laveinsten distance* untuk mendeteksi adalah penambahan karakter atau karakter yang hilang. Contoh daftar kata yang ada di kamus lokal ditunjukkan oleh Tabel 2-1. Kolom informal berisi daftar kata lokal yang sering digunakan pada *twitter*, sedangkan kolom formal berisi daftar kata baku sesuai kaidah kamus besar bahasa Indonesia (KBBI) dari kata lokal di kolom informal.

Tabel 2-1 Daftar kata pada kamus lokal.

informal	**formal**	**informal**	**formal**
takon	tanya	banget	sangat
rek	teman	buk	ibu
lyn	angkutan umum	buesar	besar
niku	itu	dadi	jadi
warnae	warnanya	gag, ga, g, gk	tidak
iki	ini	monggo	silahkan

2.4.1.3 Part of Speech (POS) Tagging

Part of speech (POS) *tagging* adalah proses untuk menentukan kelas kata (*part of speech*) dalam sebuah kalimat (Pisceldo, et al., 2009). Kelas kata *(part of speech)* merupakan bagian dari gramatikal. Bahasa Indonesia memiliki lima (5) jenis kelas kata meliputi kata kerja (*verb*), kata benda (*noun*), kata sifat (*adjective*), kata keterangan (*adverb*), kata tugas (*function*) (Pisceldo, et al., 2009). Dalam

sentiment analysis, pada umumnya kelas kata yang digunakan adalah kata sifat (*adjective*), kata keterangan (*adverb*), kata benda (*noun*), dan kata kerja (*verb*). Karena kelas kata tersebut sebagian besar mengandung sentimen (emosi), terutama kata sifat (*adjective*).

Permasalahan utama dalam *pos tagging* adalah pada kebanyakan bahasa, kata - kata berperilaku berbeda dalam konteks berbeda. Seperti kata "padat" pada kalimat "Kondisi jalanan Surabaya di siang hari sangat padat" yang memiliki kelas kata sifat (*adjective*). Sedangkan pada kalimat "Ramadhan ini, ibu walikota jarang buka bersama keluarga. Karena jadwal yang sangat padat", kata padat memiliki kelas kata sifat (*adverb*).

Oleh karena itu, tantangan dalam *pos tagging* adalah bagaimana kita bisa mengindentifikasi kelas kata dengan benar dalam konteks tertentu. *Pos tagger* untuk bahasa Indonesia masih sangat terbatas dan sifatnya belum tersedia bebas *source code*-nya. Berbeda dengan *pos tagger* bahasa Inggris yang sudah beredar luas dan banyak digunakan dalam penelitian di bidang *natural language processing* (NLP), salah satunya adalah *English Stanford Tagger*.

2.4.1.4 Filtering

Setelah didapatkan daftar kata penyusun dokumen pada proses *tokenizing*. Selanjutnya dilakukan tahapan *filtering*. *Filtering* merupakan proses pengambilan kata-kata yang penting, dimana nantinya kata-kata tersebut digunakan pada proses klasifikasi. Proses filtering bisa menggunakan algoritma *stoplist* yakni dengan membuang kata-kata yang dianggap kurang penting dan tidak memiliki pengaruh pada proses klasifikasi, atau algoritma *wordlist* yakni menyimpan kata-kata yang penting saja yang bisa menjadi penciri dari dokumen. *Stoplist / stopword* adalah kata-kata yang tidak deskriptif dan tidak memiliki berpengaruh jika dihilangkan. Yang termasuk *stopword* dsini diantaranya kata sambung, partikel, dan preposisi. Seperti kata "yang", "saya" dan "sering". Seringkali dalam dokumen, *stopword* justru lebih sering muncul dibandingkan kata yang menjadi penciri dari dokumen tersebut. Oleh karena itu, *stopword* harus dibuang agar tidak keluar sebagai kata yang mewakili dokumen. Daftar kata yang termasuk dalam *stopword* disimpan

dalam *stoplist*. Tabel 2-2 menyajikan contoh kata yang masuk dalam *stoplist* bahasa Indonesia berdasarkan (Tala, 2003).

Tabel 2-2 Daftar *stopword* Tala

stopword	**stopword**	**stopword**	**stopword**	**stopword**
yang	di	dan	itu	dengan
untuk	tidak	ini	dari	dalam
akan	pada	juga	saya	ke
karena	tersebut	bisa	ada	merela
lebih	kata	tahun	sudah	atau
saat	oleh	menjadi	orang	ia
telah	adalah	seperti	sebagai	bahwa
dapat	para	harus	namun	kita

2.4.1.5 Stemming

Proses *stemming* merupakan tahapan yang bertujuan mengubah kata ke bentuk dasarnya sesuai kamus besar bahasa Indonesia (KBBI) dengan cara membuang imbuhan. Tujuan dari *stemming* adalah mengurangi kata yang bermakna jamak dan kata tunggal (Manning 2008). Selain itu, tahapan *stemming* juga bertujuan untuk mengurangi dimensi dari data teks. Karena kata berimbuhan dikembalikan ke bentuk dasarnya, dan dianggap memiliki makna yang sama. Data teks yang digunakan dalam penelitian ini adalah teks berbahasa Indonesia maka proses *stemming* dilakukan dengan cara menghilangkan imbuhan seperti awalan (prefiks), sisipan (infiks), akhiran (suffiks) dan kombinasi awalan dan akhiran (*confiks*). Algoritma yang sering digunakan untuk proses *stemming* teks bahasa Indonesia adalah Nazief dan Adriani. Pada umumnya kata dasar pada bahasa Indonesia terdiri dari kombinasi :

Prefiks 1 + Prefiks 2 + Kata Dasar + Sufiks 3 + Sufiks 2 + Sufiks 1

Algoritma Nazief dan Adriani dibuat oleh Bobby Nazief dan Mirna Adriani. Tahapan algoritma *stemming* Nazief dan Adriani adalah :

1. Langkah awal cari kata dalam kamus kata dasar. Jika ditemukan, maka dianggap kata tersebut adalah kata dasar dan algoritma berhenti.
2. Menghapus Inflection Suffiks seperti ‘-lah’, ‘-kah’, ‘-mu’, ‘-ku', ‘-nya’, ‘-pun’. Jika ditemukan partikel ‘-lah’, ‘-kah’, ‘-nya’ setelah menghapus partikel tersebut, hendaknya mengulangi langkah ini lagi untuk mengecek apakah ada possessive pronoun seperti ‘-ku’, ‘-mu’, ‘-nya’.
3. Menghapus derivation suffiks seperti ‘-i’, ‘-an’, ‘-kan’. Setelah menghapus akhiran, lakukan pengecekan kata di kamus. Jika ditemukan algoritma berhenti. Jika tidak, maka :
   a. lakukan pengecekan huruf terakhir setelah akhiran dihapus. Jika huruf terakhir adalah ‘k’, maka hapus huruf ‘k’ dan lakukan pengecekan kembali kata di kamus. Jika ditemukan, maka algoritma berhenti. Jika tidak lakukan langkah pada 3.b
   b. kembalikan akhiran yang telah dihapus. Setelah itu lanjutkan ke langkah 4 (keempat).
4. menghapus derivation prefiks. Untuk menghapus prefiks ada beberapa pengecekan, yakni :
   a. lakukan pengecekan pada tabel kombinasi awalan-akhiran yang tidak diijinkan. Jika ditemukan, maka algoritma berhenti. Jika tidak lakukan langkah 4.b
   b. Lakukan pengecekan awalan. Jika terdapat awalan, maka hapus awalan tersebut. Lakukan pengecekan ini sebanyak 3 kali. Selanjutnya lakukan pengecekan kata pada kamus. Jika ditemukan, maka algoritma berhenti. Jika tidak, maka dianggap kata tersebut adalah kata dasar. Dan disimpan dalam kamus kata dasar.

Daftar awalan-akhiran yang tidak diijinkan ditunjukkan oleh Tabel 2-3. Salah satunya awalan *be-* dengan akhiran *–i* tidak diijinkan untuk digunakan bersama-sama.

Tabel 2-3 Daftar awalan dan akhiran yang tidak diijinkan

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

2.4.2 Proses Ekstraksi Fitur

Feature extraction atau ekstraksi fitur merupakan suatu proses pengambilan ciri / *feature* dari sebuah objek yang akan digunakan dalam proses klasifikasi selanjutnya. Sedangkan definisi fitur sendiri adalah karakteristik unik yang bisa digunakan untuk mewakili sebuah objek. Dalam *text classification* tahapan ekstraksi fitur memegang peranan penting. Karena tahapan ini berperan untuk menentukan fitur mana yang akan digunakan untuk proses klasifikasi dan fitur mana yang akan diabaikan. Untuk setiap kasus klasifikasi yang berbeda pemilihan fitur yang relevan juga berbeda. Misalnya, pada kasus klasifikasi website yang mengandung konten negatif fitur yang dianggap relevan adalah kata (*term*) yang mengandung makna konotasi negatif seperti "seks", "adult", dan "free". Sedangkan pada kasus klasifikasi sentimen pada teks, fitur yang dianggap relevan adalah kata (*term*) yang mengandung opini / sentimen seperti "bagus", "baik", dan "buruk".

Dalam klasifikasi teks, ada dua pendekatan yang bisa digunakan untuk ekstraksi fitur yakni pendekatan statistik (*statistical based method*) dan pendekatan semantik (*semantic based method*). Pada pendekatan statistik, pembobotan fitur memanfaatkan perhitungan matematis atau statistik. Seperti menggunakan perhitungan frekuensi kemunculan kata (*term*) yang biasa disebut dengan *term frekuensi* (TF) maupun *term frekuensi inverse document frequency* (TF-IDF). Pada perhitungan *term frekuensi* (TF) kata yang sering muncul dianggap sebagai ciri / fitur yang mewakili sebuah dokumen dan mendapatkan bobot yang tinggi. Sedangkan pada TF-IDF, frekuensi kemunculan kata pada sebuah dokumen juga

dibandingkan dengan kemunculan kata pada keseluruhan dokumen. Dimana semakin sering kata tersebut muncul dibanyak dokumen maka kata tersebut tidak dijadikan ciri / fitur. Hal ini dikarenakan, kata tersebut dianggap bersifat umum, sehingga tidak bisa mewakili sebuah dokumen.

Pada pendekatan semantik, proses ekstraksi fitur memanfaatkan makna dari kata penyusun kalimat. Dalam *sentiment analysis* yang menjadi fitur adalah kata yang mengandung sentimen (*opinion word*) baik sentimen positif maupun sentimen negatif. Disini *lexical resources* atau *sentiment lexicon* menjadi poin penting untuk menemukan kata (*term*) yang menjadi fitur dari kelas yang telah ditentukan. Sebuah kata dikatakan mengandung sentimen (*opinion word*), jika kata tersebut mengekspresikan perasaan yang diinginkan (sentimen positif) maupun tidak diinginkan (sentimen negatif) (Liu, 2010). Salah satu contoh kata yang mengandung sentimen positif adalah "cantik", "hebat", "bagus" dan "menakjubkan". Sedangkan contoh kata yang mengandung sentimen negatif adalah "jelek", "miskin" dan "buruk".

2.4.3 Metode Klasifikasi

Klasifikasi adalah proses menemukan model untuk membedakan kelas yang satu dengan kelas yang lain, dengan harapan model tersebut bisa digunakan untuk memprediksi kelas dari objek yang belum diketahui kelasnya. Secara garis besar metode klasifikasi yang bisa digunakan untuk menyelesaikan problem *sentiment analysis* maupun *opinion mining* adalah *rule based method*, *statistical based method* atau mengkombinasikan keduanya. *Statistical based method* memanfaatkan perhitungan matematis untuk memisahkan data sesuai kelas masing-masing atau biasa disebut dengan *machine learning*. Metode ini telah banyak digunakan dalam klasifikasi teks, begitu pula pada *sentiment analysis*. Pada pendekatan statistik metode yang sering digunakan diantaranya *naïve bayes*, *decision tree*, *support vector machine* (SVM), *maximum enthropy*, dan lain sebagainya. Masing-masing metode memiliki karakteristik dan teknik yang berbeda satu dengan lainnya dalam memisahkan data sesuai kelasnya. Misalkan pada metode SVM keunggulannya adalah cocok digunakan pada data yang berdimensi tinggi.

Sedangkan *rule based method* menggunakan koleksi aturan (*rule*) kondisi yang sesuai dengan label kelas yakni "if .. then .." dalam mengklasifikasikan data. Pendekatan *rule based* cocok jika diaplikasikan dalam klasifikasi sederhana. Tetapi ketika data yang diklasifikasikan kompleks dan bervariasi, pendekatan ini kurang cocok digunakan dan tidak mungkin untuk melakukannya dengan baik.

2.5 Definisi Lexical Resource

Terdapat tiga pendekatan yang bisa digunakan untuk mengumpulkan *sentiment lexicon*, yakni *manual approach*, *dictionary based approach* dan *corpus based approach*. Pendekatan kamus (*dictionary based approach*) memanfaatkan relasi kata (*synset*) seperti sinonim, antonim, hipernim dan hiponim pada *wordnet* untuk memperoleh *opinion word* lainnya. Sedangkan pada pendekatan corpus, memanfaatkan kumpulan *opinion word* sebagai benih dan pola sintaksis dari benih kata untuk menambang *opinion word* pada *corpus* yang besar. Pada penelitian ini kami menggunakan *Sentiwordnet* dan sentimen leksikon Indonesia untuk ekstraksi fitur, yakni kata bersentimen (*opinion word*). *Sentiwordnet* dibangun menggunakan pendekatan kamus (*dictionary based approach)*. Sedangkan untuk sentimen leksikon Indonesia, Aqsath dan Ayu (Naradhipa & Purwarianti, 2011) menggunakan pendekatan manual dan Vania dkk (Vania, et al., 2014) menggunakan pendekatan *corpus* (*corpus based approach*).

2.5.1 Sentiwordnet

Sentiwordnet adalah *lexical resource* yang didesain untuk mendukung *opinion mining* dan *sentiment analysis* dan bersifat publik (*free*) (Cernian & Sgarciu, 2015). *Sentiwordnet* merupakan perluasan dari *wordnet*, yang bertujuan mengaitkan seluruh *synset* dari *wordnet* dengan sentimen positif, negatif atau obyektif (Kreutzer & Witte, 2013). Setiap *synset* dari *wordnet* diberi label dengan nilai dari 0.0 hingga 1.0 untuk masing-masing kategori. Dan hasil penjumlahan dari ketiga (3) kategori tersebut selalu bernilai 1.0 seperti ditunjukkan oleh Persamaan 2.1. Sehingga setiap *synset* memungkinkan bernilai 0 pada salah satu atau lebih kategori. Karena *synset* bisa bernilai positif, negatif atau obyektif tergantung pada konteks dari kalimat (Kreutzer & Witte, 2013).

Sentiwordnet memperluas kegunaan dari *wordnet* di bidang *opinion mining* dan *sentiment analysis*. Dimana struktur hirarki dari *synset* dan *glosses* pada *sentiwordnet* masih mempertahankan sesuai *wordnet*. Informasi dalam *sentiwordnet* disajikan dengan format berikut ini (Cernian & Sgarciu, 2015):

- #POS : berisi kelas (*part of speech*) dari *synset* (kata).
- #ID : berisi kode unik untuk setiap *synset* dengan kelas kata tertentu.
- #PosScore : berisi nilai sentimen positif dari *synset*.
- #NegScore : berisi nilai sentimen negatif dari *synset*
- #SynsetTerms : berisi daftar sinonim dari *synset*. Setiap sinonim dipisahkan dengan spasi dan mengandung informasi indeks dari *synset* yang menandai seberapa sering *synset* dalam konteks tersebut digunakan.
- #Gloss : berisi makna dan konteks dari *synset*.

$$Objective\ score\ = 1\ -\ (Positive\ score\ +\ Negative\ score) \qquad (2.1)$$

2.5.2 Sentimen leksikon Indonesia

Sentimen leksikon merupakan sumber daya paling penting dalam *opinion mining* atau *sentiment analysis* yang menggunakan pendekatan semantik (makna). Sentimen leksikon berisi daftar kata-kata dengan kecenderungan sentimen positif atau negatif. Seperti kata “baik”, “bagus”, “aman” memiliki kecenderungan sentimen positif, sedangkan kata “buruk”, “lambat”, dan “ceroboh” memiliki kecenderungan sentimen negatif.

Berbeda dengan sentimen leksikon Inggris yang sudah beredar luas secara bebas (*free*) secara online. Sentimen leksikon Indonesia masih sangat terbatas jumlahnya dan belum beredar secara bebas (*free*) secara online. Aqsath dan Ayu (Naradhipa & Purwarianti, 2011) membangun sentimen leksikon Indonesia menggunakan pendekatan manual (*manual approach*), sedangkan Vania dkk (Vania, et al., 2014) menggunakan pendekatan corpus (*corpus based approach*).

Vania dkk (Vania, et al., 2014) mengekstrak *opinion word* sebagai benih kata (*feed seed*) menggunakan hasil *translate synset* dari *sentiwordnet* yang

memiliki *sentiment score* lebih besar dari 0.7 dan sentimen leksikon dari *opinion finder* yang memiliki *subjectivity score* tinggi. Selanjutnya benih (*feed seed*) digunakan untuk menemukan pola (pattern) dan struktur kalimat yang mengandung kata benih tersebut dan memiliki sentimen yang sama dengan sentimen benih. Selanjutnya, Vania menggunakan kata benih dan pola dari kalimat untuk mendapatkan *opinion word* lainnya di *corpus* dalam ukuran besar dengan topik tertentu.

Ekstraksi fitur menggunakan sentimen leksikon Indonesia memanfaatkan frekuensi kemunculan *opinion word* dalam setiap kalimat. Hal ini dikarenakan sentimen leksikon Indonesia belum memiliki *polarity score* seperti sentimen leksikon Inggris yakni *sentiwordnet*. Misalnya kata "baik", "senang" dan "mengagumkan" dalam sentimen leksikon Indonesia memiliki *polarity* yang sama yakni sama-sama positif. Padahal sebenarnya mereka memiliki tingkatan *polarity* yang berbeda satu dengan lainnya. *Polarity score* ketiga kata tersebut berdasarkan *sentiwordnet* adalah 0.63, 0.70, dan 0.75. Oleh karena itu, sentimen leksikon Indonesia belum bisa digunakan untuk mengklasifikasikan emosi yang terdapat pada sebuah teks. Tabel 2-4 menyajikan contoh sentimen leksikon Indonesia yang dibuat oleh Vania, dkk. Kolom positif leksikon berisi daftar kata – kata yang mengandung opini / sentimen positif, sedangkan kolom negatif leksikon berisi daftar kata – kata yang mengandung opini / sentimen negatif.

Tabel 2-4 Daftar *stopword* Tala

positif leksikon			**negatif leksikon**		
kagum	menarik	rajin	pedih	busuk	memalukan
gesit	damai	cerdik	benci	anarkis	menuduh
tangkas	cinta	indah	masam	marah	memalsukan
ramah	semangat	cantik	kasar	licik	khawatir
menenangkan	gairah	menguntungkan	pelupa	bodoh	membinasakan

2.6 Definisi Rule Based Method

Rule based method adalah metode klasifikasi yang memanfaatkan aturan-aturan (*rule*) untuk membedakan kelas yang saatu dengan kelas yang lain. *Rule* dibuat berdasarkan karakter dari masing-masing kelas dan dinotasikan dalam bentuk "IF …(*kondisi*)… THEN …(*solusi*)…". Dimana "IF" merupakan kondisi prasyarat (*rule antecedant*) yang terdiri dari satu atau lebih atribut tes, dimana tes-nya bersifat logika. Sedangkan "THEN" merupakan konsekuen (*rule consequent*) yang berisi hasil prediksi kelas. Misalnya untuk memisahkan data ke dalam kelas positif, negatif dan netral *rule* yang bisa digunakan adalah :

- Jika jumlah frekuensi kemunculan kata bersentimen positif lebih banyak dari kata bersentimen negatif, maka data digolongkan sebagai kelas positif.
- Jika jumlah frekuensi kemunculan kata bersentimen negatif lebih banyak dari kata bersentimen positif, maka data digolongkan sebagai kelas negatif.
- Jika jumlah frekuensi kemunculan kata bersentimen positif sama dengan kata bersentimen negatif, maka data digolongkan sebagai kelas netral.
- Jika tidak ditemukan kata bersentimen positif maupun kata bersentimen negatif pada data, maka digolongkan sebagai kelas netral.

