



UNIVERSITAS INDONESIA

**ANALISIS ASPIRASI DAN PENGADUAN DI SITUS LAPOR!
DENGAN MENGGUNAKAN *TEXT MINING***

SKRIPSI

CHYNTIA MEGAWATI

1106018316

**FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INDUSTRI
DEPOK
JUNI 2015**



UNIVERSITAS INDONESIA

**ANALISIS ASPIRASI DAN PENGADUAN DI SITUS LAPOR!
DENGAN MENGGUNAKAN *TEXT MINING***

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik

CHYNTIA MEGAWATI

1106018316

**FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INDUSTRI
DEPOK
JUNI 2015**

HALAMAN PERNYATAAN ORISINALITAS

**Skripsi ini adalah hasil karya saya sendiri, dan semua
sumber baik yang dikutip maupun dirujuk telah saya
nyatakan dengan benar**

Nama : Chyntia Megawati

NPM : 1106018316

Tanda Tangan :



Tanggal : 3 Juni 2015

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh,

Nama : Chyntia Megawati
NPM : 1106018316
Program Studi : Teknik Industri
Judul Skripsi : Analisis Aspirasi dan Pengaduan di Situs LAPOR!
dengan Menggunakan *Text Mining*

Telah berhasil dipertahankan di hadapan Dewan Pengaji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Teknik pada Program Studi Teknik Industri Fakultas Teknik Universitas Indonesia

DEWAN PENGUJI

Pembimbing : Prof. Ir. Isti Surjandari, M.T., M.A., PhD. ()

Pengaji 1 : Prof. Dr. Ir. Teuku Yuri M. Z., M.Eng. Sc. ()

Pengaji 2 : Dr. Ir. M. Dachyar, M. Sc. ()

Pengaji 3 : Ir. Yadrifil, M.Sc. ()

Ditetapkan di : Depok

Tanggal : 8 Juni 2015

KATA PENGANTAR

Puji syukur penulis ucapkan kepada Allah SWT atas rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi sebagai salah satu syarat untuk mendapat gelar Sarjana Teknik Universitas Indonesia. Dalam penyusunan skripsi ini banyak pihak yang telah membantu penulis baik secara langsung maupun tidak langsung. Oleh karena itu, pada kesempatan kali ini penulis ingin menyampaikan rasa terimakasih kepada:

1. Prof. Ir. Isti Surjandari, M.T., M.A., Ph.D. selaku pembimbing akademis, pembimbing skripsi, dan Kepala Laboratorium SQE yang selalu memberikan motivasi dan pelajaran hidup selama masa studi penulis hingga dapat menyelesaikan penelitian ini.
2. Bapak Ferdy Alfarizka Putra dari pihak LAPOR! yang telah membantu penulis dalam memperoleh data dan melakukan penelitian ini.
3. Willian Gozali selaku pengembang aplikasi pra proses yang tanpa beliau penelitian ini akan sulit diselesaikan.
4. Seluruh dosen Teknik Industri UI yang telah memberikan banyak ilmu yang bermanfaat kepada penulis selama masa perkuliahan.
5. Kedua orang tua penulis, yang telah memberikan, doa, kasih sayang, serta dukungan finansial dan moral kepada penulis dalam penyelesaian penelitian ini.
6. Esther Widya, Fakhru Agustriwan, Rediani Pramudita, Nurman Wibisana, Anitasari Titiani, Nina Jane, dan seluruh asisten laboratorium SQE yang telah membantu dan selalu bersedia membagi ilmu dengan penulis.
7. Fransiska, Maria, Anis, Bella, Gina, Fitri, Vina, Ica, Ratna, Dinta, Nora, Indri, Bintang, Benita, dan Valida yang telah menjadi sahabat-sahabat terbaik, serta tempat bertukar pikiran dan berbagi keluh kesah penulis selama masa perkuliahan.
8. Teman-teman Teknik Industri angkatan 2011 yang telah menjadi teman dan keluarga penulis selama perkuliahan.

9. Seluruh karyawan Departemen Teknik Industri UI yang selalu memberikan bantuan dan kebaikan kepada penulis.
10. Seluruh pihak yang berkontribusi dalam penyelesaian penelitian ini yang tidak dapat disebutkan satu per satu.

Akhir kata, penulis berharap semoga semua pihak yang telah membantu dibalas kebaikannya oleh Allah SWT. Penulis juga berharap skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan.

Depok, 3 Juni 2015



Chyntia Megawati

LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH
HALAMAN PERNYATAAN PERSETUJUAN
PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama : Chyntia Megawati
NPM : 1106018316
Program Studi : Teknik Industri
Departemen : Teknik Industri
Fakultas : Teknik
Jenis karya : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (Non-exclusive Royalty-Free Right)** atas karya ilmiah saya yang berjudul :

**Analisis Aspirasi dan Pengaduan di Situs LAPOR! dengan Menggunakan
*Text Mining***

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok
Pada tanggal : 3 Juni 2015

Yang menyatakan



(Chyntia Megawati)

ABSTRAK

Nama : Chyntia Megawati
Program Studi : Teknik Industri
Judul : Analisis Aspirasi dan Pengaduan di Situs LAPOR! dengan Menggunakan *Text Mining*

Perkembangan pesat Teknologi Informasi dan Komunikasi (TIK) telah membuatnya menjadi satu bagian penting dalam kehidupan sehari-hari. Sektor pemerintah di Indonesia merupakan salah satu pihak yang telah mencoba memanfaatkan TIK dengan membuat sebuah situs untuk berkomunikasi secara dua arah dengan masyarakat (*e-Governement*) dalam bentuk LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat). Laporan yang disampaikan masyarakat bisa menjadi masukan penting bagi pemerintah untuk membantu pembangunan dan peningkatan pelayanan publik. Oleh karena itu, penelitian ini menggunakan metode *text mining* untuk menganalisis data tekstual yang berupa opini atau keluhan dengan mengklasifikasikannya menjadi beberapa kelas dan kemudian data set setiap kelas akan dikelompokkan lagi menjadi beberapa topik khusus (*cluster*). Hasil penelitian menunjukkan bahwa laporan terkait kemiskinan memiliki jumlah terbanyak dengan topik mayoritas yang dibahas adalah mengenai beberapa jenis bantuan sosial seperti KPS (Kartu Perlindungan Sosial) dan BLSM (Bantuan Langsung Sementara Masyarakat) yang tidak didistribusikan dengan baik atau tidak tepat sasaran.

Kata Kunci:

Text Mining, Klasifikasi, Pengelompokan, *Support Vector Machine*, *Self Organizing Maps*, Pengaduan Masyarakat

ABSTRACT

Name : Chyntia Megawati
Study Program : Industrial Engineering
Title : The Analysis of Aspiration and Complaint in LAPOR!
Website Using Text Mining

The rapid development of Information and Communication Technology (ICT) has made it as one important part in daily life. The government sector in Indonesia is one of those who have tried to use ICT in order to build such a two way communication site with citizens (e-governement) by creating LAPOR! (*Layanan Aspirasi dan Pengaduan Online Rakyat*). All kind of reports that conveyed by citizens could be an important input for the government to assist the development and improvement of public services. Hence, this research analyze citizen's textual reports data by using text mining method. The textual data will be classified into several classes and then the data set in each class will be clustered into several specific topics (clusters). The results showed that poverty is the most reported category with the majority of its topics are about some kind of social aids such as KPS (*Kartu Perlindungan Sosial*) and BLSM (*Bantuan Langsung Sementara Masyarakat*) that are not well distributed or reached out the wrong target.

Keywords:

Text Mining, Classification, Clustering, Support Vector Machine, Self Organizing Maps, Citizen's Report

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERNYATAAN ORISINALITAS	iii
HALAMAN PENGESAHAN.....	iv
KATA PENGANTAR.....	v
LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH.....	vii
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI.....	x
DAFTAR TABEL	xii
DAFTAR GAMBAR.....	xiii
1. PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	5
1.3. Diagram Keterkaitan Masalah	5
1.4. Tujuan Penelitian	6
1.5. Batasan Masalah	7
1.6. Metodologi Penelitian	7
1.7. Sistematika Penelitian	9
2. STUDI PUSTAKA.....	10
2.1. <i>E-Governement</i> (Pemerintah Elektronik).....	10
2.2. LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat)	13
2.2.1. Sistem Pelaporan	14
2.2.2. Fitur LAPOR!	15
2.3. <i>Data Mining</i>	18
2.4. <i>Text Mining</i>	20
2.4.1. Pra-proses (<i>pre-processing task</i>).....	23
2.4.2. Penyusunan Vektor (<i>Representation</i>)	23
2.4.3. Ekstraksi Informasi pada <i>Text Mining</i>	25
2.5. Algoritma Penggolong Klasifikasi.....	26
2.5.1. <i>Support Vector Machine</i> (SVM)	27
2.5.2. <i>Naïve Bayes Classifier</i>	27
2.5.3. Evaluasi Model Pengklasifikasi	30
2.6. Algoritma Penggolong <i>Clustering</i>	32
2.6.1. <i>Self-Organizing Map</i> (SOM)	32
2.6.2. <i>K-means Clustering</i>	34
3. PENGUMPULAN DAN PENGOLAHAN DATA.....	37
3.1. Pengumpulan Data	37
3.2. Pra-proses Teks	39
3.2.1. Tokenization, Case Folding, Spelling Normalization, dan Filtering	41
3.2.2. Stemming	45

3.2.3. Pembuatan TF, IDF, dan SVD	47
3.3. Klasifikasi	48
3.4. Pengelompokan (<i>Clustering</i>)	50
4. ANALISIS HASIL DAN PEMBAHASAN	53
4.1. Klasifikasi Dokumen	53
4.2. Pengelompokan (<i>Clustering</i>) Dokumen	57
5. KESIMPULAN.....	65
5.1. Kesimpulan	65
5.2. Saran	66
DAFTAR PUSTAKA.....	68

DAFTAR TABEL

Tabel 3.1 Kumpulan Data Laporan.....	41
Tabel 3.2 Cuplikan Daftar Kata dan Singkatan	43
Tabel 3.3 Kumpulan Data Laporan Setelah Melewati Bagian Awal Pra-proses	45
Tabel 3.4 Kumpulan Data Laporan Setelah Melewati Proses Stemming	46
Tabel 3.5 Hasil error setiap <i>initial map size</i> pada data set pendidikan	51
Tabel 3.6 Hasil error setiap <i>initial map size</i> pada data set energi, pangan, dan maritim.....	51
Tabel 3.7 Hasil error setiap <i>initial map size</i> pada data set kesehatan	51
Tabel 3.8 Hasil error setiap <i>initial map size</i> pada data set infrastruktur.....	51
Tabel 3.9 Hasil error setiap <i>initial map size</i> pada data set kemiskinan	51
Tabel 3.10 Hasil error setiap <i>initial map size</i> pada data set pendidikan birokrasi	52
Tabel 4.1 <i>Confusion matrix</i> dan akurasi model klasifikasi tanpa stemming	54
Tabel 4.2 <i>Confusion matrix</i> dan akurasi model klasifikasi dengan stemming	55
Tabel 4.3 Jumlah anggota tiap kelas klasifikasi.....	56
Tabel 4.4 Hasil <i>cluster</i> kelas pendidikan	58
Tabel 4.5 Hasil <i>cluster</i> kelas energi, pangan, dan maritim.....	59
Tabel 4.6 Hasil <i>cluster</i> kelas kesehatan	60
Tabel 4.7 Hasil <i>cluster</i> kelas infrasruktur	61
Tabel 4.8 Hasil <i>cluster</i> kelas kemiskinan	63
Tabel 4.9 Hasil <i>cluster</i> kelas birokrasi	64

DAFTAR GAMBAR

Gambar 1.1 Laporan dan tindak lanjut pada situs LAPOR!	2
Gambar 1.2 Peta sebaran laporan pengguna LAPOR!.....	3
Gambar 1.3 Diagram Keterkaitan Masalah.....	6
Gambar 1.4 Diagram Alir Penelitian	8
Gambar 2.1 Hubungan antara EGDI (<i>E-Government Development Index</i>) dengan GNI (Gross National Income) per capita dari negara pendapatan menengah kebawah (<i>lower-middle income countries</i>).....	13
Gambar 2.2 <i>Homepage</i> LAPOR!	14
Gambar 2.3 Alur Kerja LAPOR!	16
Gambar 2.4 <i>Data mining</i> sebagai tahapan dari <i>knowledge discovery</i>	18
Gambar 2.5 Diagram venn 6 bidang terkait dan 7 area praktek <i>text mining</i>	22
Gambar 2.6 Kerangka proses analisis teks pada <i>text mining</i>	22
Gambar 2.7 <i>Margin Hyperplane SVM</i>	27
Gambar 2.8 Dua kelompok data <i>Naïve Bayes</i>	28
Gambar 2.9 Penambahan objek baru pada <i>Naïve Bayes</i>	29
Gambar 2.10 <i>Confusion Matrix</i>	31
Gambar 2.11 Ilustrasi jaringan SOM 4x4	33
Gambar 2.12 Radius <i>Best Matching Unit</i> (BMU).....	34
Gambar 2.13 Ilustrasi proses clustering dengan <i>k-means</i>	36
Gambar 3.1 Alur Pengumpulan dan Pengolahan Data	37
Gambar 3.2 <i>Spreadsheet</i> Laporan yang Disetujui	38
Gambar 3.3 <i>Spreadsheet</i> Laporan yang Diarsipkan	39
Gambar 3.4 <i>Interface</i> Alikasi Pra-proses	40
Gambar 3.5 Cuplikan Naskah Aplikasi Pra-proses.....	40
Gambar 3.6 Tokenization.....	42
Gambar 3.7 Case Folding	42
Gambar 3.8 Spelling Normalization	44
Gambar 3.9 Filtering	44
Gambar 3.10 Stemming	46
Gambar 3.11 Matriks <i>Term Frequency</i>	47
Gambar 3.12 Matriks <i>Inverse Document Frequency</i>	47
Gambar 3.13 Matriks <i>Singular Value Decomposition</i>	48
Gambar 3.14 Model Kalsifikasi	49
Gambar 3.15 Contoh model <i>clustering</i> SOM	52

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Perkembangan pesat teknologi informasi dan komunikasi, khususnya internet telah menjadi bagian penting dalam aktifitas sehari-hari. Banyak pihak yang memanfaatkan internet untuk berbagai tujuan. Sektor pemerintah merupakan salah satu pihak yang telah mencoba memanfaatkan internet dengan membuat sebuah situs untuk berkomunikasi secara dua arah dengan masyarakat (Suh, Park, & Jeon, 2010). Situs semacam ini dikenal dengan nama pemerintah elektronik/digital (*e-Government*). Di Indonesia, pemerintah membangun situs yang diberi nama LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat).

Pembangunan LAPOR! didasarkan pada fakta bahwa jumlah pengguna ponsel, internet, dan media sosial di Indonesia kian melonjak signifikan. Fakta itu kemudian ditangkap sebagai peluang sekaligus tantangan untuk menjaring seluas-luasnya suara masyarakat dan memuarakannya ke arah yang lebih konstruktif, mendorong koordinasi yang efektif-efisien di level intra dan antar instansi pemerintah dalam penyelenggaraan pelayanan publik serta pengelolaan aspirasi dan pengaduan masyarakat. Hingga saat ini LAPOR! telah berjalan secara terpadu dengan 80 Kementerian atau Lembaga dan 5 Pemerintah Daerah serta BUMN di Indonesia (LAPOR!, 2015). LAPOR! yang semula diinisiasi dan dikembangkan oleh Unit Kerja Presiden Bidang Pengawasan dan Pengendalian Pembangunan (UKP-PPP atau UKP4) itu kini dikelola oleh Kantor Staf Presiden.

Masyarakat dapat mengakses LAPOR! melalui situs, SMS, media sosial, maupun aplikasi pada perangkat seluler. Setiap laporan yang diterima akan diverifikasi oleh administrator LAPOR! untuk diteruskan kepada instansi yang bersangkutan. Setiap laporan yang diterima akan diproses menjadi tiga jenis perlakuan, yaitu disetujui, dipending, atau diarsipkan. Laporan yang disetujui merupakan laporan yang sudah jelas atau merupakan aspirasi yang bagus, sehingga akan langsung diproses dan diteruskan untuk ditindaklanjuti. Laporan yang dipending merupakan laporan yang dirasa bagus tetapi belum jelas, sehingga

perlu dilakukan konfirmasi kembali kepada pelapor. Sedangkan laporan yang diarsipkan merupakan laporan yang dirasa tidak jelas, laporan yang berulang atau sudah pernah dilaporkan sebelumnya, atau merupakan saran yang bersifat sangat umum sehingga tidak diproses lebih lanjut.

The screenshot shows a report titled "Penertiban PKL di Ramayana Palmerah". The report details a complaint from a user about illegal street vendors at a specific location. It includes sections for "LAPORAN", "LAMPIRAN", and "INFORMASI TAMBAHAN". The "TINDAK LANJUT LAPORAN" section shows two follow-up actions taken by the government:

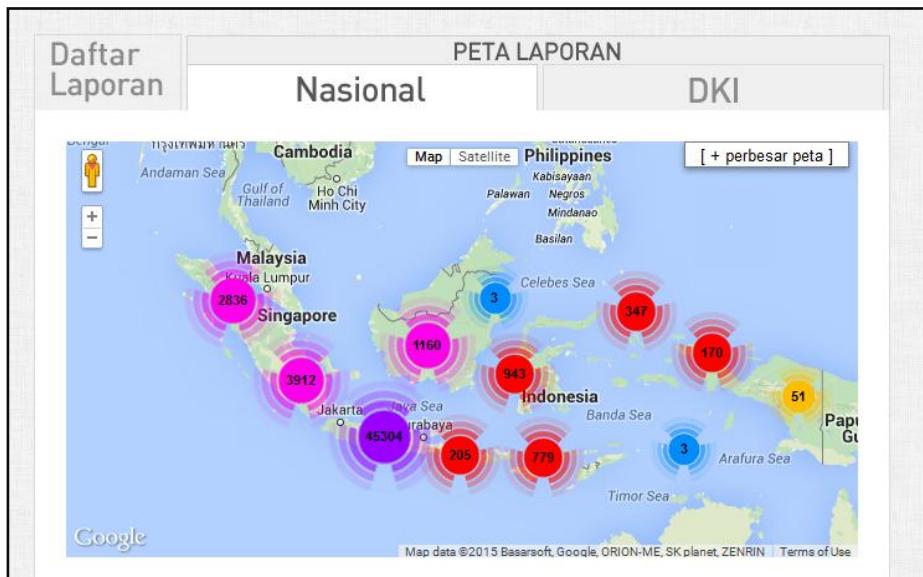
- Pemerintah Provinsi DKI Jakarta**: A tracking ID of 1322320 was issued, and the report was forwarded to the Satuan Polisi Pamong Praja (Satpol PP) in Palmerah. The status is "Didisposisikan" (Forwarded) on March 3, 2015, at 07:12:20. A note indicates that the vendor will be removed.
- Satuan Polisi Pamong Praja (Pemerintah Provinsi DKI Jakarta)**: The vendor was removed on February 26, 2015. The status is "Terlaksana" (Completed) on March 3, 2015, at 10:54:48. A note states that the removal was coordinated with the local police.

At the bottom, there are social media sharing options: Tweet (0), Share, Email, and Print.

Gambar 1.1 Laporan dan tindak lanjut pada situs LAPOR!

(Sumber : www.lapor.go.id, 2015)

LAPOR! telah menjadi suatu sarana efektif dan tepat guna bagi masyarakat untuk menyampaikan segala bentuk aspirasi dan keluhan karena dapat dibaca langsung oleh pembuat kebijakan yang bersangkutan. Jumlah laporan yang diterima LAPOR! terus meningkat setiap tahunnya seiring dengan masyarakat yang semakin mengetahui tentang LAPOR!, pada tahun 2012 tidak lebih dari 900 laporan diterima setiap bulannya dan hingga tahun 2015 rata-rata 10000-14000 laporan diterima setiap bulannya. Data laporan yang diterima juga tersebar di seluruh wilayah Indonesia, seperti ditunjukkan pada Gambar 1.2. Besarnya antusiasme masyarakat pada LAPOR! tersebut membuktikan bahwa LAPOR! saat ini merupakan salah satu media atau fasilitas penyampaian aspirasi yang semakin populer dan menjadi cikal bakal dari sistem aspirasi dan pengaduan masyarakat yang terpadu secara nasional. Masyarakat dari berbagai lapisan dapat menyampaikan opini, aspirasi, keluhan, pengaduan, hingga meminta informasi pada pemerintah. Selain dapat menyelesaikan berbagai keluhan masyarakat, pemerintah dapat menggunakan LAPOR! sebagai sarana untuk mendorong keterbukaan instansi sebagai upaya untuk meningkatkan kualitas layanan publik.



Gambar 1.2 Peta sebaran laporan pengguna LAPOR!

(Sumber : www.lapor.go.id, 2015)

Pada Peraturan Presiden Republik Indonesia Nomor 26 Tahun 2015 mengenai dasar hukum pembentukan kantor staf presiden, dijelaskan beberapa fungsi kantor staf presiden diantaranya dalam hal pengelolaan isu-isu strategis dan penyampaian analisis data serta informasi strategis dalam rangka mendukung proses pengambilan keputusan. Dalam hal ini, unit kerja LAPOR! ditunjuk sebagai alat atau sarana untuk mengelola data aspirasi dari masyarakat dan memberikan masukan kepada presiden mengenai prioritas nasional saat ini. Akan tetapi, hingga saat ini LAPOR! belum mampu menggali dan menganalisis informasi mengenai keseluruhan isu yang masuk karena banyaknya data yang masuk dan terbatasnya sumber daya.

Data hingga awal Maret 2015 menunjukkan bahwa LAPOR! telah menerima lebih dari 610.000 laporan dengan rata-rata lebih dari 900 laporan setiap harinya. Data laporan yang diterima berasal dari berbagai daerah di Indonesia dan sekitar 80-90% dari total laporan dilaporkan via sms. Dari keseluruhan 610.000 data yang diterima hanya 75.870 laporan atau sekitar 12,5% yang langsung disetujui, 6600 laporan atau 1% yang dipending, dan 527530 atau sekitar 86% diarsipkan. Hal ini mengindikasikan bahwa hanya sekitar 13-14%

laporan yang diproses dan diketahui perihalnya, sedangkan sekitar 86% laporan tidak diketahui perihalnya.

