ANALISIS FREKUENSI KATA DAN FREKUENSI PASANGAN KATA PADA SOCIAL MEDIA MENGGUNAKAN TEKNIK TEXT MINING

ANALYSIS OF WORD FREQUENCY AND WORD PAIRS FREQUENCY ON SOCIAL MEDIA USING TEXT MINING TECHNIQUES

NUR ISLAMUDDIN
2018130020

Pembimbing I: Dr. Eng. Yuyun, M.T

Pembimbing II: Dr. Eng. Ir. Zulfajri B. Hasanuddin, M.Eng

PROGRAM PASCASARJANA

STMIK HANDAYANI

MAKASSAR

2020

ANALISIS FREKUENSI KATA DAN FREKUENSI PASANGAN KATA PADA SOCIAL MEDIA MENGGUNAKAN TEKNIK TEXT MINING

Tesis

Sebagai Salah Satu Syarat untuk Mencapai Gelar Magister

Program Studi

Sistem Komputer

Disusun dan diajukan oleh

NUR ISLAMUDDIN

Kepada

PROGRAM PASCASARJANA

STMIK HANDAYANI

MAKASSAR

2020

TESIS

ANALISIS FREKUENSI KATA DAN FREKUENSI PASANGAN KATA PADA SOCIAL MEDIA MENGGUNAKAN TEKNIK TEXT MINING

Disusun dan diajukan oleh

NUR ISLAMUDDIN

2018130020

Telah dipertahankan di depan Panitia Ujian Tesis

Pada tanggal 30 September 2020

Dan dinyatakan telah memenuhi syarat

Menyetujui

Komisi penasehat,

Dr. Eng. Yuyun, M.T	**Dr. Eng. Ir. Zulfajri B. Hasanuddin, M.Eng**
Ketua	Anggota

Ketua Program Studi Sistem Komputer	Direktur Program Pascasarjana STMIK Handayani
Prof. Dr. Ir. Andani Achmad, M.T	**Dr. Eng. Yuyun, M.T**

HALAMAN PERSETUJUAN PERBAIKAN
UJIAN AKHIR MAGISTER

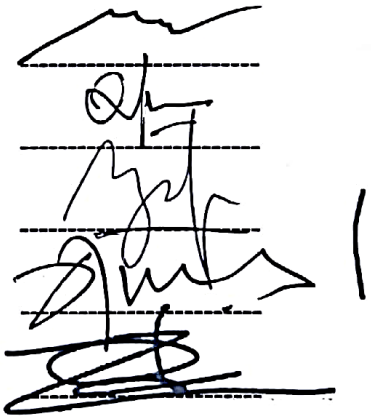
Pada hari Sabtu, tanggal 30 September 2020 telah dilaksanakan Ujian Akhir mahasiswa:

Nama Mahasiswa	: Nur Islamuddin
Nomor Pokok	: 2018130020
Jenjang Pendidikan	: S2 (Magister)
Program Studi	: Sistem Komputer
Judul Penelitian	: Analisis Frekuensi Kata dan Frekuensi Pasangan Kata Pada Social Media Menggunakan Teknik Text Mining

Hasil ujian menyepakati bahwa sebelum penggandaan tesis, yang bersangkutan harus menyempurnakan tesisnya sesuai saran dan masukan yang muncul pada ujian tersebut.

Hasil penyempurnaan tesis tersebut ditunjukkan kepada Panitia Ujian Akhir, dan dinyatakan selesai jika Panitia Ujian Akhir menandatangani persetujuan di bawah ini

Panitia Ujian Akhir			Tanda Tangan
Ketua	:	Dr. Eng. Yuyun, M.T	
Sekretaris	:	Dr. Eng. Ir. Zulfajri B. Hasanuddin, M.En	
Anggota	:	1. Prof. Dr. Ir. Syafruddin Syarif, MT	
		2. Dr. Imran Taufik. ST, M.Si	
		3. Adnan, ST, MT, Ph.D.	



Mengetahui
Ketua Program Studi,

Prof. Dr. Ir. Andani Achmad, M.T
NIP. 19601231 198703 1 022

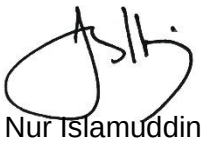
PERNYATAAN KEASLIAN TESIS / DISERTASI

Yang bertanda tangan di bawah ini:

Nama	: Nur Islamuddin
Nomor Mahasiswa	: 2018130020
Program Studi	: Sistem Komputer

Menyatakan dengan sebenarnya bahwa tesis/disertasi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan tulisan atau pemikiran orang lain. Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan tesis/disertasi ini hasil karya orang lain, saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 30 September 2020
Yang menyatakan,



Nur Islamuddin

PRAKATA

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa dengan selesainya tesis ini. Tak lupa pula shalawat dan salam kepada Nabi Muhammad SAW yang telah menyinari dunia ini dengan keindahan ilmu dan akhlak yang diajarkan kepada seluruh umatnya.

Gagasan yang melatari penelitian ini timbul dari pengamatan penulis terhadap perkembangan penggunaan media sosial dan data yang beredar dari penggunanya. Data-data tersebut dapat diolah menjadi sumber informasi yang kemudian dapat dimanfaatkan dan digunakan oleh pelaku usaha atau pihak yang membutuhkan sebagai pengambilan keputusan untuk perkembangan bisnis atau usaha yang dijalankan.

Dalam penyusunan tesis ini, penulis menyadari banyak mengalami tantangan dan hambatan namun berkat dukungan serta kerja sama yang baik dengan berbagai pihak sehingga tesis ini dapat diselesaikan. Untuk itu penulis mempersembahkan ucapan terima kasih teristimewa dan tak terhingga kepada kedua orang tua penulis yang pada masa hidupnya, dukungan dan kasih sayang senantiasa mereka berikan kepada penulis. Namun dalam masa proses penyusunan tesis ini, mereka berdua dipanggil oleh Yang Maha Kuasa. Almarhumah Ibunda yang terkasih Siti Rahmatia dan Almarhum Ayahanda Muhammad Nur Yusuf, semoga mereka diberikan tempat terbaik disisi-Nya. Doa terbaik senantiasa penulis haturkan untuk kalian berdua. Dan kepada saudara-saudari saya Nur Rahman, Nur Hasyim, Siti Dahlia Nur, Nur Ilham, Sahrul Nur dan Nur Dahniar serta kepada istri tercinta saya, Riznah Rizal dan anak-anak saya Andi Muhammad Dzaky Almer dan Aleeyza Renata Putri, terima kasih atas segala

dedikasi yang tak terhingga, kasih sayang, doa restu beserta dukungan moril maupun materil kepada penulis dalam menyelesaikan tesis ini.

Ucapan terima kasih pun penulis hanturkan kepada kepada Ketua Komisi Penasihat, Dr. Eng. Yuyun, M.T dan Dr. Eng. Ir. Zulfajri B. Hasanuddin, M.Eng., sebagai anggota komisi penasehat yang telah meluangkan waktunya kepada penulis untuk membimbing dan berkonsultasi tentang materi dalam tesis ini dan juga kepada Prof. Dr. Ir. Andani Ahmad., M.T., selaku ketua program studi Sistem Komputer serta seluruh dosen dan staf program studi Sistem Komputer Pascasarjana, STMIK Handayani Makassar yang telah membantu dalam hal keilmuan maupun administrasi pada tahap tesis ini.

Penulis juga mengucapkan terimakasih kepada teman-teman seperjuangan “Kelas Kendari” pascasarjana STMIK Handayani Makassar angkatan 2018, Ilin Sukma, Omar Wahid, Faizal Aris, Sukirno Kasau, Asmira, Nilam Kusumawati, Andi Iwan, Aris Susanto, La Ija, Laode Bakrim, Salam, Jufer, yang selalu menjadi tempat berdiskusi, mengajarkan arti kekeluargaan dan persaudaraan serta selalu memberikan dukungan dan motivasi dalam segala bentuk sehingga penulis bisa menyelesaikan tesis ini.

Makassar, Desember 2020

Nur Islamuddin

ABSTRAK

NUR ISLAMUDDIN. Analisis frekuensi kata dan frekuensi pasangan kata pada *social media* menggunakan teknik *text mining* (dibimbing oleh Yuyun dan Zulfajri B. Hasanuddin).

Text mining merupakan sebuah teknik penemuan pengetahuan yang digunakan untuk mengekstrak pola-pola menarik dan yang tidak biasa dari bahasa alami. Dengan memanfaatkan alat dan sumber daya canggih dari ilmu komputer dan linguistik komputasi, penelitian ini bertujuan untuk menganalisa informasi dari media sosial Twitter dengan menghitung frekuensi kemunculan sebuah kata dalam bentuk topik menggunakan metode *LDA (Latent Dirichlet Allocation)* berdasarkan *Topic Modeling* dan menghitung frekuensi kemunculan pasangan kata menggunakan metode *Word Pairs Frequency*.

Hasil dari penelitian ini menunjukkan perhitungan frekuensi kata dengan menggunakan metode Topic Modeling dengan LDA, terbentuk 3 topik dengan skor koherensi tertinggi yaitu 0.665045. Pada setiap topik juga diperoleh kata dengan bobot nilai tertinggi yaitu pada kata atau istilah “lagu_semoga” dengan bobot 0.129. Frekuensi pasangan kata tertinggi yaitu 215.0 pada pasangan kata "semoga" dan "semangat".

ABSTRACT

NUR ISLAMUDDIN. Analysis of word frequency and word pairs frequency on social media using text mining techniques (dibimbing oleh Yuyun dan Zulfajri B. Hasanuddin).

Text mining is a knowledge discovery technique used to extract interesting and unusual patterns from natural language. By utilizing sophisticated tools and resources from computer science and computational linguistics, this study aims to analyze information from social media Twitter by calculating the frequency of occurrence of a word in the form of a topic using the LDA (Latent Dirichlet Allocation) method based on Topic Modeling and calculating the frequency of occurrence of word pairs. using the Word Pairs Frequency method.

The results of this study indicate that the calculation of word frequency using the Topic Modeling method with LDA, formed 3 topics with the highest coherence score of 0.665045. In each topic, the word with the highest value weight is also obtained, namely the word or term "lagu_semoga" with a weight of 0.129. The highest word pair frequency is 215.0 in the word "semoga" and "semangat".

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR KODE

BAB I
PENDAHULUAN

A. LATAR BELAKANG MASALAH

Media sosial merupakan teknologi berbasis internet yang dimanfaatkan oleh banyak orang dalam satu garis waktu tertentu yang menciptakan koneksi antar penggunanya dan memungkinkan terbentuknya suatu komunitas tertentu (Ardi & Sukmawati, 2017). Teknologi media sosial juga dimanfaatkan oleh beberapa pelaku bisnis (Cakranegara & Susilowati, 2017; Juditha, 2017; Situmorang et al., 2018) dan pemerintah (Agustina, 2018; Setiawan & Santoso, 2013). Selain untuk memahami apa yang publik inginkan tentang produk dan layanan yang mereka tawarkan, juga untuk dapat mengumpulkan, mengambil, dan menyimpan semua informasi yang terkait dengan peristiwa dan perkembangan yang terjadi seiring waktu yang berjalan (Maynard et al., 2012).

Indonesia adalah salah satu negara dengan pengguna media sosial terbanyak. Berdasarkan hasil survei yang dilakukan *We Are Social* untuk Indonesia yang terbit pada tanggal 18 Februari 2020, total pengguna aktif media sosial mencapai 160 juta pengguna dengan pertambahan dari tahun sebelumnya mencapai 12 juta pengguna per April 2019 hingga Januari 2020 (We Are Social, 2020). Hasil survei juga menunjukkan *platform* media sosial yang populer digunakan oleh pengguna dengan umur antara 16 tahun hingga 64 tahun, salah satunya adalah *Twitter* dengan persentase mencapai 52% dari total pengguna media sosial di Indonesia (We Are Social, 2020).

Twitter adalah salah satu jejaring sosial yang didirikan pada Maret 2006, merupakan media yang memungkinkan orang untuk berbagi informasi maupun pendapat pribadi mereka yang menyatukan ratusan juta pengguna dengan

konsep *microblogging* yang minimalis, ditambah dengan *Application Programming Interface (API)* yang sangat terbuka, menjadikannya media yang ideal untuk menambah pengetahuan (Grandjean, 2016). Dengan layanan *microblogging,* pengguna memposting tentang cerita kehidupan sehari-hari mereka dan berdiskusi mengenai pendapat pribadi pada berbagai topik, mulai dari hal yang sederhana seperti perbincangan mengenai produk, acara, dan layanan hingga mengenai hal yang lebih kompleks yang berhubungan dengan masalah ekonomi, minat, budaya, politik, agama, penyakit, kelaparan, dan sebagainya (Rana, 2015). Perbincangan pengguna *Twitter* bukan hanya pesan biasa, tetapi dapat dijadikan sumber informasi yang bermanfaat, karena mengandung pendapat pribadi, emosi dan ekspresi tentang berbagai topik (Rana, 2015).

Kota Makassar, Indonesia, menjadi fokus dari penelitian ini. Berdasarkan sensus penduduk tahun 2010 Kota Makassar, jumlah penduduk mencapai 1,3 juta jiwa dengan persentase jumlah penduduk yang berusia antara 15 hingga 64 tahun sebanyak 68,73% (Badan Pusat Statistik, 2010). Pada tanggal 24 Agustus hingga 5 Oktober 2016, tercatat 170.595 data check-in dari pengguna *Twitter* di kota tersebut (Yuyun et al., 2017).