Aturan (*rule*) yang digunakan bisa bersifat *mutually exclusive* atau *exhaustive*. *Mutually exclusive* berarti *classifier* mengandung aturan-aturan yang bersifat independen satu sama lain, sedangkan *exhaustive* berarti *classifier* mengandung aturan-aturan yang mencatat setiap kemungkinan kombinasi nilai atribut. Dimana setiap record hanya boleh dilingkupi paling banyak satu aturan saja.

2.7 Metode Pengujian dan Validasi

Tahapan ini bertujuan untuk mengukur performansi dari metode klasifikasi yang digunakan. Pada umumnya, kita bisa mengukur performansi sebuah sistem menggunakan perhitungan *accuracy*. Dimana jumlah data prediksi yang sesuai dibandingkan dengan jumlah data secara keseluruhan seperti ditunjukkan oleh Persamaan 2.2. Semakin tinggi nilai akurasi yang didapatkan, maka metode yang digunakan semakin cocok dengan karakteristik data. Sehingga hasil prediksi semakin mendekati kondisi nyatanya. Begitu pula sebaliknya

semakin kecil nilai akurasi yang didapatkan, maka metode yang digunakan semakin tidak cocok dengan karakteristik datanya. Sehingga hasil yang diperoleh semakin tidak mendekati kondisi nyatanya.

$$Accuracy = \frac{\sum pb}{num\_data} \quad (2.2)$$

Dimana :

pb : hasil klasifikasi (prediksi) yang sesuai dengan kelas sebenarnya.

num_data : jumlah data keseluruhan

Akan tetapi, pengukuran tersebut dianggap kurang optimal untuk mengetahui data yang menyebabkan rendahnya nilai akurasi. Karena kita tidak bisa mendeteksi adanya penyimpangan data dengan melihat nilai akurasi saja. Oleh sebab itu, kami juga menggunakan perhitungan *recall* dan *precision* untuk mengukur seberapa tepat klasifikasi terhadap kelas.

Recall adalah rasio jumlah data relevan yang telah ditemukan terhadap sebuah kelas yang telah diprediksi. *Recall* menggambarkan tingkat keberhasilan sistem dalam memanggil dokumen yang relevan. Sedangkan *precision* adalah rasio jumlah data relevan yang telah ditemukan terhadap data pada kelas tertentu dari dataset. *Precision* menggambarkan tingkat keefektifan dan ketepatan sistem dalam memanggil dokumen yang relevan dari seluruh dokumen yang diambil. Persamaan 2.3 dan 2.4 digunakan untuk menghitung perolehan nilai *precision* dan *recall* pada klasifikasi.

$$recall = \frac{\sum pb}{\sum total\_kelas} \quad (2.3)$$

$$precision = \frac{\sum pb}{\sum total\_pisah} \quad (2.4)$$

Dimana :

$total\_kelas$: jumlah keseluruhan data pada kelas tertentu

$total\_pisah$: jumlah data yang dipisahkan *classifier* sebagai anggota kelas tertentu

Untuk menghitung nilai *recall* dan *precision*, kita bisa menggunakan *confussion matrix* seperti yang ditunjukkan oleh Tabel 2-5. Dalam *confussion matrix* dikenal empat istilah yang harus diketahui, yakni

1. TP (*True Positive*), yakni kelas yang diprediksi benar dan hasilnya benar.
2. TN (*True Negative*), yakni kelas yang diprediksi salah dan hasilnya salah.
3. FP (*False Positive*), yakni kelas yang diprediksi benar tetapi hasilnya salah.
4. FN (*False Negative*), yakni kelas yang diprediksi salah tetapi hasilnya benar.

Nilai *precision* dan *recall* bisa juga dihitung menggunakan Persamaan 2.5 dan 2.6. Persamaan 2.5 menunjukkan rumus yang digunakan untuk menghitung nilai *recall* pada kelas positif dengan membandingkan nilai TP (*True Positive*) dengan jumlah TP (*True Positive*) dan FN (*False Negative*). Sedangkan untuk menghitung nilai *precision* menggunakan Persamaan 2.6 dengan membandingkan nilai TP (*True Positive*) dengan jumlah TP (*True Positive*) dan FP (*False Positive*).

Tabel 2-5 *Confussion Matrix*

		Kelas Prediksi	
		TRUE	*FALSE*
Kelas Sebenarnya	*TRUE*	TP (*True Positive*)	FP (*False Positive*)
	FALSE	FN (*False Negative*)	TN (*True Negative*)

$$recall = \frac{TP}{TP + FN} \tag{2.5}$$

$$precision = \frac{TP}{TP + FP} \tag{2.6}$$

Dimana :

TP (True Positive) : Jumlah data dari kelas yang diprediksi benar dan hasilnya benar

FN (False Negative) : Jumlah data dari kelas yang diprediksi salah dan hasilnya benar

FP (False Positive) : Jumlah data dari kelas yang diprediksi salah dan hasilnya salah

2.8 Kajian Penelitian Terkait

Seiring berkembangnya media sosial, *sentiment analysis* menjadi topik yang menarik untuk diteliti. Hal ini memudahkan perusahaan dalam menganalisa respon pasar terhadap produk yang diluncurkan dari forum maupun media sosial, tanpa harus membuat survey customer secara langsung. Banyak peneliti yang sudah melakukan eksperimen di bidang ini baik menggunakan pendekatan statistik, semantik maupun menggabungkan keduanya. Selain karena manfaat dari *sentiment analysis* yang bisa dirasakan secara langsung oleh perusahaan, pemerintahan, bahkan politik. Juga dikarenakan sentimen (opini) yang bersifat subjektif, menjadi daya tarik tersendiri bagi para peneliti. Table 2-6 menyajikan daftar penelitian terkait di bidang *sentiment analysis*, dimana penelitian nomor 1 sampai dengan 3 menggunakan data pengaduan publik. Penelitian 1 (pertama) dan 2 (kedua) menganalisa sentimen pada pengaduan publik menggunakan pendekatan statistik. Hal ini bisa kita lihat dari kolom ekstraksi fitur dari penelitian 1 (pertama) dan 2 (kedua) yang memanfaatkan frekuensi kemunculan kata untuk proses ekstraksi fitur dari data pengaduan. Sedangkan penelitian 3 (ketiga), peneliti mengekstraksi topik dari pengaduan masyarakat menggunakan pendekatan statistik. Sama halnya dengan penelitian 1 (pertama) dan 2 (kedua), penelitian 3 (ketiga) memanfaatkan frekuensi kemunculan kata untuk meng-ekstraksi fitur.

Selanjutnya, penelitian 4 (keempat) sampai dengan 8 (kedelapan) merupakan penelitian mengenai *sentiment analysis* menggunakan pendekatan semantik. Hal ini bisa kita lihat dari kolom ekstraksi fitur pada Tabel 2-6, dimana penelitian 4 (keempat) sampai dengan 8 (kedelapan) memanfaatkan *lexical resource* / *sentiment lexicon* untuk proses ekstraksi fiturnya. Penelitian 4 (keempat) dan 5 (kelima) menggunakan sentimen leksikon Indonesia, sedangkan penelitian 6 (keenam) sampai dengan 8 (kedelapan) menggunakan *sentiwordnet* untuk proses ekstraksi fitur. Pada penelitian 6 (keenam), *sentiwordnet* digunakan sebagai

penunjang dalam mengekstrak fitur dari teks bahasa Indonesia. Sedangkan pada penelitian 7 (ketujuh) dan 8 (kedelapan) *sentiwordnet* digunakan untuk mengekstrak fitur dari teks bahasa Inggris.

Tabel 2-6 Daftar Penelitian Terkait

No	Judul	Data	Metode Klasifikasi	Ekstraksi Fitur
1	Sentiment analysis based on big data (Nomleni, et al., 2014)	Data pengaduan *media center* Surabaya	*Statistical based method* (SVM)	Frekuensi kemunculan kata (TF-IDF)
2	Public service satisfaction based on sentiment analysis (Susilawati, 2016)	Data pengaduan pada akun *twitter* PLN *@pln_123*	*Statistical based method* (Naïve Bayes)	Frekuensi kemunculan kata (TF-IDF)
3	Information extraction of public complaints on twitter text for Bandung government (Anggareska & Purwarianti, 2014)	Pengaduan masyarakat pada akun twitter pemkot Bandung	*Statistical based method* (SVM dan Naïve bayes)	Lexical Syntactical (*Relation Extraction*)
4	Sentiment classification for Indonesian message in social media (Naradhipa & Purwarianti, 2011)	Komentar pelanggan di akun facebook sebuah perusahaan	*Rule based method* dan *Statistical based method*	Sentimen leksikon Indonesia
5	Comparison on the rule based method and statistical based method on emotion classification for Indonesian twitter text (Atmadja & Purwarianti, 2015)	Data twitter dalam bahasa Indonesia	*Rule based method* dan *Statistical based method*	Sentimen leksikon Indonesia dan Wordnet
6	Indonesian social media sentiment analysis with sarcasm detection (Lunando & Purwarianti, 2013)	Data twitter bahasa Indonesia dengan topik politik	*Statistical based method* (SVM, Naïve bayes, ME)	Sentiwordnet
7	Opinion Mining of News Headlines using Sentiwordnet (Agarwal, et al., 2016)	*Headline* surat kabar dalam bahasa inggris	*Rule based method*	Sentiwordnet

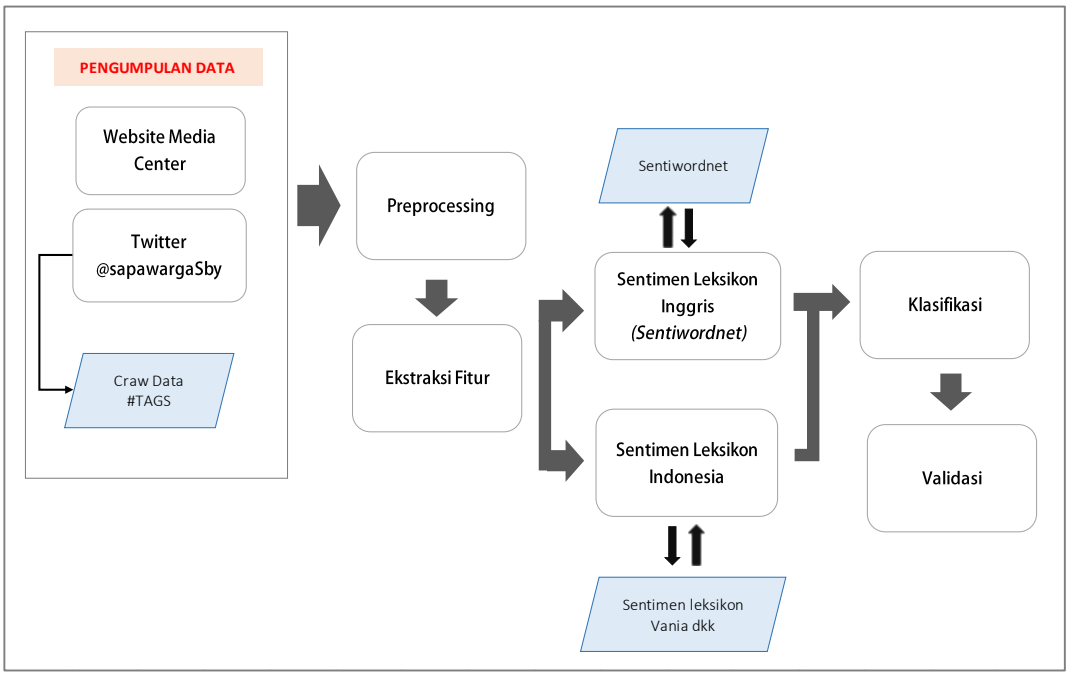
Tabel 2-6 Daftar Penelitian Terkait (Lanjutan)

No	Judul	Data	Metode Klasifikasi	Ekstraksi Fitur
8	Sentiment analysis from product reviews using sentiwordnet as lexical resources (Cernian & Sgarciu, 2015)	review produk amazon bahasa Inggris	*Rule based method*	Sentiwordnet

Halaman ini sengaja dikosongkan

BAB 3
METODOLOGI PENELITIAN

Alur dan metode yang digunakan dalam penelitian *sentiment analysis* ini disajikan secara detail oleh Gambar 3.1. Kelas *output* yang diharapkan adalah sentimen positif, sentimen negatif dan netral. Tahapan awal dari penelitian ini adalan proses pengumpulan data pengaduan masyarakat dari *twitter* resmi pemerintah kota Surabaya yakni @sapawargasby dan website resmi layanan pengaduan pemerintah kota Surabaya yakni *media center*. Untuk proses *crawling data* dari *twitter* kami menggunakan bantuan *twitter* API yakni *Twitter Archieved Google Spreadsheet* (TAGS).



Gambar 3.1 Blok diagram penelitian *sentiment analysis* pengaduan masyarakat.

Tahapan selanjutnya adalah *pre-processing* untuk menormalisasi data menjadi bentuk baku (formal) sesuai kaidah dalam kamus besar bahasa Indonesia (KBBI). *Pre-processing* diperlukan karena data yang digunakan adalah data dari *twitter* yang banyak mengandung *noise* seperti singkatan, bahasa tidak formal, bahasa lokal, *slang word*, bahkan topik yang tidak jelas (Naradhipa & Purwarianti, 2011).

Tahapan paling penting dalam penelitian ini adalah ekstraksi fitur. Karena disini kami menggunakan pendekatan semantik untuk *sentiment analysis*, maka proses ekstraksi fitur menggunakan *lexical resource* yakni sentimen leksikon Inggris (*Sentiwordnet*) dan sentimen leksikon Indonesia (Vania, et al., 2014).

Setelah fitur kata (*term*) didapatkan, maka tahapan berikutnya adalah klasifikasi. Metode klasifikasi yang digunakan dalam penelitian *sentiment analysis* ini adalah *rule based method*, dimana data dipisahkan ke dalam kelas *output* berdasarkan aturan (*rule*) yang dibuat pada tahapan klasifikasi.

Langkah terakhir adalah tahapan validasi untuk melihat tingkat keakuratan sistem yang dibangun untuk *sentiment analysis* pengaduan masyarakat melalui perhitungan nilai *accuracy* yang didapatkan. Selain itu, pada tahapan ini juga dilihat tingkat optimal dan ketepatan *classifier* dalam memisahkan data sesuai dengan kelasnya (relevan) melalui perhitungan nilai *precision* dan *recall* untuk masing-masing kelas *output*.

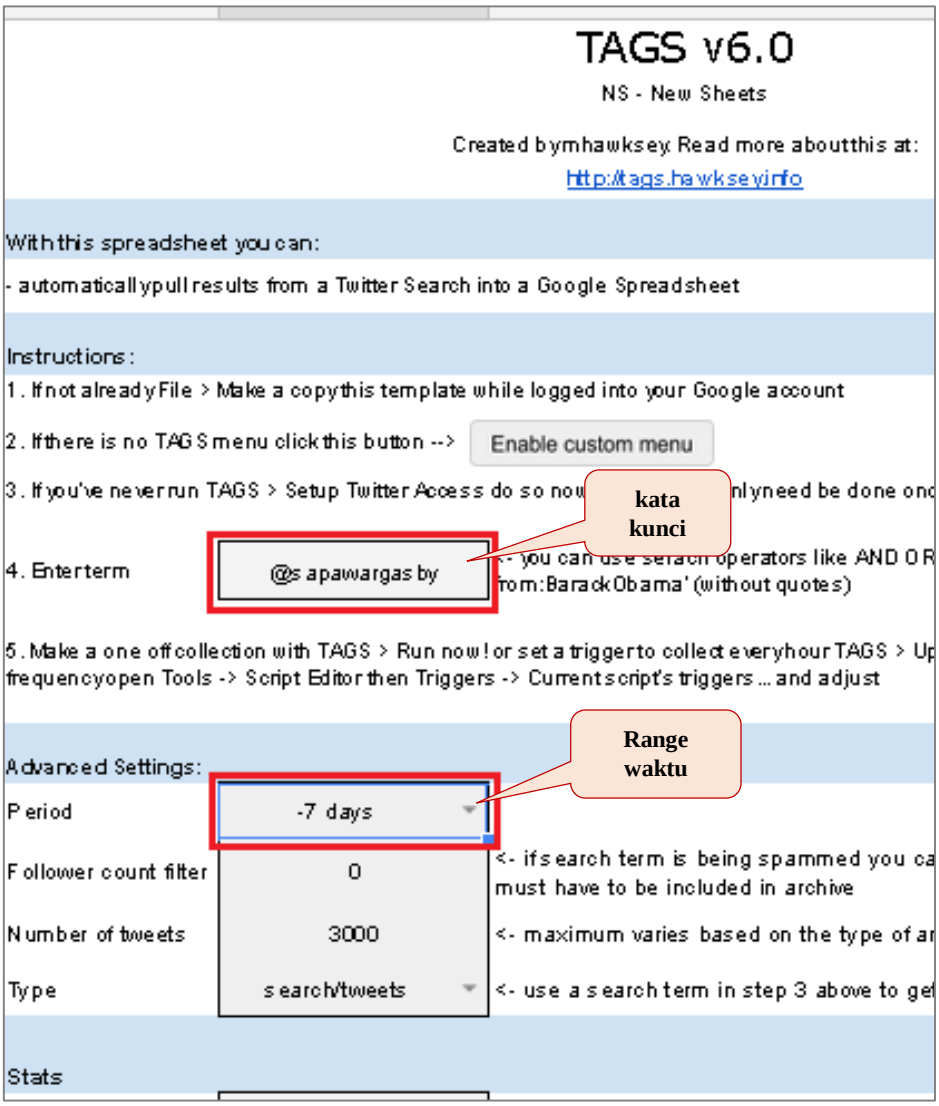
3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari data pengaduan masyarakat yang masuk pada layanan pengaduan publik pemerintah kota Surabaya. Baik melalui website resmi layanan pengaduan publik yakni *media center* maupun akun *twitter* resmi pemerintah kota Surabaya untuk menampung aspirasi dan pengaduan masyarakat yakni @sapawargasby. Jumlah data yang digunakan sebanyak 1000 pengaduan dari masing-masing sumber.

Untuk menambang data (*crawling data*) dari *twitter* @sapawargasby, kami menggunakan bantuan *API Twitter* yakni *Twitter Archieve Google Spreadsheet* (TAGS V6) seperti ditunjukkan pada Gambar 3.2. Proses penambangan data pada *twitter* dilakukan pada bulan November 2016 sampai dengan January 2017. Sedangkan data *media center* diambil dari data bulan Agustus 2015 sampai dengan Mei 2016.

Proses pelabelan data dilakukan secara manual menggunakan tiga *annotator*. Label yang digunakan adalah label hasil penentuan dengan suara terbanyak dari ketiga *annotator*. Ketika ketiga *annotator* tidak sependapat, maka

kami menggunakan *annotator* lain untuk menentukan sentimen dari pengaduan tersebut.