Saat ini LAPOR! masih belum mampu untuk mengetahui prioritas nasional yang paling mendapat sorotan dari masyarakat secara *real time* dan menyeluruh. Belum ada metode sistematis khusus yang digunakan LAPOR! untuk menganalisis jumlah data yang begitu besar selain dengan teknik manual oleh administrator. Jumlah data yang begitu besar melebihi kemampuan manusia untuk melakukannya secara cepat dan efisien. Jumlah data laporan yang begitu besar tersebut dapat didefinisikan sebagai *Big Data*. *Big Data* merupakan data yang mempunyai jumlah dan variasi besar, serta bergerak cepat, sehingga melampaui kapasitas pengolahan database konvensional (Dumbill, 2014). Dalam mengolah *Big Data*, *Data Mining* merupakan metode yang dapat mengotomatisasi proses pengolahan data untuk mengekstraksi pengetahuan dari informasi yang tidak bisa diamati hanya dengan melihat data karena terlalu rumit atau multidimensi. Pada kasus data pada LAPOR! yang merupakan data teks, jenis metode *Data Mining* yang dapat digunakan adalah *Text Mining*. *Text Mining* memegang peran penting dalam analisis *Big Data* yang bersifat tidak terstruktur seperti data teks dan dalam jumlah yang sangat besar (Xiang et al, 2015).

Text Mining merupakan salah satu bagian dari *Data Mining* yang dapat menganalisis dan memproses data teks yang bersifat semi terstruktur (*semistructured*) dan tidak terstruktur (*unstructured*). Hal ini sedikit berbeda dengan *Data Mining* yang secara umum digunakan untuk menganalisis data yang lebih bersifat kategorikal, ordinal, maupun kontinyu (Suh, Park, & Jeon, 2010). Pertama kali dikembangkan pada tahun 1980-an, *Text Mining* telah menjadi semakin efektif seiring dengan daya komputasi yang meningkat. *Text Mining* telah sukses dalam menemukan berbagai pola dan koneksi yang tidak terlihat oleh sekilas mata, seperti mengukur tingkat kebahagiaan melalui kata-kata dalam *twitter*, mengetahui hubungan senyawa tertentu dengan enzim tertentu, dan sebagainya (Belsky, 2012).

Pemerintah sebagai penyusun kebijakan sangat perlu untuk mengetahui dan memahami segala bentuk aspirasi dan keluhan masyarakat. Dengan bantuan teknik *Data Mining* dan *Text Mining* untuk LAPOR!, maka dapat dilakukan

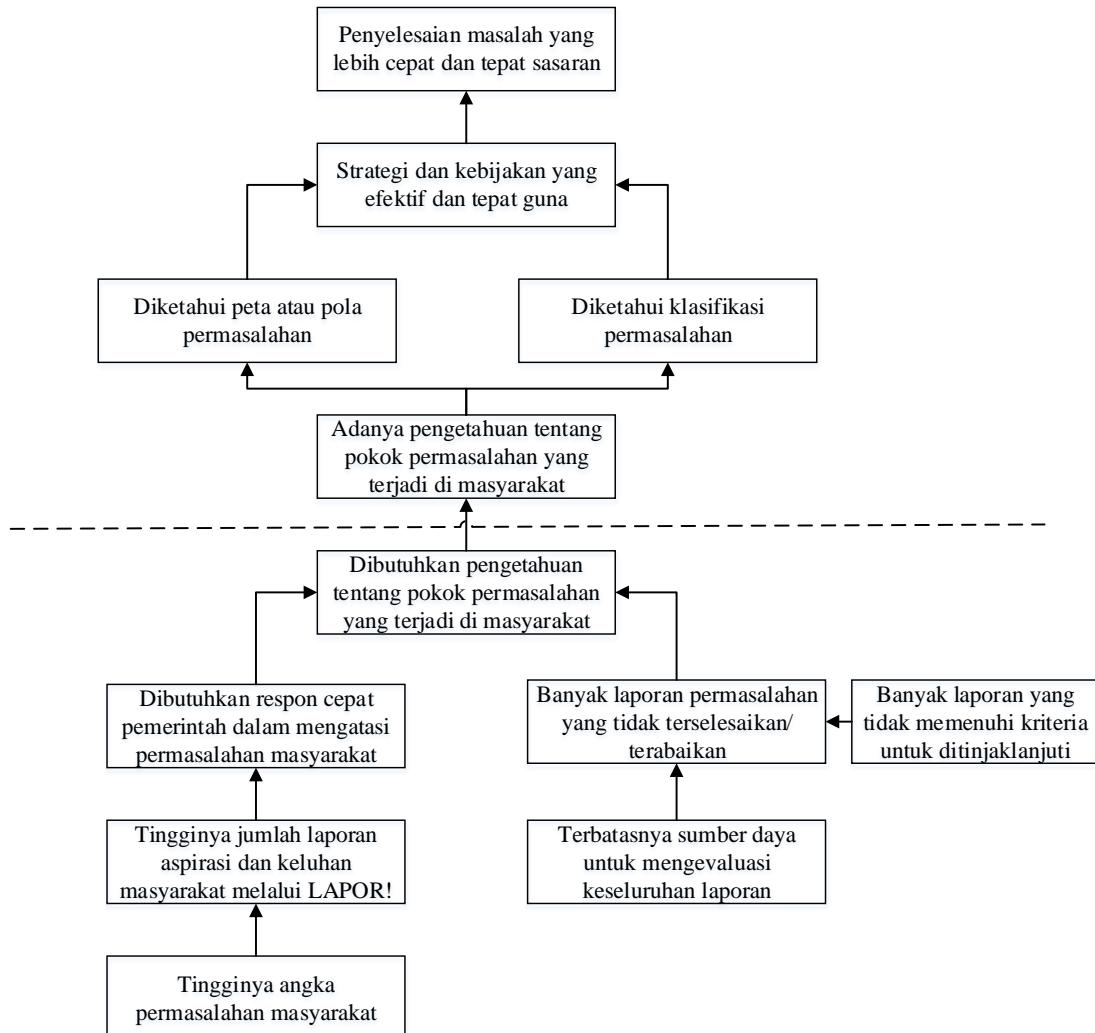
analisis data laporan dengan lebih cepat dan efisien, sehingga pemerintah dapat mengetahui peta atau pola isu dan permasalahan yang terjadi di masyarakat secara *real time*. Pemahaman mengenai segala informasi yang berasal dari masyarakat dapat menjadi masukan bagi pemerintah dalam membuat kebijakan yang lebih efektif dan tepat guna.

1.2. Rumusan Masalah

Dalam rangka pembuatan kebijakan yang efektif dan tepat guna, pemerintah perlu untuk mengetahui berbagai permasalahan yang dilaporkan masyarakat. Akan tetapi, tingginya angka laporan permasalahan yang dilaporkan masyarakat pada LAPOR! tidak mampu dianalisis semua secara manual sehingga banyak laporan yang tidak dianalisis lebih lanjut.

1.3. Diagram Keterkaitan Masalah

Peningkatan jumlah laporan aspirasi dan keluhan pada LAPOR! terjadi seiring dengan meningkatnya angka permasalahan masyarakat. Masyarakat membutuhkan respon cepat pemerintah untuk mengatasi segala permasalahan yang ada, tetapi jumlah sumber daya manusia yang terbatas tidak mampu mengevaluasi seluruh laporan yang ada. Pemerintah perlu mengetahui dan memahami pokok-pokok permasalahan yang terjadi melalui analisis lebih mendalam dari laporan yang disampaikan masyarakat untuk menyusun strategi dan kebijakan yang efektif dan tepat guna. Gambar 1-3 menunjukkan ilustrasi diagram keterkaitan masalah pada penelitian ini.



Gambar 1.3 Diagram Keterkaitan Masalah

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mendapatkan analisis pengaduan dari laporan yang diterima melalui LAPOR! yang berupa pola permasalahan yang terjadi di masyarakat dengan cara membangun model klasifikasi (model yang dapat membantu mengklasifikasikan dokumen) dan melakukan pengelompokan (*clustering*) terhadap dokumen. Pendekatan yang digunakan adalah *Text Mining* dengan metode *Support Vector Machine* (SVM) untuk mendapatkan klasifikasi, dan *Self Organizing Map* (SOM) untuk memetakan kelompok atau *cluster* dari setiap kelas klasifikasi.

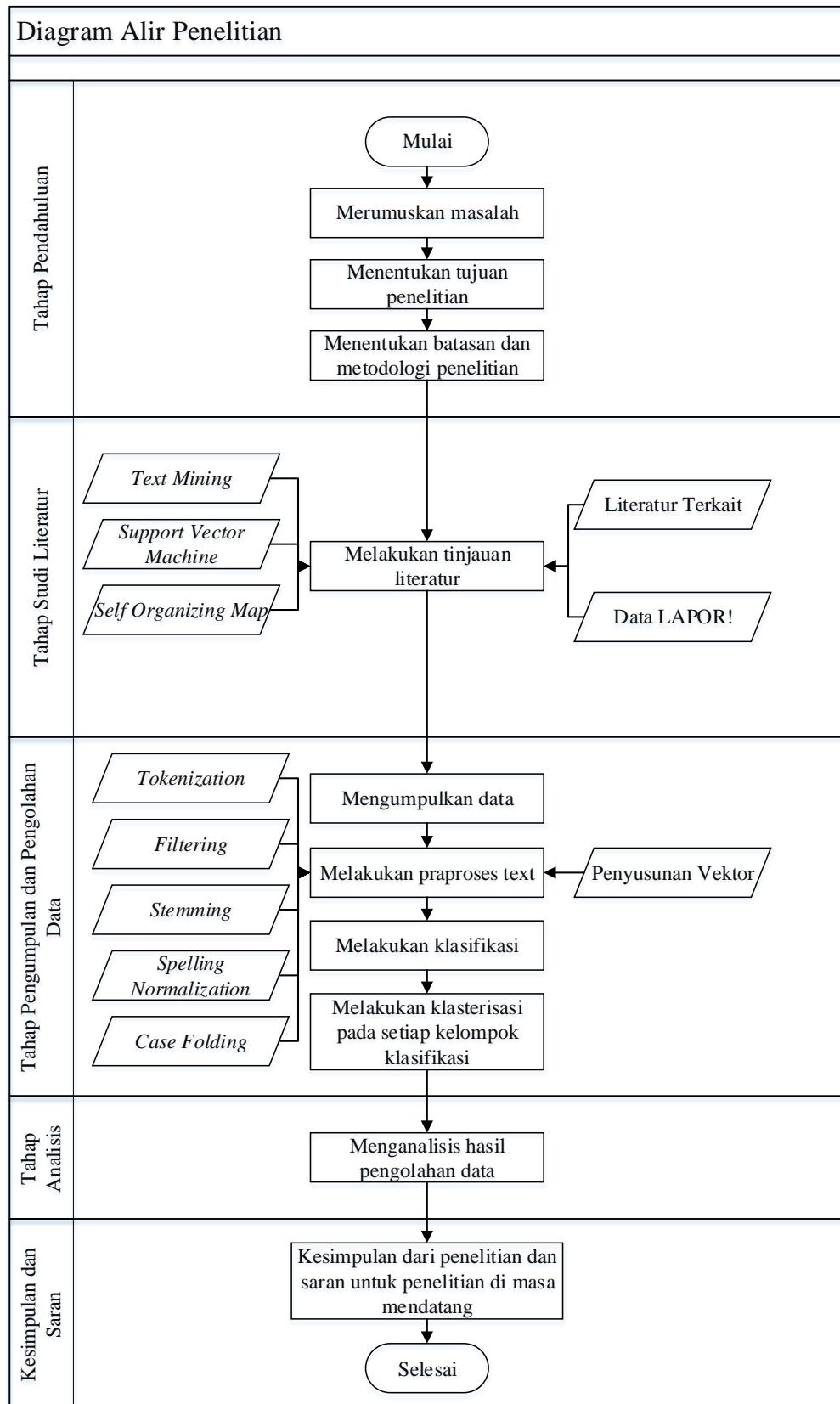
1.5. Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data laporan masyarakat yang digunakan adalah data laporan yang disampaikan melalui LAPOR!
2. Data yang digunakan adalah data historis laporan mulai dari bulan Oktober 2014 - Maret 2015

1.6. Metodologi Penelitian

Metodologi penelitian ini dibagi menjadi lima tahapan seperti ditunjukkan pada gambar 1.4. Secara umum, pada tahap pertama dilakukan perumusan masalah hingga menentukan batasan serta metodologi masalah. Pada tahap kedua dilakukan tinjauan literatur untuk mempelajari dan memahami segala dasar teori dan metode yang digunakan dalam penelitian. Selanjutnya data dikumpulkan dan diolah menggunakan metode yang telah ditentukan di tahap ketiga dan hasil pengolahan data dianalisis di tahap keempat. Tahap terakhir melakukan penarikan kesimpulan dari penelitian yang dilakukan dan memberikan saran untuk kemungkinan penelitian di masa mendatang.

**Gambar 1.4 Diagram Alir Penelitian**

1.7. Sistematika Penelitian

Penulisan skripsi ini terdiri dari lima bab, yaitu:

1. Bab pertama merupakan bab pendahuluan yang membahas mengenai latar belakang dilaksanakannya penelitian, tujuan, batasan masalah, metodologi, dan sistematika penulisan penelitian.
2. Bab kedua merupakan bab yang menjelaskan dasar teori dalam penelitian. Tinjauan literatur mengenai LAPOR! sebagai objek penelitian dalam penelitian dan metode *Data Mining* dan *Text Mining* yang akan digunakan dalam penelitian ini akan dibahas dalam bab dua.
3. Bab ketiga merupakan bagian yang membahas mengenai pengumpulan dan pengolahan data. Pada bab ini akan dijelaskan mengenai keseluruhan langkah dalam penelitian mulai dari pengambilan data, pembersihan data, hingga pengolahan data menjadi hasil akhir yang diinginkan.
4. Bab keempat merupakan bab yang membahas mengenai analisis dari hasil pengolahan data yang dilakukan.
5. Bab kelima merupakan bab yang berisi kesimpulan dan saran. Kesimpulan akan menjelaskan rangkuman hasil penelitian berdasarkan tujuan penelitian dan saran akan membahas mengenai pengembangan penelitian yang bisa dilakukan di masa depan.

BAB 2 **LANDASAN TEORI**

Dewasa ini pemerintah semakin serius dalam meningkatkan mutu pembangunan dan pelayanan publik. Salah satu langkah mewujudkannya adalah dengan mengembangkan *e-Governement* di Indonesia dengan maksud untuk menciptakan interaksi dengan masyarakat melalui LAPOR!. Berbagai masukan yang berasal dari masyarakat memegang peranan penting bagi peningkatan dan kemajuan dalam berbagai aspek. Pada penelitian ini, data tekstual yang berupa masukan atau komentar dari masyarakat akan dijadikan sebagai topik analisis. Dengan menggunakan teknik analisis *text mining*, data tekstual tersebut akan memberikan pengetahuan mengenai berbagai isu dan masalah yang mendapat banyak perhatian atau sorotan dari masyarakat.

2.1. *E-Governement* (Pemerintah Elektronik)

Menurut *Global E-Government Readiness Report* 2004, *e-Government* adalah penggunaan teknologi informasi dan komunikasi dan aplikasinya oleh pemerintah untuk menyediakan informasi dan pelayanan publik kepada masyarakat. Menurut Palvia dan Sharma (2007), *e-Government* merupakan sebuah istilah yang digunakan untuk mendeskripsikan layanan berbasis web dari instansi pemerintah baik pemerintah nasional, negara bagian, atau daerah dimana pemerintah menggunakan teknologi informasi terutama internet untuk mendukung operasi pemerintah yang melibatkan masyarakat, dan menyediakan layanan pemerintah yang terbuka. Lebih luas lagi, *e-Government* dapat didefinisikan sebagai penggunaan dan penerapan teknologi informasi dalam administrasi publik untuk mempersingkat dan mengintegrasikan alur kerja dan proses; secara efektif mengelola data dan informasi; meningkatkan pelayanan publik; serta memperluas saluran komunikasi dalam rangka keterlibatan dan pemberdayaan masyarakat (UN, 2014). *E-Government* memfasilitasi penyediaan informasi yang relevan dalam bentuk elektronik kepada masyarakat secara *real time*; pelayanan yang lebih baik kepada masyarakat; pemberdayaan masyarakat melalui akses informasi tanpa birokrasi; dan partisipasi dalam kebijakan publik melalui pengambilan

keputusan bersama. Secara keseluruhan, terdapat empat konsep *e-Government* (Palvia & Sharma, 2007) yaitu:

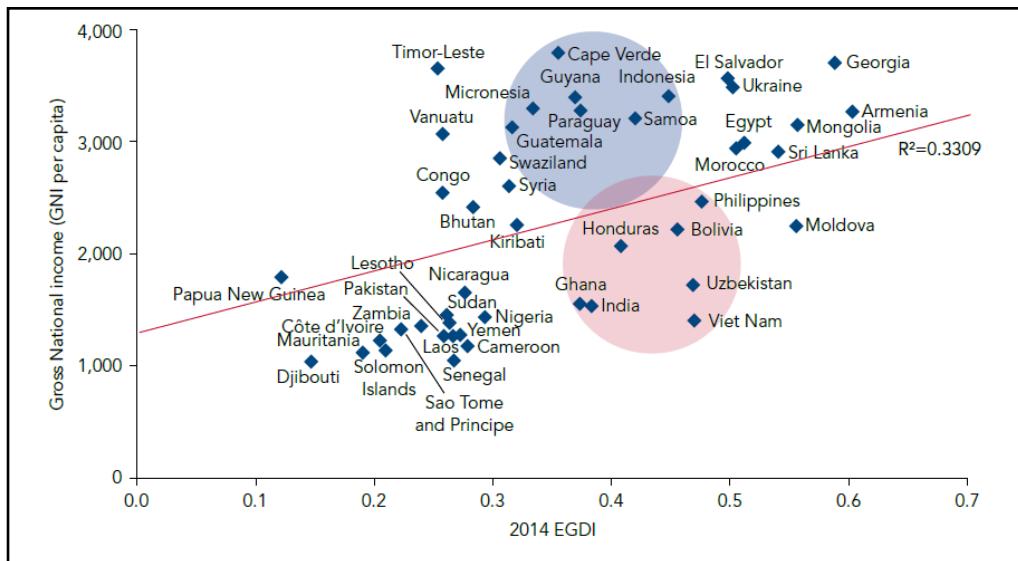
- G2C (*Government to Citizen*), merupakan berbagai jenis kegiatan dimana pemerintah menyediakan akses informasi dan layanan secara *online* kepada masyarakat. Aplikasi G2C memungkinkan masyarakat untuk mengajukan pertanyaan kepada instansi pemerintah dan menerima jawaban; mengurus segala urusan pajak secara *online*; memperbarui SIM dan membayar tiket lalu lintas *online*; pemerintah dapat menyebarkan berbagai informasi kebijakan juga secara *online* di web; menyediakan formulir yang dapat diunduh secara online; membantu masyarakat mencari pekerjaan; memberikan informasi pariwisata dan rekreasi; memberikan pengetahuan tentang isu-isu kesehatan dan keselamatan; dan masih banyak lainnya
- G2B (*Government to Business*), merupakan interaksi pemerintah dengan bisnis. G2B biasanya merupakan transaksi dua arah dimana terdapat G2B itu sendiri dan B2G (*Business to Government*). Kegiatan yang dilakukan biasanya mencangkup proses interaksi dan transaksi mengenai penyediaan dan jual beli pemerintah dengan bisnis.
- G2G (*Government to Government*), merupakan kegiatan-kegiatan yang terjadi antar organisasi pemerintah/lembaga yang berbeda. Banyak dari kegiatan ini bertujuan untuk meningkatkan efisiensi dan efektifitas operasi pemerintah secara keseluruhan. Misalnya integrasi antara kementerian dan BUMN dan lainnya.
- *Governement to Constituents (E-Democracy)*, merupakan kegiatan yang mengacu pada aktivitas *online* pemerintah, partai politik, dan warga negara dalam lingkup proses demokrasi. Kegiatan yang dimaksud mencangkup *polling*, *voting*, dan kampanye *online* dimana terjadi diskusi urusan politik dan konsultasi *online* antara masyarakat dengan kandidat politik. Misalnya selama pemilihan umum presiden Amerika Serikat tahun 2004 dan 2006, kedua kandidat partai besar memiliki portal informasi mereka sendiri dan mereka juga mengirimkan pesan *e-mail* kepada pemilih potensial. Di Korea Selatan, karena banyaknya pengguna

web yang jarang membaca koran atau menonton televisi, politisi harus bergantung pada Internet untuk merekrut pemilih.

E-Government memegang peranan penting dalam perkembangan berkelanjutan global (UN, 2014). Tantangan global yang meliputi kemiskinan, perubahan iklim, perdamaian dan keamanan, dan lainnya menyebabkan tidak ada satu aktor pemerintah atau satu kementerian tunggal yang mampu menanganinya sendiri. Kolaborasi yang efektif antar instansi di semua tingkat pemerintahan dan juga instansi non-pemerintahan menjadi sangat penting untuk memastikan tata pemerintahan yang baik dan hasil pembangunan yang baik pula. Sektor publik harus memberikan keadilan dan efisiensi layanan untuk memenuhi kebutuhan masyarakat, memberikan peluang untuk pertumbuhan ekonomi, serta memfasilitasi keterlibatan warga negara dalam pengambilan kebijakan dan pelayanan publik dan pelayanan. Peluang yang ditawarkan oleh perkembangan dunia digital beberapa tahun terakhir, baik melalui layanan *online*, *big data*, media sosial, aplikasi mobile, hingga *cloud computing*, memperikan kesempatan bagi *e-government* untuk berkembang.

Menurut *United Nations E-Government Survey 2014*, perkembangan *e-Government* tidak lepas dari tingkat pendapatan nasional suatu negara dimana akses infrastruktur dan literatur teknologi informasi dan komunikasi dan penyediaan tingkat pendidikan bisa dikatakan sangat terkait dengan tingkat pendapatan suatu negara. Namun, pendapatan nasional bukan merupakan satu-satunya penentu perkembangan *e-Government*. Beberapa negara mempunyai tingkat perkembangan *e-Government* yang tinggi meskipun pendapatan nasional mereka relatif rendah dan beberapa negara dengan tingkat pendapatan yang relatif lebih tinggi mempunyai peringkat yang sama atau bahkan lebih rendah. Negara dengan pendapatan lebih tinggi tersebut sebenarnya merupakan negara yang mempunyai potensi yang sangat bagus untuk perkembangan di masa depan. Diantara negara-negara pendapatan menengah kebawah (*low-middle income*), negara dengan potensi tersebut misalnya Cape Verde, Guatemala, Guyana, Mikronesia, Paraguay, Samoa dan Indonesia, seperti yang ditunjukkan pada Gambar 2.1 dengan tanda biru. Walaupun Indonesia mempunyai tingkat pendapatan nasional yang lebih tinggi daripada Filipina, tetapi tingkat

perkembangan *e-Government* Filipina sedikit lebih tinggi daripada Indonesia. Hal ini mengindikasikan bahwa sebenarnya Indonesia mempunyai potensi yang bagus untuk melakukan perkembangan *e-Government* kedepannya.