Penelitian ini berfokus untuk menaganalisa frekuensi teks dengan memanfaatkan teknik Text Mining dari sejumlah data pada media sosial Twitter dengan menggunakan metode Topic Modeling dengan Latent Dirichlet Allocation (LDA) dan metode Word Pairs Frequency.

Beberapa penelitian yang berkaitan dengan masalah di atas telah dilakukan sebelumnya. Penelitian yang dilakukan Mehmet F. Dicle (2018) mengungkapkan tentang pentingnya mendistribusi frekuensi setiap kata dari

informasi yang dapat diperoleh di internet (Dicle & Dicle, 2018). Analisis frekuensi kata telah banyak digunakan pada bidang-bidang tertentu, seperti dalam ilmu politik dengan membandingkan efisiensi metode tradisional dengan metode analisis frekuensi kata (Laver et al., 2003) dan menganalisa bahasa politik menggunakan analisis frekuensi kata (Buchanan & Padfield, 2019). Pada bidang psikologi, analisis frekuensi kata memberikan ringkasan tentang bagaimana sistem otomatis komputer mengidentifikasi psikopatologi, kejujuran, status, jenis kelamin, atau usia (Pennebaker & Chung, 2013) dan menunjukkan fokus perhatian, emosionalitas, hubungan sosial, gaya berpikir, dan perbedaan individu (Tausczik & Pennebaker, 2009).

Menganalisa frekuensi kata juga dilakukan dengan menggunakan metode seperti data mining dan corpus linguistics (Lijffijt et al., 2011). Ni Wayan Sri Arini dan Ida Bagus Putu Widja (2019) pada penelitiannya mengungkapkan bahwa kegiatan pemilahan artikel ilmiah yang biasanya dilakukan secara manual dapat dilakukan secara komputasi dengan memanfaatkan hasil analisa frekuensi kata dengan menerapkan algoritma string similarity, yaitu dengan mencari kata-kata kunci yang terdapat dalam sebuah karya ilmiah (Sri Arini et al., 2019).

Menurut penelitian Jelodar dkk (2019), topic modeling merupakan salah satu pendekatan pada Text Mining yang cukup handal dalam melakukan penemuan data-data teks. Salah satunya yaitu menghitung frekuensi kata berdasarkan topik (Jelodar et al., 2019). Ada berbagai metode yang dapat digunakan untuk pemodelan topik, salah satunya yaitu metode Latent Dirichlet Allocation (LDA) yang merupakan salah satu metode pemodelan topik yang paling populer (Jelodar et al., 2019).

Analisis trend pada media sosial dapat diukur dengan menggunakan pemodelan topik dengan LDA (Lau et al., 2012). Pemodelan topik dengan LDA juga dapat mengidentifikasi tema artikel ilmiah pada Organizational Research Methods (Piepenbrink & Gaur, 2017), mendeteksi topik dalam pelacakan konten perbincangan (Yeh et al., 2016) dan mampu berkerja dengan baik untuk dokumen dengan konten yang panjang maupun dokumen dengan konten yang pendek (Tong & Zhang, 2016).

Penelitian tentang penambangan teks juga dilakukan oleh Agatha Deolika dkk (2019) dengan melakukan analisa pada kata dengan pembobotan dan klasifikasi. Dengan klasifikasi Naïve bayes lebih baik dari pembobotan TF.IDF dan WIDF dengan nilai Accuracy 98,67%, Precision 93,81%, dan Recall 96,67% dan berkesimpulan bahwa klasifikasi naïve bayes dapat digunakan untuk mengelompokan atau klasifikasi text mining dengan baik (Deolika et al., 2019).

Penelitian oleh Edi Surya Negara (2016) tentang ekstraksi data *Twitter* dan analisis data geospasial yang dilakukan terhadap isu publik yang sedang berkembang menggunakan empat tahapan proses yaitu, *crawling, storing, analyzing* dan *visualizing,* didapatkan hasil bahwa berdasarkan data yang ditarik, dapat diketahui informasi pengguna *Twitter* berupa lokasi, negara asal, jenis kelamin dan lain-lain (Negara et al., 2016).

Penelitian yang dilakukan oleh S. Siddharth (2018) menggunakan algoritma *Machine Learning* untuk mengekstraksi pendapat dan sentimen pengguna *Twitter* dengan bahasa pemrograman *Python* dengan fokus pada tren bahasa *tweet* yang berbeda (Siddharth et al., 2018). Hasil evaluasi menunjukkan bahwa mengklasifikasi data dengan menerapkan algoritma

Machine Learning pada bahasa pemrograman Python yang diusulkan lebih efisien dan berkinerja lebih baik dalam hal akurasi dan waktu (Siddharth et al., 2018).

Sampai saat ini kami belum menemukan penelitian tentang Pembentukan Pasangan Kata untuk menentukan topik dari sejumlah dokumen. Kebanyakan yang ada yaitu analisis pada sebuah kata. Kami berpendapat bahwa teks dengan dua kata yang saling berhubungan lebih mudah dipahami dibandingkan dengan teks dengan satu kata. Tujuan dari penelitian ini adalah untuk menganalisa frekuensi kata di social media *twitter* dengan teknik topic modeling dan frekuensi pasangan kata menggunakan metode Word Pairs Frequency.

B. RUMUSAN MASALAH

Berdasarkan latar belakang di atas, maka dapat disimpulkan rumusan masalahnya sebagai berikut:

1. Bagaimana menganalisis frekuensi kata postingan text pada sosial media Twitter menggunakan metode Topic Modeling?
2. Bagaimana menganalisis frekuensi pasangan kata postingan text pada sosial media Twitter menggunakan metode Word Pairs Frequency?

C. BATASAN MASALAH

Dari permasalah yang diuraikan di atas, batasan masalah dalam penelitian ini adalah:

1. Media sosial yang digunakan adalah Aplikasi Media Sosial Twitter.
2. Penelitian ini tidak membahas mengenai kata penghubung, kata keterangan, tanda baca, imbuhan, simbol-simbol dan angka.

D. TUJUAN PENELITIAN

Berdasarkan latar belakang dan rumusan masalah di atas, tujuan penelitian ini antara lain untuk:

1. Melakukan analisis pada postingan teks data Twitter dengan Topic Modeling untuk mengetahui frekuensi kata.
2. Melakukan analisis pada postingen teks data Twitter dengan Word Pairs Frequency untuk mengetahui frekuensi pasangan kata.

E. MANFAAT PENELITIAN

Manfaat yang dapat diperoleh dari penelitian ini adalah:

1. Dapat mengetahui frekuensi kata dari data Twitter dengan menggunakan metode Topic Modeling.
2. Dapat mengetahui frekuensi pasangan kata dari data Twitter dengan menggunakan metode Word Pairs Frequency.

F. RUANG LINGKUP PENELITIAN

Penelitian ini terbatas hanya pada proses menganalisa frekuensi kata dan pasangan kata pada sosial media Twitter di Kota Makassar menggunakan teknik Text Mining.

G. SISTEMATIKA PENULISAN

Sistematika penulisan pada penelitian ini adalah:

BAB I Pendahuluan

Bab I berisi tentang latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup penelitian dan sistematika penulisan.

BAB II Landasan Teori

Bab II berisi tentang teori apa saja yang digunakan serta gambaran tentang kerangka konseptual dari penelitian ini.

BAB III Metode Penelitian

Bab III merupakan penjelasan tentang metode yang digunakan dalam penelitian ini yang dijabarkan pada poin rancangan penelitian, lokasi dan waktu, populasi dan teknik sampel, instrumen pengumpulan data serta analisis data.

BAB IV Hasil Penelitian dan Pembahasan

Bab IV ini berisi penjelasan hasil dan pembahasan penelitian serta implikasi dari penelitian yang dilakukan serta analisa dari hasil yang diperoleh.

BAB V Penutup

Bab V ini berisi penjelasan ringkasan temuan, rangkuman kesimpulan yang berkaitan dengan analisa dan optimalisasi sistem berdasarkan yang telah diuraikan pada bab-bab sebelumnya dan saran dari penulis untuk tahap pengembangan selanjutnya.

BAB II
TINJAUAN PUSTAKA

A. TWITTER

Twitter adalah salah satu jejaring sosial yang didirikan pada Maret 2006, merupakan media yang memungkinkan orang untuk berbagi informasi maupun pendapat pribadi mereka yang menyatukan ratusan juta pengguna dengan konsep *microblogging* yang minimalis. (Grandjean, 2016). Pada Januari 2019, pengguna aktif *Twitter* mencapai 326 juta pengguna yang tersebar di seluruh dunia (We Are Social, 2020).

Pesan teks di *Twitter* terdiri dari 140 karakter yang disebut dengan *Tweet* memungkinkan pengguna untuk menyampaikan pendapat dan pemikiran secara efisien, ditambah dengan *Application Programming Interface (API)* yang sangat terbuka, menjadikannya media yang ideal untuk menambah pengetahuan (Grandjean, 2016; Hu, 2013).

Twitter menyediakan akses *programatik* melalui *Application Programming Interface (API)* data *Twitter* kepada pihak pengembang dan pengguna yang membutuhkan informasi secara global (Twitter, 2020).

Informasi yang disajikan merupakan cara program komputer “berbicara” sebagai cara untuk dapat meminta dan memberikan informasi. Proses tersebut dilakukan dengan memberikan akses kepada aplikasi perangkat lunak yang disebut dengan *endpoint* atau alamat yang terkait dengan informasi yang disajikan berupa nomor unik milik pengguna seperti nomor telepon (Twitter, 2020).

B. TEXT MINING

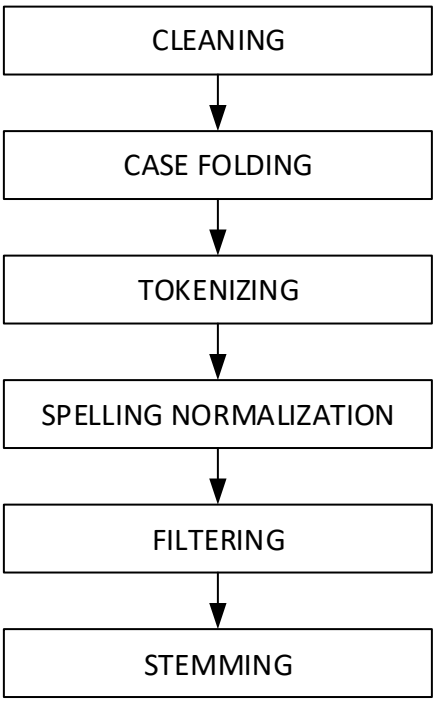
Text mining merupakan sebuah teknik penemuan pengetahuan yang digunakan untuk mengekstrak pola-pola menarik dan yang tidak biasa dari bahasa alami, yang terdiri dari bidang multidisiplin, seperti pencarian informasi, teks analisis, pemrosesan bahasa alami dan klasifikasi informasi berdasarkan pola yang logis dan tidak biasa dari kumpulan data yang besar (Irfan et al., 2015).

Teknik *Text Mining* memiliki beberapa tahapan yang penting, salah satunya yaitu tahap *pre-processing* (Irfan et al., 2015). Tahap *pre-processing* mengatur dokumen sesuai dengan kategori yang telah ditentukan, dengan menjamin keberhasilan dalam analisis teks, tetapi membutuhkan waktu proses yang cukup besar (Irfan et al., 2015).

Secara umum, text mining memiliki dua tahap, yaitu text pre-processing dan *feature selection* (Feldman & Sanger, 2007). Berikut penjelasan dari tahapan-tahapan tersebut:

1. Text Preprocessing

Text processing merupakan tahapan pertama yang digunakan untuk memproses teks menjadi data yang akan diolah (Feldman & Sanger, 2007) yang dikutip oleh (Indranandita, Amelia, & dkk, 2008). Langkah-langkah yang dilakukan dari tahapan *text preprocessing* ini dapat dilihat pada gambar 1.



Gambar 1. Tahapan Text Processing

Tahapan *preprocessing* dimulai dari proses *cleaning*, yaitu proses untuk membersihkan *tweet* dari kata-kata yang tidak diperlukan untuk mengurangi *noise*. Kata atau karakter yang akan dihilangkan adalah karakter atau simbol, HTML, hashtag (#), username atau mention (@username), link url (http://situs.com), emoticon, dan RT (tanda retweet). Setelah proses cleaning dilakukan, tahapan selanjutnya ialah case folding. Dimana proses ini melakukan penyeragaman bentuk huruf dengan mengubah semua huruf menjadi huruf besar atau huruf kecil, kemudian hanya menggunakan huruf a sampai z. Pada proses case folding ini juga akan menghilangkan tanda baca dan angka.

Kemudian dilakukan proses tokenizing. Pada proses ini tweet atau kalimat akan dipecah menjadi sebuah kata dari sekumpulan data, dengan

memisahkan kata tersebut dan menentukan struktur sintaksis setiap kata tersebut.

Selanjutnya dilakukan proses spelling normalization. Tahapan ini adalah tahapan yang mengidentifikasi penulisan kata berlebihan dan kata silang kemudian diganti dengan kata kamus KBBI (Rosdiansyah, 2014). Setiap kata yang dijumpai dan penggunaan hurufnya berlebihan dan tidak baku akan diubah.