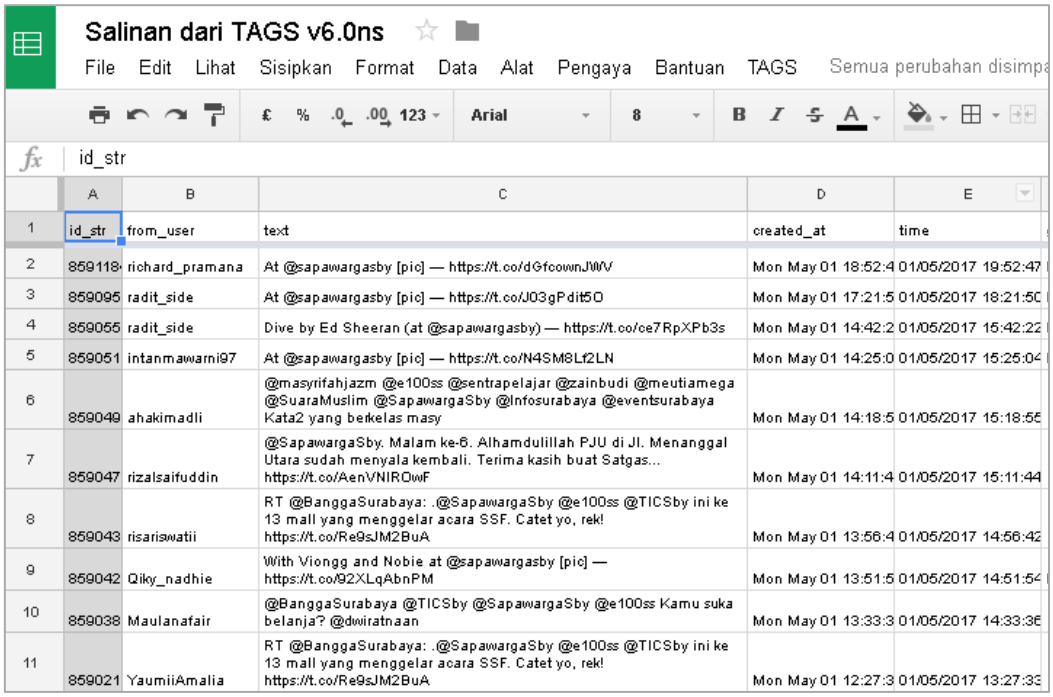


Gambar 3.2 *Twitter Archieved Google Spreadsheet (TAGS) API.*

Berdasarkan Gambar 3.2, untuk menambang data (*crawling*) dari *twitter* menggunakan *twitter API TAGS* terlebih dahulu ditentukan kata kunci yang ingin dicari, pada penelitian ini kami menggunakan kata kunci nama akun dari *twitter* pengaduan pemkot Surabaya itu sendiri yakni @sapawarga, dengan tujuan untuk mengambil semua *tweet* yang masuk di akun tersebut. Kemudian kata kunci tersebut kita masukkan pada kolom *Enter term*. Selain itu, kita juga bisa mengatur *range* waktu pengambilan data dengan mengganti pilihan hari pada kolom *Period*.

Contoh hasil *crawling data twitter* ditunjukkan oleh Gambar 3.3. Dimana data hasil penambangan (*crawling*) *twitter* mengandung informasi sebagai berikut :

- *id_str*, berisi kode unik dari masing-masing *tweet*
- *from_user*, berisi informasi akun user yang memposting *tweet*
- *text*, berisi informasi isi dari *tweet* yang diposting oleh user
- *created_at*, berisi informasi tanggal *tweet* diposting
- *time*, berisi informasi waktu dari *tweet* diposting
- *geo_coordinates*, berisi informasi koordinat user ketika memposting *tweet*
- *user_lang*, berisi informasi default setting bahasa *twitter* yang digunakan *user*



Salinan dari TAGS v6.0ns

File Edit Lihat Sisipkan Format Data Alat Pengaya Bantuan TAGS Semua perubahan disimpa

Arial 8

fx id_str

	A	B	C	D	E
1	id_str	from_user	text	created_at	time
2	859118	richard_pramana	At @sapawargasby [pic] — https://t.co/dGfcownJWV	Mon May 01 18:52:4	01/05/2017 19:52:47
3	859095	radit_side	At @sapawargasby [pic] — https://t.co/J03gPdit5O	Mon May 01 17:21:5	01/05/2017 18:21:50
4	859055	radit_side	Dive by Ed Sheeran (at @sapawargasby) — https://t.co/ce7RpXPb3s	Mon May 01 14:42:2	01/05/2017 15:42:22
5	859051	intanmawarni97	At @sapawargasby [pic] — https://t.co/N4SM8Lf2LN	Mon May 01 14:25:0	01/05/2017 15:25:04
6	859049	ahakimadli	@masyrifahjazm @e100ss @sentrapelajar @zainbudi @meutiamega @SuaraMuslim @SapawargaSby @Infosurabaya @eventsurabaya Kata2 yang berkelas masy	Mon May 01 14:18:5	01/05/2017 15:18:55
7	859047	rizalsaifuddin	@SapawargaSby. Malam ke-6. Alhamdulillah PJU di Jl. Menanggal Utara sudah menyala kembali. Terima kasih buat Satgas... https://t.co/AenVNIROwF	Mon May 01 14:11:4	01/05/2017 15:11:44
8	859043	risariswatii	RT @BanggaSurabaya: .@SapawargaSby @e100ss @TICSby ini ke 13 mall yang menggelar acara SSF. Catet yo, rek! https://t.co/Re9sJM2BuA	Mon May 01 13:56:4	01/05/2017 14:56:42
9	859042	Qiky_nadhie	With Viongg and Nobie at @sapawargasby [pic] — https://t.co/92XLqAbnPM	Mon May 01 13:51:5	01/05/2017 14:51:54
10	859038	Maulanafair	@BanggaSurabaya @TICSby @SapawargaSby @e100ss Kamu suka belanja? @dwiratnaan	Mon May 01 13:33:3	01/05/2017 14:33:36
11	859021	YaumiiAmalia	RT @BanggaSurabaya: .@SapawargaSby @e100ss @TICSby ini ke 13 mall yang menggelar acara SSF. Catet yo, rek! https://t.co/Re9sJM2BuA	Mon May 01 12:27:3	01/05/2017 13:27:33

Gambar 3.3 *Twitter Archieved Google Spreadsheet (TAGS) API.*

Pada penelitian ini, data yang digunakan adalah isi *tweet* dan tanggal posting dari *tweet* tersebut. Sedangkan informasi lainnya diabaikan karena tidak memiliki pengaruh dalam penelitian *sentiment analysis* pengaduan publik ini. Contoh data pengaduan publik dari media *twitter* dan *media center* disajikan pada Tabel 3-1. Kolom pengaduan menyajikan data pengaduan / aspirasi yang masuk dari masyarakat baik dari media *twitter* maupun *media center*. Sedangkan kolom media menyajikan asal sumber dari data pengaduan apakah media *twitter* atau *media center*.

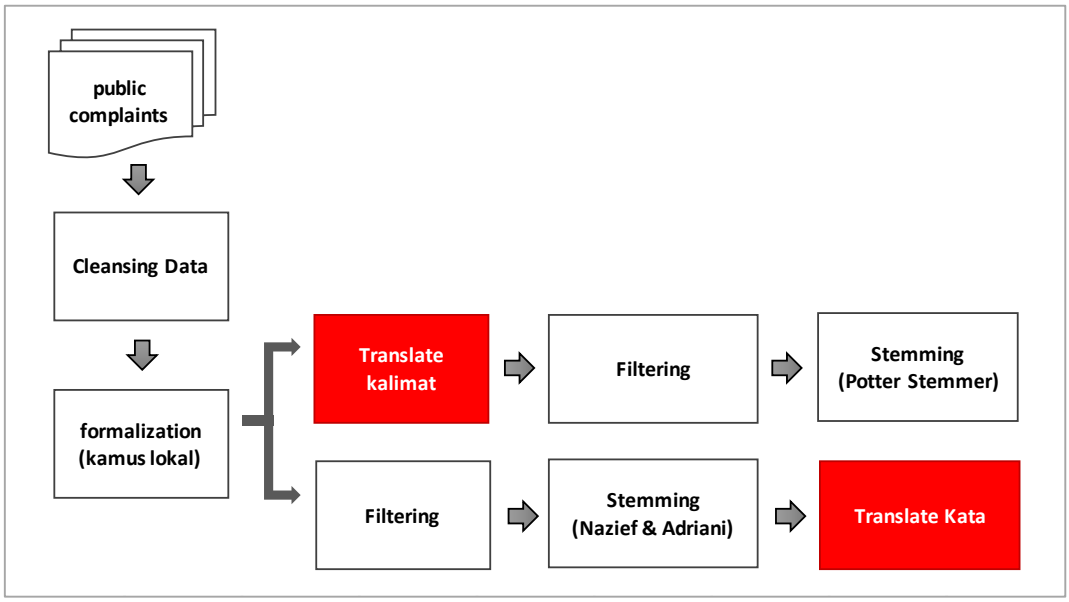
Tabel 3-1 Contoh pengaduan masyarakat.

No	Pengaduan	Media
1	RT @SapawargaSby: Kedamaian tidak bisa dijaga dengan paksaan. Kedamaian hanya bisa dicapai dengan saling pengertian"- Albert Enstein	*twitter*
2	@SapawargaSby terima kasih :-)	*twitter*
3	Kom blm ada action bos, tuk pohon di jl Pulo Wonokromo 292/ II, itu ada 2 phn Sono besar2	*twitter*
4	@SapawargaSby jalan disekolahan SD Bendul merisi (sampai bebek maryem) sangat parah mohon dilakukan pengaspalan #keluhan	*twitter*
5	@SapawargaSby min, mau tanya, persyaratan apa saja yg diperlukan utk program sertifikasi tanah massal tahun 2016 ini, suwun infone	*twitter*
6	@SapawargaSby masih banyak kelurahan nakal nih buk di Surabaya ,pengurusan prona Dan sertifikatkan tanah di persulit sangaat mahal	*twitter*
7	WIFI DI TAMAN PELANGI TIDAK BISA DIGUNAKAN SEJAK TGL 17 NOVEMBER 2015	*media center*
8	Selamat siang,mau tanya untuk di area Surabaya apakah msh ada operator taxi baru yg akan beroperasi diluar yg operator yg sdh beroperasi saat ini ? terima kasih sebelumnya	*media center*
9	jalan pacuan kuda banyak PKL menggunakan trotoar untuk berjualan baik siang hari ataupun malam hari, sehingga kendaraan pembelinya parkir dipinggir jalan yang mengakibatkan kemacetan. terima kasih	*media center*
10	Lampu PJU di Jl. Ngagel Tama Selatan kalau siang nyala kalau malam padam. Mohon segera diperbaiki	*media center*
11	Syarat apa saja yang dibutuhkan untuk mendaftar BLC?	*media center*

3.2 Tahapan Pre-processing Data

Pada tahapan ini dilakukan proses pembersihan data dari *noise* dan mengubah data ke bentuk dasar-nya (kata dasar). Sesuai tujuan dari tahapan *pre-processing* yakni menormalisasi data dan mereduksi dimensi dari data, maka pada tahapan ini dilakukan proses *cleansing*, *formalization* dengan menggunakan kamus lokal, serta menghapus nama jalan. Dengan harapan dapat meningkatkan akurasi.

Pada penelitian ini dilakukan tiga (3) eksperimen untuk membandingkan tingkat kesesuaian penerapan sentimen leksikon bahasa Inggris yakni *Sentiwordnet* dan sentimen leksikon Indonesia terhadap data pengaduan masyarakat. Eksperimen pertama dan kedua menggunakan *Sentiwordnet* dalam proses ekstraksi fitur, yakni menghitung *sentiment score* kata penyusun kalimat. Karena *Sentiwordnet* menggunakan bahasa Inggris, sedangkan data pengaduan dalam bahasa Indonesia. Maka diperlukan proses *translate* menggunakan *Google API Translate* seperti ditunjukkan oleh Gambar 3.4.

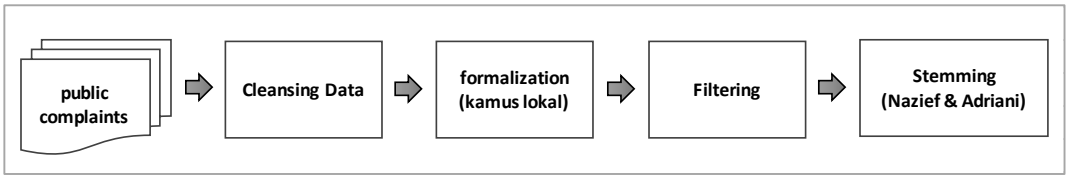


Gambar 3.4 *Preprocessing* pada eksperimen menggunakan *Sentiwordnet*.

Pada eksperimen pertama proses translate dilakukan setelah proses pembuangan *stopword* dan stemming bahasa Indonesia dengan algoritma nazief dan adriani serta kamus besar bahasa Indonesia (KBBI). Eksperimen kedua juga menggunakan *Sentiwordnet* untuk ekstraksi fitur, hanya saja proses *translate* dilakukan setelah proses formalisasi kata ke bentuk baku. Jika pada eksperimen

pertama *translate* dilakukan terhadap kata dasar, eksperimen kedua *translate* dilakukan terhadap kalimat utuh. Proses *stemming* pada eksperimen kedua menggunakan algoritma *Potter Stemmer* dan *Wordnet*.

Eksperimen ketiga menggunakan sentimen leksikon bahasa Indonesia yang dibuat oleh Vania, dkk (Vania, et al., 2014) untuk ekstraksi fitur. Bobot fitur diperoleh dari jumlah frekuensi kemunculan kata bersentimen sesuai sentimen leksikon Indonesia yang ada di kalimat tersebut. Sedangkan tahapan *pre-processing* untuk eksperimen ketiga ditunjukkan oleh Gambar 3.5 yang terdiri dari *cleansing data*, *formalization* dengan menggunakan kamus lokal, *filtering*, dan *stemming*. Berbeda dengan eksperimen pertama dan kedua, pada eksperimen ketiga ini tidak siperlukan tahapan *translate*, karena *lexical resource* / *sentiment lexicon* yang digunakan untuk proses ekstraksi fitur dalam bahasa Indonesia.



Gambar 3.5 *Preprocessing* pada eksperimen menggunakan sentimen leksikon Indonesia.

3.2.1 Tahapan Cleansing Data

Tahapan dalam *cleansing data* diawali dengan proses *retweet removal*. Hal ini dilakukan untuk menghindari adanya duplikasi data. Karena *retweet* (RT) bertujuan memposting atau *share* kembali informasi dari *tweet* orang lain. Proses selanjutnya melakukan *hashtag*, *mention*, dan *html removal*. Karena *hashtag* (#), *mention*(@), dan *html* (http://www) termasuk kata yang tidak memiliki makna dan tidak berpengaruh dalam proses *sentiment analysis*. Sama halnya dengan *hashtag*, *mention*, dan *html*. Angka decimal dan romawi juga tidak memiliki pengaruh dalam proses klasifikasi sentimen. Oleh karena itu perlu dilakukan pembuangan angka decimal dan romawi, baik angka yang berdiri sendiri maupun angka yang terdapat di awal atau akhir sebuah kata seperti :

- kom blm ada action bos tuk pohon di jl pulo wonokromo 292 ii itu ada 2 phn sono besar2.

- yuk daur ulang smpah2 plastik jdi brng yg lbh berguna.
- kamis 17 11 2016 10 00 14 00 quest hotel surabaya cc
- almat perumahan graha indah jln medayu utara 17 no d 9 rt 07 rw iii kel medokan ayu kec rungkut an bpk bambang.

Setelah semua proses tersebut dilakukan, langkah selanjutnya adalah mengubah semua huruf menjadi huruf kecil untuk menyamakan kode ASCII dari kata penyusun kalimat dan mereduksi jumlah fitur dengan menggabungkan fitur yang sama. Kata "rusak" dan "Rusak" dikenali sebagai kata yang berbeda oleh komputer, meskipun memiliki makna yang sama. Hal ini disebabkan huruf "r" dengan "R" memiliki kode ASCII yang berbeda.

Contoh kalimat :

"kapan bos pohon di jalan Pulo Wonokromo depan Rumah no 295/II dipotong? Pohon itu bener2 membahayakan, karena dia tinggi besar"

Hasil : Tabel 3-2 menyajikan hasil *tokenizing* tanpa mengubah semua huruf menjadi huruf kecil. Sehingga diperoleh kata sebanyak 17 buah dengan kata "pohon" dan "Pohon" dianggap berbeda, padahal memiliki makna yang sama. Sedangkan pada Tabel 3-3 proses *tokenizing* disertai pengubahan semua huruf menjadi huruf kecil, dan berhasil mereduksi hasil perolehan kata menjadi 16 buah.

Tabel 3-2 Tanpa mengubah semua huruf menjadi huruf kecil.

kata	**kata**	**kata**	**kata**	**kata**	**kata**
kapan	bos	pohon	jalan	Wonokromo	depan
rumah	no	dipotong	Pohon	itu	bener
membahayakan	karena	dia	tinggi	besar	

Tabel 3-3 Dengan mengubah semua huruf menjadi huruf kecil.

kata	**kata**	**kata**	**kata**	**kata**	**kata**
kapan	bos	pohon	jalan	wonokromo	depan
rumah	no	dipotong	itu	bener	membahayakan
karena	dia	tinggi	besar		

Langkah terakhir dalam tahapan *pre-processing* ini adalah memecah setiap kata penyusun sebuah kalimat (*tokenizing*). Akan tetapi sebelum melakukan proses *tokenizing*, perlu dilakukan pengecekan sekali lagi apakah ada pengaduan yang sama atau tidak. Hal ini untuk menghindari adanya duplikasi data.

3.2.2 Tahapan Formalization

Tahapan ini dilakukan untuk mengubah kata ke bentuk baku sesuai kaidah dalam Kamus Besar Bahasa Indonesia (KBBI). Seperti kita ketahui, data *twitter* banyak mengandung *noise* seperti singkatan, menggunakan kata informal bahkan kata dalam bahasa lokal yang tidak sesuai dengan kamus bahasa Indonesia. Proses *formalization* kata menggunakan dua pendekatan yakni algoritma *laveinstein distance* dan kamus lokal yang kami buat secara manual untuk memfasilitasi kata lokal yang tidak ada di kamus bahasa Indonesia.

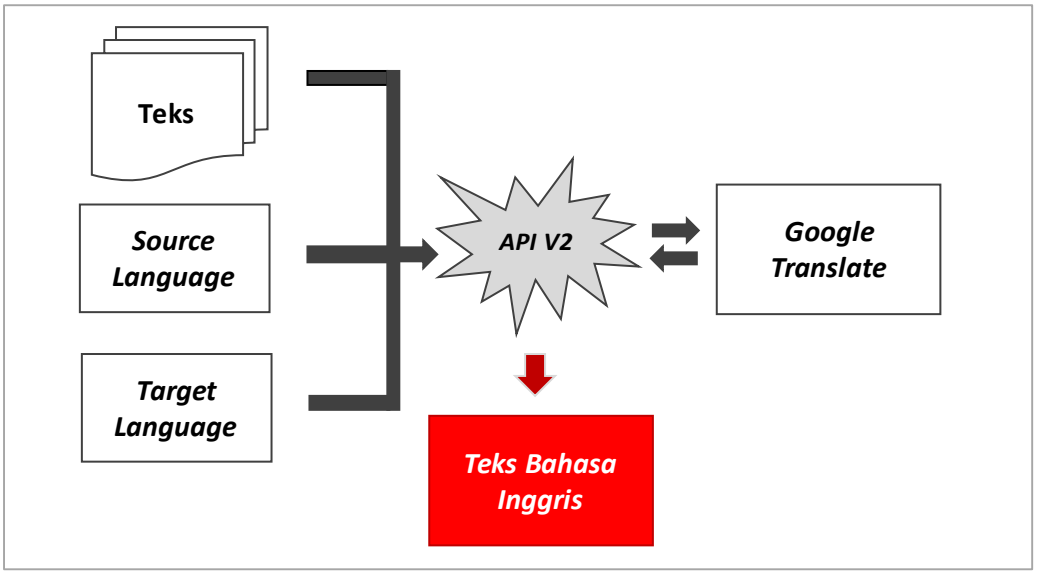
Algoritma *Laveinstein distance* digunakan untuk mengubah kata singkatan ke bentuk aslinya sesuai kamus, seperti : "jln" menjadi "jalan", "blm" menjadi "belum", "bhy" menjadi "bahaya" dan "phn" menjadi "pohon". Akan tetapi, kelemahan dari algoritma *laveinstein distance* adalah hanya akan mengeksekusi kata yang ada di kamus bahasa Indonesia saja. Kata yang tidak terdapat dikamus seperti kata berimbuhan dan kata lokal tidak akan dieksekusi, seperti kata "djdikan", "dtnami", "manfaatkn", "takon", "bocoen", "mbois". Oleh karena itu, untuk melengkapi kekurangan tersebut penulis membuat kamus lokal untuk mengubah kata dalam bahasa lokal ke kata baku.

3.2.3 Tahapan Translate

Tahapan *translate* diperlukan untuk mengubah data dari bahasa Indonesia menjadi bahasa Inggris. Tahapan ini dibutuhkan oleh eksperimen pertama dan kedua, karena kedua eksperimen tersebut menggunakan *sentiwordnet* untuk eksrtaksi fitur. Dalam penelitian ini proses *translate* menggunakan bantuan *API Google Translate* versi 2.