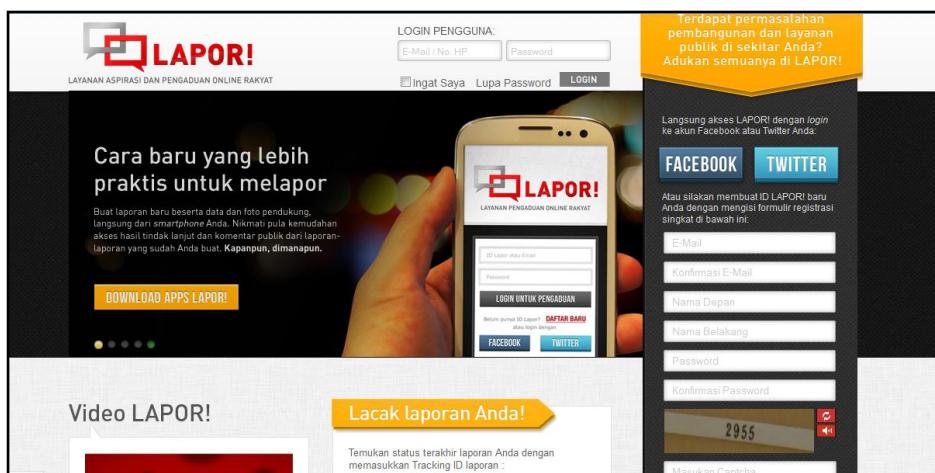
Gambar 2.1 Hubungan antara EGDI (*E-Government Development Index*) dengan GNI (Gross National Income) per capita dari negara pendapatan menengah kebawah (*lower-middle income countries*)

(Sumber: *United Nations E-Government Survey 2014*)

2.2. LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat)

Portal LAPOR! merupakan salah satu portal di Indonesia yang bisa dikategorikan sebagai *e-Government*. Pasal 36 dan 37 Undang-Undang Nomor 25 tahun 2009 tentang Pelayanan Publik mengamanatkan pemerintah wajib memberikan akses seluas luasnya kepada masyarakat untuk memberikan masukan atas pemberian layanannya. Sebagai tindak lanjut amanat Undang-Undang Nomor 25 tahun 2009, telah diterbitkan Peraturan Presiden Nomor 76 Tahun 2013 Tentang Pengelolaan Pengaduan Pelayanan Publik, yang mengisyaratkan dibentuknya Sistem Pengelolaan Pengaduan Pelayanan Publik Nasional (SP4N) yang merupakan integrasi pengelolaan pengaduan pelayanan publik secara berjenjang pada setiap penyelenggara dalam kerangka sistem informasi pelayanan publik dan LAPOR! sebagai alat atau sarana pengaduan yang disediakan.

LAPOR! yang semula diinisiasi dan dikembangkan oleh Unit Kerja Presiden Bidang Pengawasan dan Pengendalian Pembangunan (UKP-PPP atau UKP4) itu kini dikelola oleh Kantor Staf Presiden. LAPOR! bertujuan untuk memfasilitasi masyarakat yang mengalami kesulitan dan kebingungan ketika ingin menyampaikan keluh kesahnya. Hingga saat ini LAPOR! telah mencangkup 80 Kementerian atau Lembaga dan 5 Pemerintah Daerah serta BUMN di Indonesia (LAPOR!, 2015). LAPOR! akan mendisposisikan laporan masyarakat kepada Kementerian/Lembaga atau Pemerintah Daerah terkait untuk ditindaklanjuti. LAPOR! juga berpotensi untuk memudahkan koordinasi antar instansi, tidak hanya koordinasi kewenangan antar Kementerian/Lembaga namun juga koordinasi antara Pemerintah Pusat, Pemerintah Provinsi, dan Pemerintah Kabupaten/Kota.



Gambar 2.2 Homepage LAPOR!

(Sumber : www.lapor.go.id, 2015)

2.2.1. Sistem Pelaporan

LAPOR! adalah sistem aplikasi pengelolaan pengaduan berbasis media sosial pertama yang mengedepankan prinsip mudah dan terpadu. Dalam platform LAPOR!, masyarakat dapat menyampaikan aspirasi dan pengaduannya melalui berbagai kanal, antara lain: SMS 1708, situs www.lapor.ukp.go.id, mobile apps Android dan Blackberry, serta media sosial Twitter @LAPOR1708 dan Facebook LAPOR!. Setiap laporan yang masuk akan diverifikasi terlebih dahulu oleh administrator LAPOR! mengenai kejelasan dan kelengkapan dan akan diteruskan ke instansi terkait secara digital, cepat, dan tepat paling lambat 3 hari kerja setelah

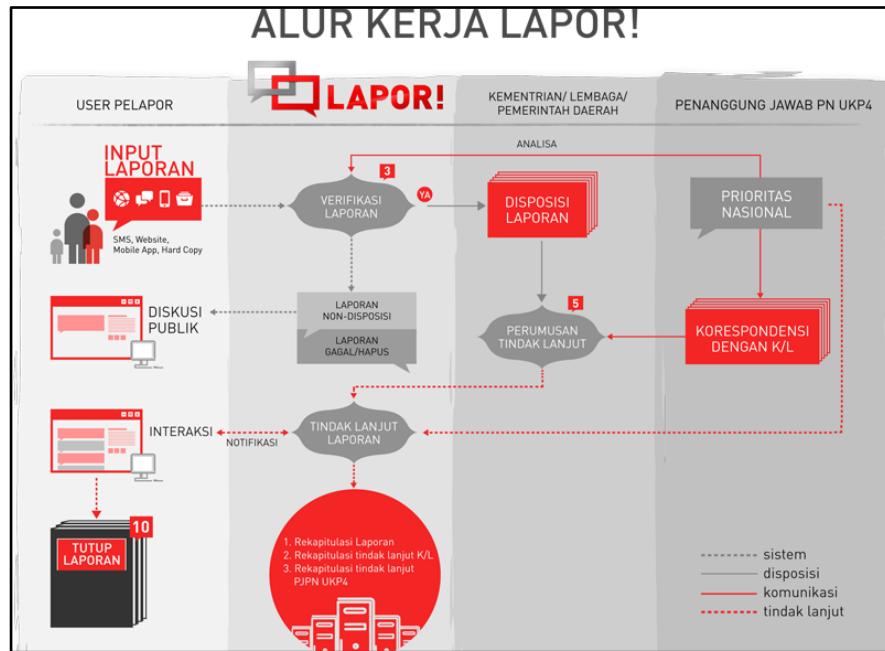
pelaporan dilakukan kemudian dipantau tindak-lanjutnya secara interaktif. Secara umum alur kerja pelaporan ditunjukkan pada Gambar 2-2.

Setiap laporan yang diterima diproses menjadi tiga jenis perlakuan, disetujui, dipending, atau diarsipkan. Laporan yang disetujui merupakan laporan yang sudah jelas atau merupakan aspirasi yang bagus, sehingga akan langsung diproses dan diteruskan untuk ditindaklanjuti. Laporan yang dipending merupakan laporan yang dirasa bagus tetapi belum jelas, sehingga perlu dilakukan konfirmasi kembali kepada pelapor. Sedangkan laporan yang diarsipkan merupakan laporan yang dirasa tidak jelas, laporan yang berulang atau sudah pernah dilaporkan sebelumnya, atau merupakan saran yang bersifat sangat umum sehingga tidak diproses lebih lanjut.

LAPOR! akan mempublikasikan setiap laporan yang sudah disetujui dan diteruskan sekaligus memberikan notifikasi kepada pelapor. Instansi Kementerian/Lembaga/Daerah diberikan waktu paling lambat 5 hari kerja untuk melakukan koordinasi internal dan tindak lanjut laporan yang dilaporkan oleh masyarakat. Jika sudah ada tindak lanjut terkait laporan tersebut, maka instansi yang bersangkutan akan memberikan informasi kepada pelapor pada halaman tindak lanjut laporan. Kemudian laporan akan ditutup atau dianggap selesai jika sudah terdapat tindak lanjut dari instansi terkait dan telah berjalan 10 hari kerja setelah tindak lanjut dilakukan tanpa adanya balasan dari pelapor maupun administrator LAPOR! di halaman tindak lanjut (LAPOR!, 2015).

2.2.2. Fitur LAPOR!

LAPOR! mempunyai komitmen untuk menjaga transparansi dan akuntabilitas pengelolaan pengaduan, oleh karena itu sistem LAPOR! didesain dengan mekanisme terbuka. Hal ini diwujudkan dengan dukungan berbagai fitur yang disediakan oleh LAPOR! untuk memperbaiki kemudahan dan kenyamanan masyarakat dalam menggunakan LAPOR!.



Gambar 2.3 Alur Kerja LAPOR!

(Sumber : www.lapor.go.id, 2015)

Beberapa fitur yang disediakan antara lain (LAPOR!, 2015):

- Tracking ID LAPOR!, yaitu sebuah kode unik yang akan secara otomatis melengkapi setiap laporan yang dipublikasikan pada situs LAPOR!. Tracking ID dapat digunakan pengguna untuk melakukan penelusuran atas suatu laporan.
- Anonim dan Rahasia, anonim merupakan fitur yang diperuntukkan bagi pelapor yang ingin merahasiakan identitasnya, sedangkan fitur rahasia digunakan untuk membatasi akses atas laporan hanya bagi pelapor dan instansi terlapor. Biasanya kedua fitur ini diperuntukkan bagi pelaporan isu-isu sensitif dan sangat privat.
- Unggah data pendukung, fitur ini merupakan fitur yang bisa digunakan bagi pelapor untuk melengkapi laporannya. Data pendukung biasanya berupa foto, dokumen, atau berbagai bukti kejadian.
- Peta dan Kategorisasi, fitur ini dapat digunakan untuk melabeli setiap laporan dengan dengan lokasi geografis, topik, status ketuntasan laporan, dan institusi terkait sehingga pemerintah maupun masyarakat dapat

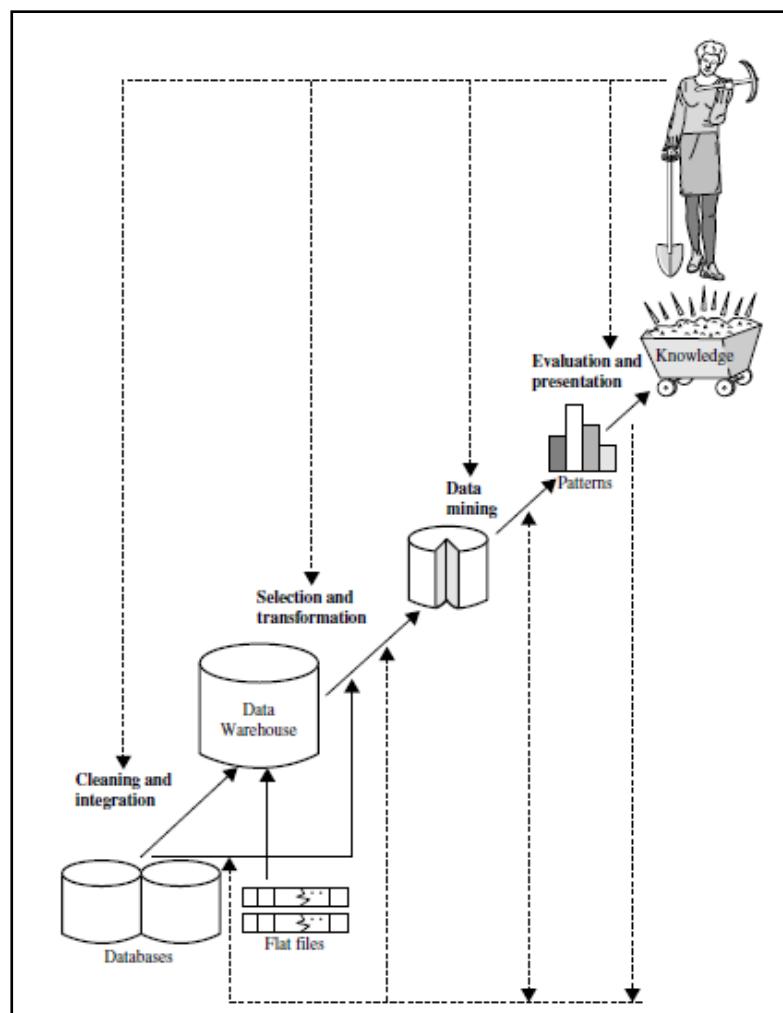
memonitor isu dalam berbagai skala dan sudut pandang. Salah satu contohnya peta LAPOR! dipergunakan sebagai pusat informasi banjir pada saat bencana banjir besar Jakarta di tahun 2012 dan 2014 sebagai rujukan dalam rangka penyaluran bantuan kepada para korban. Pada fitur kategorisasi, beberapa kategori topik yang digunakan oleh LAPOR! pada situsnya yaitu mengenai:

1. Reformasi Birokrasi dan Tata Kelola
 2. Pendidikan
 3. Kesehatan
 4. Kemaritiman
 5. Pertanian
 6. Energi dan Sumber Daya Alam
 7. Infrastruktur
 8. Lingkungan Hidup dan Penanggulangan Bencana
 9. Bidang Politik, Hukum, dan Keamanan
 10. Bidang Perekonomian
 11. Bidang Kesejahteraan Masyarakat
- Opini kebijakan, fitur ini merupakan fitur yang dapat digunakan oleh instansi pemerintah yang terhubung sebagai sarana jajak pendapat masyarakat. Salah satu contoh jajak pendapat yang telah dilakukan melalui fitur ini diantaranya tentang Badan Penyelenggara Jaminan Sosial Kesehatan dan Rencana Implementasi Kurikulum Baru Pendidikan 2013.

Kedepannya, LAPOR! diharapkan akan mampu memberikan pengukuran kinerja bagi pemerintah dan masukan berarti bagi pemerintah dalam langkah-langkah pengambilan keputusan, seperti pengoptimalan APBN sesuai dengan apa yang dibutuhkan rakyat, dan lain sebagainya.

2.3. Data Mining

Data mining didefinisikan sebagai proses komputasi untuk menganalisis data dalam jumlah besar dengan mengekstrak pola dan informasi yang berguna (Gullo, 2015). Dalam beberapa dekade terakhir, *data mining* telah banyak mendapat sebutan lain seperti *knowledge discovery*, *business intelligence*, *predictive modeling*, *predictive analytics*, dan beberapa lainnya (Linoff & Berry, 2011). Tetapi, tidak sedikit orang yang mendefinisikan *data mining* sebagai sinonim dari istilah populer lainnya yaitu *knowledge discovery from data* (KDD) dan yang lain melihat *data mining* hanya sebagai salah satu tahapan dari *knowledge discovery* (Jiawei, Kamber, & Pei, 2012).



Gambar 2.4 Data mining sebagai tahapan dari *knowledge discovery*

(Sumber : Jiawei, Kamber, & Pei, 2012)

Pada proses *knowledge discovery* seperti ditunjukkan pada Gambar 2.4, terdapat beberapa tahapan proses yang dilakukan yaitu:

- *Cleaning* data, yaitu proses untuk mengeliminasi *noise* (pengganggu) dan data yang tidak konsisten)
- Integrasi data, yaitu proses penggabungan data jika data diperoleh dari berbagai sumber
- Seleksi data, yaitu proses pemilihan data yang benar-benar berguna untuk dianalisis
- Transformasi data, yaitu proses transformasi data menjadi bentuk yang sesuai untuk dilakukan proses data mining
- *Data mining*, yaitu proses dimana metode-metode khusus diaplikasikan untuk mengekstrak informasi dan pola data
- *Pattern evaluation*, yaitu proses untuk mengidentifikasi pola-pola dan informasi menarik yang didapatkan dari data

Pentingnya *data mining* saat ini terutama didorong oleh banyaknya data yang dikumpulkan dan disimpan dengan berbagai aplikasi terkemuka terkini, seperti data web, data *e-commerce*, data pembelian, transaksi bank, dan sebagainya. Data yang dihasilkan oleh aplikasi-aplikasi tersebut umumnya merupakan jenis *Big Data* dimana data tersebut sulit diolah atau dimengerti secara sederhana. *Big Data* merupakan data yang mempunyai tiga karakteristik yaitu jumlah (*volume*) dan variasi (*variety*) besar, serta bergerak cepat (*velocity*), sehingga melampaui kapasitas pengolahan database konvensional (Dumbill, 2014). Hingga saat ini, *data mining* telah banyak diakui sebagai suatu alat analisis data serbaguna yang bisa diaplikasikan untuk menganalisis *big data* dalam berbagai bidang, tidak hanya dalam bidang teknologi informasi tetapi juga dalam dunia pengobatan klinis, sosiologi, fisika, dan banyak lainnya.

Penggunaan data mining dibedakan menjadi dua jenis fungsi yaitu prediktif dan deskriptif (Gullo, 2015). Penggalian prediktif mengacu pada pembangunan model yang berguna untuk memprediksi perilaku atau nilai-nilai di masa depan. Tugas deskriptif meliputi klasifikasi dan prediksi, tugas yang dilakukan seperti membangun beberapa model (atau fungsi) yang

menggambarkan kelas atau konsep data oleh satu set objek data yang label kelasnya diketahui (*training set*), sehingga dapat memprediksi kelas yang labelnya tidak diketahui; deteksi penyimpangan, yaitu berurusan dengan penyimpangan data, yang didefinisikan sebagai perbedaan antara nilai yang terukur dan nilai referensi; analisis evolusi, yaitu, mendeteksi dan menggambarkan pola yang teratur dalam data yang perilakunya berubah dari waktu ke waktu. Sedangkan tujuan penggalian deskriptif yaitu membangun model untuk mendeskripsikan data menjadi bentuk yang mudah dimengerti, efektif, dan efisien. Contoh dari tugas deskriptif diantaranya karakterisasi data, yang tujuan utamanya adalah untuk meringkas karakteristik umum atau fitur dari kelas target data; *association rule*, yaitu, menemukan aturan yang menunjukkan kondisi atribut-nilai yang sering muncul bersama-sama dalam himpunan data; dan *clustering*, yang bertujuan untuk membentuk kelompok yang memiliki kohesif tinggi dan terpisahkan dengan baik dari satu set objek data.

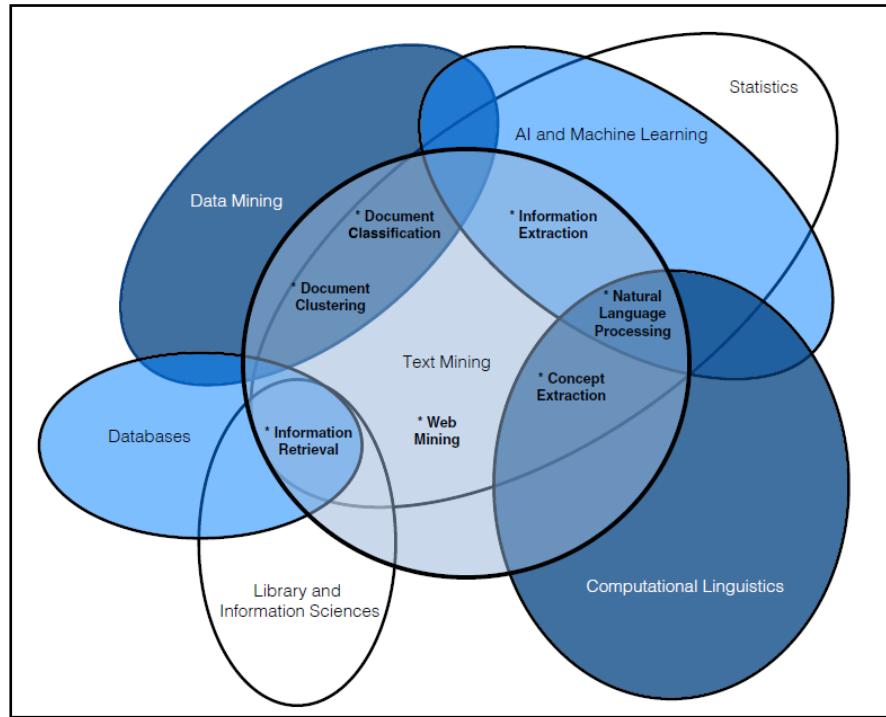
2.4. Text Mining

Text mining atau *text analytics* adalah istilah yang mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, hal inilah yang membedakannya dengan *data mining* dimana *data mining* mengolah data yang sifatnya terstruktur. Pada dasarnya, *text mining* merupakan bidang interdisiplin yang mengacu pada perolehan informasi (*information retrieval*), *data mining*, pembelajaran mesin (*machine learning*), statistik, dan komputasi linguistik (Jiawei, Kamber, & Pei, 2012). Secara umum konsep pekerjaan *text mining* mirip dengan *data mining*, yaitu penggalian prediktif dan penggalian deskriptif. *Text mining* mengekstrak indeks numerik yang bermakna dari teks dan kemudian informasi yang terkandung dalam teks akan diakses dengan menggunakan berbagai algoritma *data mining* (statistik dan *machine learning*) (Miner et al, 2012).

Beberapa tahun terakhir, penggunaan dan penelitian mengenai *text mining* telah banyak mendapat perhatian dan aktif dilakukan seiring dengan semakin banyaknya data teks yang diperoleh dari berbagai jaringan sosial, web, dan aplikasi lainnya. Sebagian besar informasi teks yang disimpan tersebut seperti misalnya artikel berita, makalah, buku, perpustakaan digital, pesan email, blog,

dan halaman web. *Text mining* dapat menganalisis dokumen, mengelompokkan dokumen berdasarkan kata-kata yang terkandung di dalamnya, serta menentukan kesamaan di antara dokumen untuk mengetahui bagaimana mereka berhubungan dengan variabel lainnya (Statsoft, 2015). Aplikasi yang paling umum dilakukan *text mining* saat ini misalnya penyaringan spam, analisis sentimen, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topik penelitian, dan banyak lainnya. Menurut Miner et al (2012), pekerjaan *text mining* dikelompokkan menjadi 7 daerah praktik yang diilustrasikan seperti Gambar 2-4.