2. Transformation

Tahapan transformation dilakukan dengan memberikan nilai kemunculan dari setiap kata pada setiap dokumen yang diproses. Pada text transformation ini yang dilakukan adalah menentukan *Term Frequency* (TF), yaitu penentuan jumlah frekuensi kemunculan kata pada dokumen yang diolah (Hadna & Paulus Insap Santosa, 2016).

3. Penggalian 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, dkk, 2012 dikutip oleh Chyntia, 2015).

a. Klasifikasi (*Clasification*)

   Klasifikasi adalah bentuk analisis data yang mengekstrak model untuk menggambarkan kelas data (Jiawei, Kamber, & Pei, 2012 dikutip oleh Chyntia, 2015).

b. Pengelompokan (*Clustering*)

Pada model clustering pengelompokan data dilakukan dengan menggunakan algoritma yang sudah ditentukan dan data akan diproses oleh algoritma untuk dikelompokkan menurut karakteristik alaminya.

c. Asosiasi (*Asociation*)

Asosiasi merupakan proses pencarian hubungan antar elamen data. Dalam dunia industri retail, analisis asosiasi biasanya disebut market Basket Analysis (Miner, dkk, 2012 dikutip oleh Chyntia, 2015).

d. Analisis Tren

Tujuan dari analisis tren yaitu untuk mencari perubahan suatu objek atau kejadian oleh waktu (Miner, dkk, 2012 dikutip oleh Chyntia, 2015).

C. TOPIC MODELING

Konsep Topic modeling menurut Blei terdiri dari entitas-entitas yaitu “kata”, “dokumen”, dan “corpora”. “Kata” memiliki pengertian sebagai sederetan huruf yang berada di antara dua spasi dan memiliki sebuah arti. “Kata” juga dianggap sebagai unit dasar dari data diskrit dalam dokumen, didefinisikan sebagai item dari kosa kata yang diberi indeks untuk setiap kata unik pada dokumen. “Dokumen” adalah susunan N kata-kata. Sebuah corpus adalah kumpulan M dokumen dan corpora merupakan bentuk jamak dari corpus. Sementara “topik” adalah distribusi dari beberapa kosakata yang bersifat tetap. Secara sederhana, setiap dokumen dalam corpus mengandung proporsi tersendiri dari topik-topik yang dibahas sesuai kata-kata yang terkandung di dalamnya.

Ide dasar dari Topic modeling adalah bahwa sebuah topik terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen

memiliki kemungkinan terdiri dari beberapa topik dengan probabilitas masing-masing. Namun secara pemahaman manusia, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik, distribusi topik per-dokumen, dan penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi, maka dari itu Topic modeling bertujuan untuk menemukan topik dan kata-kata yang terdapat pada topik tersebut.

D. LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) merupakan metode Topic modeling dan topik analisis yang paling populer saat ini. LDA muncul sebagai salah satu metode yang dipilih dalam melakukan analisis pada dokumen yang berukuran sangat besar. LDA dapat digunakan untuk meringkas, melakukan klasterisasi, menghubungkan maupun memproses data yang sangat besar karena LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen. Adapun distribusi yang digunakan untuk mendapatkan distribusi topik per-dokumen disebut distribusi Dirichlet, kemudian dalam proses generatif untuk LDA, hasil dari Dirichlet digunakan untuk mengalokasikan kata-kata pada dokumen untuk topik yang berbeda. Dalam LDA, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik, distribusi topik perdokumen, penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi3 , maka dari itu, Algoritma ini dinamakan Latent Dirichlet Allocation (LDA).

E. HIPOTESIS

Hipotesis pada penelitian ini adalah:

H_1 = Dalam postingan status media sosial terdapat kata dengan frekuensi tinggi, sedang dan sedikit bahkan hanya sekali.

H_2 = Dengan pendekatan Word Pairs, terdapat pasangan kata yang saling berhubungan maupun tidak berhubungan.

F. DEFENISI OPERASIONAL VARIABEL

Tabel 1. Defenisi Operasional Variabel

No	Variabel	Defenisi Operasional
1	Kata	Merupakan kumpulan huruf yang diapit oleh dua spasi yang telah melalui tahap pra-pemrosesan data.
2	Istilah	Merupakan kumpulan dua kata atau lebih yang dipisahkan oleh simbol garis bawah yang telah melalui tahap pra-pemrosesan data.
3	Pasangan Kata	Merupakan perpaduan dari dua kata atau istilah yang dipisahkan oleh sebuah spasi yang telah melalui tahap pra-pemrosesan data.

G. PENELITIAN TERKAIT

Pada tabel berikut dapat dilihat beberapa penelitian sebelumnya tentang *Text Mining Data Twitter*:

Tabel 2. Penelitian Terkait

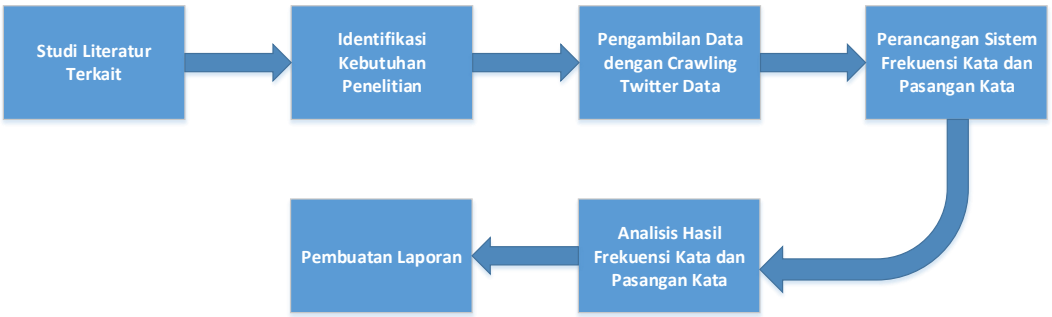
No	Judul	Penulis dan Tahun	Metode	Hasil
1	Content analysis: Frequency distribution of words	Mehmet F. Dicle; Betul Dicle (2018)	*Wordfreq (Word Frequency)*	Analisis konten dengan metode Wordfreq menjadi acuan utama dalam analisis frekuensi kata.
2	Analisis Frekuensi Kata untuk Mengekstrak Kata Kunci dari Artikel Ilmiah Berbahasa Indonesia	Ni Wayan Sri Arini; Ida Bagus Putu Widja (2019)	Analisa Frekuensi Kata	Kata kunci sesuai dengan kata-kata yang sering muncul

No	Judul	Penulis dan Tahun	Metode	Hasil
3	Analisis Pembobotan Kata pada Klasifikasi Text Mining	Agatha Deolika; Kusrini; Emha Taufiq Luthfi (2019)	Klasifikasi Text Mining	Klasifikasi algoritma naïve bayes dapat digunakan untuk klasifikasi Text Mining
4	Analysis of Urban Population Using Twitter Distribution Data: Case Study of Makassar City, Indonesia	Yuyun Wabula; B.J. Dewancker (2016)	*Collection Data Twitter* dan Kuisioner	Adanya korelasi antara geolokasi *Twitter* dan data kuisioner
5	Location reference identification from tweets during emergencies: A deep learning approach	Abhinav Kumar; Jyoti Prakash Singh (2019)	*Convolutional Neural Network (CNN)*	Skor pencocokan di atas 92% untuk *tweet* yang terkait dengan gempa bumi
6	Analisis Data Twitter: Ekstraksi dan Analisis Data G eospasial	Edi Surya Negara; Ria Andryani; Prihambodo Hendro Saksono (2016)	*Social Media Analytics*	Diketahui informasi pengguna berupa lokasi, negara asal, jenis kelamin dan rentang usia

BAB III
METODE PENELITIAN

A. TAHAPAN PENELITIAN

Penelitian Identifikasi Teks dan Lokasi Populer berdasarkan Data Twitter ini terdiri dari beberapa tahapan penelitian, sebagai berikut:



Gambar 2. Diagram Tahapan Penelitian

Tahapan penelitian pada Gambar 2 diuraikan sebagai berikut:

1. Studi literatur. Pada studi literatur, dilakukan pencarian penelitian-penelitian terkait analisis data twitter menggunakan metode N-Gram dan Topic Model LDA (Latent Dirchlect Allocation). Pada tahap ini juga dilakukan pencarian dokumentasi hasil penelitian-penelitian sebelumnya.
2. Identifikasi kebutuhan penelitian. Pada tahap ini, dilakukan penetapan berbagai kebutuhan penelitian untuk menunjang kegiatan penelitian.
3. Pengambilan data dilakukan dengan mengambil data twitter.
4. Perancangan Sistem Frekuensi Kata dan Pasangan Kata. Pada tahap ini dirancang sebuah sistem untuk menampilkan frekuensi kemunculan sebuah kata berdasarkan topik menggunakan Topic Modeling dengan Latent Dirichlet Allocation (LDA) dan frekuensi kemunculan pasangan kata menggunakan teknik Word Pairs Frequency.

5. Analisis Hasil Frekuensi Kata dan Pasangan Kata. Pada tahap ini dilakukan analisis dan pengujian hipotesis.
6. Pembuatan laporan. Setelah melewati semua tahapan, proses akhir adalah menuliskan laporan penelitian menyeluruh sebagai bahan publikasi dan penyusunan naskah tugas akhir magister.

B. LOKASI DAN WAKTU PENELITIAN

1. Lokasi Penelitian

Penelitian ini dilakukan di Kota Makassar, Sulawesi Selatan, Indonesia.

2. Waktu Penelitian

Penelitian dilaksanakan selama 7 bulan dimulai bulan Maret 2020 hingga bulan September 2020.

C. JENIS PENELITIAN

Jenis penelitian ini merupakan penelitian eksperimental yang bersifat analisis sehingga dari ruang lingkup masalah dapat dilakukan dengan metode studi pustaka (library research), metode pengumpulan data (field research) dan perancangan sistem serta analisis.

D. SUMBER DATA

Dalam penelitian ini, data latih yang digunakan sebanyak 2000 data tweet yang bersumber dari data pengguna media sosial *Twitter* yang berada di Kota Makassar tanggal 8 Agustus 2020 hingga tanggal 21 Agustus 2020.

E. INSTRUMEN PENELITIAN

1. Software

a. Windows 10 Pro 64-bit

b. JupyterLab 1.0

c. Microsoft Office Excel

2. Hardware

Laptop Asus Model X550VX Nvidia Geforce GTX 950M dengan Processor Intel Core i7-6700HQ, RAM 4GB.

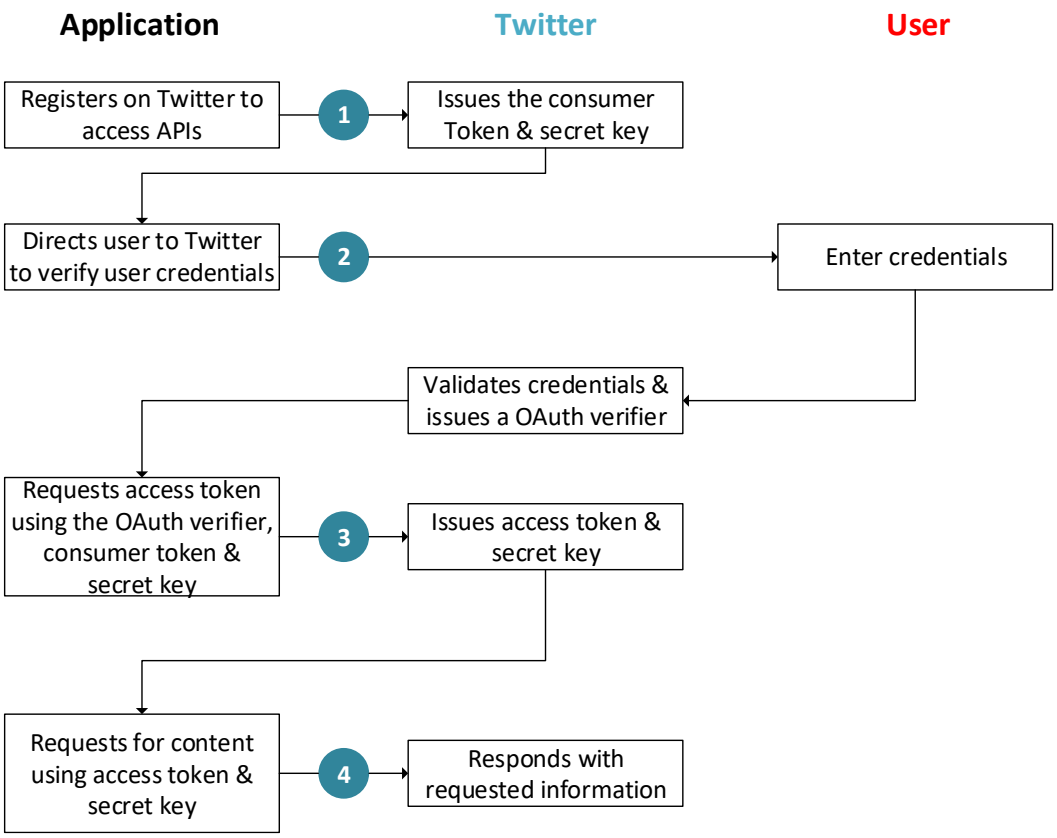
F. RANCANGAN PENELITIAN

Alat, komponen, teknik dan metode digunakan sebagai solusi untuk penyelesaian permasalahan yang dijabarkan pada bab 1 di atas. Tahapan yang harus dilakukan untuk mengidentifikasi data *Twitter* yang dibutuhkan yaitu:

1. *Data Collection* (Mengumpulkan data)

Proses ini dilakukan untuk mengumpulkan data teks berdasarkan *tweet* yang di *posting* oleh pengguna *Twitter* dengan memanfatkan *Application Programming Interface* (*API*) yang telah disediakan oleh *Twitter*. Data lokasi berupa titik koordinat juga dihasilkan dari proses penarikan data berdasarkan proses *check-in location* oleh pengguna *Twitter*. Pada penelitian ini, proses penarikan data menggunakan bahasa pemrograman *Python* dan Microsoft Office Excel untuk tempat penyimpanan data yang telah di kumpulkan.