API Google Translate sangat sederhana dan mudah diimplementasikan, karena dibangun menggunakan bahasa html serta dipadukan dengan *Jquery*. Sebelum menggunakan *API Google Translate*, poin utama yang harus didapatkan

adalah *Google API key*. Kode kunci tersebut digunakan *google* untuk mengenali dan melakukan validasi pengguna fungsi API-nya. Parameter yang diperlukan untuk mengakses fungsi *translate* pada *API Google Translate* diantaranya teks yang ingin di *translate* (*input teks*), bahasa asal yang digunakan oleh teks (*source language*), bahasa yang menjadi tujuan *translate* (*target language*). Proses akses *google translate* menggunakan API ditunjukkan oleh Gambar 3.6, dimana parameter inputan yang dibutuhkan adalah teks yang ingin di *translate*, bahasa yang digunakan oleh teks inputan dan bahasa yang diharapkan sebagai *output*. Ketiga parameter tersebut dikirim menggunakan fungsi API untuk ditranslate oleh aplikasi *Google Translate*. Hasil *translate* dikembalikan melalui fungsi API dalam bentuks teks sesuai bahasa tujuan.



Gambar 3.6 Alur dari tahapan *translate* menggunakan *API Google Translate*

3.2.4 Tahapan Part of Speech (POS) Tagging

POS Tagging dilakukan untuk memperoleh kelas (jenis) kata dalam sebuah kalimat. Kelas (jenis) kata yang dikenali di *sentiwordnet* antara lain kata benda (*noun*), kata sifat (*adjective*), kata kerja (*verb*), dan kata keterangan (*adverb*). Jenis kata ini diperlukan pada eksperimen pertama dan kedua untuk mengekstrak *sentiment score* sebuah kata di *Sentiwordnet*. Pada eksperimen pertama jenis kata diperoleh dari kamus bahasa Indonesia, sedangkan eksperimen kedua menggunakan algoritma dari Bill Tagger yang diimplementasikan oleh Mark

Watson ke beberapa bahasa pemrograman. Pada eksperimen pertama, *POS Tagging* dilakukan pada kata dasar setelah melalui tahapan stemming sesuai kaidah kamus besar bahasa Indonesia (KBBI). Sedangkan pada eksperimen kedua, *POS Tagging* dilakukan pada kalimat hasil translate sebelum melalui tahapan *stopword removal* dan stemming sesuai kaidah bahasa Inggris.

Misalkan :

"lampu penerangan jalan umum di jl ngagel tama selatan kalau siang nyala kalau malam padam mohon segera diperbaiki"

Translate :

"Street lighting lights on the street if the day lights up when night goes out please fix it"

Tabel 3-4 menyajikan hasil ekstrak kelas kata dari kalimat contoh sesuai kaidah Kamus Besar Bahasa Indonesia (KBBI). Kolom kata berisi daftar kata dasar hasil stemming dalam bahasa Indonesia. Sedangkan Tabel 3-5 menyajikan contoh hasil *POS Tagging* dalam bahasa Inggris, dimana kolom kata berisi daftar kata penyusun kalimat hasil *translate* yang belum melalui tahapan *stemming*.

Tabel 3-4 *POS Tagging* menggunakan kamus besar bahasa Indonesia.

kata	**kelas kata**	**kata**	**kata**
terang	kata sifat (*adj*ective)	jalan	kata benda (*noun*)
siang	kata benda (*noun*)	malam	kata benda (*noun*)
padam	kata sifat (*adjective*)	mohon	kata kerja (*verb*)
baik	kata sifat (*adjective*)		

Tabel 3-5 *POS Tagging* pada kalimat hasil *translate*.

kata	**kelas kata**	**kata**	**kata**
street	kata benda (*noun*)	lighting	kata kerja (*verb*)
light	kata benda (*noun*)	night	kata benda (*noun*)
day	kata benda (*noun*)	please	kata kerja (*verb*)
goes	kata kerja (*verb*)	immediately	kata keterangan (*adverb*)
fix	kata kerja (*verb*)		

3.2.5 Tahapan Filtering

Tahapan *filtering* bertujuan untuk membuang kata yang tidak berpengaruh terhadap proses klasifikasi atau biasa disebut dengan istilah *stopword*. Pada eksperimen pertama dan ketiga *stopword* acuan yang digunakan adalah *stopword list* Indonesia dari Tala (Tala, 2003). Karena pada tahapan *filtering* kata masih menggunakan bahasa Indonesia. Sedangkan eksperimen kedua menggunakan NLTK *English stopword*. Karena pada tahapan *filtering* kata sudah di *translate* ke dalam bahasa Inggris.

Misalkan :

"lampu penerangan jalan umum di jl ngagel tama selatan kalau siang nyala kalau malam padam mohon segera diperbaiki"

Stopword removal

- eksperimen pertama dan ketiga

  "lampu penerangan jalan jalan siang nyala malam padam mohon diperbaiki"
- eksperimen kedua

translate : Street lighting lights on the street if the day lights up when night goes out please fix it

filtering : street lighting lights street day lights night goes please fix immediately

Selain membuang kata yang termasuk dalam *stopword list*, karena dianggap tidak berpengaruh dalam proses klasifikasi. Pada tahapan ini juga dilakukan pembuangan nama jalan. Hal ini dikarenakan ada beberapa nama jalan di Surabaya yang menggunakan kata bersentiment seperti "Dharmahusada indah", "Kalijudan Asri", "Medokan Ayu", "Rungkus Asri" dan "Nirwana Indah". Kata "indah", "asri", dan "ayu" tersebut ketika di *translate* ke dalam bahasa Inggris menghasilkan kata bersentimen positif dengan *score* yang cukup tinggi yakni "*beautifull*". Hal ini dapat mengubah *sentiment score* dari kalimat dan menyebabkan terjadinya kesalahan klasifikasi. Untuk membuang nama jalan yang terdapat pada kalimat pengaduan, kami menggunakan daftar nama jalan yang ada di Surabaya. Disini daftar nama jalan yang kami gunakan hanya nama jalan yang mengandung kata bersentimen saja.

3.2.6 Tahapan Stemming

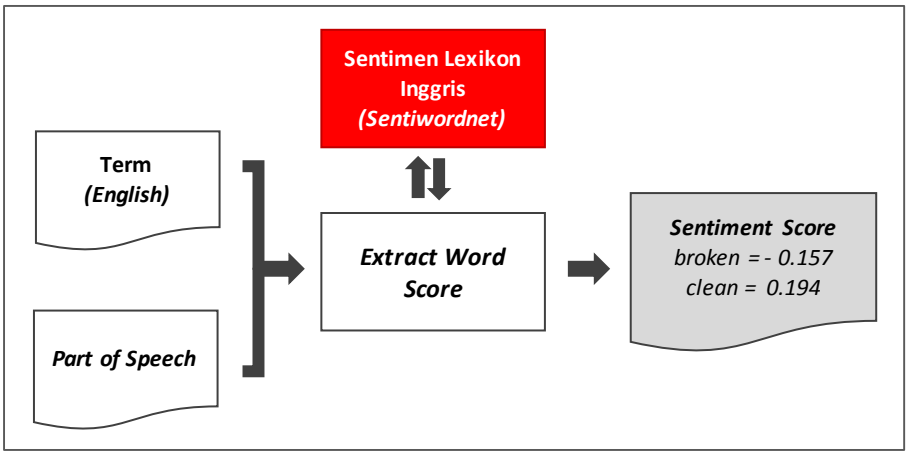
Tujuan dari tahapan *stemming* adalah mengembalikan kata ke bentuk dasarnya sesuai kamus. Pada penelitian ini, eksperimen pertama dan ketiga menggunakan algoritma nazief dan adriani untuk proses *stemming*. Karena kata inputan untuk proses *stemming* dalam bahasa Indonesia. Sedangkan pada eksperimen kedua, proses *stemming* menggunakan algoritma *potter stemmer* dan *wordnet*. Karena kata inputan untuk proses *stemming* dalam bahasa Inggris.

3.3 Proses Ekstraksi Fitur

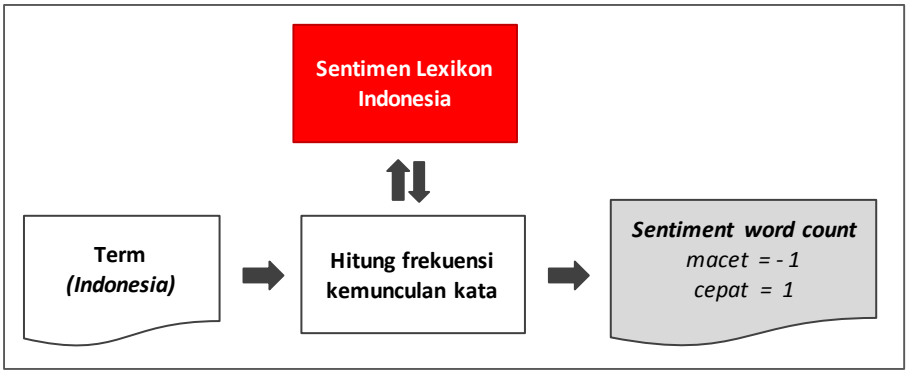
Tahapan ini memegang peranan penting dalam *text classification* dan *sentiment analysis*. Penelitian ini menggunakan pendekatan semantik, sehingga proses ekstraksi fitur menggunakan *lexical resource* seperti kamus sentimen. Fitur yang digunakan dalam penelitian ini adalah *term* (*unigram*) yang mengandung sentimen positif atau sentimen negatif. *Sentiment* dari kalimat ditentukan oleh *sentiment* kata penyusunnya.

Pada eksperimen pertama dan kedua ekstraksi fitur menggunakan sentimen leksikon Inggris yakni *Sentiwordnet*. Tahapan proses ekstraksi fitur menggunakan *sentiwordnet* ditunjukkan oleh Gambar 3.7. Untuk mendapatkan *score* sebuah kata dari *sentiwordnet*, parameter inputan yang diperlukan adalah kata (*unigram*) dalam bahasa Inggris dan kelas (*part of speech*) dari kata tersebut. Jika ditemukan *synset* yang sesuai dengan kata tersebut maka dilakukan perhitungan *score* menggunakan Persamaan 3.1, 3.2, dan 3.3. Hasil perhitungan *score* tersebut dijadikan bobot kata yang disebut sebagai *sentiment score*.

Sedangkan eksperimen ketiga menggunakan sentimen leksikon Indonesia buatan Vania, dkk (Vania, et al., 2014). Tahapan proses ekstraksi fitur menggunakan sentimen leksikon Indonesia ditunjukkan oleh Gambar 3.8. Berbeda dengan *sentiwordnet* yang membutuhkan kelas (*part of speech*) kata untuk meng-ekstrak *score*, pada sentimen leksikon Indonesia ini parameter inputan yang dibutuhkan hanyalah kata (*unigram*) dalam bahasa Indonesia. Selanjutnya parameter inputan dicocokkan dengan daftar kata dalam sentimen leksikon Indonesia. Bobot kata diperoleh dari perhitungan frekuensi kemunculan kata yang sesuai dengan sentimen leksikon Indonesia.



Gambar 3.7 Proses ekstraksi fitur menggunakan sentimen leksikon Inggris.



Gambar 3.8 Proses ekstraksi fitur menggunakan sentimen leksikon Indonesia.

3.3.1 Ekstraksi Fitur Menggunakan *Sentiwordnet*

Pada eksperimen yang menggunakan *sentiwordnet*, fitur yang diambil adalah *term* (*unigram*) yang ada di *synset* dari *Sentiwordnet*. Karena *sentiwordnet* menggunakan bahasa Inggris, sedangkan data pada penelitian ini menggunakan bahasa Indonesia. Maka diperlukan proses *translate*, disini kami menggunakan bantuan *API Google Translate*. Selain itu, untuk mengekstrak *sentiment score* dari *Sentiwordnet* juga dibutuhkan *part-of-speech (POS)* dari kata tersebut.

Pada eksperimen pertama dan kedua, kami juga menambahkan proses deteksi kalimat pertanyaan menggunakan daftar kata tanya. Kalimat pertanyaan tersebut dianggap sebagai kelas netral. Hal ini dimaksudkan untuk mengurangi kesalahan klasifikasi pada kalimat pertanyaan yang mengandung kata bersentimen. Misalnya, “saya dapet info klo tgl 20 ada gangguan pdam di surabaya apa itu benar

dan didaerah mana surabayanya ". Pada kalimat pengaduan tersebut secara keseluruhan termasuk kalimat pertanyaan (netral). Akan tetapi, kalimat tersebut mengandung kata bersentimen yakni kata "gangguan" (*interruption*) dengan *score* -0.115 dan kata "benar" (*true*) dengan *score* 0.205. Jika tidak dilakukan deteksi kalimat pertanyaan menggunakan daftar kata Tanya, maka pengaduan tersebut dianggap kalimat yang mengandung sentimen.

Persamaan 3.1 merupakan rumus perhitungan bobot *score* sebuah *synset*, dimana untuk mendapatkan *score* dari sebuah *synset* (kata) dengan mencari selisih antara *positive score* dengan *negative score* dari *synset* (kata) tersebut. Selanjutnya dibandingkan dengan total keseluruhan index dari *synset* tersebut di *sentiwordnet*. Sedangkan Persamaan 3.2 digunakan untuk menghitung *sentiment score* dari sebuah kalimat. Yakni dengan menjumlahkan *sentiment score* dari semua kata penyusunnya

$$Score = \sum \frac{(PosScore - NegScore)}{tot\_index} \quad (3.1)$$

$$Sentence_{score} = \sum Score \quad (3.2)$$

Dimana :

PosScore	: *positivity score* dari *synset* (kata)
NegScore	: *negativity score* dari *synset* (kata)
tot_index	: jumlah indek dari *synset* yang dipisahkan oleh karakter #
Score	: bobot *score* dari kata (term)
$Sentence_{score}$	: *sentiment score* sebuah kalimat

Contoh perhitungan *score* dari kata "mati" (*dead*) pada *sentiwordnet* sebagai berikut. Pada *sentiwordnet* kata "*dead*" memiliki beberapa *score* seperti ditunjukkan oleh Tabel 3-6, dimana setiap *score* ditentukan oleh konteks kata tersebut terhadap kalimat seperti ditunjukkan oleh kolom *Gloss* pada Tabel 3-6. Kolom *PosScore* berisi skor positif dari sebuah *synset* (kata), dan kolom *NegScore*

berisi skor negatif dari sebuah *synset* (kata). Sedangkan kolom POS berisi kelas kata hasil *POS Tagging* dari *synset* (kata).

Tabel 3-6 Daftar *score* dari *synset* (kata) "*dead*" pada *sentiwordnet*.

POS	Pos Score	Neg Score	Synset Terms	Gloss
a	0.000	0.750	dead#1	no longer having or seeming to have or expecting to have life
a	0.000	0.625	dead#2	not showing characteristics of life especially the capacity to sustain life
a	0.000	0.625	dead#3	(followed by `to') not showing human feeling or sensitivity; unresponsive
a	0.625	0.000	dead#4	unerringly accurate;

Berdasarkan Tabel 3-6, untuk menghitung *sentiment score* dari *synset* (kata) "*dead*" dengan kelas kata sifat (*adjective*) yang disimbolkan dengan "*a*" seperti ditunjukkan oleh kolom POS pada Tabel 3-6 digunakan Persamaan 3.1 dan didapatkan *sentiment score* sebesar – 0.343 dengan rincian sebagai berikut :

$$Score = \sum \frac{(PosScore - NegScore)}{tot\_index}$$

$$Score = \frac{((0 - 0.75) + (0 - 0.625) + (0 - 0.625) + (0.625 - 0))}{4}$$

$$Score = -0.343$$

3.3.2 Ekstraksi Fitur Menggunakan Sentimen Leksikon Indonesia

Sama halnya dengan tahapan pada *Sentiwordnet*, fitur yang diambil adalah *term* (*unigram*) yang ada di sentimen leksikon Indonesia. Selanjutnya, perhitungan bobot dari fitur tersebut menggunakan frekuensi kemunculan dari kata bersentimen positif maupun negatif pada sebuah kalimat. Selanjutnya fitur tersebut akan digunakan untuk proses klasifikasi, seperti ditunjukkan oleh Persamaan 3.3. Jika kata memiliki *polarity* negatif maka frekuensi yang didapatkan disimbolkan sebagai angka negatif, begitu pula sebaliknya ketika *polarity* dari kata positif maka disimbolkan sebagai angka positif. Sedangkan Persamaan 3.4 digunakan untuk

menghitung *sentiment score* dari kalimat, yakni dengan menjumlahkan *sentiment score* setiap kata penyusunnya.

$$Score = \sum frek_{Opini\_word} \quad (3.3)$$

$$Sentence_{score} = \sum_{i=0}^{n} Score_i \quad (3.4)$$

Dimana :

$frek_{Opini\_word}$: frekuensi kemunculan kata yang mengandung opini / sentimen

3.4 Proses Klasifikasi

Output kelas yang diharapkan dalam penelitian ini adalah sentimen positif, sentimen negatif dan netral. Kami menggunakan algoritma *rule based method* untuk klasifikasi sentimen pada pengaduan masyarakat dengan membandingkan fitur yang didapatkan pada setiap kalimat pengaduan.

3.4.1 Proses Klasifikasi Pada Eksperimen Menggunakan *Sentiwordnet*

Sentiment score dari sebuah kalimat diperoleh dari perhitungan *sentiment score* kata penyusunnya dan jumlah fitur (term) yang mengandung opini / sentimen yang terdapat pada kalimat tersebut (Cernian & Sgarciu, 2015). Persamaan 3.5 menyajikan rumus perhitungan *sentiment score* dari sebuah kalimat.

$$Sentence_{score} = \frac{\sum_{i=0}^{n} Score_i}{num\_words} \quad (3.5)$$

Dimana :

num_words : jumlah kata yang mengandung opini / sentimen dalam sebuah kalimat

Rule yang digunakan untuk menentukan sentimen dari kalimat pengaduan menggunakan *Sentiwordnet* sebagai berikut :

- Jika *sentiment score* dari kalimat lebih besar dari batas atas nilai *threshold* kelas netral. Maka kalimat diklasifikasikan ke dalam kelas positif.
- Jika *sentiment score* dari kalimat lebih kecil dari batas bawah nilai *threshold* kelas netral. Maka kalimat diklasifikasikan ke dalam kelas negatif.
- Jika *sentiment score* dari kalimat berada di dalam *range* nilai *threshold* kelas netral. Maka kalimat diklasifikasikan ke dalam kelas netral.

3.4.2 Proses Klasifikasi Pada Eksperimen Dengan Sentimen Leksikon Indonesia

Berbeda dengan teknik perhitungan *sentiment score* sebuah kalimat pada *Sentiwordnet*, dimana *sentiment score* dari kata penyusunnya dijumlahkan. Pada sentimen leksikon Indonesia, sentimen dari kalimat diperoleh dengan membandingkan frekuensi kemunculan kata bersentimen positif dan negatif. *Rule* yang digunakan untuk menentukan sentimen dari kalimat pengaduan menggunakan sentimen leksikon Indonesia sebagai berikut :

- Jika jumlah frekuensi kemunculan kata bersentimen positif lebih besar dari kata bersentimen negatif. Maka kalimat diklasifikasikan ke dalam kelas positif.
- Jika jumlah frekuensi kemunculan kata bersentimen negatif lebih besar dari kata bersentimen positif. Maka kalimat diklasifikasikan ke dalam kelas negatif.
- Jika jumlah frekuensi kemunculan kata bersentimen positif sama dengan kata bersentimen negatif. Maka kalimat diklasifikasikan ke dalam kelas netral.
- Jika kalimat tidak ditemukan kata bersentimen positif maupun kata bersentimen negatif. Maka kalimat diklasifikasikan ke dalam kelas netral.

3.5 Tahapan Pengujian dan Validasi

Pengukuran kinerja dari sistem / metode klasifikasi dilakukan dengan menghitung nilai *accuracy*, *recall* dan *precision*. *Accuracy* diperoleh dengan membandingkan jumlah data hasil klasifikasi (prediksi) yang sesuai dengan jumlah keseluruhan data. Semakin tinggi nilai akurasi yang diperoleh, maka hasil

klasifikasi semakin baik. Akan tetapi, jika hanya melihat nilai akurasi saja tidak bisa mendeteksi adanya penyimpangan data. Oleh karena itu kami juga menghitung nilai *recall* dan *precision*.