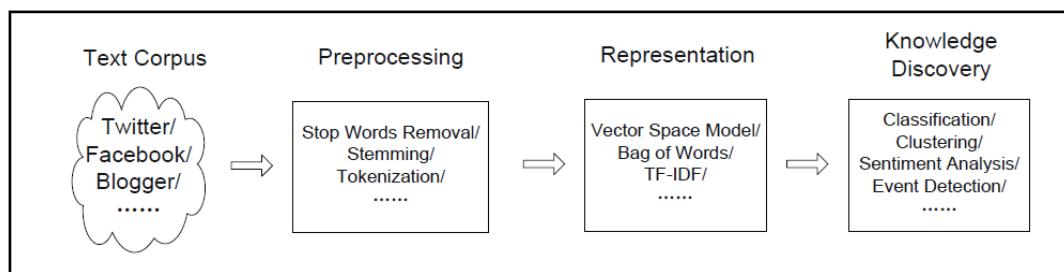
- Pencarian dan perolehan informasi (*search and information retrieval*), yaitu penyimpanan dan penggalian dokumen teks misalnya dalam mesin pencarian (*search engine*) dan pencarian kata kunci (*keywords*)
- Pengelompokan dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode klaster (*clustering*) *data mining*.
- Klasifikasi dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode klasifikasi (*classification*) *data mining* berdasarkan model terlatih yang sudah memiliki label.
- *Web mining*, yaitu penggalian informasi dari internet dengan skala fokus yang spesifik.
- Ekstraksi informasi (*information extraction*), yaitu mengidentifikasi dan mengekstraksi informasi dari data yang sifatnya semi-terstruktur atau tidak terstruktur dan mengubahnya menjadi data yang terstruktur.
- *Natural language processing* (NLP), yaitu pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia.
- Ekstraksi konsep, yaitu pengelompokan kata atau frase ke dalam kelompok yang mirip secara semantik



Gambar 2.5 Diagram venn 6 bidang terkait dan 7 area praktik *text mining*

(Sumber: Miner et al, 2012)

Untuk memperoleh tujuan akhir dari *text mining*, diperlukan beberapa tahapan proses yang harus dilakukan seperti ditunjukkan pada Gambar 2.6. Data terpilih yang akan dianalisis pertama akan melewati tahap Pra-proses dan representasi teks, hingga akhirnya dapat dilakukan *knowledge discovery*.



Gambar 2.6 Kerangka proses analisis teks pada *text mining*

(Sumber : Zhai & Aggarwal, 2012)

2.4.1. Pra-proses (*pre-processing task*)

Data yang diinput perlu melewati fase pra-proses terlebih dahulu agar dapat dimengerti oleh sistem pengolahan *text mining* dengan baik. Fase pra-proses merupakan fase yang penting untuk menentukan kualitas proses selanjutnya (proses klasifikasi dan pengelompokan). Tujuan utama fase pra-proses adalah untuk mendapatkan bentuk data siap olah untuk diproses oleh *data mining* dari data awal yang berupa data tekstual. Fitur-fitur fase pra-proses terdiri dari beberapa tahap sebagai berikut:

- Pemilihan dokumen yang digunakan (dokumen yang mengandung ancaman, caci maki, SARA, dan pornografi dihilangkan).
- *Tokenization*, merupakan proses pemisahan teks menjadi potongan kalimat dan kata yang disebut *token*.
- *Filtering*, merupakan proses membuang kata-kata serta tanda-tanda yang tidak bermakna secara signifikan, seperti hashtag (#), url, tanda baca tertentu (*emoticon*), dan lainnya.
- *Stemming*, merupakan proses pengambilan akar kata. Misalnya kata memakai, dipakai, pemakai, dan pemakaian akan memiliki akar kata yang sama yaitu “pakai”.
- *Spelling normalization*, merupakan perbaikan kata-kata yang salah eja atau disingkat dengan bentuk tertentu. Misalnya kata “tidak” memiliki banyak bentuk penulisan seperti tdk, gak, nggak, enggak, dan banyak lainnya.
- *Case Folding*, merupakan proses pengubahan huruf dalam dokumen menjadi satu bentuk, misalnya huruf kapital menjadi huruf kecil dan sebaliknya.

2.4.2. Penyusunan Vektor (*Representation*)

Proses operasi algoritma belajar (*learning algorithms*) tidak bisa langsung memproses dokumen teks dalam bentuk aslinya. Oleh karena itu, setelah tahap *pre-processing*, dokumen diubah menjadi representasi yang lebih mudah dikelola. Biasanya, dokumen akan diwakili oleh vektor (Feldman & Sanger, 2007). Model vektor dibangun dari dokumen dengan mengubah *token-token* dalam dokumen menjadi vektor numerik yang akan dioperasikan berdasarkan operasi aljabar linear

(Zhai & Aggarwal, 2012). Dalam rangka membangun model vektor, perlu dilakukan proses pembobotan. Skema pembobotan yang paling banyak digunakan adalah skema *term frequency-inverse document frequency* (TF-IDF). *Term frequency* (TF) didefinisikan sebagai jumlah kemunculan suatu kata/istilah dalam suatu dokumen. Misalnya TF pada dokumen pertama untuk kata/istilah “jalan” adalah 2, karena kata/istilah tersebut muncul 2 kali dalam dokumen pertama. Pada asumsi pembobotan dibalik TF-IDF, kata-kata dengan nilai TF yang tinggi akan mendapat bobot yang tinggi kecuali jika jumlah dokumen yang mengandung kata tersebut juga tinggi (*inverse document frequency* (IDF)). Misalnya kata “yang” memiliki jumlah kemunculan yang tinggi tetapi jumlah dokumen yang mengandung kata “yang” juga tinggi, sehingga kata tersebut akan memiliki bobot yang rendah. Skema persamaan TF-IDF ditunjukkan oleh persamaan berikut (Zhai & Aggarwal, 2012).

$$tfidf(w) = tf \times \log \frac{N}{df(w)} \quad (2.1)$$

Keterangan:

$tf(w)$ = *Term frequency* (jumlah kemunculan suatu kata dalam suatu dokumen)

$df(w)$ = *Document frequency* (jumlah dokumen yang mengandung suatu kata)

N = Jumlah dokumen

Setelah melewati skema TF-IDF, akan didapatkan hasil yang berupa matriks. Matriks yang didapatkan adalah matriks yang merepresentasikan dokumen dalam baris dan *token-token* atau kata yang sudah dipisah-pisahkan dalam kolom. Walaupun sudah memiliki bentuk yang sudah sesuai dan mampu diolah lebih lanjut menggunakan algoritma pembelajaran, tetapi hasil matriks yang didapatkan masih memiliki dimensi yang sangat tinggi. Oleh karena itu dibutuhkan satu tahapan lagi yaitu *Singular Value Decomposition* (SVD) untuk menurunkan dimensi matriks. *Singular Value Decomposition* (SVD) merupakan suatu metode untuk menurunkan dimensi matriks dengan cara menemukan struktur dan korelasi tersembunyi dari matriks (Miner et al, 2012). Pada dasarnya, metode SVD menggunakan dasar teorema aljabar linier, dimana matriks persegi

panjang A dapat diuraikan menjadi tiga jenis matriks, yaitu matriks ortogonal U, matriks diagonal S, dan transpose matriks diagonal V ((Baker, 2013). Persamaan tersebut dapat ditulis sebagai berikut.

$$A_{mn} = U_{mn} S_{mn} V^T_{mn} \quad (2.2)$$

Keterangan :

- U = Vektor eigen ortonormal dari AA^T
- V = Vektor eigen ortonormal dari $A^T A$
- S = Matriks diagonal dimana nilai diagonalnya merupakan akar dari nilai U dan V yang disusun dengan urutan menurun berdasarkan besarnya nilai (nilai singular dari A)
- A_{mn} = Matriks yang mewakili m jumlah dokumen dan n jumlah kata pada dokumen

2.4.3. Ekstraksi Informasi pada *Text Mining*

Tahap akhir penggalian informasi pada *text mining* yaitu ekstraksi ilmu pengetahuan (*knowledge discovery*), dimana terdapat beberapa jenis kategori utama yang bisa dilakukan sebagai berikut (Miner et al, 2012).

- Klasifikasi/prediksi,

Klasifikasi adalah bentuk analisis data yang mengekstrak model untuk menggambarkan kelas data (Jiawei, Kamber, & Pei, 2012). Model yang dibangun meliputi pengklasifikasian dan prediksi kategori label kelas. Klasifikasi data mempunyai dua tahapan proses, yaitu tahap pembelajaran (*learning step*) dimana model klasifikasi dibangun berdasarkan label yang sudah diketahui sebelumnya dan tahapan klasifikasi (*classification step*) dimana model digunakan untuk memprediksi label kelas dari data yang diberikan (Miner et al, 2012). Klasifikasi memiliki berbagai aplikasi, termasuk deteksi penipuan, penargetan marketing, prediksi kinerja, manufaktur, diagnosis medis, dan banyak lainnya. Sebagai contoh, kita dapat membangun sebuah model klasifikasi untuk mengkategorikan apakah suatu aplikasi pinjaman bank termasuk aman atau berisiko. Karena pada

awal pembangunan model label kelas dari data telah diketahui, klasifikasi juga disebut sebagai metode *supervised learning*.

- Pengelompokan (*clustering*)

Tidak seperti klasifikasi, kelompok label kelas pada model *clustering* tidak diketahui sebelumnya dan tugas *clustering* adalah untuk mengelompokkannya (Linoff & Berry, 2011). Menurut Linof & Berry (2011), *clustering* adalah proses pengelompokan satu set data objek menjadi beberapa kelompok atau klaster sehingga objek dalam sebuah klaster memiliki kemiripan yang tinggi satu sama lain, tetapi sangat berbeda dengan objek dalam kelompok lainnya.

- Asosiasi

Asosiasi merupakan proses pencarian hubungan antar elemen data. Dalam dunia industri retail, analisis asosiasi biasanya disebut *Market Basket Analysis* (Miner et al, 2012). Asosiasi tersebut dihitung berdasarkan ukuran *support* (presentase dokumen yang memuat seluruh konsep suatu produk A dan B) dan *confidence* (presentase dokumen yang memuat seluruh konsep produk B yang berada dalam subset yang sama dengan dokumen yang memuat seluruh konsep produk A).

- Analisis Tren

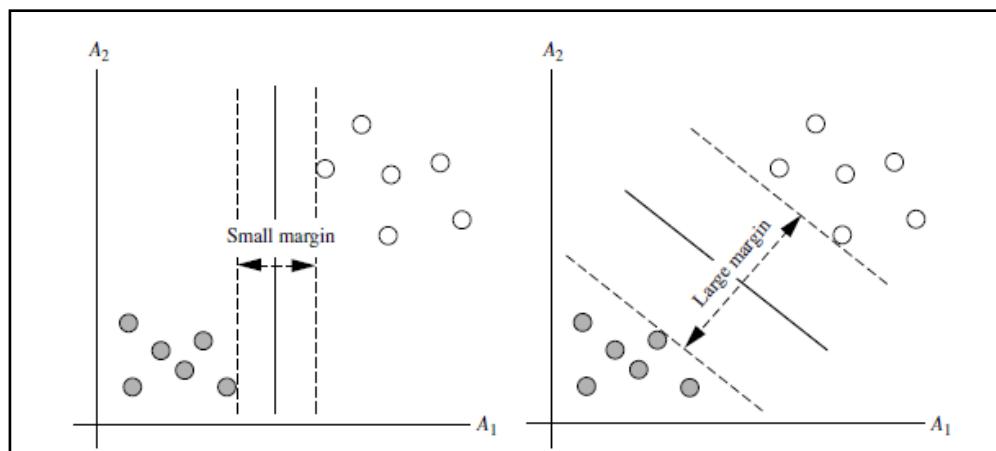
Tujuan dari analisis tren yaitu untuk mencari perubahan suatu objek atau kejadian oleh waktu (Miner, et al). Salah satu aplikasi analisis tren yaitu kegiatan identifikasi evolusi topik penelitian pada artikel akademis..

2.5. Algoritma Penggolong Klasifikasi

Pada studi klasifikasi, proses pembelajaran dilakukan berdasarkan prinsip *machine learning*. *Machine learning* merupakan suatu metode yang menyelidiki bagaimana komputer belajar mengenai data (Jiawei, Kamber, & Pei, 2012). Dalam *machine learning*, *training model* (model latihan) akan dipelajari dengan menggunakan berbagai algoritma yang ditentukan untuk mendapatkan model pengklasifikasi yang dapat digunakan untuk mengklasifikasikan dokumen lainnya yang belum mempunyai kategori sebelumnya. Algoritma yang biasanya digunakan untuk melakukan pengklasifikasian antara lain adalah *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier*.

2.5.1. Support Vector Machine (SVM)

SVM merupakan algoritma klasifikasi yang memiliki tujuan untuk menemukan fungsi pemisah (*hyperplane*) dengan margin paling besar, sehingga dapat memisahkan dua kumpulan data secara optimal (Jiawei, Kamber, & Pei, 2012). Gambar 2.7 menunjukkan dua *hyperplane* yang mungkin untuk memisahkan dua kelompok data. Kedua *hyperplane* dapat mengklasifikasikan semua tupel data yang diberikan, tetapi *hyperplane* dengan margin yang lebih besar mempunyai tingkat akurasi lebih tinggi dalam melakukan klasifikasi karena dapat memisahkan kumpulan data yang satu dengan lainnya dengan mencari tingkat pemisah yang paling jauh antar kelompok.



Gambar 2.7 Margin Hyperplane SVM

(Sumber: Jiawei, Kamber, & Pei, 2012)

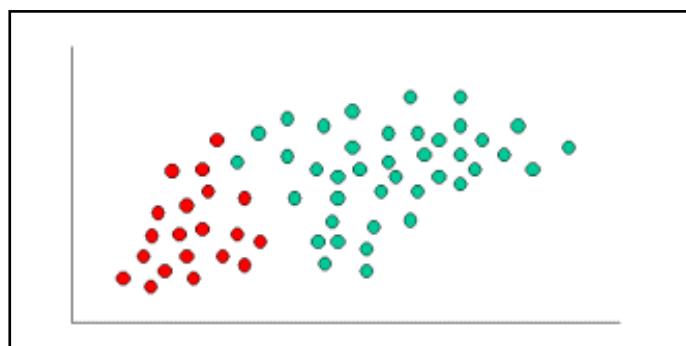
SVM pada awalnya digunakan untuk klasifikasi data numerik, tetapi ternyata SVM juga sangat efektif dan cepat untuk menyelesaikan masalah-masalah data teks. Data teks cocok untuk dilakukan klasifikasi dengan algoritma SVM karena sifat dasar teks yang cenderung mempunyai dimensi yang tinggi, dimana terdapat beberapa fitur yang tidak relevan, tetapi akan cenderung berkorelasi satu sama lain dan umumnya akan disusun dalam kategori yang terpisah secara linear (Zhai & Aggarwal, 2012).

2.5.2. Naïve Bayes Classifier

Naïve Bayes Classifier merupakan pengklasifikasi probabilistik sederhana yang didasarkan pada teorema Bayes, yang menyatakan bahwa kemungkinan

terjadinya suatu peristiwa sama dengan probabilitas intrinsik (dihitung dari data yang tersedia sekarang) dikalikan probabilitas bahwa hal serupa akan terjadi lagi di masa depan (berdasarkan pengetahuan yang terjadinya di masa lalu) (Miner et al, 2012). Pengkalsifikasian naïve bayes memiliki asumsi bahwa efek dari suatu nilai atribut tertentu tidak bergantung (independent) terhadat nilai atribut lainnya (Zhai & Aggarwal, 2012). Asumsi tersebut disebut class conditional independence. Hal tersebut dilakukan untuk menyederhanakan perhitungan dan karenanya asumsi tersebut dianggap “naif”. Meskipun dengan desain sederhana dan asumsi naif (yang hampir tidak pernah terjadi di dunia nyata), pengklasifikasi ini dapat sangat efisien dan akurat, terutama ketika jumlah variabel yang tinggi.

Contoh sederhana dalam perhitungan *naïve bayes* misalnya terlihat pada Gambar 2.8 terdapat dua kumpulan data yaitu hijau dan merah (Statsoft, 2015). Data baru akan ditambahkan dan akan ditentukan data baru tersebut merupakan bagian dari kelas yang mana. Karena jumlah data hijau dua kali lebih banyak daripada merah, maka diasumsikan bahwa data yang baru memiliki probabilitas menjadi anggota hijau dua kali lebih besar dari merah. Dalam analisis Bayesian, keyakinan ini dikenal sebagai probabilitas prior. Probabilitas prior didasarkan pada pengalaman sebelumnya, dalam hal ini persentase data hijau dan merah.



Gambar 2.8 Dua kelompok data *Naïve Bayes*

(Sumber: Statsoft, 2015)

$$\text{Probabilitas prior untuk hijau} = \frac{\text{Jumlah data hijau}}{\text{Jumlah keseluruhan data}}$$

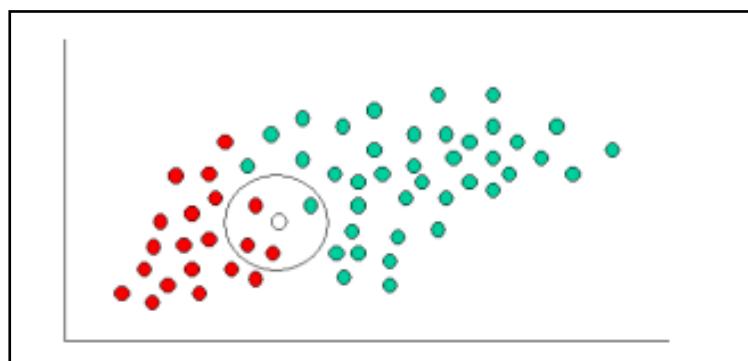
$$\text{Probabilitas prior untuk merah} = \frac{\text{Jumlah data merah}}{\text{Jumlah keseluruhan data}}$$

Dengan asumsi jumlah keseluruhan data yaitu 60, dengan 20 merah dan 40 hijau, maka probabilitas prior untuk keanggotaan kelas yaitu:

$$\text{Probabilitas prior untuk hijau} = \frac{40}{60}$$

$$\text{Probabilitas prior untuk merah} = \frac{20}{60}$$

Setelah menentukan probabilitas prior, objek baru (lingkaran putih) akan ditentukan keanggotaanya seperti terlihat pada Gambar 2.9. Diasumsikan bahwa semakin banyak suatu kelompok data tertentu (hijau atau merah) di sekitar objek baru tersebut, maka kemungkinan objek baru mempunyai keanggotaan sesuai kelompok data tersebut semakin besar.



Gambar 2.9 Penambahan objek baru pada *Naïve Bayes*

(Sumber: Statsoft, 2015)

Untuk mengukur kemungkinan tersebut, digambarkan lingkaran di sekitar objek baru X (putih), kemudian jumlah poin dalam lingkaran milik masing-masing label kelas akan dihitung. Dari sini kita didapatkan:

$$\text{Kemungkinan objek baru adalah hijau} = \frac{\text{Jumlah hijau di sekitar } X}{\text{Jumlah data hijau}}$$

$$\text{Kemungkinan objek baru adalah merah} = \frac{\text{Jumlah merah di sekitar } X}{\text{Jumlah data merah}}$$

Sehingga,

$$\text{Kemungkinan objek baru adalah hijau} = \frac{1}{40}$$

$$\text{Kemungkinan objek baru adalah merah} = \frac{3}{20}$$

Meskipun probabilitas prior sebelumnya menunjukkan bahwa X mungkin merupakan anggota hijau (dimana jumlah hijau dua kali lebih banyak dibandingkan dengan merah) kemungkinan setelahnya menunjukkan hal yang

sebaliknya; bahwa keanggotaan kelas X adalah merah (dimana terdapat lebih banyak merah di sekitar X daripada hijau). Dalam analisis Bayesian, klasifikasi akhir yang dihasilkan adalah penggabungan kedua sumber informasi tersebut.

$$\text{Probabilitas posterior hijau} = \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$

$$\text{Probabilitas posterior merah} = \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Sehingga, dapat disimpulkan bahwa objek baru tersebut merupakan bagian dari kelompok data merah karena memiliki probabilitas posterior merah yang lebih besar.

2.5.3. Evaluasi Model Pengklasifikasi

Model klasifikasi yang dibangun perlu dievaluasi untuk mengetahui seberapa bagus model tersebut dalam melakukan klasifikasi yang diinginkan. Dalam mengevaluasi kinerja pengklasifikasi khususnya klasifikasi teks umumnya dilakukan dengan *accuracy* atau dengan *precision and recall* (Miner, et al, 2012). Nilai *accuracy* merepresentasikan seberapa banyak keseluruhan dokumen diklasifikasikan dengan benar. Semakin tinggi nilai *accuracy* yang dihasilkan maka semakin bagus dan akurat model tersebut dalam melakukan klasifikasi. Persamaan untuk mendapatkan nilai *accuracy* adalah sebagai berikut:

$$\textbf{Accuracy} = \frac{\text{Total dokumen yang diklasifikasikan dengan benar}}{\text{Total dokumen}} \quad (2.3)$$

Pada kasus ketidakseimbangan kelas/kategori pada data latihan dimana terdapat kelas data mayoritas dan minoritas, seringkali nilai *accuracy* kurang bisa merepresentasikan performa model secara signifikan (Jiawei, Kamber, & Pei, 2012). Misalnya pada kasus klasifikasi deteksi kanker, jumlah data kasus yang dideteksi kanker jauh lebih sedikit atau jarang daripada jumlah kasus yang dideteksi bukan kanker. Nilai *accuracy* memberikan nilai 97%, dengan nilai ini model klasifikasi bisa dikatakan sudah sangat akurat. Namun, bisa saja 97% tersebut hanya mendeteksi kelas bukan kanker secara benar, dan 3% sisanya salah mendeteksi seluruh kelas kanker. Untuk mengatasinya, pengukuran *precision and*

recall biasa dilakukan dalam mengevaluasi model klasifikasi. Selain mampu menunjukkan keakuratan model secara keseluruhan, pengukuran ini juga mampu menunjukkan bagaimana performa model pada setiap kelas.

Pengukuran *precision* dan *recall* merupakan metrik evaluasi yang paling sering digunakan pada kasus klasifikasi teks (Sokolova & Lapalme, 2009). Misalnya terdapat dua kelas A dan B, *precision* yaitu jumlah sampel berkategori A yang ditebak dengan benar sebagai A dibanding dengan jumlah total data yang ditebak sebagai A, sedangkan *recall* yaitu jumlah sampel berkategori A yang ditebak dengan benar dibandingkan dengan jumlah total sampel A. Dalam melakukan pengukuran ini biasanya dibangun *confusion matrix* yang merupakan sebuah tabel yang terdiri atas banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi (Gambar 2.10).