Pada gambar 3 memperlihatkan proses *crawling* data twitter secara rinci dengan mengakses ke API Twitter.



Gambar 3. Proses Akses Twitter APIs

(Sumber: https://image.slidesharecdn.com/twitterapi-160308173613/95/twitter-api-7-638.jpg?cb=1457458688)

2. *Pre-Processing* (Pra-pemrosesan)

Setelah data dikumpulkan dari *Twitter*, langkah selanjutnya adalah pra-pemrosesan yang diimplementasikan menggunakan bahasa pemrograman *Python*. Tahap ini untuk membersihkan dan mengurai data *tweet*.

Tahapan yang dilakukan pada Twitter Data *Pre-Processing* sebagai berikut:

a. Basic Data Cleaning

Tahap ini dilakukan untuk menghapus bagian yang tidak penting dari *tweet*. Pembersihan *Data Twitter* antara lain:

- Alamat *Uniform Resource Locators (URLs)*,
- Jumlah spasi yang berlebihan dan menggantinya dengan spasi tunggal,
- Tanda baca, angka-angka, dan karakter khusus, dan
- Konversi semua huruf kapital menjadi huruf kecil.

b. Removal of Stopwords

Tahap Removal of Stopwords digunakan sebagai kata-kata acuan yang untuk menemukan kata-kata tersebut pada dokumen yang disusun berdasarkan penelitian Fadillah Z Tala. Beberapa contoh daftar stopwords yang tersimpan dalam daftar yang dimaksud (Tala, 2003) tercantum pada Tabel 3.

Tabel 3. Contoh daftar stopwords berdasarkan penelitian Fadillah Z Tala

Daftar Stopwords
'bagaimana', 'mana', 'agak', 'cukup', 'ada', 'bahwa', 'cuma', 'demikian', 'amatlah', 'dapat', 'entah', 'hal', 'hendak', 'ialah', 'guna', 'saat', 'ini', 'jika', 'juga', 'kira', 'lalu', 'namun', 'memang', 'oleh', 'pasti', 'para', 'sangat', 'tanpa', 'tiap', 'yakni', 'yaitu'

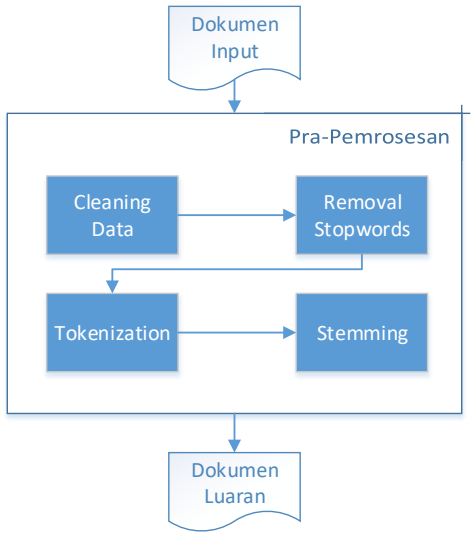
c. Tokenization

Tokenization adalah aktivitas atau proses memisahkan deretan kata di dalam kalimat atau paragraf menjadi potongan kata tunggal atau termmed word. Proses tokenization bertujuan untuk mempersiapkan dokumen untuk proses berikutnya, yaitu proses stopwords dan Stemming dapat dilakukan.

d. Stemming

Stemming digunakan untuk mengganti bentuk dari sebuah kata berimbuhan dan berakhiran menjadi kata dasar. Imbuhan (*affixes*) yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan kombinasi dari awalan dan akhiran (*confixes*) pada kata dasar merupakan bagian yang dihilangkan pada tahap stemming. Stemming perlu dilakukan untuk membentuk data tesk menjadi kata dasar agar tidak terdapat kata yang sama yang berbeda arti karena adanya imbuhan (*affixes*). Adapun proses Stemming yang digunakan pada penelitian ini menggunakan library Sastrawi, yaitu library Stemmer bahasa indonesia dengan lisensi MIT yang memanfaatkan kamus kata dasar dari Kateglo sebagai acuan.

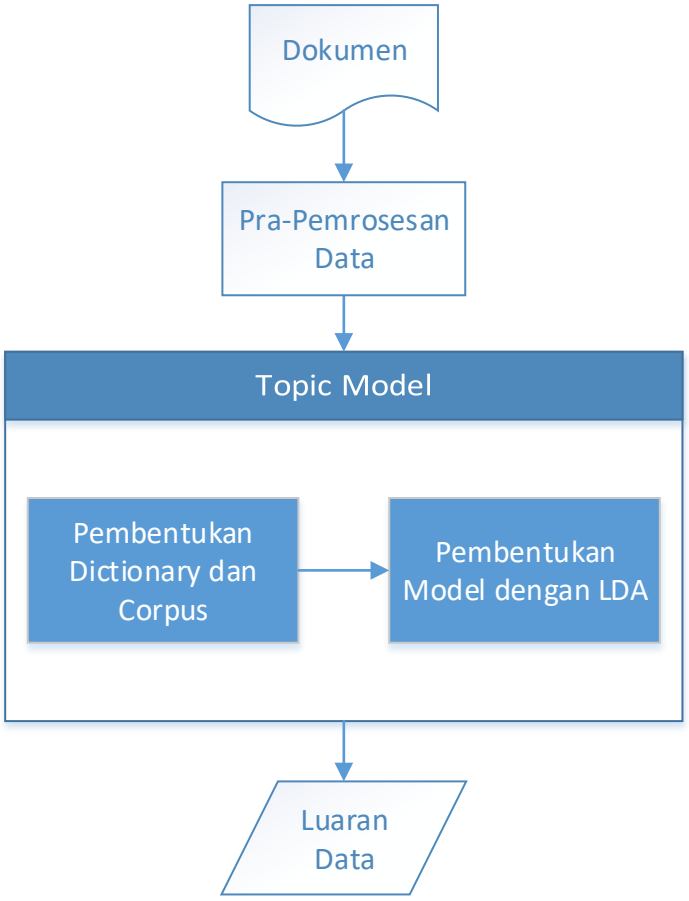
Adapun sub-aktifitas dari tahap pra-pemrosesan data digambarkan seperti gambar 4 berikut:



Gambar 4. Sub-Aktifitas Tahap Pra-Pemrosesan Data

3. Topic Modeling dengan Latent Dirichlet Allocation (LDA)

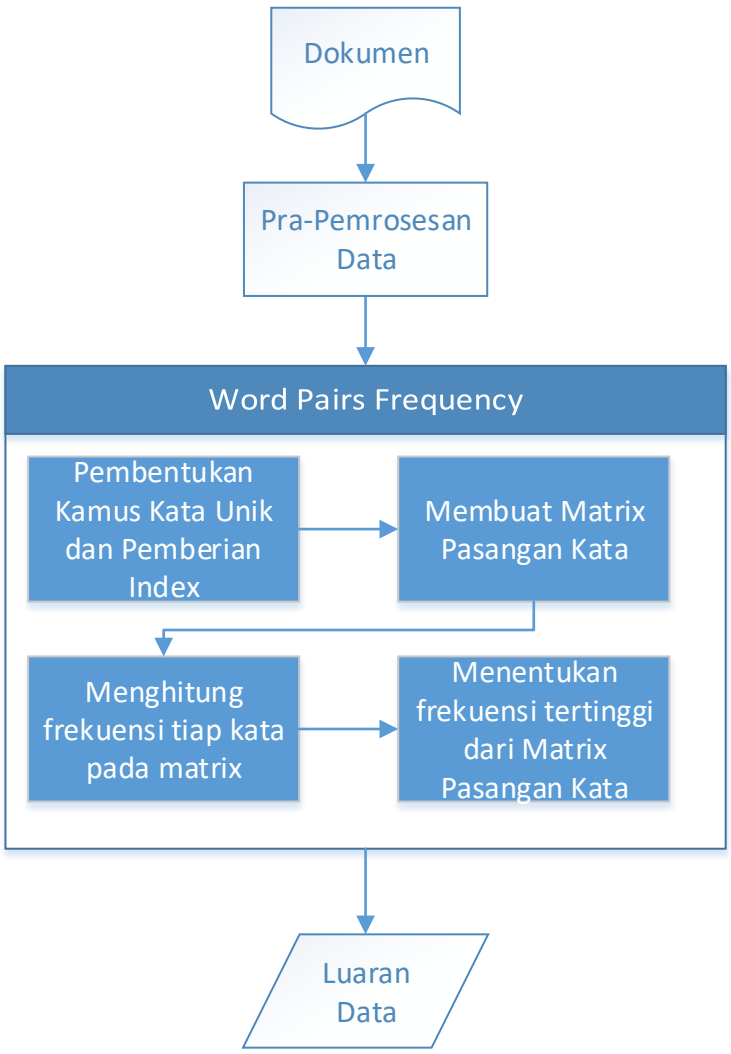
Tahap *Topic modeling* dengan *Latent Dirichlet Allocation (LDA)* terhadap dokumen yang berasal dari data Twitter terdiri dari beberapa tahap, yang diperlihatkan pada gambar 5 berikut:



Gambar 5. Tahap Topic modeling *dengan* Latent Dirichlet Allocation (LDA)

4. Word Pairs Frequency

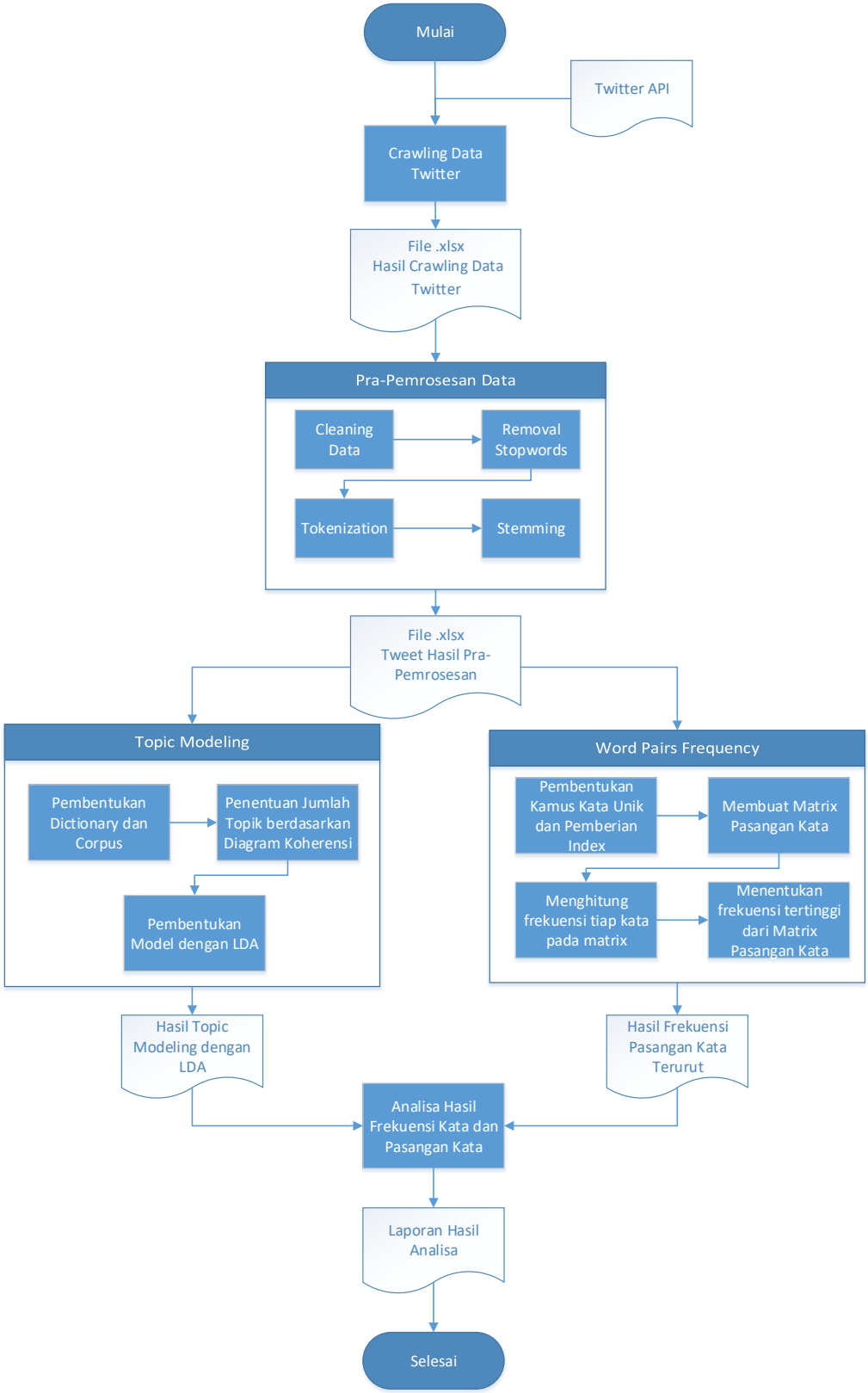
Word Pairs Frequency merupakan metode untuk menentukan frekuensi kemunculan kata dalam bentuk berpasangan. Tahapan proses yang dilakukan sebagai berikut:



Gambar 6. Tahap Pembentukan Word Pairs Frequency

G. MODEL PENELITIAN

Pada penelitian ini dirancang sebuah model penelitian yang memenuhi setiap alur proses yang dilakukan untuk menganalisa frekuensi kata menggunakan teknik Teks Mining dengan Topic Modeling dan Word Pairs Frequency.