Recall diperoleh dengan membandingkan jumlah data hasil klasifikasi yang relevan dan total data yang dianggap relevan. Sedangkan *precision* diperoleh dengan membandingkan jumlah data hasil klasifikasi yang relevan dan total jumlah data yang ditemukan pada kelas tertentu. Perhitungan yang digunakan untuk menghitung nilai *accuracy*, *recall*, dan *precision* ditunjukkan oleh Persamaan 3.6, Persamaan 3.7, dan Persamaan 3.8.

$$Accuracy = \frac{\sum pb}{num\_data} \tag{3.6}$$

$$recall = \frac{\sum pb}{\sum total\_kelas} \tag{3.7}$$

$$precision = \frac{\sum pb}{\sum total\_pisah} \tag{3.8}$$

Halaman ini sengaja dikosongkan

BAB 4
HASIL DAN PEMBAHASAN

Bab ini membahas hasil penelitian *sentiment analysis* terhadap data pengaduan masyarakat pada media *twitter* dan website resmi layanan pengaduan pemerintah kota Surabaya menggunakan pendekatan *semantic* (makna kata) sesuai tahapan yang sudah dijelaskan pada bab sebelumnya. Bab ini menyajikan apakah sentimen leksikon dalam bahasa Inggris yakni *Sentiwordnet* sesuai jika diterapkan pada data pengaduan masyarakat Indonesia terutama data media sosial yang sebagian besar menggunakan bahasa informal.

4.1 Data Eksperimen

Data diambil dari *twitter* @sapawargasby dan *media center* surabaya sebanyak 1000 data untuk masing-masing sumber. Setelah melalui tahap normalisasi data dan membuang data yang sama (*duplicate*) diperoleh data sebanyak 685 pengaduan pada *twitter*, dan 672 pengaduan pada *media center*.

Tahapan awal setelah mendapatkan data hasil normalisasi adalah melakukan pelabelan data sesuai kelas yang diharapkan. Dalam penelitian ini, *output* kelas yang diharapkan adalah kelas sentimen positif, negatif dan netral. Dalam proses pelabelan data, kriteria yang digunakan sebagai berikut :

- Pengaduan dilabeli kelas sentimen positif, jika kalimat pengaduan tersebut mengandung kata-kata yang bermakna sentimen (emosi) positif. Begitu pula ketika jumlah kata dengan sentimen (emosi) positif lebih banyak dibandingkan jumlah kata dengan sentimen (emosi) negatif.
- Pengaduan dilabeli kelas sentimen negatif, jika kalimat pengaduan tersebut mengandung kata-kata yang bermakna sentimen (emosi) negatif. Begitu pula ketika jumlah kata dengan sentimen (emosi) negatif lebih banyak dibandingkan jumlah kata dengan sentimen (emosi) positif.
- Pengaduan dilabeli kelas sentimen netral, jika kalimat tersebut tidak mengandung kata bersentimen. Begitu pula ketika jumlah kata dengan sentimen (emosi) positif sama dengan jumlah kata dengan sentimen (emosi) negatif.

Pelabelan data dilakukan secara manual menggunakan tiga (3) *annotator* seperti disajikan dalam Tabel 4-1. Label dari data diambil dari suara terbanyak sesuai kriteria pelabelan diatas. Ketika semua *annotator* berbeda pendapat, maka kami menggunakan pendapat *annotator* lain sebagai pembanding.

Tabel 4-1 Data Penelitian Pengaduan Masyarakat

Sentiment	**Source Data**	
	Twitter	***Media Center***
Positif	133	74
Negatif	271	384
Netral	281	214
Total	685	672

Berdasarkan Tabel 4-1 diperoleh informasi bahwa jumlah pengaduan yang memiliki sentimen positif lebih sedikit dibandingkan pengaduan dengan sentimen negatif dan netral, baik pada media *twitter* maupun *media center* yakni sekitar 15% saja. Sedangkan pengaduan dengan sentimen negatif mencapai 48%, sisanya sebanyak 36% termasuk pengaduan dengan sentimen netral. Hal ini dikarenakan sebagian besar masyarakat menggunakan media tersebut sebagai alat untuk menyampaikan kekesalan dan ketidak puasan mereka terhadap fasilitas umum dan pelayanan publik. Disamping itu, media tersebut juga dijadikan masyarakat sebagai alat untuk berkomunikasi dengan pemerintah ketika ada kebijakan maupun prosedur yang kurang jelas buat mereka.

4.2 Hasil Tahapan Pre-processing

Sebelum dilakukan *pre-processing* data pengaduan mengandung banyak *noise* terutama data pada media *twitter*, seperti duplikasi data karena pesan di *retweet* (RT) atau di *sharing* kembali, mengandung informasi yang tidak berguna untuk klasifikasi seperti *hashtag* (#), *mention* (@), *html* (http://www), angka, penggunaan bahasa lokal serta singkatan seperti disajikan oleh Tabel 4-2. Serta banyak terdapat duplikasi data, karena *tweet* yang sama diposting kembali oleh orang lain (RT). Sedangkan data pengaduan pada *media center* tidak banyak

mengandung *noise* seperti data *twitter*, hanya saja kadang kala ditemukan penggunaan bahasa lokal dan informal meskipun jumlahnya tidak banyak seperti ditunjukkan oleh Tabel 4-3.

Tabel 4-2 Data pengaduan sebelum *pre-processing* pada media *twitter*

No	from_user	text	time
1	rafika nilasari	Banyaaaaak pengendara tanpa helm yg urakan di jalan. Tapii tidak ada polantas sama sekali. @e100ss @SbyTrafficServ @SapawargaSby @RTMCJatim	11/19/2016 3:55:39 PM
2	oyest28	RT @bonekstudent27: Hilang satu tumbuh seribu spanduk perlawanan di surabaya #BonekMelawan @SapawargaSby	11/18/2016 6:39:48 PM
3	v14nzone	@SapawargaSby saya dapet info klo tgl 20 ada gangguan PDAM di Surabaya apa itu benar? Dan didaerah mana surabayanya?? #tanya	11/18/2016 2:45:44 PM
4	ceritasby	RT @KORANSINDO_JTIM: Koran Sindo Jatim, 18 Nov 2016, Jatim Dirikan SMA Unggulan @jatimpemprov @dindik_jatim @Kemdikbud_RI @SapawargaSby @ce…	11/18/2016 2:08:16 PM
5	opqmaulana	Selamat untuk Kota Surabaya atas diraihnya Juara III Lomba Penerbitan Media Humas Internal Kategori untuk K/D/kab dan kota. @SapawargaSby	11/18/2016 1:50:33 PM
6	cemplukAe	RT @bonekcasuals: Nah kan @SapawargaSby @satpolppsby https://t.co/8hobLtEpNk	11/18/2016 1:32:04 PM
7	ardiano_js	@SapawargaSby Min mw tny utk bantuan prmhnan cat, serta material utk perbaikan kampung bs y? Klo bs sbg syaratny apa saja??? Terima kasih.	11/18/2016 11:39:11 AM
8	brahmanluqman	RT @bonekstudent27: Hilang satu tumbuh seribu spanduk perlawanan di surabaya #BonekMelawan @SapawargaSby	11/18/2016 10:10:18 AM
9	thegaruda45	RT @bonekcasuals: Passing @SapawargaSby @satpolppsby semoga dibaca https://t.co/jxmUR82pMf	11/18/2016 9:49:03 AM
10	Tatra Romadhoni	Respon yg sangat cepat! Terimaksih satgas DPUBMP Pemkot Sby @SapawargaSby @e100ss https://t.co/RQxZeeDZdZ	11/18/2016 2:57:45 AM

Tabel 4-3 Data pengaduan sebelum *pre-processing* pada *media center*

No	Topik	Jenis	Pengaduan
1	mati lampu	Keluhan	Lampu PJU di Jl Ngagel mati
2	jalan rusak	Keluhan	keluhan dari TPKPM
3	mati lampu	Keluhan	Lampu PJU di Jl. ngagel tama selatan kalau siang nyala kalau malam padam. Mohon segera diperbaiki
4	BLC	Permohonan informasi	Syarat apa saja yang dibutuhkan untuk mendaftar BLC ?
5	macet	Keluhan	(Sumber : KORAN SURYA) TARIF PARKIR TIDAK SESUAI KARCIS. Kepada Dinas terkait, mohon untuk menertibkan tukang parkir yang ada di depan Siloam Hospital, depan BRI Surabaya gubeng. Masak tarif parkir gak sesuai dengan yang tertera di karcis Rp 3000. Masak tarif Rp. 5000 sampai Rp 10.000. Kalau gak mau bayar disuruh pergi. Itupun mereka sangat-sangat memaksa kayak preman. Saya bayar sesuai yg tertera di karcis dia marah gak mau. Bagaimana ini?
6	mati lampu	Keluhan	Pagi pak, PJU di sepanjang jalan Kelurahan Babat Jerawat samapai Pakal banyak yang mati, mohon segera diperbaiki. Terima kasih.
7	PJU	Keluhan	Mohon di perbaiki PJU yg mati di tempat kami tinggal tepatnya d RT 06 - RW 08 Karang asem utara kelurahan ploso kecamatan Tambaksari.menginggat PJU tersebut berada d area vasum lapangan volley jadi d tempat tersebut terasa gelap.apalagi bersamaan pembangunan selokan d daerah kami.mungkin bisa membahayan bagi pengguna jalan di area tersebut.terima kasih atas kerjasamanya.

4.2.1 Hasil Tahapan Cleansing Data

Tahapan *cleansing data* bertujuan membuang karakter yang tidak memberikan pengaruh terhadap proses klasifikasi sentimen seperti *hashtag* (#), *mention* (@), *html* (http://www), karakter selain huruf, angka desimal dan romawi. Kemudian data diubah menjadi huruf kecil. Disamping itu, tahapan ini juga

membersihkan duplikasi data yang berasal dari tindakan *retweet* (RT). Tabel 4-4 menampilkan contoh pengaduan pada media *twitter* sebelum di *cleansing*. Sedangkan Tabel 4-5 menyajikan data setelah proses *cleansing*. Dari Tabel 4-4 didiketahui data masih mengandung *noise* seperti *hashtag* (#), *mention* (@), *html* (http://www), serta *retweet* (RT). Akan tetapi setelah dilakukan proses *cleansing*, data yang dihasilkan bersih dari *noise*, duplikasi data, dan hanya berisi huruf yang nantinya menjadi fitur untuk proses klasifikasi selanjutnya.

Tabel 4-4 Data Pengaduan Sebelum Proses *Cleansing*

No	Pengaduan
1	RT @SapawargaSby: Kedamaian tidak bisa dijaga dengan paksaan. Kedamaian hanya bisa dicapai dengan saling pengertian"- Albert Enstein
2	RT @SapawargaSby: Orang yang kuat adalah orang yang tinggi kesabarannya dan tinggi kesalehannya. ~Ir Joko Widodo. #morningquote
3	@brainwash18 @SapawargaSby oke, maturnuwun infone mas :))
4	@SapawargaSby jalan disekolahan SD Bendul merisi (sampai bebek maryem) sangat parah mohon dilakukan pengaspalan #keluhan
5	Surabaya Punya Lapangan Hockey-Softball Internasional @PemkotSurabaya @DisporaSby @KEMENPORA_RI @ceritasby… https://t.co/kP7CV5tCkp
6	@SapawargaSby min, mau tanya, persyaratan apa saja yg diperlukan utk program sertifikasi tanah massal tahun 2016 ini, suwun infone
7	Kumuhnya Surabaya....daerah pacar keling. https://t.co/2BpXl98TM1

Tabel 4-5 Data Pengaduan Setelah Proses *Cleansing*

No	Pengaduan
1	oke maturnuwun infone mas
2	jalan disekolahan sd bendul merisi sampai bebek maryem sangat parah mohon dilakukan pengaspalan
3	surabaya punya lapangan hockey softball internasional
4	min mau tanya persyaratan apa saja yg diperlukan utk program sertifikasi tanah massal tahun ini suwun infone
5	kumuhnya surabaya daerah pacar keling

4.2.2 Hasil Tahapan Formalization

Pada tahapan ini dilakukan pemecahan kata (*tokenizing*) dan *term* (kata) yang tidak sesuai kamus diubah ke bentuk baku sesuai kamus besar bahasa Indonesia (KBBI). Tahapan ini memegang peranan yang cukup penting. Karena pada proses *translate*, jika kata tidak sesuai bentuk baku sesuai Kamus Besar Bahasa Indonesia (KBBI) maka hasil *translate*-nya juga tidak akan sesuai dengan makna kata aslinya. Pemecahan kata (*tokenizing*) menggunakan delimiter spasi (*space*), karena ketika proses *cleansing* karakter selain huruf digantikan dengan spasi. Selanjutnya, untuk mengembalikan kata singkatan ke bentuk aslinya sesuai kamus bahasa Indonesia menggunakan algoritma *laveinstein distance*. Contoh penerapan algoritma *laveinstein distance* disajikan oleh Tabel 4-6. Kolom Inputan menyajikan data hasil proses *cleansing*, dimana sudah tidak ditemukan karakter *noise* dalam data. Akan tetapi data masih mengandung kata singkatan seperti "jd", "jl", "kpn" dan "tnmn". Sedangkan kolom hasil menampilkan data setelah di normalisasi menggunakan algoritma *laveinstein distance* seperti kata "jl" menjadi "jalan", dan "tnmn" menjadi "tanaman". Langkah terakhir, digunakan bantuan kamus lokal untuk mengubah kata informal dan bahasa lokal yang tidak ada di kamus bahasa Indonesia ke bentuk baku sesuai kamus bahasa Indonesia seperti ditunjukkan oleh Tabel 4-7. Sama halnya dengan Tabel 4-6, kolom inputan pada Tabel 4-7 mengandung kata informal atau bahasa lokal seperti kata "gag", "duwek", "cpoti", dan "pancali". Dan kolom hasil menampilkan hasil normalisasi menggunakan bantuan kamus lokal seperti kata "gag" dan "gak" diubah menjadi "tidak" dan "duwek" diubah menjadi "uang".

Pada tahapan ini juga dilakukan proses *filtering* kalimat pertanyaan sebagai kelas netral, dengan menggunakan daftar kata tanya seperti "apa", "mana", "tanya", "dimana", "gimana" dan "apakah". Jika kalimat mengandung kata Tanya, maka secara otomatis kalimat tersebut digolongkan sebagai kelas netral tanpa melalui tahapan *scoring*. Hal ini dilakukan untuk menghindari terjadinya kesalahan klasifikasi terhadap kalimat pertanyaan yang didalamnya mengandung kata bersentimen, yang nantinya bisa mengubah sentimen dari kalimat tersebut.

Tabel 4-6 Penerapan Algoritma *laveinstein distance*

No	Inputan	Hasil
1	jd sepanjang jl gunungsari arah mastrip macet krn byk anak brutal usai konser lalu polantas kemana	jadi sepanjang jalan gunungsari arah mastrip macet karna banyak anak brutal usai konser lalu polantas kemana
2	kpn bos pohon di jln pulo wonokromo dpn rumah no dipotong phn itu bener membahayakan karena dia tinggi besar	kapan bos pohon di jalan pulo wonokromo depan rumah no dipotong pohon itu bener membahayakan karena dia tinggi besar
3	tumbuhan toga yg stu ini sering dianggp tnmn liar biasa pdhl manfaatnya luarrr biasa loh ada yg tau tanaman apa i	tumbuhan toga yang satu ini sering dianggp tanaman liar biasa padahal manfaatnya luarrr biasa loh ada yang tau tanaman apa
4	barusan dpt info via wassap ttg aturan sertifikasi tnh yg msh suratijo petokd apkh info tsb benar	barusan dapat info via wassap tentang aturan sertifikasi tanah yang masih suratijo petokd apakah info tsb benar
5	cc mhon kanan amp kiri jln tsb diberi got air hujan selalu tergenang penyebab aspal rusak jalan jg g	cc mohon kanan amp kiri jalan tsb diberi got air hujan selalu tergenang penyebab aspal rusak jalan juga

Tabel 4-7 Normalisasi menggunakan kamus lokal

No	Inputan	Hasil
1	spanduke parpol kok gag dilucuti itu cukup merusak keindahan kota	spanduknya parpol kok tidak dicopoti itu cukup merusak keindahan kota
2	ngunu lak gak dikei duwek spanduk prjuangan d cpoti	itu kalau tidak dikasih uang spanduk perjuangan d lepas
3	ketemu di pancali ae justru sg g sneng iku seng dadi kaki tgane mafia	bertemu di ditendang saja justru yang tidak suka itu yang jadi kaki tangan mafia
4	bocoen iku ma risma ojok tanduran ae di urus wargamu urusen	bacalah itu ma risma jangan tanaman saja di urus warga kamu uruskan
5	bangga dadi wong suroboyo cuk repost	bangga jadi orang surabaya cuk repost

4.2.3 Hasil TahapanTranslate

Proses *translate* diperlukan untuk mengubah data dari bahasa Indonesia ke bahasa Inggris. Disini kami menggunakan bantuan *Google API Translate* V.2 untuk proses *translate*-nya. Kata hasil translate tersebut nantinya digunakan untuk mengekstrak *sentiment score* pada *Sentiwordnet*. Tahapan *translate* ini diperlukan oleh eksperimen pertama dan kedua saja, karena kedua eksperimen tersebut memanfaatkan *Sentiwordnet* untuk proses ekstraksi fiturnya. Sedangkan eksperimen ketiga menggunakan sentimen leksikon Indonesia untuk proses ekstraksi fitur, sehingga tidak memerlukan tahapan *translate*.

Pada eksperimen pertama, proses *translate* dilakukan pada kata dasar setelah proses *stopword removal* dan *stemming* seperti ditunjukkan oleh Tabel 4-8. Kolom normalisasi pada Tabel 4-8 berisi data hasil normalisasi pada tahapan *formalization*, dan kolom *stemming* berisi daftar kata dasar hasil proses *stopword removal* dan *stemming*. Terakhir kolom *translate* berisi kata hasil *translate*. Sedangkan pada eksperimen kedua, proses *translate* dilakukan pada kalimat setelah proses normalisasi seperti ditunjukkan oleh kolom *translate* pada Tabel 4-9. Dan kolom normalisasi pada Tabel 4-9 berisi kalimat *output* dari tahapan *formalization*.

Tabel 4-8 Proses *translate* pada eksperimen pertama

No	Normalisasi	*Stemming*	*Translate*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	kok belum tindak bos pohon jalan pulo pohon sono	why not yet acts boss tree street pulo tree sono
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	jalan sekolah bebek maryem parah mohon aspal	street school duck maryem severe please asphalt
3	bapak ingin tanya persyaratan apa saja yang diperlukan untuk program sertifikasi tanah massal tahun ini infonya	syarat sertifikasi tanah massal info	requirement certification soil bulk info
4	kumuhnya surabaya daerah	kumuh surabaya daerah	slum surabaya area
5	airnya mati lebih jam tolong hidupkan	air mati tolong hidup	water die please life

Tabel 4-9 Proses *translate* pada eksperimen kedua

No	Normalisasi	*Translate*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	why no boss action for the tree on the road pulo there is a big sono tree
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	road to school until duck maryem very bad please do asphalt
3	bapak ingin tanya persyaratan apa saja yang diperlukan untuk program sertifikasi tanah massal tahun ini infonya	You want to ask what conditions are required for this year's mass land certification program
4	kumuhnya surabaya daerah	kumuhnya surabaya area
5	airnya mati lebih jam tolong hidupkan	the water dies over hours please turn it on

4.2.4 Hasil Part of Speech (POS) Tagging

POS Tagging bertujuan mendapatkan jenis (kelas) kata dalam sebuah kalimat. Kelas (jenis) kata yang ada di *Sentiwordnet* diantaranya kata benda (*noun*), kata sifat (*adjective*), kata kerja (*verb*) dan kata keterangan (*adverb*). Selanjutnya kelas kata tersebut akan digunakan untuk ekstraksi *sentiment score* di *Sentiwordnet*. *Sentiwordnet* digunakan untuk proses ekstraksi fitur dalam eksperimen pertama dan kedua, akan tetapi *POS Tagging* hanya diterapkan pada eskperimen kedua saja. Sedangkan eksperimen pertama kelas kata diperoleh dari Kamus Besar Bahasa Indonesia (KBBI) tanpa *POS Tagging*. Hal ini dikarenakan proses *translate* pada eksperimen pertama dilakukan pada kata dasar setelah proses *stemming* berdasarkan kaidah Kamus Besar Bahasa Indonesia (KBBI). Sehingga jenis (kelas) dari kata diambil dari informasi sesuai Kamus Besar Bahasa Indonesia (KBBI). Algoritma *POS Tagging* yang digunakan pada eksperimen kedua adalah Algoritma Bill Tagger.