Data class	Classified as <i>pos</i>	Classified as <i>neg</i>
<i>pos</i>	true positive (<i>tp</i>)	false negative (<i>fn</i>)
<i>neg</i>	false positive (<i>fp</i>)	true negative (<i>tn</i>)

Gambar 2.10 Confusion Matrix

(Sumber: Sokolova & Lapalme, 2009)

Precision dan *recall* dapat merepresentasikan nilai keakuratan model pada setiap kelas. Sedangkan untuk mengetahui akurasi secara keseluruhan digunakan pengukuran *FI* yang merupakan pengukuran tunggal dari kombinasi *precision* dan *recall*. Persamaan *precision*, *recall*, dan *F1* adalah sebagai berikut:

$$\textbf{\textit{Precision}} = \frac{tp}{tp+fp} \quad (2.4)$$

$$\textbf{\textit{Recall}} = \frac{tp}{tp+fn} \quad (2.5)$$

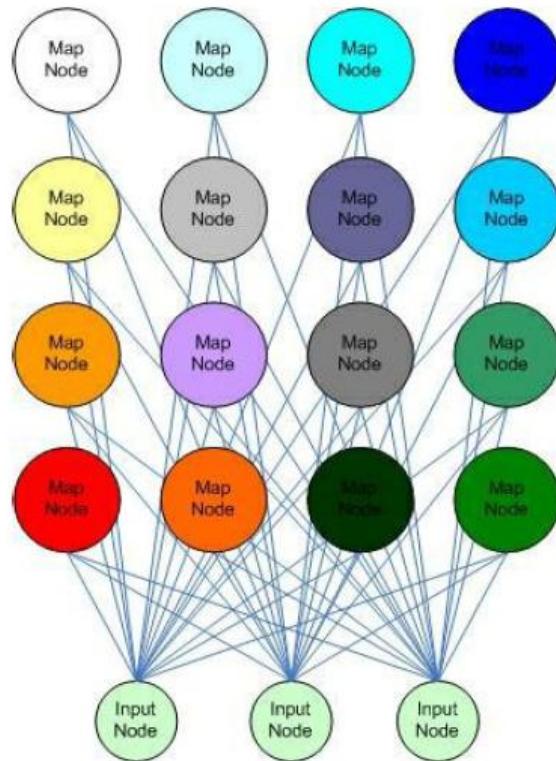
$$\textbf{\textit{F1}} = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \quad (2.6)$$

2.6. Algoritma Penggolong *Clustering*

Berbeda dengan algoritma klasifikasi, proses *clustering* menggunakan skema *unsupervised learning* dimana pengelompokan data akan dilakukan tanpa model latihan (*learning model*). Pada konten *text mining*, dokumen-dokumen akan dikelompokkan dalam berbagai klaster berdasarkan konten isi dari dokumen (Suh, Park, & Jeon, 2010). Jenis metode *clustering* yang paling umum digunakan yaitu *partitioning-based* (partisi), *hierarchical-based* (hirarki), dan *kohonen neural network* atau yang sering disebut sebagai *self-organizing map* (SOM). Pada penelitian kali ini, akan digunakan dua jenis algoritma *clustering* yaitu *self-organizing map* (SOM) dan *k-means clustering* yang termasuk dalam pendekatan *partitioning-based* (partisi).

2.6.1. *Self-Organizing Map* (SOM)

Kohonen Self-Organizing Maps atau *Self-Organizing Maps* (SOM) adalah salah satu jenis model *neural network*. SOM dikembangkan pada tahun 1982 oleh professor Teuvo Kohonen. Dinamakan "Self-Organizing" karena tidak memerlukan pengawasan (*unsupervised learning*) dan disebut "Maps" karena SOM berusaha untuk memetakan bobotnya agar sesuai dengan input data yang diberikan. SOM memungkinkan visualisasi dan proyeksi dari data berdimensi tinggi ke dimensi yang lebih rendah, paling sering menjadi bidang 2-D dengan tetap mempertahankan topologi data tersebut (Feldman & Sanger, 2007). Data yang berdekatan di ruang vektor berdimensi tinggi, saat dipetakan ke ruang vektor 2-D akan terletak pada lokasi yang berdekatan pula.



Gambar 2.11 Ilustrasi jaringan SOM 4x4

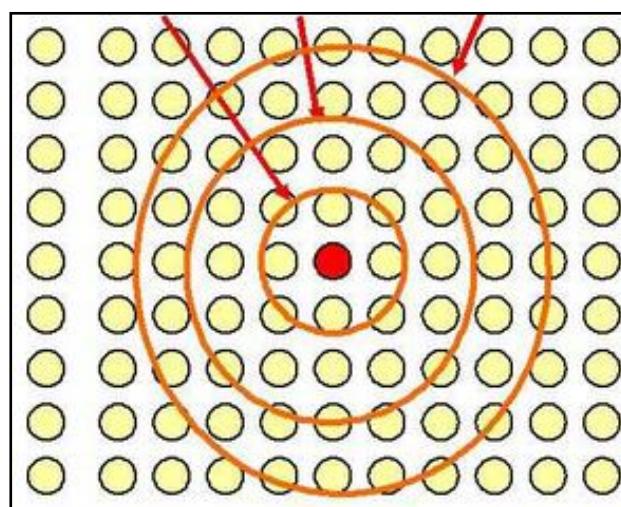
(Sumber: Guthikonda, 2005)

Pada Gambar 2.11 diilustrasikan jaringan SOM 4x4. Dari ilustrasi sederhana tersebut terdapat beberapa hal yang bisa diperhatikan yaitu setiap *map node* terhubung ke setiap *input node*, sehingga untuk jaringan node 4x4 ini terdapat $4 \times 4 \times 3 = 48$ koneksi dan setiap *map node* peta tidak terhubung satu sama lain. Node diatur dengan cara ini, sebagai bidang 2-D untuk memudahkan visualisasi hasil. Dalam konfigurasi ini, setiap *map node* memiliki koordinat (i, j) yang unik. Hal ini memudahkan untuk melihat *node* secara jelas dan menghitung jarak antara *node*. Karena koneksi hanya terjadi dengan *input node*, maka nilai *map node* tidak terpengaruh oleh *map node* lainnya. Hal ini memungkinkan sebuah *map node* hanya akan memperbarui bobotnya berdasarkan vektor input saja.

Tahapan algoritma SOM adalah sebagai berikut (Guthikonda, 2005):

1. Menginisialisasi bobot setiap node.

2. Sebuah vektor input dipilih secara acak dari data latihan dan dimasukkan ke dalam jaringan.
3. Setiap node dalam jaringan akan diperiksa bobotnya, node yang memiliki bobot paling sesuai dengan vektor input akan dipilih sebagai *best matching unit* (BMU).
4. Radius lingkup BMU akan dihitung dan biasanya dimulai dengan lingkup radius yang besar (sebesar jaringan) dan akan berkurang pada setiap kali proses dilakukan kembali seperti ditunjukkan pada Gambar 2.12.
5. Setiap node yang berada dalam radius BMU akan menyesuaikan dirinya menjadi seperti vektor input. Semakin dekat dengan BMU maka bobot akan semakin berubah.
6. Mengulangi langkah 2 hingga N iterasi



Gambar 2.12 Radius Best Matching Unit (BMU)

(Sumber: Guthikonda, 2005)

2.6.2. *K-means Clustering*

K-Means merupakan salah satu metode data *clustering* yang berusaha mempartisi N jumlah data ke dalam K jumlah kelompok/klaster. *K-means* melakukan partisi data ke dalam kelompok/klaster sehingga data yang memiliki karakteristik yang sama akan dikelompokkan ke dalam satu klaster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain (Jiawei, Kamber, & Pei, 2012). Tujuan dari metode *clustering*

ini yaitu untuk meminimalisasikan *objective function* yang diset dalam proses *clustering*, dimana *objective function* tersebut pada umumnya berusaha meminimalisasikan variasi di dalam suatu klaster dan memaksimalisasikan variasi antar klaster (Agusta, 2007). Dalam melakukan pengelompokan menggunakan *k-means*, langkah-langkah yang harus dilakukan adalah sebagai berikut (Jiawei, Kamber, & Pei, 2012):

1. Menentukan k jumlah klaster/kelompok.
2. Mengalokasikan data ke dalam kelompok secara acak.
3. Menghitung pusat klaster (*centroid*) dari data yang ada pada masing-masing klaster. Untuk menghitung *centroid* cluster ke- i , v_i , digunakan persamaan sebagai berikut:

$$v_i = \frac{\sum_{k=1}^{N_i} x_{kj}}{N_i} \quad (2.7)$$

N_i : Jumlah data yang menjadi anggota klaster ke- i

4. Mengalokasikan kembali masing-masing data ke *centroid* terdekat. Perhitungan jarak data dengan *centroid* dilakukan dengan menghitung jarak dua titik (*euclidean distance*) berikut:

$$D(x_2, x_1) = \sqrt{\sum_{j=1}^p |x_{2j} - x_{1j}|^2} \quad (2.8)$$

P : Dimensi data

Kemudian dalam pengalokasian data dirumuskan sebagai berikut:

$$a_{ik} = \begin{cases} 1 & d = \min\{D(x_k, v_i)\} \\ 0 & \text{lainnya} \end{cases} \quad (2.9)$$

a : Keanggotaan data ke- k ke klaster ke- i

v_i : Nilai *centroid* cluster ke- i

5. Kembali menghitung pusat klaster (*centroid*) seperti langkah 3 dan seterusnya secara berulang-ulang hingga tidak ada lagi perubahan atau anggota klaster yang berpindah dimana fungsi objektif F sudah memiliki nilai yang optimal.

$$F = \sum_{k=1}^N \sum_{i=1}^c a_{ik} D(x_k, v_i)^2 \quad (2.10)$$

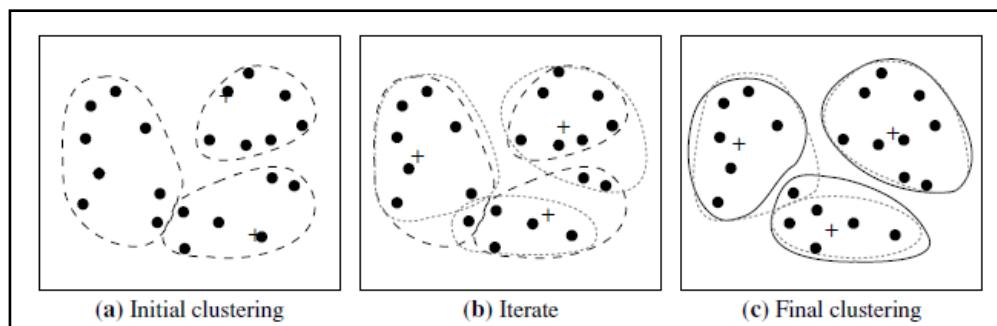
N : Jumlah data

c : Jumlah klaster

a_{ik} : Keanggotaan data ke- k ke klaster ke- i

v_i : Nilai *centroid* klaster ke- i

Pada dasarnya kegiatan pengelompokan *k-means* berlangsung dengan melakukan iterasi-iterasi untuk mendapatkan pengelompokan paling optimal seperti diilustrasikan pada Gambar 2.13.



Gambar 2.13 Ilustrasi proses clustering dengan *k-means*

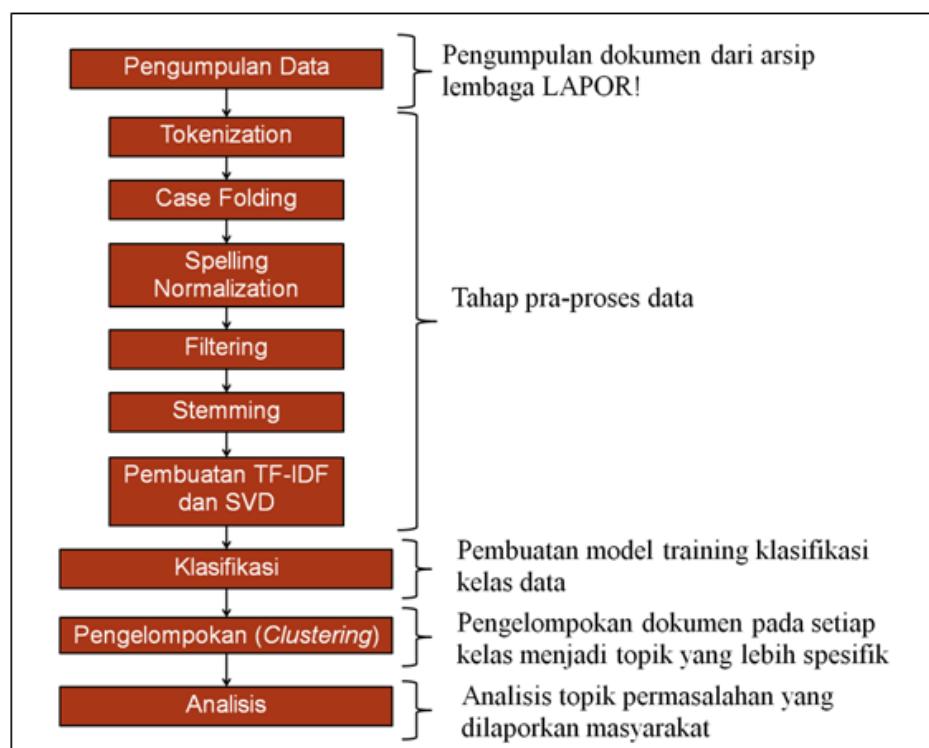
(Sumber: Jiawei, Kamber, & Pei, 2012)

Pemilihan jumlah *seeds* (jumlah kelompok) awal sangat mempengaruhi kualitas pengelompokan *k-means*. Oleh karena itu, sejumlah teknik dapat digunakan untuk memilih *seeds* awal yang untuk meningkatkan kualitas hasil. Misalnya, dengan melakukan metode pengelompokan ringan seperti teknik pengelompokan *agglomerative* untuk mengetahui jumlah *seeds* awal yang bisa dipilih (Zhai & Aggarwal, 2012).

BAB 3

PENGUMPULAN DAN PENGOLAHAN DATA

Bab ini mencakup pengumpulan dan pengolahan data yang dilakukan dalam penelitian. Laporan atau pesan yang disampaikan masyarakat melalui LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat) akan menjadi data utama yang digunakan dalam penelitian ini. LAPOR! menerima laporan tersebut dari beberapa kanal seperti sms, website, aplikasi seluler, dan media sosial. Data teks laporan yang bersifat tidak terstruktur tersebut akan diolah dengan menggunakan text mining. Secara umum, alur pengumpulan dan pengolahan data pada penelitian ini ditunjukkan pada Gambar 3.1.



Gambar 3.1 Alur Pengumpulan dan Pengolahan Data

3.1. Pengumpulan Data

Data teks laporan yang digunakan adalah data historis laporan yang dilaporkan masyarakat melalui LAPOR! pada periode Oktober 2014-Maret 2015. Laporan yang diterima adalah laporan yang berupa aspirasi, keluhan, pengaduan,

dan permintaan informasi dari masyarakat kepada pemerintah atau berbagai lembaga. Data historis laporan pada sistem LAPOR! dibagi menjadi dua jenis yaitu laporan yang disetujui (ditindaklanjuti) dan laporan yang diarsipkan. Laporan yang telah disetujui dan ditindaklanjuti sudah disunting dan ditambahkan label kategori sesuai dengan maksud dan tujuan laporan tersebut, sedangkan laporan yang diarsipkan merupakan laporan yang tidak ditindaklanjuti dan masih berbentuk teks asli yang belum disunting. Pada penelitian ini, kedua jenis data tersebut akan digunakan.

LAPOR! menerima laporan dari berbagai media yaitu sms, website, aplikasi seluler, dan sosial media. Jumlah data yang diterima pada Oktober 2014-Maret 2015 adalah 48819 laporan dimana 16143 laporan merupakan laporan yang disetujui dan 32676 merupakan laporan yang diarsipkan. Dari total jumlah data yang digunakan tersebut, data yang termasuk spam akan dieliminasi. Data yang diperoleh dari LAPOR! sudah disajikan dalam bentuk *spreadsheet* dan telah mengalami proses penyaringan dimana laporan atau pesan yang mengandung kata-kata ancaman, caci maki, SARA, dan pornografi tidak diikutsertakan atau dianggap sebagai pesan spam.

<u>id</u>	<u>tanggal</u>	<u>topik_laporan</u>	<u>tags</u>	<u>judul_laporan</u>	<u>isi_laporan</u>
1274870	10/31/2014 16:36	Kesehatan	dilarang mer	Sopir Angkot N selamat sore, seorang supir angkot kalapa-caheum berplat D 1981 BF merokok di dalam angkot dan itu	
1274868	10/31/2014 16:31	Bidang Politik, berandalan	Tempat Beran	Lapor Pak Wali, setiap sore anak-anak berandalan dengan buka baju nongkrong di warung Gamelan an	
1274815	10/31/2014 11:26	Infrastruktur	pembanguna	Permasalahan Yth. Kementerian Pembangunan Daerah Tertinggal,Kepada pemerintah daerah & pusat mohon kiranya	
1274819	10/31/2014 0:00	Topik Lainnya	cpns	Permasalahan Yth. Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi,Saya Fotuho dari pulau ni	
1273281	10/22/2014 18:42	Infrastruktur		Tarif Tiket Keru Publik trutama, "rakyat biasa" suka naik kreta api atau bis umun trutama utk kluar kota, TAPI kalo tarip	
1272985	10/21/2014 10:26	Infrastruktur		Perjalanan KRL Perjalanan KRL 1781 & 1785 dibatalkan, tidak ada pemberitahuan, sangat merugikan karena kereta tan	
1272149	10/16/2014 23:28	Infrastruktur		Perumahan se Kepada PT kereta api dan Pemda depok kok bisa bangunan berdiri dan membangun bangunan rumah y	
1274079	10/27/2014 19:46	Infrastruktur		Penerangan Di Ada Saran nih Buat Commuter Line Soal Papan Nama Petunjuk Arah Di Stasiun, Khususnya Stasiun Raw	
1274081	10/27/2014 19:48	Infrastruktur		Penerangan Di Lampu Penerangan Di Stasiun Rawu buaya Tidak Menyal, Dari Laporan Satpam Stasiun Terjadi Pelece	
1274878	10/31/2014 17:05	Energi dan Sumber aliran listrik	Pemandaman Li	Yth. PT. PLN,Terjadi pemandaman Listrik di Palembang sejak awal oktober dan yang paling parah itu 3 ha	
1271044	10/14/2014 0:16	Infrastruktur		Komersialisasi Pedagang dibawah stasiun jalan sepanjang Taman sari ,diperbolehkan Oleh PT KAI , konon menurut sa	
1271043	10/14/2014 0:10	Infrastruktur		Penertiban dai PT KAI dan IGNATIUS JONAN, jelas telah melakukan persaingan usaha tidak sehat , memonopoli ruang	
1271042	10/14/2014 0:05	Infrastruktur		Pengamanan A Ignatius Jonan Direktur PT KAI telah berlaku diskriminasi , surau(langgar di jalan juanda 1c (intidiyah)	
1270790	10/12/2014 23:18	Infrastruktur		Harga Tiket Ke PT KAI..mhn harga tiket kereta api jarak jauh dan menengah ,klau sy bilang ini ganti harga bukan naik	
1270530	10/11/2014 0:11	Infrastruktur		Sikap Layanan Hari ini, 10 Okt 2014 jam 23.35 saya menghubungi PT.KAI lewat telp. 021-121 untuk menanyakan Harga	
1263206	9/17/2014 17:10	Infrastruktur		Eskalator di St; Selamat siang, Eskalator di Stasiun Gondangdia sudah lama tidak berfungsi, mohon segera diperbaiki.	
1261977	9/13/2014 5:15	Infrastruktur		Kereta antar k Lapor ini kereta antar kota pelayanannya kurang banget. Seringkali telat sampe sejam dua jam, tolong	
1261754	9/12/2014 10:59	Infrastruktur		Menunggu Kel Ingguan atau bagaimana. Terima kasih!npir 45 merit menunggu bahkan lebih, dan penumpang memb	
1274877	10/31/2014 16:52	Kesehatan	bpjs kesehat	Permasalahan Yth. BPJS Kesehatan,Mohon dapat dicarikan solusi dan kemudahan bagi warga masyarakat yang ingin r	
1261355	9/11/2014 10:26	Infrastruktur		Petugas Stasius Yth PT KAI,Ingin melaporkan petugas a.n. Pandu di Stasiun Kediri tanggal 11 September 2014 pukul 09,	
1251041	8/11/2014 23:41	Infrastruktur		Jalur Listrik at&t jalur listrik atas di jalur KA daop 6 untuk selanjutnya di gunakan untuk KRL demik kesejahteraan bersa	
1274086	10/27/2014 20:09	Infrastruktur		Arsitektur Stas melihatnya, terima kasihMohon kepada Yth. Manajemen PT KAI, supaya arsitektur stasiun kereta api o	
1274833	10/31/2014 13:19	Pengertian Ketenaga kerja	Permasalahan	Yth. Kementerian Pariwisata dan Ekonomi KreatifSaya melihat banyaknya orang yang putus sekolah at	
1274683	10/30/2014 18:11	Kesehatan	kawasan lara	Permohonan P Kepada Yth. Pemerintah Provinsi DKI Jakarta. Saya ingin melaporkan mengenai kawasan dilarang mer	