Gambar 7. Rancangan Model Penelitian

BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

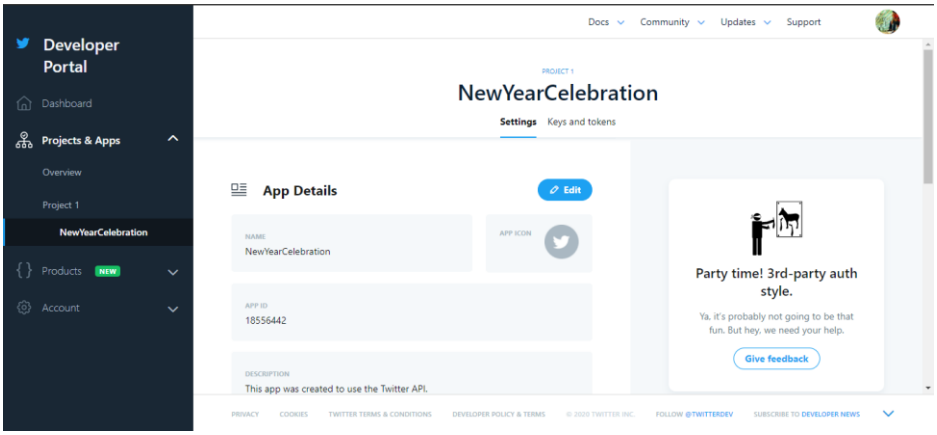
Bab ini menjelaskan bagaimana rancangan dari penelitian yang meliputi subyek dan obyek dari penelitian, pemilihan subyek dan obyek penelitian dan bagaimana penelitian akan dilakukan.

A. PENGAMBILAN DATA

Penelitian ini menggunakan data berupa tweet dari para pengguna twitter. Tweet yang digunakan pada penelitian ini berjumlah 2000 tweet yang diambil.

Untuk mendapatkannya maka digunakan Application Programming Interface (API) Twitter yang diberikan oleh pihak Twitter bagi para pengembang teknologi informasi. API twitter ini nantinya akan disiapkan dalam kode program Python versi 3.8.5 agar memperoleh data yang akan digunakan dalam penelitian ini.

Untuk mendapatkan API twitter, langkah pertama yang dilakukan adalah dengan mendaftarkan aplikasi di situs apps.twitter.com. Proses pendaftaran dapat dilakukan jika sudah memiliki akun twitter yang sudah tertaut nomor ponsel dan email pendaftaran yang sudah dikonfirmasi pemilik akun. Twitter kemudian memberikan form pendaftaran aplikasi yang wajib diisi. Gambar 8 memperlihatkan Aplication Management yang telah dibuat untuk mendapatkan access key pada situs apps.twitter.com untuk pendaftaran Twitter Apps.



Gambar 8. Aplication Management Twitter

Untuk menjalankan permintaan API dari twitter, langkah-langkah yang harus dilakukan adalah sebagai berikut:

1. Pemasangan perangkat lunak yang dibutuhkan. Perangkat lunak utama yang digunakan pada tahap ini yaitu *tweepy*
2. Otentikasi data twitter. Tahap otentikasi meliputi langkah-langkah sebagai berikut:

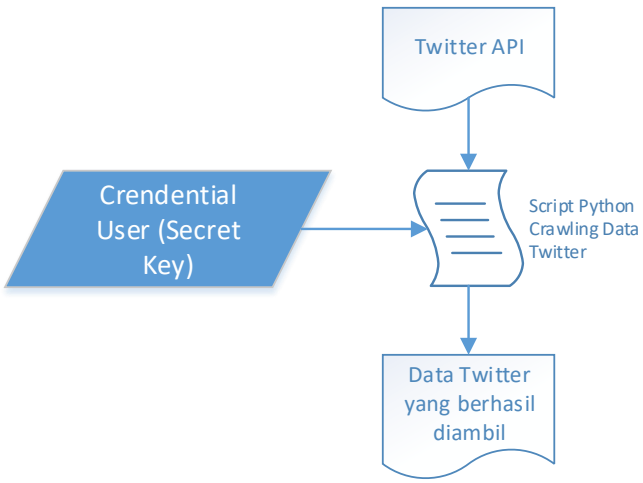
   Langkah 1: kunjungi situs web twitter dan klik tombol “buat aplikasi baru”.

   Langkah 2: isi detail di formulir yang disediakan dan kirim.

   Langkah 3: kemudian ditampilkan halaman aplikasi dimana “*consumer keys*”, “*consumer access*”, “*access token*”, dan “*access token secret*” yang dibutuhkan untuk mengakses data twitter ditampilkan.

   Langkah 4: implementasikan ke dalam python.

Proses pengunduhan data tweet dilakukan dengan crawling Twitter dengan teknik search. Teknik ini menggunakan satu atau beberapa kata kunci (keyword) untuk mengumpulkan data. Berikut ditampilkan pada gambar 9 alur proses pengunduhan data twitter.



Gambar 9. Alur Proses Pengunduhan Data Twitter

Selain API Twitter, digunakan pula twitter aoutentification yang berfungsi untuk mengakses ke API Twitter. OAuth adalah sebuah authorization framework yang memungkinkan aplikasi pihak ketiga untuk mendapatkan akses terbatas secara aman dan ringkas. Dengan OAuth, untuk melakukan request ke API Twitter, setiap aplikasi harus terlebih dahulu mendapatkan OAuth akses token. Akses token ini yang kemudian digunakan ketika menuliskan kode program, seperti yang dapat dilihat pada kode 1 berikut:

```
consumer_key = 'xxxxxxxxxxxxxxxxxxxxxxxxxxx'
consumer_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token = 'xxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_secret = ' xxxxxxxxxxxxxxxxxxxxxxxxxxx'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
api = tw.API(auth, wait_on_rate_limit=True)
try:
    api.verify_credentials()
    print("Authentication OK")
except:
    print("Error during authentication")
```

Kode 1. Akses Token API Twitter

Setelah melakukan otentikasi token API Twitter, selanjutnya melakukan proses *collection data*. *Source code* proses *collection data* ditunjukkan pada kode berikut ini:

```
xls_name = 'data_train.xlsx'
tweets = tw.Cursor(api.home_timeline,
              lang="id").items(2000)

users_locs = [[tweet.user.screen_name, tweet.text,
tweet.user.location, tweet.created_at] for tweet in tweets]
users_locs

tweet_text = pd.DataFrame(data=users_locs,
                    columns=['user', 'text', "location",
'created_at'])
tweet_text.to_excel(xls_name)
```

Kode 2. Collection Data Twitter

Pada tabel berikut diperlihatkan beberapa contoh data yang berhasil di unduh.

Tabel 4. Beberapa Hasil pengunduhan data tweet

No	User_name	Tweet
1	abcdekki	RT @ladazaa: bunsay pap foto keluarga kalian dong say, klw ini azula family https://t.co/Mc0JBKUooF
2	of_fangirls	RT @hanbinenim: Sini yg suka jbjb???????????????? Mutualan yok????????â€□â™€ï¸□ fandom ap aj????????â€□â™€ï¸□ jgn floop yah????
3	irnayaaa	RT @fucekriki: "apaan si baru deket udah sok akrab banget" kalo aku ga sokap gimana mau punya temen xixi.
4	ichaaptr96	Nungguin admin kirim link buat tes toefl bikin diriku ngantuk beraaatt
5	ichaChubidab	@ReginaRosaliani yukk yg minat produk young living bisa langsung dm yaa https://t.co/UDK9JsJhjc
6	Kakikuleeknow	RT @ladazaa: prajurit mesir: tuan, musa telah membelah laut firaun:
7	noahgariskeras	@ayayawaee_ dahlah cape sm kamu kamu ya????
8	Gearjoe2	@Virgonize Ahahaha emang bukan orang jawa?

No	User_name	Tweet
9	Dspzulvii	RT @chaeriysf: Liat cowoÂ² jalan abis jumatan itu damage nya masyaAllah banget, jadi pengen cuciin sarungnya deh
10	baca_pesan	New post (Rapat Bamus DPRD Makassar Membahas Jadwal Reses dan Sosper Berubah) has been published on Baca Pesan -â€¦ https://t.co/nU29vWoJYl
11	Taufik_A14	RT @BersihMakassar: Makin tak terurus kebersihannya. Saatnya Pak @DP_dannypomanto harus kembali urus Makassar. https://t.co/weGNH01FOz
12	yolandamarselab	Selamat datang di dunia anak gagah ku "Kairo Mahatma Pradana Bima" https://t.co/IZUN0hFL1u
13	magdaleneid	RT @Leaves1850: Sangat jarang aku temui orang tua yg seterbuka itu ttg mslh seksualitas. Bukn hanya dlm lingkup keluarga, sistem pendidikanâ€¦
14	Adrio13_	RT @txtdrmks: ngerina pertanyanna deh???? https://t.co/RcK8ySDs9G
15	fdlhtntry__	RT @Mallrmdhan: singkatnya sebuah pesan, adalah awal dari hancurnya hubungan.
16	ndiyant	RT @agakarebaa: Paagii. Apakah Stadion Mattoanging adalah bangunan cagar budaya? Cek dulu baik-baik. Secara kriteria, memang iya krn usianâ€¦
17	aletommo70	Dek @TsamaraDKI berhenti saja ngetwit. Tiap twittmu kaya muntahan kucing. Menjijikkan . https://t.co/I8cAp9WEIh
18	kyuraam0302	@akudisiniyaa butuh pulsanya buat dftr nelpon sebulan, trus buat dftr paketan juga
19	pesawatsukoii	RT @ladazaa: kemana aja gw selama ini baru tau mebel itu memek belakang ???????????????? aq kira usaha kayu say
20	BellaSianitta	Dibalik setiap kesedihan, terdapat kebahagiaan. Serahkan segala urusan kepada Tuhan. Biarkan waktu dan kehidupan berjalan.

B. PRE-PROCESSING

Setelah data terkumpul dari twitter, langkah selanjutnya adalah *pre-processing* yang juga diimplementasikan dengan menggunakan python. Ada beberapa tahapan yang terlibat dalam *pre-processing*, antara lain sebagai berikut:

1. Data Cleaning

Tahap ini dilakukan untuk menghapus bagian yang tidak penting dari *tweet*. *Source Code* yang diimplementasikan untuk *Data Cleaning* ditampilkan pada kode 2 berikut:

```
for tweet in outlist_init:
    tw_asli = tweet[0]
    tw_clean = []
    tw_clean = [ch for ch in tweet if ch not in char_remove]

    tw_clean = re.sub(URL, "", str(tw_clean))
    tw_clean = re.sub(html_tag, "",str(tw_clean))
    tw_clean = re.sub(hash_tag, "",str(tw_clean))
    tw_clean = re.sub(slash_all,"", str(tw_clean))
    tw_clean = re.sub(cont_number, "",str(tw_clean))
    tw_clean = re.sub(numbers, "",str(tw_clean))
    tw_clean = re.sub(start_pound, "",str(tw_clean))
    tw_clean = re.sub(start_quest_pound, "",str(tw_clean))
    tw_clean = re.sub(at_sign, "",str(tw_clean))
    tw_clean = re.sub("'", "",str(tw_clean))
    tw_clean = re.sub('"', "",str(tw_clean))
    tw_clean = re.sub(r'(?:^|\s)[@#].*?(?=[,;:.!?]|\s|$)', r'',
tw_clean)

    tw_filter = tw_clean.replace("RT:", "").replace("RT",
"").translate(str.maketrans('', '', string.punctuation)).strip();
    tw_clean = tw_filter
```

Kode 3. Source Code Data Cleaning

Contoh data sebelum dan sesudah melalui tahap *cleaning data* ditampilkan pada Tabel 5 berikut:

Tabel 5. Hasil Data Twitter setelah melalui proses Cleaning data

No	Tweet Hasil Crawling	Hasil Cleaning
1	RT @ladazaa: bunsay pap foto keluarga kalian dong say, klw ini azula family https://t.co/Mc0JBKUooF	bunsay pap foto keluarga kalian dong say klw ini azula family
2	RT @hanbinenim: Sini yg suka jbjb???????????????? Mutualan yok????????â€□â™€ï¸□ fandom ap aj????????â€□â™€ï¸□ jgn floop yah????	Sini yg suka jbjb Mutualan yok€™€ï¸ fandom ap aj€™€ï¸ jgn floop yah