Contoh hasil penerapan *POS Tagging* disajikan oleh Tabel 4-10, sedangkan Tabel 4-11 menyajikan contoh kelas kata dengan menggunakan Kamus Besar Bahasa Indonesia, seperti ditunjukkan oleh kolom Kelas Kata pada Tabel 4-11. Berdasarkan informasi yang disajikan oleh Tabel 4-10, proses *POS Tagging*

dilakukan pada data hasil normalisasi (*formalization*). Dan kelas kata terhadap kalimat disajikan pada kolom *POS Tagging* Tabel 4-10.

Tabel 4-10 Hasil penerapan *POS Tagging* pada eksperimen pertama

No	***Translate***	***POS Tagging***
1	why no boss action for the tree on the road pulo there is a big sono tree	why = *adverb* boss = *noun* action = *noun* tree = *noun* big = *adjective*
2	road to school until duck maryem very bad please do asphalt	road = *noun* school = *noun* very = *adverb* bad = *adjective* please = *verb* do = *verb* asphalt = *noun*
3	You want to ask what conditions are required for this year's mass land certification program	want = *verb* ask = *verb* conditions = *noun* required = *verb* year = *noun* mass = *noun* land = *noun* certification = *noun* program = *noun*

Tabel 4-11 Hasil kelas kata menggunakan kamus bahasa Indonesia

No	***Normalisasi***	***Stemming***	***Kelas Kata***
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	belum tindak bos pohon jalan pulo pohon sono	belum = keterangan tindak = kata benda bos = kata benda pohon = kata benda jalan = kata benda
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	jalan sekolah bebek maryem parah mohon aspal	jalan = kata benda sekolah = kata benda parah = kata sifat mohon = kata kerja aspal = kata benda

Tabel 4-11 Hasil kelas kata menggunakan kamus bahasa Indonesia (Lanjutan)

No	*Normalisasi*	*Stemming*	*Kelas Kata*
3	bapak ingin tanya persyaratan apa saja yang diperlukan untuk program sertifikasi tanah massal tahun ini infonya	syarat sertifikasi tanah massal info	syarat = kata benda sertifikasi = kata benda tanah = kata benda massal = kata sifat info = kata benda

4.2.5 Hasil Tahapan Filtering

Tahapan *Filtering* dilakukan untuk membuang kata-kata yang tidak berpengaruh dalam proses klasifikasi (*stopword removal*) dan hanya menyimpan kata-kata yang dianggap penting untuk tahapan klasifikasi selanjutnya. Pada eksperimen pertama dan ketiga *stopword list* yang digunakan adalah Tala (2003) (Tala, 2003) dalam bahasa Indonesia. Sedangkan pada eksperimen kedua *filtering* menggunakan *stopword list* dalam bahasa Inggris. Karena proses *translate* dilakukan pada kalimat hasil normalisasi yang belum melalui tahapan *filtering*. Untuk eksperimen ketiga kami menggunakan *stopword list* dari NLTK *English stopword*. Contoh hasil *filtering* menggunakan *stopword list* bahasa Indonesia ditunjukkan oleh Tabel 4-12, misalnya pada kalimat "kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar", maka kata yang termasuk dalam *stopword* dan harus dihilangkan adalah kata "kok", "ada", "untuk", "di", "itu", "ada", dan "besar". Karena kata-kata tersebut tidak memiliki makna yang berpengaruh terhadap proses klasifikasi.

Tabel 4-13 menyajikan hasil *filtering* menggunakan *stopword list* bahasa Inggris, karena proses *translate* dilakukan terhadap kalimat hasil normalisasi. Misalnya, pengaduan "kok belum ada tindakan bos untuk pohon dijalan pulo itu ada pohon sono besar" dengan hasil *translate* "why there is no boss action for the tree on the road pulo there is a big sono tree". Maka kata yang termasuk dalam *stopword* dan harus dihilangkan adalah kata "why", "there", "is", "for", "the", "on", and "big" karena kata-kata tersebut tidal memiliki makna yang berpengaruh dalam proses klasifikasi.

Tabel 4-12 *Filtering* menggunakan *stopword list* bahasa Indonesia

No	Normalisasi	*Filtering*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	belum tindakan bos pohon jalan pulo pohon sono
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	jalan disekolahan bebek maryem parah mohon pengaspalan
3	bapak ingin tanya persyaratan apa saja yang diperlukan untuk program sertifikasi tanah massal tahun ini infonya	persyaratan sertifikasi tanah massal infonya
4	kumuhnya surabaya daerah	kumuhnya surabaya daerah
5	airnya mati lebih jam tolong hidupkan	airnya mati tolong hidupkan

Tabel 4-13 *Filtering* menggunakan *stopword list* bahasa Inggris

No	Normalisasi	*Translate*	*Filtering*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	why there is no boss action for the tree on the road pulo there is a big sono tree	no boss action tree road pulo sono tree
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	road to school until duck maryem very bad please do asphalt	road school duck maryem bad please asphalt
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	admin and mum guardian our esteemed citizen rt rw ir we die already more ja	admin mum guardian esteemed citizen rt rw ir die ja
4	kumuhnya surabaya daerah	kumuhnya surabaya area	kumuhnya surabaya
5	airnya mati lebih jam tolong hidupkan	the water dies over hours please turn it on	the water dies over hours please turn it on

4.2.6 Hasil Tahapan Stemming

Stemming merupakan tahap terakhir dalam *pre-processing*, yang bertujuan mengembalikan kata ke bentuk dasar sesuai kamus. Dalam penelitian ini kami menggunakan dua acuan kamus, yakni kamus bahasa Indonesia untuk eksperimen

pertama dan ketiga dan *Wordnet* untuk eksperimen kedua. Tahapan *stemming* ini diperlukan, karena sebagian besar sentimen leksikon di *Sentiwordnet* dan sentimen leksikon Indonesia dalam bentuk kata dasar. *Stemming* dengan acuan kamus bahasa Indonesia diterapkan pada eksperimen pertama dan ketiga menggunakan algoritma Nazief dan Adriani seperti ditunjukkan oleh Tabel 4-14 dimana hasil dari proses *filtering* kemudian dikembalikan ke bentuk dasar (*stemming*) misalnya pada kata "tindakan" diubah menjadi "tindak". Sedangkan *stemming* pada eksperimen kedua menggunakan algoritma *potter stemmer* dan *wordnet* seperti terlihat pada Tabel 4-15 dimana *stemming* dilakukan pada data hasil *filtering*. Seperti kata "esteemed" dikembalikan ke bentuk dasar menjadi "esteem".

Tabel 4-14 *Stemming* menggunakan acuan kamus bahasa Indonesia

No	***Filtering***	***Stemming***
1	belum tindakan bos pohon jalan pulo pohon sono	belum tindak bos pohon jalan pulo pohon sono
2	jalan disekolahan bebek maryem parah mohon pengaspalan	jalan sekolah bebek maryem parah mohon aspal
3	persyaratan sertifikasi tanah massal infonya	syarat sertifikasi tanah massal info
4	kumuhnya surabaya daerah	kumuh surabaya daerah
5	airnya mati tolong hidupkan	air mati tolong hidup

Tabel 4-15 *Stemming* dengan algoritma *potter stemmer* dan *wordnet*

No	***Filtering***	***Stemming***
1	no boss action tree road pulo sono tree	no boss action tree road pulo sono tree
2	road school duck maryem bad please asphalt	road school duck maryem bad please asphalt
3	admin mum guardian esteemed citizen rt rw ir die ja	admin mum guardian esteem citizen rt rw ir die ja
4	kumuhnya surabaya	kumuhnya surabaya
5	water dies hours please	water die hours please

4.3 Hasil Proses Ekstraksi Fitur

Tahapan ini memegang peranan penting dalam klasifikasi. Fitur yang digunakan dalam penelitian ini adalah *sentiment word*. *Sentiment word* adalah kata yang mengandung emosi (sentimen), baik emosi positif maupun emosi negatif. Salah satu contohnya kata "buruk" yang mengandung emosi (sentimen) negatif dan kata "indah" yang mengandung emosi (sentimen) positif.

Karena penelitian ini menggunakan pendekatan semantik (makna kata), kami menggunakan sentimen leksikon bahasa Inggris yakni *Sentiwordnet* dan sentimen leksikon bahasa Indonesia yang dibuat oleh Vania dkk (Vania, et al., 2014) untuk mengekstraksi fitur untuk proses klasifikasi sentimen teerhadap data pengaduan masyarakat.

4.3.1 Hasil Ekstraksi Fitur Menggunakan *Sentiwordnet*

Sentimen leksikon (*synset*) yang terdapat dalam *Sentiwordent* mencapai 207.000 kata. *Synset* tersebut diadopsi dari *Wordnet* dengan memperhatikan sentimen (emosi) yang dikandung kata tersebut dalam konteks tertentu dan dinyatakan dalam bentuk angka dengan *range* antara 0.0 sampai dengan 1.0 untuk masing-masing kategori. Sentimen negatif dinyatakan dalam *range* 0.0 sampai dengan (-1.0) dan sentimen positif dinyatakan dalam *range* 0.0 sampai dengan 1.0. Seperti pada kata "rusak" (*damaged*) memiliki *positive score* 0.0 dan *negative score* 0.75 untuk konteks "dirugikan atau terluka". Sedangkan kata "rusak" (*damaged*) memiliki *positive score* 0.375 dan *negative score* 0.5 untuk konteks "menjadi tidak adil dibawa ke dalam keburukan".

Dalam penelitian ini *Sentiwordnet* digunakan pada eksperimen pertama dan kedua untuk ekstraksi fitur. Tabel 4-16 menyajikan contoh penerapan *Sentiwordnet* pada eksperimen pertama. Tabel 4-17 menyajikan contoh penerapan *Sentiwordnet* pada eksperimen kedua. Kolom *translate* berisi hasil *translate* kata atau kalimat pengaduan, dan kolom *sentiment score* berisi daftar kata (*term*) yang terdapat di *sentiwordnet* beserta *score*-nya. Selanjutnya *score* dari kata penyusun kalimat tersebut yang akan digunakan untuk menentukan *sentiment score* dari kalimatnya.

Tabel 4-16 Contoh penerapan *Sentiwordnet* pada eksperimen pertama

No	Normalisasi	*Translate*	*Sentiment Score*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	not yet acts boss tree street pulo tree sono	not (*r*) = - 0.625 act (*v*) = 0.043 boss (*n*) = 0 tree (*n*) = 0 street (*n*) = 0.014
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	street school duck maryem severe please asphalt	street (*n*) = 0.014 school (*n*) = 0 severe (*a*) = - 0.41 please (*v*) = 0.34 asphalt (*n*) = 0
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	admin guardian respect citizens die	admin (*n*) = 0 guardian (*n*) = 0 respect (*a*) = 0 citizen (*n*) = 0 die (*v*) = - 0.175
4	kumuhnya surabaya daerah	slum surabaya area	slum (*a*) = 0 area (*n*) = 0
5	airnya mati lebih jam tolong hidupkan	water die please life	water (*n*) = 0.023 die (*v*) = - 0.175 please (*v*) = 0.34 life (*v*) = 0

Tabel 4-17 Contoh penerapan *Sentiwordnet* pada eksperimen kedua

No	Normalisasi	*Translate*	*Sentiment Score*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	no boss action for the tree on the road pulo there is a big sono tree	boss (*n*) = 0 action (*n*) = 0.02 tree (*n*) = 0 road (*n*) = 0
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	road to school until duck maryem very bad please do asphalt	road (*n*) = 0 school (*n*) = 0 duck (*n*) = - 0.02 maryem (*n*) = 0 bad (*a*) = - 0.57 please (*v*) = 0.34 asphalt(*n*) = 0

Tabel 4-17 Contoh penerapan *Sentiwordnet* pada eksperimen kedua (Lanjutan)

No	Normalisasi	*Translate*	*Sentiment Score*
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	admin and mum guardian our esteemed citizen rt rw ir we die already more ja	admin (*n*) = 0 mum (*a*) = 0 guardian (*n*) = 0 esteem (*v*) = 0.33 citizen (*n*) = 0 die (*v*) = - 0.175
4	kumuhnya surabaya daerah	kumuhnya surabaya area	kumuhnya (*n*) = 0 surabaya (*n*) = 0
5	airnya mati lebih jam tolong hidupkan	the water dies over hours please turn it on	water (*n*) = 0.022 die (*v*) = - 0.175 hours (*n*) = 0 please (*v*) = 0.34

4.3.2 Hasil Ekstraksi Fitur Menggunakan Sentimen Leksikon Indonesia

Sentimen leksikon Indonesia berisi daftar kata yang mengandung sentimen (emosi) positif maupun negatif. Berbeda dengan *sentiwordnet*, kata-kata dalam sentimen leksikon Indonesia tidak memiliki *score* hanya *polarity* (kecenderungan) positif atau negatif saja. Dalam penelitian ini, kami menggunakan sentimen leksikon buatan Vania dkk (2014) (Vania, et al., 2014) yang terdiri dari 415 kata sentimen positif dan 581 kata sentimen negatif. Penggunaan sentimen leksikon Indonesia ini diaplikasikan pada eksperimen ketiga. Fitur yang digunakan dalam eksperimen ketiga ini adalah kata yang mengandung opini / sentimen saja. Perhitungan bobot dari fitur menggunakan frekuensi kemunculan kata bersentimen dalam sebuah kalimat. Contoh ekstraksi fitur menggunakan sentimen leksikon Indonesia ditunjukkan oleh Tabel 4-18.

Berbeda dengan eksperimen pertama dan kedua, pada eksperimen ketiga ini tidak diperlukan kelas (jenis) kata untuk proses ekstraksi fitur. Seperti terlihat pada Tabel 4-18, fitur yang diambil hanyalah kata yang terdapat dalam sentimen leksikon Indonesia saja, selain itu dibuang. Dan kolom *word count* pada Tabel 4-18 berisi daftar kata yang diambil sebagai fitur beserta bobot dari kata tersebut. Sedangkan kolom normalisasi pada Tabel 4-18 berisi kalimat hasil normalisasi,

dimana sebagian besar kata sudah diubah ke bentuk baku sesuai kaidah Kamus Besar Bahasa Indonesia (KBBI).

Tabel 4-18 Contoh ekstraksi fitur dengan sentimen leksikon Indonesia

No	*Normalisasi*	*Word count*
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	belum = -1
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	parah = -1
3	admin dan ibu wali yang terhormat kami warga dupak rukun rt rw ir kami mati sudah lebih ja	hormat = 1 mati = -1
4	kumuhnya surabaya daerah pacar keling	kumuh = -1
5	airnya mati lebih jam tolong hidupkan	mati = -1

4.4 Hasil Klasifikasi Data

Kelas *output* yang diharapkan pada penelitian *sentiment analysis* pengaduan masyarakat ini adalah kelas positif, kelas negatif dan netral. Dikatakan kelas positif jika kalimat pengaduan mengandung kata bersentimen positif. Begitu pula ketika kalimat pengaduan mengandung kata bersentimen negatif, maka dikatakan kelas negatif. Sedangkan kelas netral jika kalimat tersebut tidak mengandung kata bersentimen, atau nilai kata bersentimen positif sama dengan nilai kata bersentimen negatif. Metode klasifikasi yang digunakan dalam penelitian ini adalah *rule based method*, dengan *rule* sebagai berikut :

- Jika nilai kata (*term*) positif lebih besar dari nilai kata (*term*) negatif, maka kalimat pengaduan tersebut diklasifikasikan sebagai kelas positif.
- Jika nilai kata (*term*) negatif lebih besar dari nilai kata (*term*) positif, maka kalimat pengaduan tersebut diklasifikasikan sebagai kelas negatif.
- Jika nilai kata (*term*) positif sama dengan nilai kata (*term*) negatif, maka kalimat pengaduan tersebut diklasifikasikan sebagai kelas netral.

Tabel 4-19 menyajikan contoh klasifikasi menggunakan *rule based method* dengan perhitungan bobot kata dari *Sentiwordnet*, dan Tabel 4-20 menyajikan contoh klasifikasi dengan perhitungan bobot dari jumlah frekuensi

kemunculan kata sentimen sesuai sentimen leksikon Indonesia. Dari Tabel 4-19 diketahui kelas dari pengaduan diperoleh dengan membandingkan *negative score* dengan *positive score* dalam kalimat pengaduan tersebut. Sedangkan pada eksperimen 3, cara penentuan kelas diperoleh dengan membandingkan frekuensi kemunculan kata bersentimen seperti dilihat pada kolom hasil dari Tabel 4-20

Tabel 4-19 Klasifikasi *Rule Based Method* dengan *Sentiwordnet*

No	**Normalisasi**	***Sentiment Score***	***Hasil***
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	*negative score* = - 0.0202 *positivity score* = 0	kelas negatif
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	*negative score* = - 0.59 *positivity score* = 0.34	kelas negatif
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	*negative score* = - 0.175 *positivity score* = 0.33	kelas positif
4	kumuhnya surabaya daerah	*negative score* = - 0.175 *positivity score* = 0.33	kelas netral
5	airnya mati lebih jam tolong hidupkan	*negative score* = - 0.175 *positivity score* = 0.362	kelas positif

Tabel 4-20 Klasifikasi *Rule Based Method* dengan sentimen leksikon Indonesia

No	**Normalisasi**	***Frekuensi Word***	***Hasil***
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	*negative word* = 1 *positivite word* = 0	kelas negatif
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	*negative word* = 1 *positivite word* = 0	kelas negatif
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	*negative word* = 1 *positivite word* = 1	kelas netral
4	kumuhnya surabaya daerah	*negative word* = 1 *positivite word* = 0	kelas negatif
5	airnya mati lebih jam tolong hidupkan	*negative word* = 1 *positivite word* = 0	kelas negatif

4.5 Hasil Pengujian dan Validasi

Dalam penelitian *sentiment analysis* pengaduan masyarakat ini, proses validasi menggunakan perhitungan pencacahan kesalahan klasifikasi (*error*) secara manual seperti ditunjukkan pada Tabel 4-21 dan Tabel 4-22. Kolom label kelas pada Tabel 4-21 dan Tabel 4-22 menyatakan kelas yang sebenarnya dan kolom prediksi menyatakan kelas hasil klasifikasi berdasarkan perolehan *sentiment score* atau frekuensi kemunculan kata sentimen. Penentuan adanya kesalahan klasifikasi (*error*) dengan membandingkan label kelas yang sebenarnya dan hasil kelas prediksi. Jika kelas prediksi sama dengan label kelas sebenarnya, maka hasil klasifikasi dikatakan benar (*correct*). Akan tetapi, ketika kelas prediksi tidak sama dengan label kelas sebenarnya, maka hasil klasifikasi dikatakan salah (*error*).

Tabel 4-21 Pencacahan kesalahan klasifikasi pada *Sentiwordnet*

No	***Normalisasi***	***Label Kelas***	***Score***	***Kelas Prediksi***	***Error***
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	neg	0.0202	pos	E
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	neg	- 0.2497	neg	C
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	neg	0.1580	pos	E
4	kumuhnya surabaya daerah	neg	0	net	E
5	airnya mati lebih jam tolong hidupkan	neg	0.3636	pos	E

Setelah diperoleh jumlah terjadinya *error*, maka digunakan Persamaan 2.2 untuk menghitung nilai akurasi dari sistem. Berikut ini disajikan contoh perhitungan nilai akurasi dari eksperimen kedua untuk data *twitter*.

$$Akurasi = \frac{\sum prediksi\ benar}{jumlah\ data} \times 100\%$$

$$Akurasi = \frac{(Jumlah\ Data - Jumlah\ Error)}{Jumlah\ Data} \times 100\%$$

$$Akurasi = \frac{(685 - 363)}{685} \times 100\%$$

$$Akurasi = 47\%$$

Dari perhitungan diatas diperoleh nilai akurasi sebesar 47% untuk *sentiment analysis* menggunakan *Sentiwordnet* pada data pengaduan masyarakat di *twitter*. Perolehan nilai akurasi dari ketiga eksperimen menggunakan Persamaan 5 disajikan pada Tabel 4-23, dimana baik data *twitter* maupun data *media center* nilai akurasi tertinggi diperoleh dari eksperimen ketiga menggunakan sentimen leksikal Indonesia yakni 65.4% dan 81.4%. Sedangkan untuk *Sentiwordnet* nilai akurasi pada eksperimen kedua memperoleh nilai akurasi lebih tinggi dibandingkan eksperimen pertama untuk data *media center* yakni 56.85%, untuk data *twitter* perolehan nilai akurasi antara eksperimen pertama dan kedua hampir sama yakni 47%.