Gambar 3.2 Spreadsheet Laporan yang Disetujui

nid	tanggal	isiLaporan	tanggal_hapus	kategoriHapus	
1272671	10/20/2014 6:10	JAMKESMAS sangat menolong rakyat miskin untuk berobat,mohon tidak hapus se	10/20/2014 10:16	Input merupakan saran, masukan umum	
1272672	10/20/2014 6:14	SELAMAT UNTUK PEMERINTAHAN BARU program kesehatan dan pangan masyarakat	10/21/2014 10:39	Input merupakan saran, masukan umum	
1272673	10/20/2014 6:39	Asalamualaikum Wb wb pak mohon ya pak tetap bantu kami guru kami guru selalu	10/20/2014 10:15	Input merupakan saran, masukan umum	
1272682	10/20/2014 8:04	Hari ini, Pak SBY melepas tugasnya sbg RI 1, kami dg rasa hormt & sedih menganta	10/20/2014 10:15	Input merupakan saran, masukan umum	
1272683	10/20/2014 8:11	tim lpr yth semua program pemmerintah itu bagus mengatasnamakan rakyat api.j	10/21/2014 14:14	Input merupakan saran, masukan umum	
1272685	10/20/2014 8:18	Semoga slogan Jokowi-JK yi Indonesia Hebat bs jd kenyata an, tidak apert slogan Be	10/21/2014 10:40	Input merupakan saran, masukan umum	
1272688	10/20/2014 8:51	Pak wagub.mhn di tindak terkait di laporkanya kasudin dan kasi pemberitaan ke l	11/11/2014 10:16	Input merupakan saran, masukan umum	
1272697	10/20/2014 9:20	Slmt pagi. YTH BPK Dr. SBY Presiden RI. Info. Kita bersyukur dan Banyak selamat ke	10/20/2014 10:07	Input merupakan saran, masukan umum	
1272700	10/20/2014 9:36	SLMT ATAS PLANTIKAN PRESIDEN KE 7 INDONESIA SMG KDPN DPT MMBUAT NEGRI	10/20/2014 10:05	Input merupakan saran, masukan umum	
1272703	10/20/2014 9:38	SELAMAT BERSAHABAT ERAT PARA PEJABAT DAN APARATUR NEGARA SERTA MASY	10/20/2014 10:00	Input merupakan saran, masukan umum	
1272732	10/20/2014 11:25	Yth.Bpk.Susilo Bambang Yudhoyono, kami sangat bangga bapak sangat luar biasa r	10/21/2014 11:59	Input merupakan saran, masukan umum	
1272733	10/20/2014 11:31	YTH! BUMN/BUM D/BU.SWASTA perlu diminta utk membudayakan pemberian so	10/21/2014 11:58	Input merupakan saran, masukan umum	
1272735	10/20/2014 11:33	Terimakasih sudah memimpin negara Indonesia	10/21/2014 11:58	Input merupakan saran, masukan umum	
1272736	10/20/2014 11:35	Aslm yang mulia bpk SBY, terma kasih atas pengadian bpk selama 10 thn memimp	10/21/2014 11:58	Input merupakan saran, masukan umum	
1272738	10/20/2014 11:43	Bpk SBY dan Bpk Budiono aku mengucapkan Terima Kasih telah memimpin RI jadi	10/21/2014 11:57	Input merupakan saran, masukan umum	
1272739	10/20/2014 11:44	Bpk SBY dan Bpk Budiono aku mengucapkan Terima Kasih telah memimpin RI jadi	10/21/2014 15:53	Input merupakan saran, masukan umum	
1272742	10/20/2014 11:59	Selama SBY masih hidup. Saya akan tetap memanggil bapak Presiden SBY.	10/21/2014 11:56	Input merupakan saran, masukan umum	
1272744	10/20/2014 12:11	Kami dari sungai jernihbaso kab agam sumbar mengucap kan selamat kepada yth	10/21/2014 15:53	Input merupakan saran, masukan umum	
1272750	10/20/2014 12:50	TRIMS ATAS KEPEMIMPINAN BPK SBY YG TELAH BANYAK MEMBERIKAN KESEJAHTER	10/21/2014 11:54	Input merupakan saran, masukan umum	
1272751	10/20/2014 12:51	JANGAN LUPA JANJI PENDIDIKAN GRATIS KESEHATAN GRATIS BLT 1 JUTA/KK JANJI	10/21/2014 15:57	Input merupakan saran, masukan umum	
1272758	10/20/2014 13:17	Pak Sby ku terima kasih buat kerja keras, perjuangan, pengabdian di 10 tahun tera	10/21/2014 11:52	Input merupakan saran, masukan umum	
1272759	10/20/2014 13:18	Mohon maaf jika selama ini kami sering mengkritik bpk. Kami menghargai kerend	10/21/2014 11:51	Input merupakan saran, masukan umum	
1272760	10/20/2014 13:30	at melalui pengamanan pelantikan pak jokowi, pak jk dan juga pesta rakyat dimor	10/21/2014 17:00	Input merupakan saran, masukan umum	
1272765	10/20/2014 13:50	KOK BELUM ADA PIDATO DEMENSIONER KABINET IP JILID 2.....UTK MENYEMPURNA	10/21/2014 11:50	Input merupakan saran, masukan umum	

Gambar 3.3 Spreadsheet Laporan yang Diarsipkan

3.2. Pra-proses Teks

Pra-proses teks merupakan tahap yang dilakukan untuk mentransformasi data tekstual yang bersifat tidak terstruktur menjadi model yang terstruktur. Model yang terstruktur diperlukan agar data bisa diolah dan dianalisis dengan menggunakan teks mining. Pra-proses teks terdiri dari berberapa tahap, dimana setiap tahap dapat dilakukan secara manual. Akan tetapi, dikarenakan data laporan mempunyai jumlah yang sangat besar dan memerlukan waktu yang cukup banyak untuk melakukan seluruh tahap pra-proses secara manual, maka sebuah aplikasi khusus dikembangkan untuk melakukan automasi tahapan pra-proses. Aplikasi tersebut dikembangkan dengan menggunakan bahasa pemrograman C++ dimana naskah didalamnya khusus dimodifikasi sesuai dengan kebutuhan penelitian. Dengan menggunakan aplikasi tersebut, tahapan pra-proses dapat dilakukan secara lebih singkat sesuai dengan keinginan dan kebutuhan penelitian.

Terdapat beberapa tahap proses yang dilakukan dalam pra-proses teks. Secara umum, tahap pra-proses teks dibagi menjadi dua bagian utama yaitu proses tokenization-case folding-spelling normalization-filtering dan proses stemming. Pada *interface* aplikasi yang digunakan, kedua tahap tersebut dijalankan sesuai

Tabel 3.1 Kumpulan Data Laporan

Isi Laporan
Yth. Kementerian Dalam Negeri, Saya mau tanya kalau untuk membuat Kartu Keluarga baru, KTP baru dan Akte kelahiran biayanya berapa ya? Mohon informasinya, terima kasih.
<p>Yth. Kepolisian RI,</p> <p>Mengapa Polantas di setiap melakukan penilangan tidak pernah menanyakan kepada pengendara apakah ingin mengikuti sidang karena tidak mengakui kesalahan atau membayar langsung tilang ke bank karena mengakui kesalahan?</p> <p>Selama 20 tahun saya berkendara, saya pernah ditilang 2x, namun tidak pernah diberi lembar biru untuk membayar langsung ke bank. Saya selalu di beri lembar merah untuk ikut sidang padahal saya sudah mengaku bersalah.</p> <p>Sebagai tambahan bukti materi, coba lihat di youtube siaran polisi yang menilang, Polantas selalu memberikan surat tilang merah.</p> <p>Mohon penjelasannya secara lengkap bagaimana sebenarnya aturan penggunaan slip merah dan slip biru ini ketika penilangan?</p> <p>Terima kasih.</p>
<p>Lapor!... saya penerima KKS, ingin bertanya kapan dana bantuan nya akan keluar, dan berapa jumlah dananya, berapa kali pengambilannya dalam satu tahun...</p> <p>lampa2 PJU dari Jln.Margacinta sampe Derwati perlu diganti,selain banyak mati juga kurang terang.</p> <p>359mn22552001 Rumah tangga saya tidak menerima bantuan sekolah untuk anak saya</p>

3.2.1. Tokenization, Case Folding, Spelling Normalization, dan Filtering

Proses yang paling awal dilakukan yaitu tokenization. Pada prinsipnya, tokenization adalah proses pemisahan teks menjadi potongan kata yang disebut *token*. Tokenization dilakukan untuk mendapatkan token atau potongan kata yang akan menjadi entitas yang memiliki nilai dalam penyusunan matriks dokumen pada proses selanjutnya. Langkah transformasi proses tokenization ditunjukkan pada Gambar 3.6.

Penerima dana BLSM di Desa Sumendi Kec Tongas Kab Probolinggo Jatim. Tdk ada tambhan, padahal lewat nmr ini tlh saya kirimkan data orang-orang miskin yg belum dapat dana BLSM



Penerima dana BLSM di Sumendi Kec Tongas Kab Probolinggo Jatim.	Tdk ada tambhan, padahal lewat nrmr ini tlah saya kirimkan	data orang-orang miskin yg belum dapat dana BLSM
--	---	---

Gambar 3.6 Tokenization

Setelah melalui proses tokenization, selanjutnya dilakukan case folding. Case folding merupakan proses pengubahan huruf dalam dokumen menjadi satu bentuk, misalnya huruf kapital menjadi huruf kecil dan sebaliknya. Perubahan yang terjadi dalam proses case folding ditunjukkan pada Gambar 3.7.

Penerima dana **BLSM** di Desa **Sumendi** **Kec** Tongas **Kab** **Probolinggo** **Jatim**. Tdk ada tambhan, padahal lewat nmr ini tlh saya kirimkan data orang-orang miskin yg belum dapat dana **BLSM**



penerima dana blsm di sumendi kec tongas kab probolinggo jatim.	tdk ada tambhan, padahal lewat nmor ini tlah saya kirimkan	data orang-orang miskin yg belum dapat dana blsm
--	---	---

Gambar 3.7 Case Folding

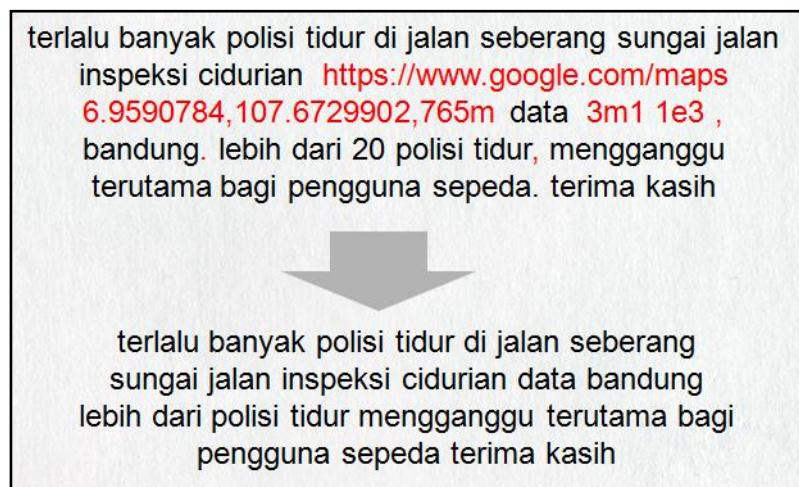
Setelah seluruh kata dibubah menjadi satu bentuk, selanjutnya dilakukan spelling normalization. Proses ini merupakan proses perbaikan atau substitusi kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Substitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya mempunyai maksud dan arti yang sama tetapi akan dianggap sebagai entitas yang berbeda proses penyusunan matriks. Proses ini dilakukan dengan menyusun kamus kata dan singkatan pada aplikasi C++ yang digunakan. Contoh cuplikan daftar kata dan singkatan yang disusun ditunjukkan pada Tabel 3.2. Pada Gambar 3.8 ditunjukkan contoh masukan dan luaran yang dihasilkan dari proses ini.

Tabel 3.2 Cuplikan Daftar Kata dan Singkatan

ad~ada	blum~belum	mnjadi~menjadi
adany~adanya	blz~balas	mnjd~menjadi
adek~adik	bnar~benar	mnjdi~menjadi
adlh~adalah	bndg~bandung	mnjdi~menjadikan
aer~air	bner~benar	mnrima~menerima
aj~saja	bneran~benar	mnrm~menerima
aja~saja	bnget~sangat	mnurma~menerima
ajah~saja	bngks~bungkus	mnrt~menurut
aje~saja	bngng~bingung	mnta~minta
ajh~saja	bngt~sangat	mnujukan~menunjukkan
ak~aku	bngun~bangun	mnum~minum
akn~akan	bngunan~bangunan	mo~mau
alesan~alasan	bnjr~banjir	moga~semoga
almt~alamat	bnr~benar	montor~motor

**Gambar 3.8 Spelling Normalization**

Proses berikutnya yaitu filtering, dimana kata dan tanda baca yang tidak memiliki arti yang signifikan atau termasuk *noise* (pengganggu) akan dieliminasi. Kata atau frase yang tidak bermakna secara signifikan, misalnya hashtag (#), url, tanda baca tertentu (*emoticon*), dan lainnya. Laporan banyak diterima lewat sms, sehingga menyebabkan banyaknya tanda baca dan frase yang masuk pada penarikan laporan pada sistem LAPOR! yang tidak bisa diproses atau mengurangi performa pengolahan data pada tahap selanjutnya. Oleh karena itu, kata atau frase tersebut perlu dieliminasi. Contoh eliminasi yang dilakukan ditunjukkan pada Gambar 3.9.

**Gambar 3.9 Filtering**

Hasil akhir dari proses ini ditunjukkan pada Tabel 3.3 yang merupakan transformasi dari Tabel 3.1. Setelah melewati keempat tahap pada bagian awal pra-proses tersebut, data sudah bisa dikatakan bersih dan siap olah. Bagian pra-proses selanjutnya (stemming) sebenarnya merupakan pilihan, akan tetapi pada penelitian ini proses stemming dilakukan untuk melihat apakah proses stemming mempunyai pengaruh yang berarti pada performa hasil penelitian.

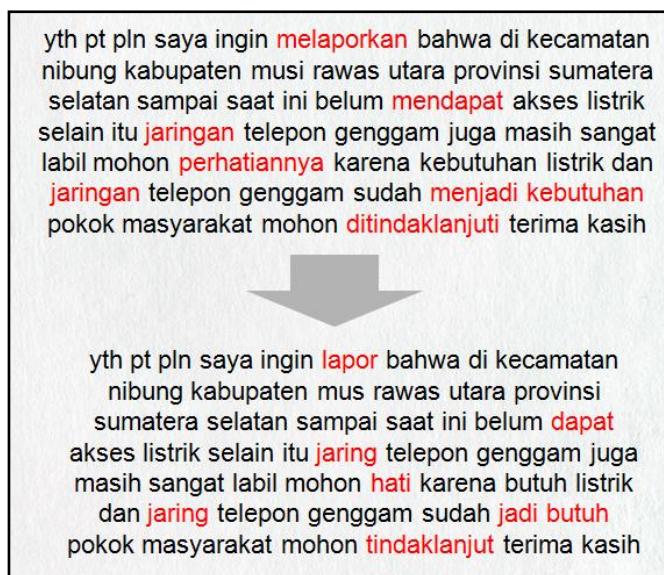
Tabel 3.3 Kumpulan Data Laporan Setelah Melewati Bagian Awal Pra-proses

Isi Laporan
yth kementerian dalam negeri saya mau tanya kalau untuk membuat kartu keluarga baru kartu tanda penduduk baru dan akte kelahiran biayanya berapa ya mohon informasinya terima kasih
yth kepolisian ri mengapa polantas di setiap melakukan penilangan tidak pernah menanyakan kepada pengendara apakah ingin mengikuti sidang karena tidak mengakui kesalahan atau membayar langsung tilang ke bank karena mengakui kesalahan selama tahun saya berkendara saya pernah ditilang x namun tidak pernah diberi lembar biru untuk membayar langsung ke bank saya selalu di beri lembar merah untuk ikut sidang padahal saya sudah mengaku bersalah sebagai tambahan bukti materi coba lihat di youtube siaran polisi yang menilang polantas selalu memberikan surat tilang merah mohon penjelasannya secara lengkap bagaimana sebenarnya aturan penggunaan slip merah dan slip biru ini ketika penilangan terima kasih
lapor saya penerima kks ingin bertanya kapan dana bantuan nya akan keluar dan berapa jumlah dananya berapa kali pengambilannya dalam satu tahun
lampa pju dari jln margacinta sampai derwati perlu diganti selain banyak mati juga kurang terang
rumah tangga saya tidak menerima bantuan sekolah untuk anak saya

3.2.2. Stemming

Pada bagian ini dilakukan proses untuk menemukan akar kata atau kata dasar dari sebuah kata. Proses *stemming* dilakukan dengan menghilangkan semua imbuhan (afiks) baik yang terdiri dari awalan (prefiks) sisipan (infiks) maupun akhiran (sufiks) dan kombinasi dari awalan dan akhiran (konfiks). Stemming digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar sesuai dengan struktur morfologi bahasa indonesia yang baik dan benar. Pengambilan akar kata dilakukan untuk mengurangi dimensi matriks yang dihasilkan oleh kemunculan entitas berbeda yang sebenarnya mempunyai akar kata yang sama.

Terdapat banyak algoritma yang telah dikembangkan untuk melakukan proses stemming, khususnya stemming bahasa indonesia. Pada naskah aplikasi pra-proses penelitian ini digunakan algoritma porter stemmer Bahasa Indonesia yang merupakan versi Bahasa Indonesia dari algoritma porter stemmer Bahasa Inggris yang dikembangkan oleh Martin Porter pada 1980. Gambar 3.10 mengilustrasikan proses stemming yang dilakukan. Hasil stemming dari Tabel 3.3 ditunjukkan pada Tabel 3.4.



Gambar 3.10 Stemming

Tabel 3.4 Kumpulan Data Laporan Setelah Melewati Proses Stemming

Isi Laporan
yth menteri dalam negeri saya mau tanya kalau untuk buat kartu keluarga baru kartu tanda duduk baru dan akte lahir biaya berapa ya mohon informasi terima kasih
yth polisi r apa polantas di tiap laku ilang tidak pernahanya kepada endara apakah ingin ikut sidang karena tidak aku salah atau bayar langsung tilang ke bank karena aku salah lama tahun saya kendaraanya pernah tilang x namun tidak pernah beri lembar biru untuk bayar langsung ke bank saya selalu di beri lembar merah untuk ikut sidang padahal saya sudah a sa bagai tambah bukti materi coba lihat di youtube siar polisi yang ilang polantas selalu ikan surat tilang merah mohon jelas cara lengkap bagaimana benar atur guna slip merah dan slip biru ini ketika ilang terima kasih
lapor saya erima kks ingin ta kapan dana bantu nya akan keluar dan berapa jumlah dana berapa kali ambil dalam satu tahun
lampu pju dari jln margacinta sampai derwat perlu ganti selain banyak mati juga kurang terang
rumah tangga saya tidak erima bantu sekolah untuk anak saya

3.2.3. Pembuatan TF, IDF, dan SVD

Pada tahap ini, hasil data bersih dari proses sebelumnya akan ditransformasi dari data teks menjadi data numerik dalam bentuk matriks. Proses transformasi data dilakukan dengan melalui tiga prosedur, yaitu TF (*term frequency*), IDF (*inverse document frequency*), dan SVD (*singular value decomposition*). Pada proses *term frequency*, potongan kata atau *token* akan diberi nilai sesuai dengan jumlah kemunculannya dalam satu dokumen dan akan didapatkan matriks *term frequency* seperti ditunjukkan pada Gambar 3.11. Setelah melalui proses *term frequency*, dilakukan prosedur *inverse document frequency* dimana potongan kata diberi bobot atau nilai berdasarkan frekuensi kemunculannya pada seluruh dokumen. Matriks *inverse document frequency* ditunjukkan pada Gambar 3.12.

	1 Var1	2 account	3 ada	4 adalah	5 agar	6 air	7 aju	8 akan	9 anak	10 anggota	11 antara	12 apa
1	yth erintah provinsi dk jakarta saya ingin lapor bahwa honor kami bagai											1
2	rumah tangga saya hanya erima kip padahal ada anak lajar di rumah s		1								1	
3	subsidi dari pmrinth untuk siswa sd ca das rtajaya camat telagasar ka											
4	yth menteri didi dan budaya rujuk pada buku tunjuk teknis tentang bar											
5	keluarga saya belum erima kip Hasan ali untuk anak anak saya yang										2	
6	kami dapat kps tetapi anak saya kelas smp tidak dapat bsm padahal							1	1			
7	saya erima kps dan saya punya anak sekolah smp namun saya tidak									1		
8	asalamualaikum yang hormat bapak ibu yang kerja di kps mohon jela:									1		
9	rumah tangga di desa saya tidak erima kip untuk siswa kurang mamp											
10	kang emil kenapa bantu kenapa bantu sekolah di kota bandung di hap											
11	yth menteri didi smk muhammadiyah simo boyolalilamat jl madu ngr											
12	anak saya tidak ada yang dapat kip	1								1		
13	kepada yang hormat prov dk jakarta kami guru guru dk jakarta sampai											1
14	saya tidak erima kartu indonesia pintar kip untuk dapat bantu siswa k									1		
15	anak saya tidak dapat kip wahyud yang nama arida lutfian tidak dapat kip kartu indon									1		

Gambar 3.11 Matriks *Term Frequency*

	1 Var1	2 account	3 ada	4 adalah	5 agar	6 air	7 aju	8 akan	9 anak	10 anggota	11 antara	12 apa
1	yth erintah provinsi dk jakarta saya ingin lapor bahwa honor kami bagai											2.50899
2	rumah tangga saya hanya erima kip padahal ada anak lajar di rumah s	1.43443								2.07647		
3	subsidi dari pmrinth untuk siswa sd ca das rtajaya camat telagasar ka											
4	yth menteri didi dan budaya rujuk pada buku tunjuk teknis tentang bar											
5	keluarga saya belum erima kip Hasan ali untuk anak anak saya yang sekolah muham									3.51577		
6	kami dapat kps tetapi anak saya kelas smp tidak dapat bsm padahal sudah aju syara							3.73826	2.07647			
7	saya erima kps dan saya punya anak sekolah smp namun saya tidak erima kip sedai								2.07647			
8	asalamualaikum yang hormat bapak ibu yang kerja di kps mohon jelas berapa bulan s								2.07647			
9	rumah tangga di desa saya tidak erima kip untuk siswa kurang mampu erima kks kks											
10	kang emil kenapa bantu kenapa bantu sekolah di kota bandung di hapus											
11	yth menteri didi smk muhammadiyah simo boyolalilamat jl madu ngrn simo boyolal											
12	anak saya tidak ada yang dapat kip	1.43443							2.07647			2.50899
13	kepada yang hormat prov dk jakarta kami guru guru dk jakarta sampai saat ini ingin ta											
14	saya tidak erima kartu indonesia pintar kip untuk dapat bantu siswa kurang mampu bi								2.07647			
15	anak saya tidak dapat kip wahyud yang nama arida lutfian tidak dapat kip kartu indon								2.07647			
16	di rumah tangga saya hanya erima kip padahal ada anak yang masih seko	1.43443							2.07647			
17	maaf mau tanya kenapa bsm di sekolah kami sampai sekarang belum cair deng seko									3.51577		
18	anak anak saya sekolah semua sd smp smk tetapi ko keluarga saya tidak erima kip									2.07647		
19	rumah tangga di desa saya tidak erima kartu kip dang saya punya anak yang masih s									2.07647		

Gambar 3.12 Matriks *Inverse Document Frequency*

Setelah melalui prosedur TF-IDF, matriks yang dihasilkan masih memiliki dimensi yang besar sehingga perlu dilakukan reduksi dimensi matriks dengan melakukan prosedur SVD (*singular value decomposition*). Dari prosedur tersebut didapatkan matriks *singular value decomposition* (Gambar 3.13) yang tidak lagi memuat seluruh entitas kata melainkan sejumlah konsep yang merepresentasikan hubungan antar entitas kata dalam dokumen.