No	Tweet Hasil Crawling	Hasil Cleaning
3	RT @fucekriki: "apaan si baru deket udah sok akrab banget" kalo aku ga sokap gimana mau punya temen xixi.	apaan si baru deket udah sok akrab banget kalo aku ga sokap gimana mau punya temen xixi
4	Nungguin admin kirim link buat tes toefl bikin diriku ngantuk beraaatt	Nungguin admin kirim link buat tes toefl bikin diriku ngantuk beraaatt Nungguin
5	@ReginaRosaliani yukk yg minat produk young living bisa langsung dm yaa https://t.co/UDK9JsJhjc	yukk yg minat produk young living bisa langsung dm yaa
6	RT @ladazaa: prajurit mesir: tuan, musa telah membelah laut firaun:	prajurit mesir tuan musa telah membelah laut firaun
7	@ayayawaee_ dahlah cape sm kamu kamu ya????	dahlah cape sm kamu kamu ya
8	@Virgonize Ahahaha emang bukan orang jawa?	Ahahaha emang bukan orang jawa
9	RT @chaeriysf: Liat cowoÂ² jalan abis jumatan itu damage nya masyaAllah banget, jadi pengen cuciin sarungnya deh	Liat cowoÂ² jalan abis jumatan itu damage nya masyaAllah banget jadi pengen cuciin sarungnya deh
10	New post (Rapat Bamus DPRD Makassar Membahas Jadwal Reses dan Sosper Berubah) has been published on Baca Pesan - â€¦ https://t.co/nU29vWoJYl	New post Rapat Bamus DPRD Makassar Membahas Jadwal Reses dan Sosper Berubah has been published on Baca Pesan â€¦ New
11	RT @BersihMakassar: Makin tak terurus kebersihannya. Saatnya Pak @DP_dannypomanto harus kembali urus Makassar. https://t.co/weGNH01FOz	Makin tak terurus kebersihannya Saatnya Pak harus kembali urus Makassar
12	Selamat datang di dunia anak gagah ku "Kairo Mahatma Pradana Bima" https://t.co/IZUN0hFL1u	Selamat datang di dunia anak gagah ku Kairo Mahatma Pradana Bima Selamat
13	RT @Leaves1850: Sangat jarang aku temui orang tua yg seterbuka itu ttg mslh seksualitas. Bukn hanya dlm lingkup keluarga, sistem pendidikanâ€¦	Sangat jarang aku temui orang tua yg seterbuka itu ttg mslh seksualitas Bukn hanya dlm lingkup keluarga sistem pendidikanâ€¦
14	RT @txtdrmks: ngerina pertanyanna deh???? https://t.co/RcK8ySDs9G	ngerina pertanyanna deh
15	RT @Mallrmdhan: singkatnya sebuah pesan, adalah awal dari hancurnya hubungan.	singkatnya sebuah pesan adalah awal dari hancurnya hubungan

No	Tweet Hasil Crawling	Hasil Cleaning
16	RT @agakarebaa: Paagii. Apakah Stadion Mattoanging adalah bangunan cagar budaya? Cek dulu baik-baik. Secara kriteria, memang iya krn usianâ€¦	Paagii Apakah Stadion Mattoanging adalah bangunan cagar budaya Cek dulu baikbaik Secara kriteria memang iya krn usianâ€¦
17	Dek @TsamaraDKI berhenti saja ngetwit. Tiap twittmu kaya muntahan kucing. Menjijikkan . https://t.co/I8cAp9WEIh	Dek berhenti saja ngetwit Tiap twittmu kaya muntahan kucing Menjijikkan Dek
18	@akudisiniyaa butuh pulsanya buat dftr nelpon sebulan, trus buat dftr paketan juga	butuh pulsanya buat dftr nelpon sebulan trus buat dftr paketan juga
19	RT @ladazaa: kemana aja gw selama ini baru tau mebel itu memek belakang ???????????????? aq kira usaha kayu say	kemana aja gw selama ini baru tau mebel itu memek belakang aq kira usaha kayu say
20	Dibalik setiap kesedihan, terdapat kebahagiaan. Serahkan segala urusan kepada Tuhan. Biarkan waktu dan kehidupan berjalan.	Dibalik setiap kesedihan terdapat kebahagiaan Serahkan segala urusan kepada Tuhan Biarkan waktu dan kehidupan berjalan Dibalik

2. Tokenization

Proses tokenization bertujuan untuk mempersiapkan dokumen untuk proses berikutnya, yaitu proses stopwords dan Stemming. *Source Code* yang diimplementasikan untuk *Tokenization* ditampilkan pada kode 3 berikut:

```
tw_clean = lmtzr.lemmatize(str(tw_clean))
tw_clean_lst = re.findall(r'\w+', str(tw_clean))
    tw_filter = tw_clean
    tw_token = word_tokenize(tw_filter.lower())
```

Kode 4. Source Code tahap *Tokenization*

Contoh data sebelum dan sesudah melalui tahap Tokenization ditampilkan pada Tabel 6.

Tabel 6. Contoh Hasil Data Twitter setelah melalui proses Tokenizing

No	Hasil Cleaning	Hasil Tokenizing
1	bunsay pap foto keluarga kalian dong say klw ini azula family	['bunsay', 'pap', 'foto', 'keluarga', 'kalian', 'dong', 'say', 'klw', 'ini', 'azula', 'family']
2	Sini yg suka jbjb Mutualan yok€™€ï¸ fandom ap aj€™€ï¸ jgn floop yah	['sini', 'yg', 'suka', 'jbjb', 'mutualan', 'yok€™€ï¸', 'fandom', 'ap', 'aj€™€ï¸', 'jgn', 'floop', 'yah']
3	apaan si baru deket udah sok akrab banget kalo aku ga sokap gimana mau punya temen xixi	['apaan', 'si', 'baru', 'deket', 'udah', 'sok', 'akrab', 'banget', 'kalo', 'aku', 'ga', 'sokap', 'gimana', 'mau', 'punya', 'temen', 'xixi']
4	Nungguin admin kirim link buat tes toefl bikin diriku ngantuk beraaatt Nungguin	['nungguin', 'admin', 'kirim', 'link', 'buat', 'tes', 'toefl', 'bikin', 'diriku', 'ngantuk', 'beraaatt', 'nungguin']
5	yukk yg minat produk young living bisa langsung dm yaa	['yukk', 'yg', 'minat', 'produk', 'young', 'living', 'bisa', 'langsung', 'dm', 'yaa']
6	prajurit mesir tuan musa telah membelah laut firaun	['prajurit', 'mesir', 'tuan', 'musa', 'telah', 'membelah', 'laut', 'firaun']
7	dahlah cape sm kamu kamu ya	['dahlah', 'cape', 'sm', 'kamu', 'kamu', 'ya']
8	Ahahaha emang bukan orang jawa	['ahahaha', 'emang', 'bukan', 'orang', 'jawa']
9	Liat cowoÂ² jalan abis jumatan itu damage nya masyaAllah banget jadi pengen cuciin sarungnya deh	['liat', 'cowoâ²', 'jalan', 'abis', 'jumatan', 'itu', 'damage', 'nya', 'masyaallah', 'banget', 'jadi', 'pengen', 'cuciin', 'sarungnya', 'deh']
10	New post Rapat Bamus DPRD Makassar Membahas Jadwal Reses dan Sosper Berubah has been published on Baca Pesan â€¦ New	['new', 'post', 'rapat', 'bamus', 'dprd', 'makassar', 'membahas', 'jadwal', 'reses', 'dan', 'sosper', 'berubah', 'has', 'been', 'published', 'on', 'baca', 'pesan', 'â€¦', 'new']
11	Makin tak terurus kebersihannya Saatnya Pak harus kembali urus Makassar	['makin', 'tak', 'terurus', 'kebersihannya', 'saatnya', 'pak', 'harus', 'kembali', 'urus', 'makassar']
12	Selamat datang di dunia anak gagah ku Kairo Mahatma Pradana Bima Selamat	['selamat', 'datang', 'di', 'dunia', 'anak', 'gagah', 'ku', 'kairo', 'mahatma', 'pradana', 'bima', 'selamat']
13	Sangat jarang aku temui orang tua yg seterbuka itu ttg mslh seksualitas Bukn hanya dlm lingkup keluarga sistem pendidikanâ€¦	['sangat', 'jarang', 'aku', 'temui', 'orang', 'tua', 'yg', 'seterbuka', 'itu', 'ttg', 'mslh', 'seksualitas', 'bukn', 'hanya', 'dlm', 'lingkup', 'keluarga', 'sistem', 'pendidikanâ€¦']
14	ngerina pertanyanna deh	['ngerina', 'pertanyanna', 'deh']

No	Hasil Cleaning	Hasil Tokenizing
15	singkatnya sebuah pesan adalah awal dari hancurnya hubungan	['singkatnya', 'sebuah', 'pesan', 'adalah', 'awal', 'dari', 'hancurnya', 'hubungan']
16	Paagii Apakah Stadion Mattoanging adalah bangunan cagar budaya Cek dulu baikbaik Secara kriteria memang iya krn usianâ€¦	['paagii', 'apakah', 'stadion', 'mattoanging', 'adalah', 'bangunan', 'cagar', 'budaya', 'cek', 'dulu', 'baikbaik', 'secara', 'kriteria', 'memang', 'iya', 'krn', 'usianâ€¦']
17	Dek berhenti saja ngetwit Tiap twittmu kaya muntahan kucing Menjijikkan Dek	['dek', 'berhenti', 'saja', 'ngetwit', 'tiap', 'twittmu', 'kaya', 'muntahan', 'kucing', 'menjijikkan', 'dek']
18	butuh pulsanya buat dftr nelpon sebulan trus buat dftr paketan juga	['butuh', 'pulsanya', 'buat', 'dftr', 'nelpon', 'sebulan', 'trus', 'buat', 'dftr', 'paketan', 'juga']
19	kemana aja gw selama ini baru tau mebel itu memek belakang aq kira usaha kayu say	['kemana', 'aja', 'gw', 'selama', 'ini', 'baru', 'tau', 'mebel', 'itu', 'memek', 'belakang', 'aq', 'kira', 'usaha', 'kayu', 'say']
20	Dibalik setiap kesedihan terdapat kebahagiaan Serahkan segala urusan kepada Tuhan Biarkan waktu dan kehidupan berjalan Dibalik	['dibalik', 'setiap', 'kesedihan', 'terdapat', 'kebahagiaan', 'serahkan', 'segala', 'urusan', 'kepada', 'tuhan', 'biarkan', 'waktu', 'dan', 'kehidupan', 'berjalan', 'dibalik']

3. Removal of Stopwords.

Stopwords adalah kata-kata yang sering digunakan dalam suatu bahasa dan hanya memiliki sedikit makna. Sebagian besar merupakan kata ganti, misalnya, kata-kata seperti, "adalah", "dan", "itu", dan lain-lain. *Source Code* yang diimplementasikan untuk tahap *Removal of Stopwords* ditampilkan sebagai berikut:

```
tw_clean_lst = [tw.lower() for tw in tw_clean_lst if tw.lower()
not in stopwords.words('indonesian')]
    tw_clean_lst_e = [tw.lower() for tw in tw_clean_lst if
tw.lower() not in stopwords.words('english')]
    tw_stopword = tw_clean_lst
    tw_stopword_e = tw_clean_lst_e
```

Kode 5. Source Code tahap Removal of Stopwords

Contoh data sebelum dan sesudah melalui tahap *Removal of Stopwords* ditampilkan pada Tabel 7 berikut:

Tabel 7. Contoh Hasil Data Twitter setelah melalui proses Removal of Stopwords

No	Hasil Tokenizing	Hasil Removal of Stopwords
1	['bunsay', 'pap', 'foto', 'keluarga', 'kalian', 'dong', 'say', 'klw', 'ini', 'azula', 'family']	['bunsay', 'pap', 'foto', 'keluarga', 'say', 'klw', 'azula', 'family']
2	['sini', 'yg', 'suka', 'jbjb', 'mutualan', 'yok€™€ï¸', 'fandom', 'ap', 'aj€™€ï¸', 'jgn', 'floop', 'yah']	['yg', 'suka', 'jbjb', 'mutualan', 'yok', 'ï', 'fandom', 'ap', 'aj', 'ï', 'jgn', 'floop', 'yah']
3	['apaan', 'si', 'baru', 'deket', 'udah', 'sok', 'akrab', 'banget', 'kalo', 'aku', 'ga', 'sokap', 'gimana', 'mau', 'punya', 'temen', 'xixi']	['si', 'deket', 'udah', 'sok', 'akrab', 'banget', 'kalo', 'ga', 'sokap', 'gimana', 'temen', 'xixi']
4	['nungguin', 'admin', 'kirim', 'link', 'buat', 'tes', 'toefl', 'bikin', 'diriku', 'ngantuk', 'beraaatt', 'nungguin']	['nungguin', 'admin', 'kirim', 'link', 'tes', 'toefl', 'bikin', 'diriku', 'ngantuk', 'beraaatt', 'nungguin']
5	['yukk', 'yg', 'minat', 'produk', 'young', 'living', 'bisa', 'langsung', 'dm', 'yaa']	['yukk', 'yg', 'minat', 'produk', 'young', 'living', 'langsung', 'dm', 'yaa']
6	['prajurit', 'mesir', 'tuan', 'musa', 'telah', 'membelah', 'laut', 'firaun']	['prajurit', 'mesir', 'tuan', 'musa', 'membelah', 'laut', 'firaun']
7	['dahlah', 'cape', 'sm', 'kamu', 'kamu', 'ya']	['dahlah', 'cape', 'sm', 'ya']
8	['ahahaha', 'emang', 'bukan', 'orang', 'jawa']	['ahahaha', 'emang', 'orang', 'jawa']
9	['liat', 'cowoâ²', 'jalan', 'abis', 'jumatan', 'itu', 'damage', 'nya', 'masyaallah', 'banget', 'jadi', 'pengen', 'cuciin', 'sarungnya', 'deh']	['liat', 'cowoâ²', 'jalan', 'abis', 'jumatan', 'damage', 'nya', 'masyaallah', 'banget', 'pengen', 'cuciin', 'sarungnya', 'deh']
10	['new', 'post', 'rapat', 'bamus', 'dprd', 'makassar', 'membahas', 'jadwal', 'reses', 'dan', 'sosper', 'berubah', 'has', 'been', 'published', 'on', 'baca', 'pesan', 'â€¦', 'new']	['new', 'post', 'rapat', 'bamus', 'dprd', 'makassar', 'membahas', 'jadwal', 'reses', 'sosper', 'berubah', 'published', 'baca', 'pesan', 'â', 'new']
11	['makin', 'tak', 'terurus', 'kebersihannya', 'saatnya', 'pak', 'harus', 'kembali', 'urus', 'makassar']	['terurus', 'kebersihannya', 'urus', 'makassar']
12	['selamat', 'datang', 'di', 'dunia', 'anak', 'gagah', 'ku', 'kairo', 'mahatma', 'pradana', 'bima', 'selamat']	['selamat', 'dunia', 'anak', 'gagah', 'ku', 'kairo', 'mahatma', 'pradana', 'bima', 'selamat']
13	['sangat', 'jarang', 'aku', 'temui', 'orang', 'tua', 'yg', 'seterbuka', 'itu', 'ttg', 'mslh', 'seksualitas', 'bukn', 'hanya', 'dlm', 'lingkup', 'keluarga', 'sistem', 'pendidikanâ€¦']	['jarang', 'temui', 'orang', 'tua', 'yg', 'seterbuka', 'ttg', 'mslh', 'seksualitas', 'bukn', 'dlm', 'lingkup', 'keluarga', 'sistem', 'pendidikanâ']
14	['ngerina', 'pertanyanna', 'deh']	['ngerina', 'pertanyanna', 'deh']

No	Hasil Tokenizing	Hasil Removal of Stopwords
15	['singkatnya', 'sebuah', 'pesan', 'adalah', 'awal', 'dari', 'hancurnya', 'hubungan']	['singkatnya', 'pesan', 'hancurnya', 'hubungan']
16	['paagii', 'apakah', 'stadion', 'mattoanging', 'adalah', 'bangunan', 'cagar', 'budaya', 'cek', 'dulu', 'baikbaik', 'secara', 'kriteria', 'memang', 'iya', 'krn', 'usianâ€¦']	['paagii', 'stadion', 'mattoanging', 'bangunan', 'cagar', 'budaya', 'cek', 'baikbaik', 'kriteria', 'iya', 'krn', 'usianâ']
17	['dek', 'berhenti', 'saja', 'ngetwit', 'tiap', 'twittmu', 'kaya', 'muntahan', 'kucing', 'menjijikkan', 'dek']	['dek', 'berhenti', 'ngetwit', 'twittmu', 'kaya', 'muntahan', 'kucing', 'menjijikkan', 'dek']
18	['butuh', 'pulsanya', 'buat', 'dftr', 'nelpon', 'sebulan', 'trus', 'buat', 'dftr', 'paketan', 'juga']	['butuh', 'pulsanya', 'dftr', 'nelpon', 'sebulan', 'trus', 'dftr', 'paketan']
19	['kemana', 'aja', 'gw', 'selama', 'ini', 'baru', 'tau', 'mebel', 'itu', 'memek', 'belakang', 'aq', 'kira', 'usaha', 'kayu', 'say']	['kemana', 'aja', 'gw', 'tau', 'mebel', 'memek', 'aq', 'usaha', 'kayu', 'say']
20	['dibalik', 'setiap', 'kesedihan', 'terdapat', 'kebahagiaan', 'serahkan', 'segala', 'urusan', 'kepada', 'tuhan', 'biarkan', 'waktu', 'dan', 'kehidupan', 'berjalan', 'dibalik']	['dibalik', 'kesedihan', 'kebahagiaan', 'serahkan', 'urusan', 'tuhan', 'biarkan', 'kehidupan', 'berjalan', 'dibalik']

4. Stemming

Stemming adalah teknik untuk menghapus imbuhan kata dan menyisahkan kata dasarnya saja. *Source Code* yang diimplementasikan untuk tahap *Stemming* ditampilkan sebagai berikut:

```
tw_clean_lst = re.findall(r'\w+', str(tw_clean_lst))
    tw_clean_lst = [replace_all(word, repl_dict) for word in
tw_clean_lst]
    tweet_clean_fin.append(list(tw_clean_lst))

    tw_stemming = lmtzr.lemmatize(str(tw_clean_lst))
```

Kode 6. Source Code tahap Stemming

Tabel 8 berikut merupakan contoh data sebelum dan sesudah melalui tahap *Stemming*.

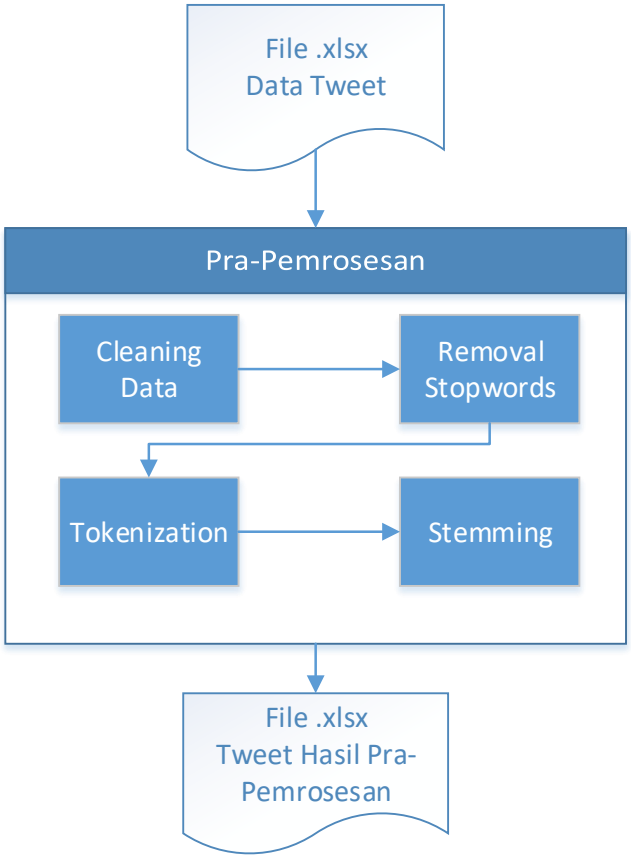
Tabel 8. Contoh Hasil Data Twitter setelah melalui proses Stemming

No	Hasil Removal of Stopwords	Hasil Stemming
1	['bunsay', 'pap', 'foto', 'keluarga', 'say', 'klw', 'azula', 'family']	['bunsay', 'foto', 'keluarga', 'azula', 'family']
2	['yg', 'suka', 'jbjb', 'mutualan', 'yok', 'i', 'fandom', 'ap', 'aj', 'i', 'jgn', 'floop', 'yah']	['suka', 'jbjb', 'mutualan', 'fandom', 'floop']
3	['si', 'deket', 'udah', 'sok', 'akrab', 'banget', 'kalo', 'ga', 'sokap', 'gimana', 'temen', 'xixi']	['deket', 'udah', 'akrab', 'banget', 'kalo', 'sokap', 'gimana', 'temen', 'xixi']
4	['nungguin', 'admin', 'kirim', 'link', 'tes', 'toefl', 'bikin', 'diriku', 'ngantuk', 'beraaatt', 'nungguin']	['nungguin', 'admin', 'kirim', 'link', 'toefl', 'bikin', 'diriku', 'ngantuk', 'beraaatt', 'nungguin']
5	['yukk', 'yg', 'minat', 'produk', 'young', 'living', 'langsung', 'dm', 'yaa']	['yukk', 'minat', 'produk', 'young', 'living', 'langsung']
6	['prajurit', 'mesir', 'tuan', 'musa', 'membelah', 'laut', 'firaun']	['prajurit', 'mesir', 'tuan', 'musa', 'membelah', 'laut', 'firaun']
7	['dahlah', 'cape', 'sm', 'ya']	['dahlah', 'cape']
8	['ahahaha', 'emang', 'orang', 'jawa']	['ahahaha', 'emang', 'orang', 'jawa']
9	['liat', 'cowoâ²', 'jalan', 'abis', 'jumatan', 'damage', 'nya', 'masyaallah', 'banget', 'pengen', 'cuciin', 'sarungnya', 'deh']	['liat', 'cowoâ²', 'jalan', 'abis', 'jumatan', 'damage', 'masyaallah', 'banget', 'pengen', 'cuciin', 'sarungnya']
10	['new', 'post', 'rapat', 'bamus', 'dprd', 'makassar', 'membahas', 'jadwal', 'reses', 'sosper', 'berubah', 'has', 'been', 'published', 'on', 'baca', 'pesan', 'â', 'new']	['new', 'post', 'rapat', 'bamus', 'dprd', 'makassar', 'membahas', 'jadwal', 'reses', 'sosper', 'berubah', 'been', 'published', 'baca', 'pesan', 'new']
11	['terurus', 'kebersihannya', 'urus', 'makassar']	['terurus', 'kebersihannya', 'urus', 'makassar']
12	['selamat', 'dunia', 'anak', 'gagah', 'ku', 'kairo', 'mahatma', 'pradana', 'bima', 'selamat']	['selamat', 'dunia', 'anak', 'gagah', 'kairo', 'mahatma', 'pradana', 'bima', 'selamat']
13	['jarang', 'temui', 'orang', 'tua', 'yg', 'seterbuka', 'ttg', 'mslh', 'seksualitas', 'bukn', 'dlm', 'lingkup', 'keluarga', 'sistem', 'pendidikanâ']	['jarang', 'temui', 'orang', 'seterbuka', 'mslh', 'seksualitas', 'bukn', 'lingkup', 'keluarga', 'sistem', 'pendidikanâ']
14	['ngerina', 'pertanyanna', 'deh']	['ngerina', 'pertanyanna']
15	['singkatnya', 'pesan', 'hancurnya', 'hubungan']	['singkatnya', 'pesan', 'hancurnya', 'hubungan']
16	['paagii', 'stadion', 'mattoanging', 'bangunan', 'cagar', 'budaya', 'cek', 'baikbaik', 'kriteria', 'iya', 'krn', 'usianâ']	['paagii', 'stadion', 'mattoanging', 'bangunan', 'cagar', 'budaya', 'baikbaik', 'kriteria', 'usianâ']
17	['dek', 'berhenti', 'ngetwit', 'twittmu', 'kaya', 'muntahan', 'kucing', 'menjijikkan', 'dek']	['berhenti', 'ngetwit', 'twittmu', 'kaya', 'muntahan', 'kucing', 'menjijikkan']

No	Hasil Removal of Stopwords	Hasil Stemming
18	['butuh', 'pulsanya', 'dftr', 'nelpon', 'sebulan', 'trus', 'dftr', 'paketan']	['butuh', 'pulsanya', 'dftr', 'nelpon', 'sebulan', 'trus', 'dftr', 'paketan']
19	['kemana', 'aja', 'gw', 'tau', 'mebel', 'memek', 'aq', 'usaha', 'kayu', 'say']	['kemana', 'mebel', 'memek', 'usaha', 'kayu']
20	['dibalik', 'kesedihan', 'kebahagiaan', 'serahkan', 'urusan', 'tuhan', 'biarkan', 'kehidupan', 'berjalan', 'dibalik']	['dibalik', 'kesedihan', 'kebahagiaan', 'serahkan', 'urusan', 'tuhan', 'biarkan', 'kehidupan', 'berjalan', 'dibalik']

5. Hasil Pra-Pemrosesan Data Twitter

Berikut ini merupakan alur proses dari keseluruhan pra-pemrosesan data twitter yang telah berhasil didapatkan:



Gambar 10. Alur Pra-Pemrosesan Data

Hasil Pra-pemrosesan data tweet ditampilkan pada tabel 9 berikut:

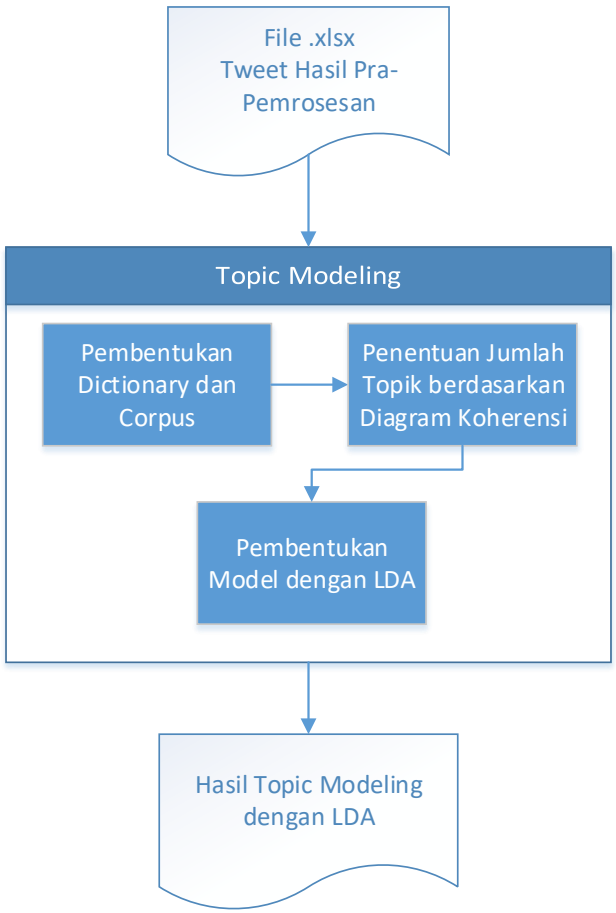
Tabel 9. Hasil Pra-Pemrosesan Data Tweet