Tabel 4-22 Pencacahan kesalahan klasifikasi pada sentimen leksikon Indonesia

No	***Normalisasi***	***Label Kelas***	***Frek***	***Kelas Prediksi***	***Error***
1	kok belum ada tindakan bos untuk pohon di jalan pulo itu ada pohon sono besar	neg	0	net	E
2	jalan disekolahan sd sampai bebek maryem sangat parah mohon dilakukan pengaspalan	neg	-1	neg	C
3	admin dan ibu wali yang terhormat kami warga rt rw ir kami mati sudah lebih ja	neg	-2	neg	C
4	kumuhnya surabaya daerah	neg	-1	neg	C
5	airnya mati lebih jam tolong hidupkan	neg	-1	neg	C

Tabel 4-23 Perbandingan nilai akurasi penelitian

Media Data	Eksperimen 1	Eksperimen 2	Eksperimen 3
Twitter	*47.74 %*	*47.00 %*	*65.4 %*
Media center	*45.98 %*	*56.85 %*	*81.4 %*

Semakin tinggi nilai akurasi yang didapatkan, maka semakin baik pula *classifier* yang digunakan dalam memisahkan data sesuai kelas masing-masing. Akan tetapi, jika hanya melihat dari nilai akurasi saja kita tidak bisa melihat adanya penyimpangan data yang menyebabkan terjadinya kesalahan dalam proses klasifikasi. Oleh karena itu, kami juga melihat nilai *precision* dan *recall* yang didapatkan dalam penelitian ini. Berikut ini disajikan data hasil klasifikasi yang diperoleh dalam penelitian *sentiment analysis* menggunakan *sentiwordnet* pada data *twitter* dalam bentuk *confussion matrix* oleh Tabel 4-24. *Confussion matrix* menggambarkan persebaran data pada kelas tertentu, yakni data yang berhasil diklasifikasikan sesuai dengan kelasnya (*True Positive*, *True Negative*) dan data yang tidak berhasil diklasifikasikan dengan kelasnya (*False Positive*, *False Negative*) atau data yang mengalami penyimpangan. Berdasarkan Tabel 4-24 pada kelas positif data yang berhasil diklasifikasikan ke dalam kelas positif sebanyak 87 data, sedangkan penyimpangan data kelas positif yang klasifikasinya tidak sesuai sebanyak 38 data dianggap kelas negative dan 8 data dianggap kelas netral.

Tabel 4-24 *Confusion matrix Sentiwordnet pada data Twitter*

Actual	*Prediksi*			*Total*
	Positif	Negatif	Netral	
Positif	*87*	*38*	*8*	*133*
Negatif	*91*	*134*	*46*	*271*
Netral	*119*	*61*	*101*	*281*

Perhitungan *precision* dan *recall* memanfaatkan data yang disajikan dalam *confussion matrix*. Berikut ini disajikan perhitungan *precision* dan *recall* untuk data pada kelas positif sesuai Persamaan 2.5 dan Persamaan 2.6.

$$Precision = \frac{87}{(87 + 91 + 119)}$$

$$= 0.29$$

$$Recall = \frac{87}{(87 + 38 + 8)}$$

$$= 0.65$$

Nilai *precision* dan *recall* yang didapatkan oleh kelas positif adalah 0.29 dan 0.65 yang berarti ketepatan *classifier* dalam mengenali data positif masih rendah, sedangkan kemampuan *classifier* dalam mengenali kembali data positif cukup bagus. Tabel 4-25 menyajikan nilai *precision* dan *recall* yang diperoleh pada eksperimen data *twitter* dan *media center* menggunakan *sentiwordnet*. Pada kelas positif baik data *twitter* maupun *media center* memperoleh nilai *precision* yang cukup rendah yakni 0.29 dan 0.23, yang berarti ketepatan *classifier* dalam mengenali data positif kurang bagus. Sedangkan nilai *recall* yang diperoleh kelas positif cukup bagus yakni 0.65 untuk data *twitter* dan 0.61 untuk data *media center*. Hal ini berarti kemampuan *classifier* dalam menemukan kembali data positif cukup tinggi. Kondisi sebaliknya terjadi pada kelas netral, dimana nilai *precision* yang diperoleh data netral dari *twitter* cukup tinggi yakni 0.65, sedangkan nilai *recall* yang didapatkan cukup rendah yakni 0.36. Hal ini menyatakan bahwa kemampuan *classifier* dalam menemukan kembali data netral masih rendah, meskipun nilai *precision* yang didapatkan cukup tinggi. Bisa jadi data netral yang dikenali oleh *classifier* hanya sedikit.

Tabel 4-25 *Precision* dan *Recall* pada *Sentiwordnet*

Media	Kelas	Precision	Recall
Twitter	Pos	*0.29*	*0.65*
	Neg	*0.58*	*0.49*
	Net	*0.65*	*0.36*
Media center	Pos	*0.23*	*0.61*
	Neg	*0.79*	*0.52*
	Net	*0.63*	*0.64*

Selanjutnya nilai *precision* dan *recall* yang didapatkan oleh eksperimen menggunakan sentimen leksikon Indonesia pada data *twitter* dan *media center* disajikan oleh Tabel 4-26. Secara keseluruhan perolehan nilai *precision* dan *recall* pada eksperimen ketiga menggunakan sentimen leksikon Indonesia lebih tinggi jika dibandingkan perolehan nilai *precision* dan *recall* pada eksperimen kedua menggunakan *sentiwordnet*. Pada kelas negatif nilai *precision* dan *recall* yang diperoleh data *twitter* dan *media center* paling tinggi dibandingkan kelas positif dan

netral, yakni 1 untuk *precision*, 0.86 untuk *recall* data *media center*, dan 0.6 untuk *recall* data *twitter*. Hal ini berarti *classifier* bekerja dengan sangat baik dalam menemukan data kelas negatif menggunakan sentimen leksikon Indonesia. Pada eksperimen ketiga ini data positif memperoleh nilai *precision* yang paling rendah yakni 0.54 untuk data *twitter* dan 0.51 untuk data *media center*. Hal ini juga terjadi pada eksperimen kedua, dimana data positif hanya memperoleh nilai *precision* sebesar 0.29 untuk data *twitter* dan 0.23 untuk data *media center*.

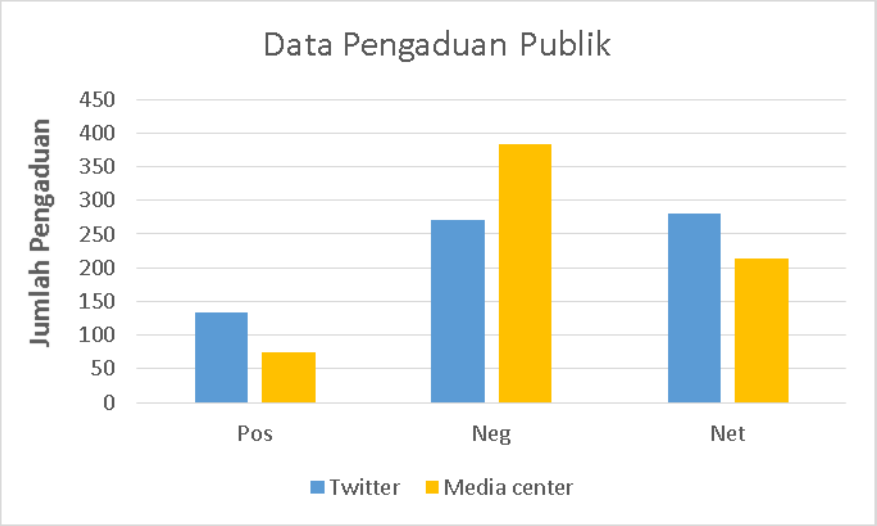
Tabel 4-26 *Precision* dan *Recall* pada sentimen leksikon Indonesia

Media	Kelas	Precision	Recall
Twitter	Pos	*0.54*	*0.6*
	Neg	*1*	*0.6*
	Net	*0.55*	*0.76*
Media center	Pos	*0.51*	*0.96*
	Neg	*1*	*0.86*
	Net	*0.72*	*0.68*

4.6 Analisa Hasil

Data pengaduan yang masuk, baik pada media *twitter* maupun *media center* masih didominasi oleh pengaduan yang bersentimen negatif dengan prosentase sebesar 57% pada data *media center* dan 40% pada data *twitter*. Sedangkan pengaduan yang mengandung sentimen positif masih sangat rendah hanya mencapai angka 11% untuk data *media center* dan 19% untuk data *twitter*. Seperti disajikan oleh Gambar 4.1. Hal ini menyiratkan bahwa masyarakat belum terbiasa menyampaikan apresiasi terhadap kinerja pemerintah melalui wadah yang sudah disediakan oleh pemerintah. Masyarakat cenderung menyampaikan keluhan dan permohonan informasi saja di wadah yang disediakan pemerintah.

Tabel 4-27 menyajikan data perbandingan nilai akurasi menggunakan *Sentiwordnet* pada data pengaduan masyarakat di media *twitter* dan *media center*. Disini *sentiwordnet* digunakan untuk proses ekstraksi fitur pada eksperimen pertama dan kedua.



Gambar 4.1 Hasil klasifikasi pada *Sentiwordnet*

Tabel 4-27 Perbandingan nilai akurasi menggunakan *Sentiwordnet*

Media Data	Eksperimen 1	Eksperimen 2
Twitter	*47.74 %*	*47.00 %*
Media center	*45.98 %*	*56.85 %*

Dari Tabel 4-27, kita memperoleh informasi bahwa pada data pengaduan di *twitter* nilai akurasi yang didapatkan sama antara eksperimen pertama dan kedua yakni 47%. Hal ini menunjukkan tidak ada perbedaan yang signifikan antara proses *translate* dilakukan pada kata hasil *stemming* maupun pada kalimat hasil normalisasi. Salah satu penyebabnya adalah data *twitter* banyak mengandung kalimat yang tidak terstruktur, sehingga tidak ada perbedaan antara hasil *translate* kata setelah *stemming* dengan kalimat hasil normalisasi. Selain itu, penggunaan kata *informal* pada data *twitter* menyebabkan hasil *translate* yang didapatkan tidak sesuai dengan makna kalimat aslinya. Kesalahan hasil *translate* sangat berpengaruh dalam proses ekstraksi *sentiment score* dari kata penyusun kalimat.

Sedangkan pada data pengaduan di *media center* nilai akurasi yang diperoleh eksperimen kedua lebih tinggi yakni 56.85% dibandingkan eksperimen pertama sebesar 45.98%. Hal ini disebabkan, pengaduan pada *media center* lebih terstruktur susunan kalimatnya. Sehingga terdapat perbedaan hasil ketika *translate* dilakukan pada kata hasil *stemming* dan kalimat hasil normalisasi. Dimana hasil

translate pada kalimat setelah dinormalisasi lebih mendekati makna dari kalimat aslinya. Selanjutnya, perbandingan nilai akurasi antara penggunaan *Sentiwordnet* dan sentiment leksikon Indonesia disajikan dalam Tabel 4-28. Kami menggunakan eksperimen kedua untuk mewakili eksperimen dengan *Sentiwordnet*, karena memberikan nilai akurasi yang lebih tinggi pada data *media center*. Selain itu hasil *translate* pada eksperimen kedua lebih mendekati makna kalimat aslinya dalam bahasa Indonesia.

Tabel 4-28 Nilai akurasi antara *Sentiwordnet* dan sentimen leksikon Indonesia

Media Data	Eksperimen 2	Eksperimen 3
Twitter	*47 %*	*65.4 %*
Media center	*56.85 %*	*81.4 %*

Berdasarkan Tabel 4-28, *Sentiwordnet* memperoleh nilai akurasi lebih kecil dibandingkan sentimen klasifikasi Indonesia, yakni 47% untuk data dari *twitter* dan 56.85% untuk data dari *media center*. Sedangkan nilai akurasi yang diperoleh sentimen leksikal Indonesia mencapai 65.4% untuk data dari *twitter* dan 81.4% untuk data dari *media center*.

Rendahnya nilai akurasi ketika menggunakan *Sentiwordnet* disebabkan oleh lima alasan. Pertama, adanya perbedaan karakteristik dalam menyampaikan pengaduan (*complaints*) antara orang Indonesia dan orang barat. Sebagian besar orang Indonesia menyampaikan pengaduan menggunakan bahasa yang sopan meskipun sebenarnya mereka merasa kesal dan kecewa. Hal ini membuat kalimat tersebut mengandung sentimen lebih dari satu, atau adanya *ambiguous sentiment*. Dan hal ini dapat merubah sentimen kalimat yang sebenarnya. Seperti pada pengaduan "admin dan ibu wali yang terhormat. Kami warga dupak rukun rt 2 rw 2 ir kami mati sudah lebih 24 jam. Tolong dibenarkan". Penggunaan kata "terhormat" dan "tolong" akan menyebabkan adanya *ambiguous sentiment*, dan bisa merubah sentimen kalimat dari negatif menjadi positif.

Alasan yang kedua, adanya kalimat yang secara keseluruhan menunjukkan sebuah sentimen. Akan tetapi tidak ditemukan adanya kata bersentimen pada kata penyusun kalimat tersebut. Atau sebaliknya, dimana kalimat secara keseluruhan tidak menunjukkan adanya sentimen (netral). Akan tetapi, pada kata penyusunnya

ditemukan adanya kata bersentimen. Kedua faktor tersebut bisa membuat terjadinya kesalahan klasifikasi sentimen. Contohnya, "Lho kemarin katanya siang, sekarang pagi toh". Pengaduan tersebut secara keseluruhan menunjukkan sentimen negatif. Akan tetapi jika dilihat dari kata penyusunnya, tidak mengandung kata bersentimen sama sekali. Contoh berikutnya "Eh ketemu mimin di acara Surabaya membara hari ini", secara keseluruhan pengaduan ini tidak mengandung sentimen (netral). Akan tetapi jika dilihat dari kata penyusunnya, ditemukan penggunaan kata bersentimen yakni "membara" yang memiliki sentimen negatif. Sehingga membuat kelimat tersebut dianggap negatif.

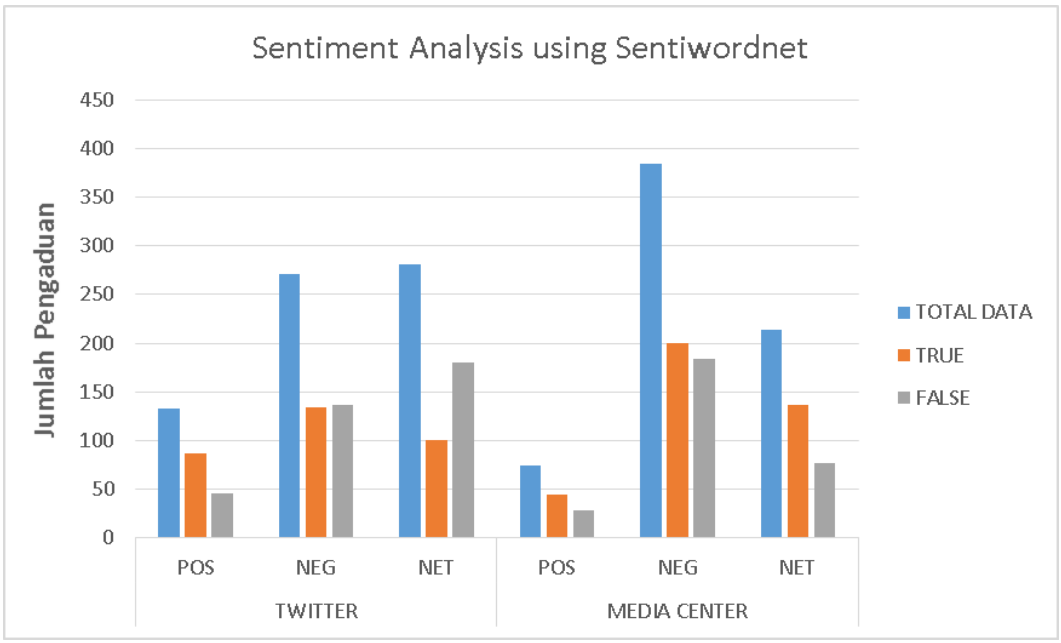
Alasan yang ketiga, beberapa kata dalam bahasa Indonesia mengandung sentiment. Tetapi, ketika di *translate* ke bahasa Inggris termasuk dalam daftar *stopword* yang harus dibuang seperti "mati" (*off*), "turun" (*down*), "bagus" (*good*), dan "semakin baik" (*better*). Hal ini bisa mempengaruhi sentiment dari kalimat, seperti pengaduan "perjuangan menyambung hidup untuk masa depan yang lebih baik, tetap berjuang ya dek!!!" setelah dilakukan translate "*the struggle of survival for the future better keep fighting ya dek*". Secara keseluruhan kalimat ini mengandung sentimen positif. Akan tetapi karena kata "*better*" dianggap sebagai *stopword*, maka sentiment dari kalimat tersebut dianggap negative.

Alasan keempat, banyak ditemukan kalimat sindiran (*sarcasm*) pada data pengaduan terutama data dari *twitter*. Seperti kita ketahui, kalimat sindiran (*sarcasm*) memiliki makna yang berlawanan dengan kalimat yang sebenarnya. Hal ini mengakibatkan terjadinya kesalahan klasifikasi sentiment. Seperti pada pengaduan "min tolong sampaikan ke bu risma itu jembatan semolowaru butuh berapa juta tahun lagi selesainya". Kalimat tersebut sebenarnya mengandung sentiment negatif, akan tetapi karena menggunakan bahasa sindiran. Maka hasil klasifikasi menunjukkan sebaliknya, yakni mengandung sentiment positif.

Penyebab terakhir rendahnya nilai akurasi ketika menggunakan *Sentiwordnet* adalah penggunaan *slang word* dan struktur kalimat yang tidak formal membuat hasil *translate* tidak sesuai dengan makna kalimat aslinya. Fenomena ini banyak ditemukan pada data pengaduan dari *twitter*. Salah satu contohnya "spanduke ndang dicopoti" hasil *translate* "e banner was quickly removed". Kalimat tersebut pada dasarnya mengandung sentimen negatif, akan tetapi karena

penggunaan *slang word* dan struktur kalimat yang tidak formal. Membuat hasil translate seolah kalimat mengandung sentimen positif dengan adanya kata “quick”.

Dari kelima penyebab rendahnya nilai akurasi pada *Sentiwordnet* diatas, poin paling besar dikarenakan factor pertama dan terakhir. Terutama pada pengaduan dari *twitter*.

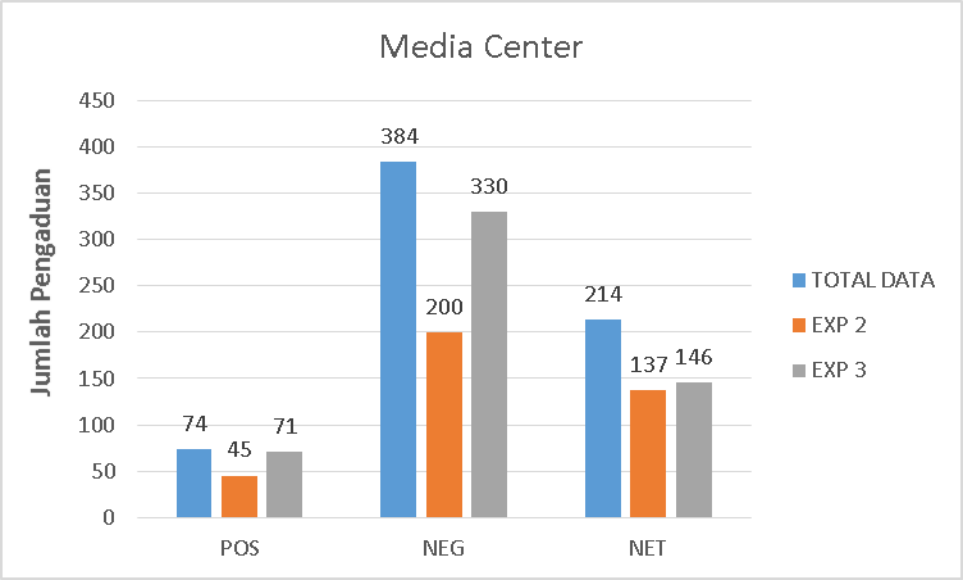


Gambar 4.2 Hasil klasifikasi pada *Sentiwordnet*

Gambar 4.2 menunjukkan perbandingan hasil klasifikasi pada data *twitter* dan *media center* menggunakan *Sentiwordnet*. Pada data sentimen positif jumlah klasifikasi yang benar (*correct*) lebih banyak dibandingkan jumlah klasifikasi yang salah (*error*) yakni 65% pada data *twitter* dan 60% *media center*. Sedangkan pada data sentimen negatif jumlah klasifikasi yang benar (*correct*) dan salah (*error*) hampir sama yakni 49.5% klasifikasi yang benar (*correct*) dan 50% salah (*error*) pada data *twitter*. Untuk data *media center* data klasifikasi benar (*correct*) 52% dan klasifikasi salah (*error*) 48%. Hal ini dikarenakan banyaknya kalimat yang mengandung sentimen lebih dari satu (*ambigous sentiment*) dan kalimat sindiran (*sarcasm*) yang dapat merusak sentimen dari kalimat.