	1 Var1	2 Concept1	3 Concept2	4 Concept3	5 Concept4	6 Concept5	7 Concept6	8 Concept7	9 Concept8
1	yth erintah provinsi dk jakarta saya ingin lapor bahwa honor kami bagi guru bantu di smp yapi	0.01731	-0.01271	-0.00253	-0.00153	0.00134	-0.00537	0.01031	0.00006
2	rumah tangga saya hanya erima kip padahal ada anak lajar di rumah saya mohon di urus	0.00550	-0.00259	0.01393	0.01187	-0.00810	0.00167	0.01764	0.00627
3	subsidi dari pmrinth untuk siswa sd ca das rtajaya camat telagasar kab ka rawang jabar dian t:	0.01171	-0.00441	0.01337	0.00354	0.00294	0.00421	0.00345	-0.00434
4	yth menteri didi dan budaya rujui pada buku turunjuk teknik tentang bantu embang smk ruju non	0.01259	0.00127	0.00276	-0.01587	-0.00012	0.00069	-0.00228	0.01774
5	keluarga saya belum erima kip Hasan ali untuk anak anak saya yang sekolah muhammad syahri	0.00453	-0.00041	0.01389	0.01217	-0.01368	0.00147	0.02171	0.00739
6	kami dapat kps tetapi anak saya kelas smp tidak dapat bsm padahal sudah aju syarat di smp	0.01195	-0.00333	0.02157	0.01262	-0.00629	-0.00229	0.01835	-0.00919
7	saya erima kps dan saya punya anak sekolah smp namun saya tidak erima kip sedang kan er	0.00749	-0.00217	0.02487	0.01891	-0.01418	0.00463	0.02639	0.00388
8	asalmaulaikum yang hormat bapak ibu yang kerja di kps mohon jelas berapa bulan sekali bar	0.00822	-0.00421	0.00988	0.00692	-0.00158	0.00030	0.00932	-0.00637
9	rumah tangga di desa saya tidak erima klp untuk siswa kurang mampu erima kks kks rumah ti	0.00813	-0.00516	0.02613	0.01946	0.00263	0.01243	0.00833	0.01346
10	kang emil kenapa bantu kenapa bantu sekolah di kota bandung di hapus	0.00348	-0.00185	0.00764	0.00465	-0.00013	0.00015	0.00527	-0.00252
11	yth menteri didi smk muhammadiyah simo boyolalalamat jl madu ngren simo boyolal kode pos	0.00525	0.00120	-0.00035	-0.00868	-0.00240	0.00149	0.00159	0.00635
12	anak saya tidak ada yang dapat kip	0.00282	-0.00088	0.00732	0.00868	-0.00564	0.00196	0.01084	0.00234
13	kepada yang hormat prov dk jakarta kami guru guru dk jakarta sampai saat ini ingin ta apa san	0.01508	-0.01248	-0.00961	-0.00008	-0.00753	-0.00023	0.00740	0.00189
14	saya tidak erima kartu indonesia pintar kip untuk dapat bantu siswa kurang mampu bsm saya	0.00836	-0.00321	0.02898	0.02144	-0.02253	0.00163	0.03152	0.00506
15	anak saya tidak dapat kip walaupun nama arida lutfian tidak dapat kip kartu indonesia nanti	0.00501	0.00028	0.01572	0.01284	-0.01684	0.00101	0.01763	0.00511

Gambar 3.13 Matriks *Singular Value Decomposition*

Data teks yang tidak terstruktur telah ditransformasi menjadi data numerik yang terstruktur dengan dihasilkannya matriks-matriks tersebut. Proses TF, IDF, dan SVD dilakukan pada setiap data set yang akan diolah lebih lanjut. Proses klasifikasi akan mengolah dua jenis data set yaitu data set yang tidak melalui stemming dan data set yang melalui stemming, sedangkan proses pengelompokan (*clustering*) akan mengolah enam data set yang merupakan kategori kelas dari data laporan.

3.3. Klasifikasi

Proses klasifikasi dilakukan untuk menggolongkan data laporan menjadi beberapa kategori kelas yang ditentukan. Dalam penelitian ini, LAPOR! memiliki enam prioritas kategori kelas untuk mengklasifikasikan dokumen. Tujuan dari proses klasifikasi ini adalah untuk membangun model prediktif yang mampu mengklasifikasikan dokumen secara otomatis sehingga diketahui kategori kelas dari sejumlah besar data secara efektif dan efisien. Proses klasifikasi menggunakan prinsip *machine learning* dimana pembuatan model klasifikasi dilakukan dengan mempelajari sejumlah data latihan (*data training*) dan mengujinya dengan sejumlah data uji (*data test*). Data latihan dan data uji telah

Setelah membangun model klasifikasi, didapatkan *confusion matrix* yang akan digunakan untuk mengevaluasi performa model yang dibangun. Tahap membangun model dilakukan dua kali, yaitu pada data set yang melalui proses stemming dan data set tanpa stemming. Kedua model tersebut kemudian akan dibandingkan keakuratan performanya. Model dengan akurasi lebih besar akan dipilih untuk mengklasifikasikan data yang belum terkласifikasi.

3.4. Pengelompokan (*Clustering*)

Pengelompokan (*clustering*) dilakukan pada enam kelas yang dihasilkan dari proses klasifikasi. Pada setiap kelas, dokumen yang menjadi anggota kelas tersebut akan dikelompokkan lagi menjadi beberapa topik. *Clustering* dilakukan dengan menggunakan metode SOM (*self organizing map*) dan dengan bantuan *software* Statistica 10. Pada penerapan SOM, setiap set data yang akan dikelompokkan akan dibagi menjadi data latihan (*data training*) dan data uji (*data test*).

Pada penelitian ini, ditetapkan jumlah data latihan (*data training*) adalah 70% dan data uji (*data test*) adalah 30% serta *training cycle* 1000 kali. Data input yang digunakan adalah data set setiap kelas yang didapatkan dari hasil klasifikasi dan sudah dalam bentuk matriks SVD (*singular value decomposition*). Jumlah data set kelas yang akan dikelompokkan adalah enam data set, yaitu kelas akses pendidikan yang berkeadilan, energi pangan dan maritim, kesehatan publik, pengentasan kemiskinan, pembangunan infrastruktur, dan reformasi birokrasi.

Dokumen anggota setiap kelas akan dikelompokkan menjadi beberapa kelompok atau *cluster*. Jumlah cluster ditentukan oleh *initial map size* atau ukuran peta SOM, yaitu ukuran geometris SOM yang akan menggambarkan banyaknya jumlah *cluster* yang dibentuk. Penentuan initial map size dilakukan dengan *trial and error* hingga mendapatkan ukuran dengan angka error terkecil. Pada penelitian ini jumlah kelompok atau *cluster* yang diinginkan adalah kurang dari 10 atau maksimal 9. Hal ini dikarenakan pihak LAPOR! ingin mengetahui topik prioritas dari masyarakat, apabila jumlah *cluster* yang dibentuk semakin banyak maka jumlah anggota setiap *cluster* akan semakin sedikit sehingga tidak bisa menggambarkan ukuran suara mayoritas masyarakat. *Trial and error* untuk menentukan jumlah *cluster* dilakukan pada setiap data set kelas.

Tabel 3.5 Hasil error setiap *initial map size* pada data set pendidikan

<i>Initial map size</i>	Jumlah cluster	Nilai error
2x2	4	0.205707
2x3	6	0.190347
2x4	8	0.176853
3x3	9	0.168435

Tabel 3.6 Hasil error setiap *initial map size* pada data set energi, pangan, dan maritim

<i>Initial map size</i>	Jumlah cluster	Nilai error
2x2	4	0.357660
2x3	6	0.326461
2x4	8	0.311581
3x3	9	0.297299

Tabel 3.7 Hasil error setiap *initial map size* pada data set kesehatan

<i>Initial map size</i>	Jumlah cluster	Nilai error
2x2	4	0.295529
2x3	6	0.251187
2x4	8	0.240036
3x3	9	0.214586

Tabel 3.8 Hasil error setiap *initial map size* pada data set infrastruktur

<i>Initial map size</i>	Jumlah cluster	Nilai error
2x2	4	0.240111
2x3	6	0.229734
2x4	8	0.196381
3x3	9	0.196249

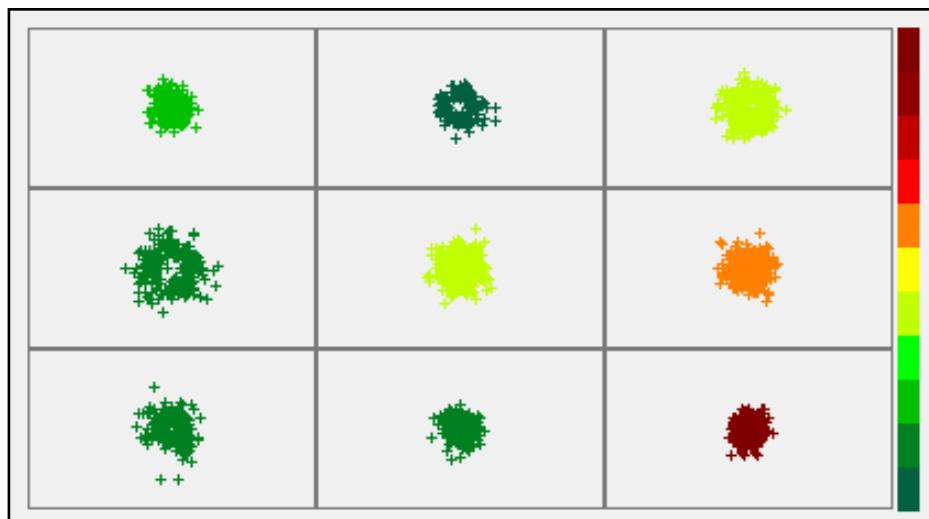
Tabel 3.9 Hasil error setiap *initial map size* pada data set kemiskinan

<i>Initial map size</i>	Jumlah cluster	Nilai error
2x2	4	0.290953
2x3	6	0.273844
2x4	8	0.257320
3x3	9	0.248564

Tabel 3.10 Hasil error setiap *initial map size* pada data set birokrasi

<i>Initial map size</i>	Jumlah cluster	Nilai error
2x2	4	0.308628
2x3	6	0.278890
2x4	8	0.255541
3x3	9	0.239329

Pada proses *trial and error* di setiap data set didapatkan bahwa semua kelas memiliki nilai error terkecil pada *initial map size* 3x3 atau maksimal jumlah *cluster* 9. Setelah nilai *initial map size* ditetapkan, maka data diolah dan selanjutnya akan diperoleh hasil pengelompokan (*clustering*). Gambar 3.15 menunjukkan contoh model *clustering* yang akan diperoleh dari teknik *clustering* SOM..

**Gambar 3.15 Contoh model *clustering* SOM**

BAB 4

ANALISIS HASIL DAN PEMBAHASAN

Dalam rangka menjawab tujuan penelitian, proses penarikan informasi dari data teks telah dilakukan pada bab sebelumnya dengan klasifikasi dan *clustering*. Pada bab ini, akan dibahas hasil dari pengolahan data dari bab sebelumnya.

4.1. Klasifikasi Dokumen

Klasifikasi dokumen dilakukan untuk mengklasifikasikan dokumen sesuai dengan kelas yang diinginkan. Prinsip klasifikasi adalah dengan membangun model pengklasifikasi dari sejumlah data set yang dijadikan sebagai data pembelajaran dan kemudian model yang dibangun akan mampu mengklasifikasikan data lainnya yang belum terkласifikasi secara otomatis. Pada penelitian ini, model klasifikasi perlu dibuat untuk mengklasifikasi data yang jumlahnya sangat banyak yang sulit dilakukan secara manual.

Model klasifikasi dibangun dengan prinsip algoritma SVM (*support vector machine*). Dari proses membangun model klasifikasi, didapatkan *confusion matrix* yang dapat digunakan untuk mengevaluasi keakuratan model. Dari proses pengembangan model klasifikasi didapatkan dua *confusion matrix*, yaitu *confusion matrix* untuk data set yang melalui proses stemming dan *confusion matrix* untuk data set yang tidak melalui proses stemming. Tabel 4.1 menunjukkan *confusion matrix* dan akurasi pada model klasifikasi tanpa stemming dan Tabel 4.2 menunjukkan *confusion matrix* dan akurasi pada model klasifikasi dengan stemming. Dari kedua hasil akurasi model tersebut didapatkan bahwa model klasifikasi dengan data set yang melalui proses stemming mempunyai akurasi yang lebih besar daripada model klasifikasi dengan data set yang tidak melalui proses stemming.

Tabel 4.1 Confusion matrix dan akurasi model klasifikasi tanpa stemming

		Predicted							Jumlah
		Pendidikan	Energi pangan maritim	Kesehatan	Infrastruktur	Kemiskinan	Birokrasi		
Observed	Pendidikan	674	24	39	31	223	9	1000	
	Energi pangan maritime	5	578	52	36	45	8	724	
	Kesehatan	140	38	591	108	55	68	1000	
	Infrastruktur	19	44	38	306	396	197	1000	
	Kemiskinan	158	27	34	48	644	89	1000	
	Birokrasi	14	39	259	254	185	249	1000	
Jumlah		1010	750	1013	783	1548	620	5724	
Pendidikan	Precision	0.67							
	Recall	0.67							
	F1	0.67							
Energi pangan maritim	Precision	0.77							
	Recall	0.80							
	F1	0.78							
Kesehatan	Precision	0.58							
	Recall	0.59							
	F1	0.59							
Infrastruktur	Precision	0.39							
	Recall	0.31							
	F1	0.34							
Kemiskinan	Precision	0.42							
	Recall	0.64							
	F1	0.51							
Birokrasi	Precision	0.40							
	Recall	0.25							
	F1	0.31							
F1 Measure		0.53							
Accuracy		0.53							

Tabel 4.2 Confusion matrix dan akurasi model klasifikasi dengan stemming

		Predicted							Jumlah
		Pendidikan	Energi pangan maritim	Kesehatan	Infrastruktur	Kemiskinan	Birokrasi		
Observed	Pendidikan	970	0	11	3	2	14	1000	
	Energi pangan maritime	1	556	70	29	35	33	724	
	Kesehatan	13	20	638	250	59	20	1000	
	Infrastruktur	10	16	41	808	111	14	1000	
	Kemiskinan	12	32	42	520	133	261	1000	
	Birokrasi	15	34	261	14	12	664	1000	
Jumlah		1021	658	1063	1624	352	1006	5724	
Pendidikan	Precision	0.95							
	Recall	0.97							
	F1	0.96							
Energi pangan maritim	Precision	0.84							
	Recall	0.77							
	F1	0.80							
Kesehatan	Precision	0.60							
	Recall	0.64							
	F1	0.62							
Infrastruktur	Precision	0.50							
	Recall	0.81							
	F1	0.62							
Kemiskinan	Precision	0.38							
	Recall	0.13							
	F1	0.20							
Birokrasi	Precision	0.66							
	Recall	0.66							
	F1	0.66							
F1 Measure		0.64							
Accuracy		0.66							

Tabel 4.2 juga menunjukkan *nilai precision, recall, F1 measure, dan accuracy* dari masing-masing kelas. Berdasarkan hasil yang ada, kategori pendidikan merupakan kategori yang mempunyai nilai rata-rata tinggi dari semua pengukuran. Hal ini menunjukkan bahwa model paling bekerja secara akurat dalam mengklasifikasikan anggota kategori pendidikan dengan benar. Ketepatan klasifikasi berbasis teks dipengaruhi oleh beberapa faktor, diantaranya ukuran

fragmen teks yang diidentifikasi, jumlah data latihan yang digunakan, fitur klasifikasi, algoritma yang digunakan, dan kemiripan kata (Botha & Barnard, 2012).

Nilai yang lebih rendah dari kategori lainnya disebabkan karena banyak dokumen yang berasal dari kategori yang berbeda tetapi memiliki kemiripan satu sama lain. Jumlah data latihan pada penelitian ini cukup terbatas yaitu tidak lebih dari 1000 data setiap kelas dikarenakan jumlah salah satu kelas yang tidak mencukupi. Menurut Botha & Barnard, 2012, dengan menambah jumlah data latihan yang digunakan, umumnya tingkat akurasi yang dihasilkan akan semakin meningkat pula.

Setelah didapatkan model klasifikasi, kemudian akan didapatkan juga klasifikasi dari data lainnya yang belum mempunyai label kelas. Jumlah masing-masing anggota kelas ditunjukkan pada Tabel 4.3.

Tabel 4.3 Jumlah anggota tiap kelas klasifikasi

Kelas	Jumlah
Akses Pendidikan yang Berkeadilan	4195
Energi, Pangan, dan Maritim	2310
Kesehatan Publik	3795
Pembangunan Infrastruktur	5077
Pengentasan Kemiskinan	8347
Reformasi Birokrasi	3611

Dari hasil tersebut dapat dilihat bahwa masyarakat paling banyak menyampaikan aspirasi terkait masalah kemiskinan. Dalam aspek tersebut, masalah yang dibahas yaitu terkait berbagai bantuan sosial dan kondisi masyarakat miskin. Kelas yang memiliki jumlah terbanyak kedua yaitu mengenai infrastruktur, dalam hal ini masalah yang dilaporkan biasanya mengenai kondisi jalan, transportasi, dan berbagai proyek pembangunan pemerintah. Di sisi lain, kelas yang paling sedikit mendapat sorotan adalah mengenai energi, pangan, dan maritim, dimana masyarakat umumnya mengeluhkan masalah BBM, harga sembako, dan kondisi maritim saat ini.

4.2. Pengelompokan (*Clustering*) Dokumen

Tahap Pengelompokan (*clustering*) merupakan tahap yang dilakukan untuk memperoleh hasil akhir dari penelitian ini. Pada pengelompokan (*clustering*) teks, proses dilakukan secara *unsupervised* atau tanpa pengarahan sebelumnya untuk mengelompokan dokumen menjadi beberapa kelompok berdasarkan kesamaan dokumen. Pengelompokan (*clustering*) dilakukan untuk memperoleh kelompok topik laporan secara spesifik dari setiap kategori atau kelas data. Hal ini dilakukan untuk mendapatkan informasi yang dituhkan dalam menjawab tujuan dari penelitian.

Pada prinsipnya, jumlah kelompok atau *cluster* yang didapatkan dari teknik *clustering* bisa sangat subjektif dan tergantung kebutuhan penelitian. Akan tetapi proses mencari jumlah kelompok yang optimal secara otomatis juga sering dilakukan jika tidak ada batasan atau kebutuhan tertentu dalam mengelompokkan. Pada penelitian ini, jumlah kelompok atau *cluster* maksimal yang diinginkan adalah 9 *cluster* pada setiap kelas data. Jumlah tersebut ditetapkan karena pihak LAPOR! ingin mengetahui prioritas topik permasalahan dengan jumlah kelompok yang terbatas agar lebih fokus dalam proses analisis kedepannya dan bisa merepresentasikan mayoritas suara masyarakat.

Pada proses *trial and error* dalam menentukan jumlah cluster, hasil dari setiap kelas menunjukkan nilai error terkecil pada jumlah *cluster* yang menjadi batas maksimum yang diinginkan yaitu 9 *cluster*. Hal ini terjadi karena dengan semakin banyak jumlah kelompok yang diperoleh, jumlah anggota tiap kelompok akan semakin sedikit dan akan lebih memiliki kemiripan satu sama lain sehingga nilai error semakin mengecil. Selain mempunyai anggota kelompok yang sedikit, jumlah *cluster* yang banyak juga akan menyebabkan terjadinya dokumen yang mempunyai topik sama terkelompokkan menjadi dua *cluster* yang berbeda. Pada penelitian ini, pengelompokan dokumen dilakukan dengan prinsip kemiripan isi dokumen, dimana hanya kemunculan kata yang diperhitungkan tanpa memperhatikan kesamaan arti.

a. Pengelompokan kelas pendidikan

Tabel 4.4 Hasil *cluster* kelas pendidikan

Cluster	Judul cluster	Jumlah
Cluster 1	Laporan seputar KJP (kartu jakarta pintar)	152
Cluster 2	Keluhan belum mendapat KIP (kartu indonesia pintar) secara umum (majoritas menyebutkan asal dari luar jakarta)	375
Cluster 3	Keluhan tidak mendapat KIP tapi mendapat KIS (kartu indonesia sehat) dan KKS (kartu keluarga sejahtera)	44
Cluster 4	Laporan dan pertanyaan untuk mendapatkan KIP	203
Cluster 5	Laporan ke berbagai kementerian (khususnya kemendiknas dan kemenristek) mengenai berbagai masalah pendidikan, baik dari sisi guru, mahasiswa, ataupun siswa	234
Cluster 6	Keluhan singkat belum mendapat KIP (tanpa penjelasan lebih detil)	766
Cluster 7	Laporan pemegang kartu keluarga sejahtera yang tidak mendapat bantuan pendidikan	332
Cluster 8	Keluhan siswa miskin terhadap pembagian dana KIP di sekolah	217
Cluster 9	Keluhan rumah tangga yang anaknya belum mendapat KIP	297

Tabel 4.4. menunjukkan hasil *clustering* pada data set kelas pendidikan. Terdapat 9 *cluster* dengan jumlah yang bervariasi tiap *cluster*. Hasil yang diperoleh menunjukkan bahwa majoritas *cluster* pada kelas pendidikan membahas mengenai bantuan pendidikan atau KIP (kartu Indonesia pintar), dimana KIP (kartu indonesia pintar) merupakan kartu bantuan yang diterbitkan oleh pemerintah untuk membantu biaya sekolah siswa kurang mampu. Hanya cluster 5 yang berisi tentang permasalahan pendidikan di sekolah dari sisi siswa atau guru, hingga permasalahan pembelajaran.

Cluster yang memiliki jumlah paling banyak yaitu *cluster* 6 yang berisi berbagai laporan singkat mengenai masyarakat yang tidak mendapatkan KIP tanpa detil lebih lanjut. Walaupun mayoritas *cluster* mempunyai anggota dokumen yang membahas mengenai bantuan pendidikan dan KIP, dokumen dipisahkan menjadi beberapa kelompok berdasarkan subjek atau objek yang terdapat pada dokumen. Misalnya pada cluster 1, objek yang disebutkan adalah wilayah DKI Jakarta dan pada cluster 2 lebih menyebutkan wilayah lain. Secara

umum, cluster 3, 6, dan 9 mempunyai karakteristik yang hampir sama yaitu tidak mendapat KIP. Namun, subjek yang disebutkan berbeda, misalnya *cluster* 3 menyebutkan kata pemegang kartu KIS dan KKS, *cluster* 6 tidak menyebutkan subjek apapun, dan *cluster* 9 menyebutkan subjek rumah tangga.

b. Pengelompokan kelas energi, pangan, dan maritim

Tabel 4.5 Hasil *cluster* kelas energi, pangan, dan maritim

Cluster	Judul cluster	Jumlah
Cluster 1	Laporan mengenai kelangkaan LPG dan harga gas yang melambung di daerah tertentu (ditujukan ke pertamina)	75
Cluster 2	Laporan mengenai kelangkaan BBM (ditujukan ke pertamina)	57
Cluster 3	Laporan mengenai pemadaman listrik dan pemasangan listrik baru (ditujukan ke PLN)	81
Cluster 4	Kualitas beras miskin yang jelek	53
Cluster 5	Laporan ke kementerian kelautan dan perikanan, terutama mengenai penangkapan ikan illegal	31
Cluster 6	Laporan mengenai pelayanan PT PLN secara umum	168
Cluster 7	Subsidi BBM	185
Cluster 8	Permasalahan mengenai pasokan air oleh PDAM	28
Cluster 9	Protes kenaikan BBM	52

Tabel 4.5. menunjukkan hasil *clustering* pada data set kelas energi, pangan, dan maritim. 9 *cluster* yang dihasilkan mempunyai jumlah yang bervariasi tiap *cluster*. Setiap cluster memiliki karakteristik tersendiri yang membedakannya dengan lainnya. Pada pengelompokan kelas ini, walaupun beberapa kelas memiliki satu kata kunci yang sama, tetapi dokumen berhasil dikelompokkan berdasarkan berbagai tujuan yang berbeda. *Cluster* 7 memiliki jumlah anggota paling banyak, dimana mayoritas masyarakat melaporkan permasalahan subsidi BBM. Masalah subsidi BBM yang dilaporkan diantaranya mengenai keluhan atas adanya kebijakan baru mengenai subsidi BBM dan dampaknya pada masyarakat. *Cluster* 6 memiliki jumlah terbanyak kedua, dimana masyarakat banyak mengeluhkan tentang kinerja pelayanan PLN secara umum selain permasalahan pemadaman listrik dan pemasangan listrik baru yang sudah dipisahkan ke *cluster* 2.

c. Pengelompokan kelas kesehatan

Hasil cluster yang diperoleh pada pengelompokan kelas kesehatan ditunjukkan oleh Tabel 4.6.

Tabel 4.6 Hasil *cluster* kelas kesehatan

Cluster	Judul cluster	Jumlah
Cluster 1	Laporan mengenai kesulitan pada pendaftaran BPJS kesehatan secara online	103
Cluster 2	Keluhan mengenai prosedur pembayaran premi/iuran/tagihan BPJS kesehatan	109
Cluster 3	Keluhan mengenai integrasi pembayaran dengan BPJS online	164
Cluster 4	Laporan belum mendapat KIS (spesifik ditujukan pada kementerian koordinator bidang pembangunan manusia dan kebudayaan)	23
Cluster 5	Keluhan mengenai pelayanan Kesehatan pada pemprov DKI Jakarta, terutama mengenai buruknya pelayanan di RS atau puskesmas	58
Cluster 6	Masalah aktivasi akun pendaftaran BPJS online	466
Cluster 7	Laporan warga desa bahwa rumah tangganya atau keluarganya belum mendapat KIS	148
Cluster 8	Laporan mengenai bantuan kesehatan secara umum	421
Cluster 9	Pertanyaan seputar BPJS kesehatan	302

Hasil yang diperoleh dari pengelompokan kelas kesehatan menunjukkan bahwa cluster yang diperoleh mayoritas membahas mengenai BPJS (badan penyelenggara jaminan sosial) kesehatan dan KIS (kartu indonesia sehat). BPJS kesehatan dan KIS merupakan jenis bantuan yang diberikan oleh pemerintah untuk memfasilitasi masyarakat kurang mampu yang membutuhkan pelayanan kesehatan. Walaupun beberapa *cluster* memiliki kata kunci yang sama, namun dokumen berhasil dipisahkan berdasarkan isi permasalahan yang dilaporkan.

Cluster yang memiliki jumlah paling banyak adalah *cluster* 6 mengenai keluhan terkait proses aktivasi akun BPJS kesehatan yang bermasalah. Pada *cluster* ini masyarakat banyak mengeluhkan proses aktivasi akun BPJS yang bermasalah misalnya dikarenakan proses yang selalu gagal atau tidak mendapat konfirmasi dari pihak BPJS kesehatan. *Cluster* 8 yang mempunyai jumlah terbanyak kedua membahas mengenai berbagai bantuan kesehatan secara umum

dimana laporan yang ada mayoritas berisi laporan masyarakat miskin yang memerlukan bantuan kesehatan. *Cluster* 1, 2, 3, dan 9 sebenarnya memiliki kesamaan dengan cluster 6 yaitu terkait BPJS kesehatan, hanya saja permasalahan yang dihadapi dan dilaporkan berbeda. Hal tersebut mengindikasikan bahwa topik permasalahan mengenai BPJS kesehatan sedang mendapat banyak sorotan publik daripada permasalahan pelayanan kesehatan lainnya.

d. Pengelompokan kelas infrasruktur

Tabel 4.7 Hasil *cluster* kelas infrasruktur

Cluster	Judul cluster	Jumlah
cluster 1	Laporan mengenai transjakarta	104
cluster 2	Saluran air dan banjir	251
cluster 3	Segala bentuk pembangunan di Jakarta	190
cluster 4	Pekerjaan umum terkait, jalan, jembatan, trotoar dll	101
cluster 5	Laporan kepada pemprov DKI Jakarta terkait jalan dan trotoar	421
cluster 6	Laporan kepada pemprov DKI Jakrata terkait rambu lalu lintas	125
cluster 7	Transportasi, lalu lintas, stasiun, halte, kereta, bus	660
cluster 8	Pemprov DKI Jakarta terkait pemukiman dan bangunan liar	93
cluster 9	Pekerjaan umum terkait pembangunan desa tertinggal	105

Tabel 4.7. menunjukkan hasil pengelompokan pada kelas infrastruktur. Hasil yang diperoleh menunjukkan bahwa topik mengenai infrastruktur terkait transportasi merupakan topik yang paling banyak dilaporkan. *Cluster* 7 merupakan *cluster* yang memiliki jumlah terbanyak. Pada *cluster* ini, masyarakat banyak melaporkan kualitas pelayanan transportasi seperti kereta, bus, dan lainnya yang kurang memadai, serta kondisi fasilitas penunjangnya seperti jalan raya, halte, atau rambu lalu lintas yang rusak. *Cluster* dengan jumlah terbanyak kedua adalah *cluster* 5 yang melaporkan kondisi jalan dan trotoar yang rusak atau kurang memadai khususnya di wilayah DKI Jakarta. *Cluster-cluster* lainnya memiliki jumlah yang relatif seimbang.

Beberapa *cluster* memiliki konten yang hampir sama, seperti *cluster* 1 dan 7 yang sama-sama membahas topik terkait transportasi tetapi topik tentang transjakarta telah dipisahkan di *cluster* 1 sedangkan *cluster* 7 memuat berbagai

bentuk keluhan terkait jenis transportasi lainnya. *Cluster 3* dan *4* dibedakan berdasarkan tujuan laporan, dimana *cluster 4* ditujukan spesifik kepada pemprov DKI. *Cluster 5* dan *6* sama-sama ditujukan pada pemprov DKI, hanya saja objek yang dilaporkan berbeda. Secara umum dapat dikatakan bahwa masyarakat lebih sering mengeluhkan pelayanan infrastruktur dalam bidang transportasi dan pembangunan sarana penunjangnya.

e. Pengelompokan kelas kemiskinan

Tabel 4.8 menunjukkan hasil pengelompokan kelas kemiskinan. Pada kelas kemiskinan, masalah yang dilaporkan tidak jauh dari berbagai jenis bantuan sosial yang diberikan oleh pemerintah. Seluruh *cluster* membahas mengenai bantuan sosial, namun laporan dikelompokkan berdasarkan jenis bantuan dan masalah yang terjadi dari bantuan yang dilakukan. *Cluster* dengan anggota terbanyak adalah *cluster 9* yang mayoritas melaporkan permasalahan terkait jenis bantuan KPS (kartu perlindungan sosial), KKS (kartu keluarga sejahtera), PKH (program keluarga harapan), dan BLT (bantuan langsung tunai). KPS, KKS, PKH, dan BLT merupakan beberapa bantuan yang diberikan pemerintah untuk masyarakat miskin tetapi memiliki konten dan aturan yang berbeda. KPS dan KKS merupakan jenis kartu yang diterbitkan pemerintah sebagai penanda keluarga kurang mampu, dimana KKS merupakan versi baru dari KPS yang diterbitkan oleh pemerintah lama. PKH dan BLT merupakan jenis program perlindungan sosial melalui pemberian uang tunai kepada keluarga yang sangat miskin dan memenuhi kriteria tertentu. Pada *cluster 9*, mayoritas laporan berisi keluhan masyarakat yang tidak mendapat bantuan, prosedur kepesertaan bantuan, pembagian bantuan yang tidak merata, atau terjadinya penyimpangan dalam pelaksanaan bantuan.

Cluster yang mempunyai jumlah paling banyak kedua adalah *cluster 6*. Pada *cluster* ini banyak dilaporkan masalah pembagian BLSM (bantuan langsung sementara masyarakat) yang tidak merata atau tidak tepat sasaran. BLSM pada prinsipnya mirip dengan BLT, dimana BLSM juga memberikan bantuan dengan membagikan uang tunai pada masyarakat miskin, tetapi BLSM diberikan sebagai kompensasi khusus dalam rangka naiknya harga BBM. *Cluster 6* sangat mirip dengan *cluster 3*, yaitu sama-sama membahas topik BLSM, namun pada *cluster 3*

majoritas menyebutkan kata kunci lain berupa lokasi desa atau daerah tertentu. Hasil dari pengelompokan kelas kemiskinan mengindikasikan bahwa topik yang paling menjadi sorotan masyarakat di bidang pengentasan kemiskinan adalah mengenai program bantuan KPS dan BLSM yang masih banyak bermasalah.

Tabel 4.8 Hasil *cluster* kelas kemiskinan

Cluster	Judul cluster	Jumlah
cluster 1	Kartu perlindungan sosial tidak merata dan salah sasaran	424
cluster 2	Kartu perlindungan sosial tapi tidak mendapat bantuan untuk keperluan sekolah	219
cluster 3	Permasalahan pada pembagian BLSM (bantuan langsung sementara masyarakat) yang menyebutkan desa atau daerah tertentu yang tidak mendapat	608
cluster 4	Pertanyaan mengenai prosedur dan pembagian kartu	338
cluster 5	Permasalahan pada kartu perlindungan sosial secara umum	621
cluster 6	Permasalahan pembagian BLSM secara umum(bantuan langsung sementara masyarakat)	800
cluster 7	Permasalahan penerimaan kartu, tidak menerima salah satu jenis kartu, dsb	351
cluster 8	Permasalahan raskin (beras miskin)	301
cluster 9	Permasalahan segala jenis bantuan, seperti KKS (kartu keluarga sejahtera), PKH (program keluarga harapan, BLT (bantuan langsung tunai), raskin, dsb	1119

f. Pengelompokan kelas birokrasi

Tabel 4.9 menunjukkan hasil pengelompokan kelas birokrasi. Pada hasil *clustering* kelas birokrasi, *cluster* 8 merupakan *cluster* yang memiliki jumlah paling banyak dan jauh melebihi *cluster* lainnya. *Cluster* 8 membahas mengenai berbagai keluhan dan pertanyaan dalam mengurus akta dan surat, mulai dari akte kelahiran, surat izin mengemudi, surat tanah, dan sebagainya. *Cluster* ini terbilang besar karena mempunyai anggota yang masih bervariasi tetapi mempunyai kata kunci mengurus dan ditujukan kepada lembaga pemerintah. Pada *cluster* ini, masalah yang dikeluhkan kebanyakan terkait proses mengurus dokumen yang rumit, pelayanan yang lama, atau terjadi pungutan liar dalam prosesnya.

Cluster dengan jumlah terbanyak kedua yaitu *cluster* 6, dimana masalah yang dilaporkan khusus ditujukan pada pemerintah DKI Jakarta dengan jenis objek yang sedikit lebih bervariasi dari *cluster* 9. Topik yang dibahas pada *cluster*

6 banyak membahas proses mengurus izin bangunan, honor atau gaji pegawai DKI yang tidak tuntas, pelayanan administrasi kependudukan DKI, dan lainnya. Secara umum, hasil dari pengelompokan kelas birokrasi menunjukkan bahwa masyarakat masih banyak mengalami kesulitan dalam mengurus berbagai jenis administrasi dikarenakan sulitnya aturan birokrasi oleh pemerintah.

Tabel 4.9 Hasil cluster kelas birokrasi

Cluster	Judul cluster	Jumlah
cluster 1	Pengurusan kartu tanda penduduk	164
cluster 2	Pengurusan sertifikat tanah	90
cluster 3	Laporan mengenai Tes CPNS	186
cluster 4	Pertanyaan mengenai prosedur BLSM	85
cluster 5	Pengurusan NPWP (pajak)	151
cluster 6	Laporan mengenai segala jenis birokrasi kepada pemprov DKI Jakarta	337
cluster 7	Prosedur pembuatan kartu perlindungan sosial	170
cluster 8	Segala jenis kepengurusan lainnya seperti akte kelahiran, sim, mendirikan bangunan, stnk, catatan pernikahan, paspor, imigrasi, dsb	1142
cluster 9	Birokrasi bantuan kartu	82

BAB 5 **KESIMPULAN**

5.1. Kesimpulan

Segala bentuk masukan dan aspirasi masyarakat sangat penting bagi kemajuan dalam berbagai hal. Oleh karena itu, pemerintah membangun LAPOR! sebagai sarana yang bisa digunakan masyarakat untuk menyampaikan segala keluh kesah. Di sisi lain, pemerintah dituntut untuk memberikan respon yang cepat dan mengambil keputusan secara tepat. Dengan banyaknya jumlah laporan yang masuk di LAPOR!, proses analisis informasi secara manual akan memerlukan waktu yang cukup lama dan tidak responsif. Teknik analisis data seperti *data mining* dan *text mining* diperlukan untuk mengautomasi proses analisis data secara manual yang memerlukan proses yang cukup lama menjadi lebih singkat. Dengan memanfaatkan teknik analisis tersebut, pemerintah dapat mengetahui prioritas permasalahan yang harus diselesaikan.

Hasil yang diperoleh dari penelitian ini merupakan hasil analisis data pada periode Oktober 2014 hingga Maret 2015. Rentang periode tersebut dipilih karena Oktober merupakan awal periode pemerintahan baru. Hasil pembangunan model untuk mengklasifikasikan dokumen menunjukkan tingkat akurasi 66%, dimana hasil tersebut bisa ditingkatkan seiring dengan bertambahnya jumlah data di masa mendatang. Dari hasil proses klasifikasi ditunjukkan bahwa laporan terkait permasalahan pengentasan kemiskinan dan pembangunan infrastruktur memiliki jumlah terbanyak selama periode Oktober 2014 hingga Maret 2015. Hal ini mengindikasikan bahwa masalah pengentasan kemiskinan dan pembangunan infrastuktur merupakan jenis permasalahan yang seharusnya menjadi prioritas utama saat ini. Terkait fokus topik permasalahan dalam kelas tersebut, pada *clustering* didapatkan bahwa topik masalah mayoritas pada kelas kemiskinan adalah mengenai BLSM (bantuan langsung sementara masyarakat) dan beberapa jenis bantuan lainnya yang dirasa kurang tepat sasaran. Pada kelas infrastruktur, hasil *clustering* menunjukkan jumlah anggota *cluster* paling banyak yaitu pada

keluhan mengenai infrastruktur dalam bidang transportasi seperti kereta dan bus, serta sarana penunjangnya seperti stasiun, halte, dan beberapa sarana lalu lintas.

Lapor memiliki kelemahan dengan adanya fitur anonim. Fitur anonim membuat laporan yang ada belum dapat dinyatakan valid karena adanya kemungkinan pesan ganda atau pesan yang dikirimkan beberapa kali. Pada penelitian ini, adanya pesan ganda sudah diminimalisir dalam tahap pra-proses dengan mengeliminasi pesan yang memiliki isi sama persis dengan pesan lainnya. Akan tetapi masih ada kemungkinan pesan dikirimkan oleh orang yang sama dengan maksud yang sama tetapi dengan bentuk isi yang berbeda. Oleh karena itu, hasil dari penelitian ini masih mungkin bisa berubah jika mempertimbangkan kemungkinan tersebut.

5.2. Saran

Penelitian mengenai *text mining* merupakan salah satu penelitian yang sedang berkembang pesat saat ini seiring dengan berkembangnya teknologi digital yang banyak menghasilkan informasi berupa data tekstual. Akan tetapi, penelitian mengenai teks berbahasa Indonesia belum banyak dilakukan. Masih banyak celah yang harus diperbaiki dalam melakukan penelitian *text mining* bahasa Indonesia. Pada tahap pra-proses, hasil stemming yang didapatkan dari algoritma stemming untuk bahasa Indonesia masih memiliki keakuratan yang rendah. Padahal proses stemming tersebut dilakukan untuk mengurangi noise pada data sehingga matriks yang dihasilkan bisa optimal. Penerapan stemming kebanyakan memang dilakukan pada dokumen berbahasa Inggris dan penerapannya pada bahasa Indonesia lebih sulit dilakukan karena bahasa Indonesia mempunyai banyak bentuk kata imbuhan yang sangat bervariasi (Triawati, 2009). Penelitian lebih lanjut pada bidang ini bisa dilakukan dengan menganalisis penggunaan beberapa algoritma stemming bahasa Indonesia yang saat ini sudah dikembangkan dan membandingkan mana yang memberikan akurasi terbaik.

Data awal yang dipakai pada penelitian ini masih bercampur dengan berbagai dokumen spam, sehingga tahap pra-proses sedikit sulit dilakukan dan membutuhkan usaha lebih banyak untuk memisahkan dokumen yang berguna. Penelitian selanjutnya dapat melakukan klasifikasi awal dengan memisahkan

dokumen yang berguna dan dokumen spam. Pengembangan aplikasi pra-proses juga bisa dilakukan untuk mendapatkan hasil transformasi data yang lebih akurat.

Penelitian ini hanya mengklasifikasikan data menjadi enam kelas. Penelitian selanjutnya dapat melakukan klasifikasi dengan menambah beberapa kelas lainnya sehingga mampu memberikan hasil yang mendekati keadaan sebenarnya. Jumlah data latihan juga perlu ditambah seiring dengan bertambahnya data agar model mampu belajar dengan lebih baik dan akurasi yang didapatkan dari model akan lebih tinggi.

Dikarenakan data laporan diterima dari berbagai media, seperti sms dan website, data teks yang ada memiliki jumlah karakter dan penggunaan bahasa yang sangat bervariasi. Penelitian ini hanya menganalisis data teks berdasarkan kemunculan kata dan asosiasi kata yang sering muncul bersama, padahal bahasa yang digunakan pada media tertentu biasanya memiliki jenis penggunaan kosa kata yang berbeda walaupun sebenarnya memiliki arti yang sama. Penelitian di masa depan dapat mempertimbangkan pemilihan fitur lain untuk mengelompokkan kata misalnya dengan memperhatikan sinonim, akronim, dan prinsip kedekatan kata sehingga hasil yang didapatkan akan lebih optimal.

DAFTAR PUSTAKA

- Agusta, Y. (2007). K-Means – Penerapan, Permasalahan, dan Metode Terkait. *Jurnal Sistem dan Informatika Vol. 3*, 47-60.
- Baker, K. (2013). *Ohio State University, Departement of Linguistics*. Retrieved April 21, 2015, from Ohio State University web: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf.
- Belsky, G. (2012). *TIME Magazine*. Retrieved March 3, 2015, from TIME Magazine Web Site: <http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/>
- Botha, G. R., & Barnard, E. (2012). Factors that affect the accuracy of text-based language identification. *Computer Speech and Language Vol. 26*, 307–320.
- Dumbill, E. (2014). *Forbes Magazine*. Retrieved March 3, 2015, from Forbes Magazine Web Site: <http://www.forbes.com/sites/edddumbill/2014/05/07/defining-big-data/>
- Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.
- Gullo, F. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. *Physics Procedia* 62, 18-22.
- Guthikonda, S. M. (2005). *Kohonen Self-Organizing Maps*. Retrieved April 22, 2015, from sya.am: <http://www.shy.am/wp-content/uploads/2009/01/kohonen-self-organizing-maps-shyam-guthikonda.pdf>.
- Jiawei, H., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third Edition*. Waltham, MA: Morgan Kaufmann.
- LAPOR! (2015). *Tentang LAPOR!* Retrieved February 25, 2015, from LAPOR! Web Site: https://lapor.ukp.go.id/lapor/tentang_lapor/tentang-layanan-aspirasi-dan-pengaduan-online-rakyat.html
- Linoff, G. S., & Berry, M. J. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management Third Edition*. Indianapolis, IN: Wiley Publishing, Inc.

- Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Oxford: Elsevier.
- Palvia, S. C., & Sharma, S. S. (2007). E-Government and E-Governance: Definitions/Domain Framework and Status around the World. *Foundation of e-Government* (pp. 1-12). India: International Congress of e-Government.
- Lembaga administrasi negara. (2015). *Pengelolaan Pengaduan Layanan Publik (Permenpan No. 3 Tahun 2015)*. Retrieved April 16, 2015, from Pemerintah.net: <http://pemerintah.net/pengelolaan-pengaduan-pelayanan-publik-download-permenpan-no-3-tahun-2015/>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45 , 427-437.
- Statsoft. (2015). *Naive Bayes Clasifier Introductory Overview*. Retrieved April 22, 2015, from Statsoft Web Site: <http://www.statsoft.com/textbook/naive-bayes-classifier>
- Statsoft. (2015). *Text Mining Introductory Overview*. Retrieved April 19, 2015, from Statsoft: <http://www.statsoft.com/textbook/text-mining>
- Suh, J. H., Park, C. H., & Jeon, S. H. (2010). Applying text and data mining techniques to forecasting the trend of petitions filed to e-people. *Expert Systems with Applications* 37, 7255-7268.
- Triawati, C. (2009). *Text Mining*. Retrieved May 19, 2015, from Digital Library Telkom Institute of Technology: http://digilib.tes.telkomuniversity.ac.id/index.php?option=com_content&view=article&id=590:text-mining&catid=20:informatika&Itemid=14
- United Nation. (2014). *UNITED NATIONS E-GOVERNMENT SURVEY 2014*. New York: United Nations.
- Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guestexperience and satisfaction? *International Journal of Hospitality Management* 44, 120-130.
- Zhai, C., & Aggarwal, C. C. (2012). *Mining Text Data*. New York: Springer.