Untuk data sentimen netral jumlah klasifikasi yang benar (*correct*) lebih banyak dari klasifikasi yang salah (*error*) pada data *media center*, yakni 64% klasifikasi yang benar (*correct*). Sedangkan untuk data *twitter* jumlah klasifikasi yang salah (*error*) lebih banyak dibandingkan klasifikasi yang benar (*true*), yakni

64% klasifikasi yang salah. Salah satu faktor penyebabnya adalah beberapa kata dalam bahasa Indonesia termasuk kata yang tidak mengandung sentimen (netral), tetapi dalam *Sentiwordnet* kata tersebut termasuk kata yang mengandung sentimen. Seperti kata ”rumah sakit” (*hospital*) dan ”jalan” (*street*), dalam bahasa Indonesia kedua kata tersebut termasuk kata yang tidak memiliki sentimen (netral). Akan tetapi dalam *Sentiwordnet* kedua kata tersebut tergolong kata bersentimen positif dan memiliki *sentiment score* sebesar 0.125 dan 0.0137.

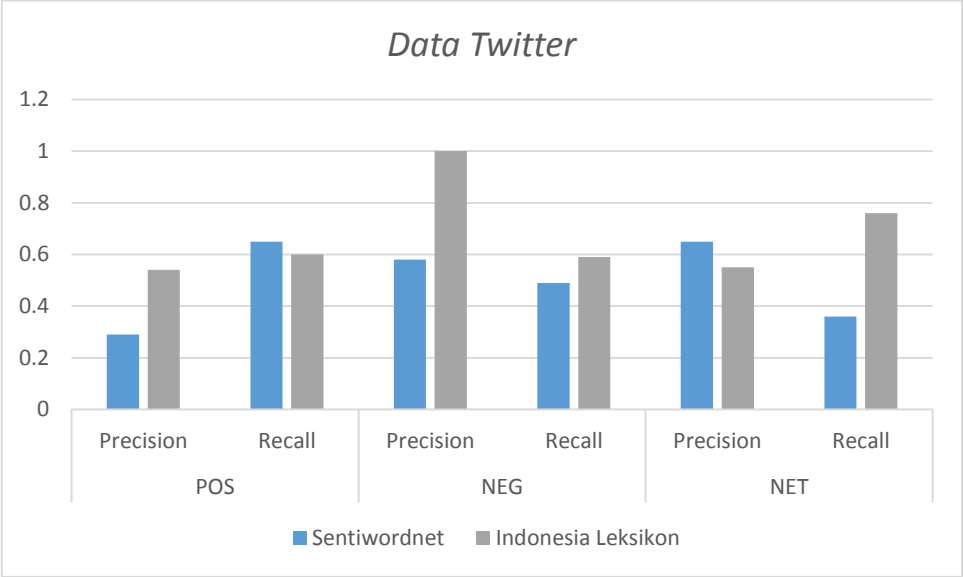


Gambar 4.3 Perbandingan hasil pada data *media center*

Perbandingan jumlah klasifikasi yang benar (*correct*) antara *Sentiwordnet* dan sentimen leksikon Indonesia pada data *media center* ditunjukkan oleh Gambar 4.3. Berdasarkan Gambar 4.3 diperoleh informasi bahwa sentimen leksikon Indonesia memiliki performansi yang lebih bagus dibandingkan *Sentiwordnet* dalam mengklasifikasikan data pengaduan masyarakat untuk data sentimen positif dan data sentimen negatif. Sedangkan pada data sentimen netral, *Sentiwordnet* dan sentimen leksikon Indonesia menghasilkan jumlah klasifikasi benar (*correct*) yang hampir sama yakni 64% untuk *Sentiwordnet* dan 68% untuk sentimen leksikon Indonesia.

Pada penelitian ini pengukuran tingkat optimal dan keefektifan sistem selain dilihat dari nilai akurasi yang diperoleh, juga dilihat dari nilai *precision* dan *recall*. Nilai *precision* dan *recall* digunakan untuk mengetahui adanya penyimpangan data yang menyebabkan rendahnya nilai akurasi. Sebuah sistem

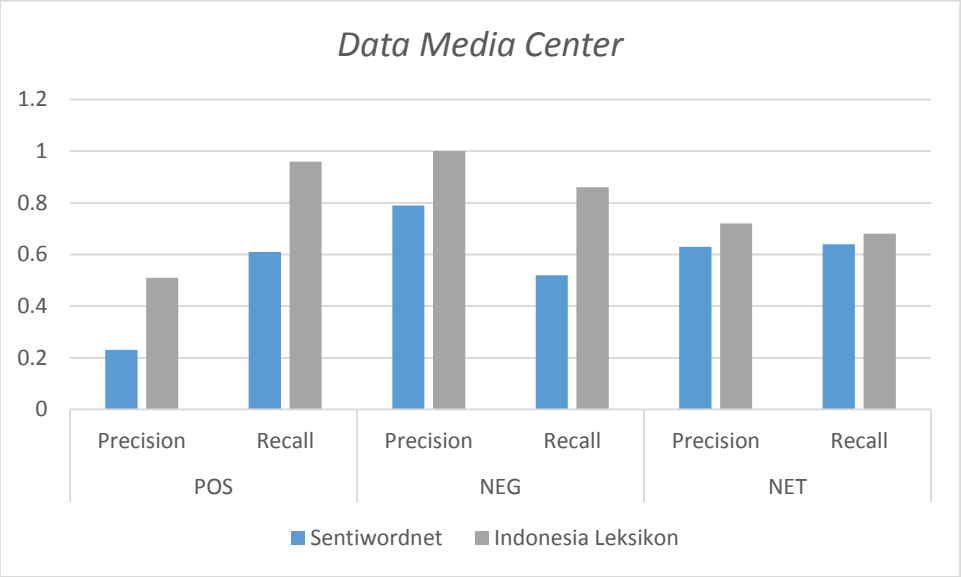
dikatakan optimal ketika nilai *precision* dan *recall* sama-sama tinggi dan seimbang. Gambar 4.4 menunjukkan perbandingan nilai *precision* dan *recall* yang didapatkan data *twitter* pada eksperimen kedua menggunakan *sentiwordnet* dan ekspreimen ketiga menggunakan sentimen leksikon Indonesia.



Gambar 4.4 Nilai *precision* dan *recall* pada data *twitter*

Informasi yang didapatkan dari Gambar 4.4 diantaranya, data kelas positif memperoleh nilai *precision* paling kecil dibandingkan kelas negatif dan netral yakni 0.29 untuk eksperimen menggunakan *sentiwordnet* dan 0.54 untuk eksperimen menggunakan eksperimen sentimen leksikon Indonesia. Hal ini menunjukkan bahwa kemampuan *classifier* dalam mengenali data kelas positif yang relevan masih rendah baik menggunakan *sentiwordnet* mapun menggunakan sentimen leksikon Indonesia. Akan tetapi bila dibandingkan dengan *sentiwordnet*, sentimen leksikon Indonesia memiliki kemampuan lebih baik dalam mengenali data kelas positif yang relevan. Penyebab rendahnya nilai *precision* pada kelas positif adalah jumlah kesalahan error data kelas negatif dan netral yang diklasifikasikan sebagai kelas positif cukup tinggi. Terutama data kelas netral pada eksperimen menggunakan *sentiwordnet*, 50% dari total data kelas netral diklasifikasikan sebagai kelas positif. Hal ini disebabkan pada data kelas netral banyak ditemukan penggunaan kata bersentimen positif yang dapat merusak sentimen dari kalimat. Seperti pada kalimat pengaduan ”Bagi yang ingin menyalurkan bantuan

kemanusiaan silahkan disini", adanya kata "bantuan" (*aid*), "kemanusiaan" (*humanitarian*) dan "silahkan" (*please*) yang memiliki *sentimen score* 0.02, 0.42, dan 0.34 membuat kalimat pengaduan tersebut diklasifikasikan sebagai kelas positif.



Gambar 4.5 Nilai *precision* dan *recall* pada data *media center*

Gambar 4.5 menyajikan perbandingan nilai *precision* dan *recall* yang didapatkan oleh data *media center* pada eksperimen kedua menggunakan *sentiwordnet* dan eksperimen ketiga menggunakan sentimen leksikon Indonesia. Sama halnya dengan data *twitter*, pada data *media center* data kelas positif memperoleh nilai *precision* paling rendah dibandingkan data kelas negatif dan netral baik untuk eksperimen kedua menggunakan *sentiwordnet* maupun eksperimen ketiga menggunakan sentimen leksikon Indonesia yakni 0.21 dan 0.51.

Hal ini dikarenakan dari 200 data yang dikenali sebagai kelas positif oleh *classifier* pada eksperimen kedua menggunakan *sentiwordnet*, hanya 22.5% saja yang benar-benar relevan sebagai kelas positif. Sedangkan 56% adalah data kelas negatif, dan 21.5% adalah data kelas netral yang mengalami penyimpangan klasifikasi menjadi kelas positif. Begitu pula pada eksperimen ketiga menggunakan sentimen leksikon Indonesia, dari 139 data yang dikenali sebagai kelas positif oleh *classifier* hanya 51% saja yang benar-benar relevan. Sedangkan 49% adalah data kelas netral yang mengalami penyimpangan klasifikasi menjadi kelas positif.

Penyebab penyimpangan data tersebut salah satunya adalah penggunaan kata bersentimen positif pada data kelas netral dan negatif. Seperti pengaduan ”Mohon bantuannya untuk diperbaiki penerangan jalan umum makam rangkah VII, karena kemaren ada pemakaman warga lampunya padam”. Penggunaan kata ”terang” dari kata penerangan dan kata ”baik” dari kata diperbaiki yang memiliki sentimen positif membuat data diklasifikasikan sebagai kelas positif pada eksperimen ketiga. Sedangkan kata ”padam” yang memiliki sentimen negatif belum terdapat di daftar sentimen leksikon Indonesia. Begitu juga pada eksperimen kedua penggunaan kata ”mohon” (*please*), ”bantuan” (*help*), dan ”penerangan” (*light*) mengubah data menjadi kelas positif. Karena *sentiment score* dari ketiga kata tersebut adalah 0.34, 0.0094, dan 0.05. Sedangkan kata ”padam” (*go out*) termasuk dalam daftar *stopword list* bahasa Inggris, sehingga dihilangkan.

Halaman ini sengaja dikosongkan

BAB 5
KESIMPULAN

Berdasarkan hasil eksperimen dan analisa yang telah dilakukan pada penelitian *sentiment analysis* pengaduan masyarakat menggunakan *rule based method* dan *lexical resources* untuk ekstraksi fitur, ada beberapa poin kesimpulan yang bisa diambil, diantaranya *sentiment analysis* pada data pengaduan masyarakat dengan pendekatan semantik memperoleh nilai *accuracy* sebesar 47% dan 65.4% untuk data pengaduan pada media *twitter*. Sedangkan untuk data pengaduan pada *media center* memperoleh nilai *accuracy* sebesar 56.85% dan 81.5% menggunakan *sentiwordnet* dan sentimen leksikon Indonesia. Rendahnya perolehan *accuracy* pada data pengaduan dari *twitter* disebabkan data *twitter* banyak mengandung kalimat yang tidak terstruktur, *slang word*, penggunaan singkatan dan kata tidak baku.

Poin selanjutnya yang didapatkan dari penelitian ini adalah Sentimen leksikon Indonesia memiliki performansi lebih baik dibandingkan *sentiwordnet* dalam proses ekstraksi fitur data pengaduan berbahasa Indonesia baik pengaduan pada media *twitter* maupun *media center*. Hal ini bisa dilihat dari nilai *accuracy* yang diperoleh sentiment leksikon Indonesia yakni 65.4% untuk data pengaduan dari *twitter* dan 81.4% untuk data pengaduan dari *media center*. Sedangkan nilai *accuracy* yang didapatkan oleh *sentiwordnet* hanya 47% untuk data pengaduan pada *twitter* dan 56.85% untuk data pengaduan pada *media center*. Salah satu penyebab rendahnya nilai *accuracy* yang diperoleh *sentiwordnet* adalah perbedaan karakteristik social dan budaya ketika mengekspresikan keluhan antara orang Indonesia dan orang luar negeri. Sebagian besar orang Indonesia menggunakan bahasa yang sopan dalam menyampaikan keluhannya, meskipun sebenarnya mereka merasa kesal. Seperti penggunaan kata "tolong" (*please*), dan "mohon" (*please*). Hal ini membuat terjadinya *ambiguous sentiment*.

Poin terakhir yang bisa diambil dari penelitian ini adalah pengaduan yang masuk baik pada media *twitter* maupun *media center* masih didominasi oleh pengaduan bersentimen negatif yakni sebesar 57% dan 40%. Sedangkan pengaduan

bersentimen positif hanya 11% untuk data pada *media center* dan 19% untuk data pada media *twitter*.

Berdasarkan hasil kesimpulan diatas diketahui bahwa pendekatan semantik memiliki performansi cukup baik dalam menganalisa sentimen yang terdapat pada teks bahasa Indonesia, terutama ketika menggunakan sentimen leksikon Indonesia. Sedangkan ketika menggunakan *sentiwordnet* hasil yang diperoleh masih rendah.

Pada penelitian selanjutnya diharapkan bisa mengatasi permasalahan *ambiguous sentiment* pada *sentiwordnet*. Dimana satu kata (*term*) bisa memiliki lebih dari satu *sentiment score* tergantung dari konteks kata tersebut dalam kalimat. Selain itu, penelitian selanjutnya diharapkan bisa meningkatkan hasil *translate* dari teks bahasa Indonesia ke bahasa Inggris. Terutama untuk data dari media *twitter*, yang banyak mengandung kata informal, singkatan dan *slang word*. Sehingga bisa meningkatkan hasil *accuracy* ketika memanfaatkan *sentiwordnet* untuk menganalisa sentimen pada teks bahasa Indonesia.

Dan poin terakhir, diharapkan penelitian selanjutnya bisa membangun sentimen leksikon Indonesia yang memiliki *polarity* score. Sehingga sentimen leksikon tersebut tidak hanya bisa digunakan untuk menganalisa sentimen pada teks bahasa Indonesia saja, tetapi bisa digunakan juga untuk menganalisa emosi yang terkandung pada teks bahasa Indonesia.

DAFTAR PUSTAKA

Agarwal, A., Sharma, V., Sikka, G. & Dhir, R., 2016. *Opinion Mining of News Headlines Using Sentiwordnet.* s.l., s.n.

Akbarisanto, R., Danar, W. & Purwarianti, A., 2016. *Analyzing Bandung Public Mood Using Twitter Data.* s.l., s.n.

Anastasia, S. & Budi, I., 2016. *Twitter Sentiment Analysis of Online Transportation Service Providers.* s.l., s.n.

Anggareska, D. & Purwarianti, A., 2014. *Information Extraction of Public Complaints on Twitter Text For Bandung Government.* s.l., s.n.

Atmadja, A. R. & Purwarianti, A., 2015. *Comparison on The Rule Based Method and Statistical Based Method on Emotion Classification for Indonesian Twitter Text.* Bali, s.n.

Cernian, A. & Sgarciu, V., 2015. *Sentiment Analysis From Product Reviews Using Sentiwordnet as Lexical Resource.* Bucharest, s.n.

Feedback Instruments Ltd., 2006. *Digital Pendulum: Control in a Matlab Environment.* Sussex, UK: Feedback Instruments Ltd..

Fiarni, C., Maharani, H. & Pratama , R., 2016. *Sentiment Analysis System for Indonesia Online Retail Shop Review Using Hierarchy Naive Bayes Technique.* s.l., s.n.

F. & Manurung, R., 2008. *Machine Learning-based Sentiment Analysis of Automatic Indonesian Translation of English Movie Reviews.* s.l., s.n.

IEEE, n.d. *IEEE Citation Reference.* [Online] Available at: www.ieee.org/documents/ieeecitationref.pdf

Informatika, D. K. d., 2015. *Dinkominfo Surabaya.* [Online] Available at: http://www. http://dinkominfo.surabaya.go.id/dki.php?hal=29 [Accessed April 2016].

Korde, V. & Mahender, C. N., 2012. Text Classification and Classifiers : Survey. *International Journal of Artificial Intellegence and Application (IJAIA),* Volume 3, p. 2.

Kreutzer, J. & Witte, N., 2013. Opinion Mining Using Sentiwordnet. *Semantic Analysis HT Uppsala University,* Volume 1, p. 14.

Liu, B., 2010. *Handbook of Natural Language Processing, Chapter Sentiment Analysis and Analysis.* s.l.:s.n.

Lunando, E. & Purwarianti, A., 2013. *Indonesian Social Media Sentiment Analysis with Sarcasm Detection.* s.l., s.n.

Masyarakat, D. H., 2015. *Hubungan Masyarakat Surabaya.* [Online] Available at: http://www.humas.surabaya.go.id/index.php?option=news&det=548 [Accessed April 2016].

Naradhipa, A. R. & Purwarianti, A., 2011. *Sentiment Classification for Indonesian Message in Social Media*. Bandung, s.n.

Pisceldo, F., Adriani, M. & Manurung, R., 2009. *Probabilistic Part Of Speech Tagging for Bahasa Indonesia.* Singapore, s.n.

Susilawati, E., 2016. *Public Services Satisfaction Based on Sentiment Analysis.* Bandung, s.n.

Tala, A. Z., 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* Netherlands: Institut for Logic, Language and Computation. Universiteit Van Amsterdam.

Tanaka, K. & Sugeno, M., 1992. Stability analysis and design of fuzzy control. *Fuzzy Sets and Systems,* Volume 45, pp. 135-156.

Vania, C., Ibrahim, M. & Adriani, M., 2014. Sentiment Lexicon Generation for an Under-Resourced Languange. *IJCLA,* Jan-Jun, Volume 5, pp. 59-72.

Wicaksono, A. F., Vania, C., T., B. D. & Adriani, M., 2014. *Automatically Building a Corpus for Sentiment Analysis on INdonesian Tweets.* Bandung, s.n., pp. 185-194.

BIOGRAFI PENULIS

Masfulatul Lailiyah, dilahirkan di Sidoarjo pada tanggal 04 Agustus 1986. Penulis merupakan anak pertama dari tiga bersaudara. Penulis menyelesaikan pendidikan sekolah dasar di SDN Terungkulon I pada tahun 1998, dan sekolah menengah lanjutan di SLTPN 1 Krian pada tahun 2001, serta sekolah menengah atas di SMAN 1 Sidoarjo dan lulus pada tahun 2004. Selanjutnya, penulis melanjutkan pendidikan ke jenjang diploma 3 dan mengambil Jurusan Teknik Informatika di STT Telkom Bandung hingga tahun 2007. Setelah itu penulis mengambil program lintas jalur sarjana di Jurusan Teknik Informatika Institut Teknologi Sepuluh Nopember pada tahun 2007. Setelah merampungkan pendidikan S1, penulis bekerja di RSU Haji Surabaya sebagai staf bidang Sistem Informasi Manajemen (SIM). Saat ini penulis menyelesaikan program Magister di Fakultas Teknologi Elektro dengan bidang keahlian Telematika-CIO yang merupakan program kerjasama antara Kementrian Komunikasi dan Informatika (KOMINFO) dengan Institut Teknologi Sepuluh Nopember (ITS).