No	Tweet Hasil Crawling	Hasil Pra-pemrosesan
1	RT @ladazaa: bunsay pap foto keluarga kalian dong say, klw ini azula family https://t.co/Mc0JBKUooF	bunsay foto keluarga azula family
2	RT @hanbinenim: Sini yg suka jbjb???????????????? Mutualan yok????????â€□â™€ï¸□ fandom ap aj????????â€□â™€ï¸□ jgn floop yah????	suka jbjb mutualan fandom floop
3	RT @fucekriki: "apaan si baru deket udah sok akrab banget" kalo aku ga sokap gimana mau punya temen xixi.	deket udah akrab banget kalo sokap gimana temen xixi
4	Nungguin admin kirim link buat tes toefl bikin diriku ngantuk beraaatt	nungguin admin kirim link toefl bikin diriku ngantuk beraaatt nungguin
5	@ReginaRosaliani yukk yg minat produk young living bisa langsung dm yaa https://t.co/UDK9JsJhjc	yukk minat produk young living langsung
6	RT @ladazaa: prajurit mesir: tuan, musa telah membelah laut firaun:	prajurit mesir tuan musa membelah laut firaun
7	@ayayawaee_ dahlah cape sm kamu kamu ya????	dahlah cape
8	@Virgonize Ahahaha emang bukan orang jawa?	ahahaha emang orang jawa
9	RT @chaeriysf: Liat cowoÂ² jalan abis jumatan itu damage nya masyaAllah banget, jadi pengen cuciin sarungnya deh	liat cowoâ² jalan abis jumatan damage masyaallah banget pengen cuciin sarungnya
10	New post (Rapat Bamus DPRD Makassar Membahas Jadwal Reses dan Sosper Berubah) has been published on Baca Pesan - â€¦ https://t.co/nU29vWoJYl	new post rapat bamus dprd makassar membahas jadwal reses sosper berubah been published baca pesan new
11	RT @BersihMakassar: Makin tak terurus kebersihannya. Saatnya Pak @DP_dannypomanto harus kembali urus Makassar. https://t.co/weGNH01FOz	terurus kebersihannya urus makassar
12	Selamat datang di dunia anak gagah ku "Kairo Mahatma Pradana Bima" https://t.co/IZUN0hFL1u	selamat dunia anak gagah kairo mahatma pradana bima selamat

No	Tweet Hasil Crawling	Hasil Pra-pemrosesan
13	RT @Leaves1850: Sangat jarang aku temui orang tua yg seterbuka itu ttg mslh seksualitas. Bukn hanya dlm lingkup keluarga, sistem pendidikanâ€¦	jarang temui orang seterbuka mslh seksualitas bukn lingkup keluarga sistem pendidikanâ
14	RT @txtdrmks: ngerina pertanyanna deh???? https://t.co/RcK8ySDs9G	ngerina pertanyanna
15	RT @Mallrmdhan: singkatnya sebuah pesan, adalah awal dari hancurnya hubungan.	singkatnya pesan hancurnya hubungan
16	RT @agakarebaa: Paagii. Apakah Stadion Mattoanging adalah bangunan cagar budaya? Cek dulu baik-baik. Secara kriteria, memang iya krn usianâ€¦	paagii stadion mattoanging bangunan cagar budaya baikbaik kriteria usianâ
17	Dek @TsamaraDKI berhenti saja ngetwit. Tiap twittmu kaya muntahan kucing. Menjijikkan . https://t.co/I8cAp9WEIh	berhenti ngetwit twittmu kaya muntahan kucing menjijikkan
18	@akudisiniyaa butuh pulsanya buat dftr nelpon sebulan, trus buat dftr paketan juga	butuh pulsanya dftr nelpon sebulan trus dftr paketan
19	RT @ladazaa: kemana aja gw selama ini baru tau mebel itu memek belakang ???????????????? aq kira usaha kayu say	kemana mebel memek usaha kayu
20	Dibalik setiap kesedihan, terdapat kebahagiaan. Serahkan segala urusan kepada Tuhan. Biarkan waktu dan kehidupan berjalan.	dibalik kesedihan kebahagiaan serahkan urusan tuhan biarkan kehidupan berjalan dibalik

C. TOPIC MODELING

Tahap pembentukan Topic Modeling menggunakan metode LDA dapat dilihat pada gambar 11 berikut:



Gambar 11. Tahap Pembentukan Topic Modeling dengan Metode LDA

Sebelum melakukan pemodelan topik, data twitter yang telah melalui pra-pemrosesan data dibentuk dalam sebuah list data. *Source code* untuk membentuk list data tersebut sebagai berikut:

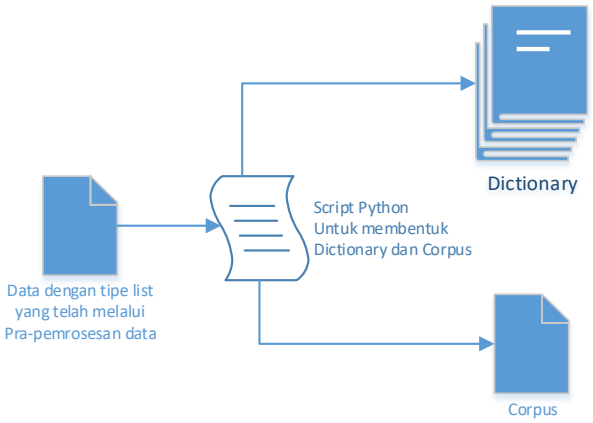
```
fo = pd.ExcelFile('clean-data_tweet.xlsx')
df = pd.read_excel(fo, 'Sheet1')
text = df['text']
text_list =  [i.split() for i in text]
print(len(text_list))
print(text_list)
```

Kode 7. Pembuatan List Data

1. Pembentukan Dictionary dan Corpus

Dictionary merupakan format data yang mengandung himpunan kata unik yang diberi indeks, sehingga dapat memudahkan dalam menampilkan kata

yang termasuk dalam model. Corpus merupakan format data yang berbentuk dokumen term-matrix, digunakan dalam melakukan eksperimen pembentukan model nantinya.



Gambar 12. Alur Pembentukan Dictionary dan Corpus

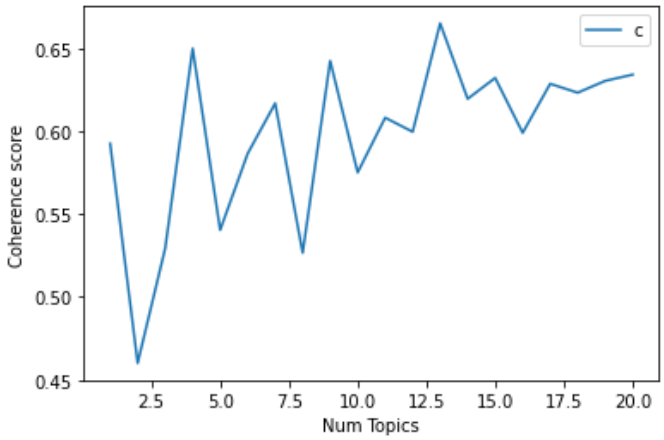
Source Code yang diimplementasikan untuk pembentukan *Dictionary* dan *Corpus* ditampilkan pada kode 8 berikut:

```
import gensim
from gensim.models import Phrases
bigram = Phrases(text_list, min_count=10)
trigram = Phrases(bigram[text_list])
for idx in range(len(text_list)):
    for token in bigram[text_list[idx]]:
        if '_' in token:
            text_list[idx].append(token)
    for token in trigram[text_list[idx]]:
        if '_' in token:
            text_list[idx].append(token)

from gensim import corpora, models
dictionary = corpora.Dictionary(text_list)
dictionary.filter_extremes(no_below=5, no_above=0.2)
print(dictionary)

doc_term_matrix = [dictionary.doc2bow(doc) for doc in text_list]
print(len(doc_term_matrix))
print(doc_term_matrix[2])

tfidf = models.TfidfModel(doc_term_matrix)
corpus_tfidf = tfidf[doc_term_matrix]
```

Kode 8. Source Code Pembentukan Dictionary dan Corpus

2. Penentuan Jumlah Topik

Penentuan banyaknya model topik dilakukan dengan cara melihat visualisasi pada grafik skor koherensi. Koherensi skor adalah ukuran yang digunakan untuk mengevaluasi Topic Modeling, model yang baik akan menghasilkan topik dengan skor koherensi topik yang tinggi. Penentuan jumlah topik yang akan digunakan dilakukan dengan melakukan eksperimen pada jumlah topik. Penentuan jumlah topik diawali dengan memberikan nilai mula-mula sebesar 50, kemudian jumlah topik ditentukan sebanyak 5 kali yaitu, 10, 20, 30, 40 dan 50 topik. Berdasarkan eksperimen penentuan jumlah topik yang akan digunakan, skor koherensi pada jumlah topik sebanyak 20 topik menghasilkan skor koherensi topik terbaik. Grafik skor koherensi dengan jumlah topik sebanyak 20 topik diperlihatkan pada gambar 13.

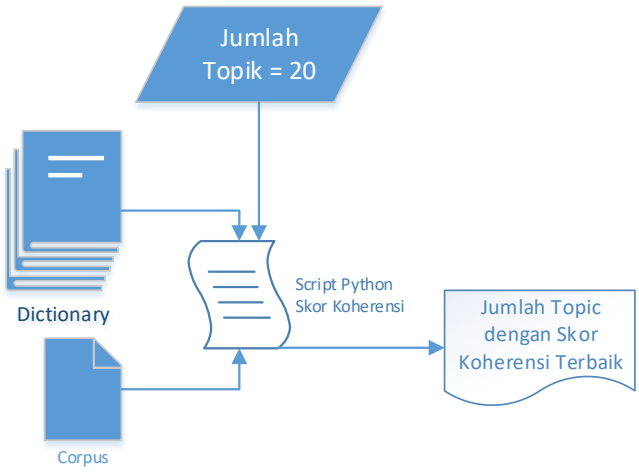


Gambar 13. Grafik Skor Koherensi terhadap Jumlah Topik

```
Num Topics = 1  has Coherence Value of 0.592502
Num Topics = 2  has Coherence Value of 0.460138
Num Topics = 3  has Coherence Value of 0.52927
Num Topics = 4  has Coherence Value of 0.649874
Num Topics = 5  has Coherence Value of 0.540408
Num Topics = 6  has Coherence Value of 0.586494
Num Topics = 7  has Coherence Value of 0.616821
Num Topics = 8  has Coherence Value of 0.526628
Num Topics = 9  has Coherence Value of 0.642377
Num Topics = 10  has Coherence Value of 0.575114
Num Topics = 11  has Coherence Value of 0.608143
Num Topics = 12  has Coherence Value of 0.599644
Num Topics = 13  has Coherence Value of 0.665045
Num Topics = 14  has Coherence Value of 0.619517
Num Topics = 15  has Coherence Value of 0.632134
Num Topics = 16  has Coherence Value of 0.598998
Num Topics = 17  has Coherence Value of 0.62859
Num Topics = 18  has Coherence Value of 0.623216
Num Topics = 19  has Coherence Value of 0.630353
Num Topics = 20  has Coherence Value of 0.634123
```

Gambar 14. Topik tertinggi berdasarkan skor koherensi

Berdasarkan gambar 14, skor koherensi tertinggi yang dihasilkan pada 20 topik adalah sebesar 0.665045 yang dihasilkan oleh topik nomor 13, skor koherensi 0.649874 dihasilkan oleh topik nomor 4 dan topik nomor 9 dengan skor koherensi 0.642377. Berdasarkan hasil terbaik pada nilai skor koherensi tersebut, jumlah topik yang dihasilkan tersebut dijadikan acuan dalam membuat model, sehingga pada penelitian ini terdapat tiga topik yang terbaik.



Gambar 15. Alur Penentuan Jumlah Topik

Source Code yang diimplementasikan untuk penentuan jumlah topik ditampilkan pada kode 9 berikut:

```
start=1
limit=21
step=1
model_list, coherence_values =
compute_coherence_values(dictionary, corpus=corpus_tfidf,

texts=text_list, start=start, limit=limit, step=step)

import matplotlib.pyplot as plt
x = range(start, limit, step)
plt.plot(x, coherence_values)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

Kode 9. Source Code Penentuan Jumlah Topik

3. Topic Modeling dengan LDA

Pada tahapan proses Topic modeling dengan LDA, langkah utama yang dilakukan adalah membentuk model dengan menggunakan library gensim. Dalam membentuk model eksperimen dilakukan pada input parameter. Hasil dari pencarian model akan digunakan untuk mendapatkan topik apa saja yang muncul dari analisis pada dokumen.

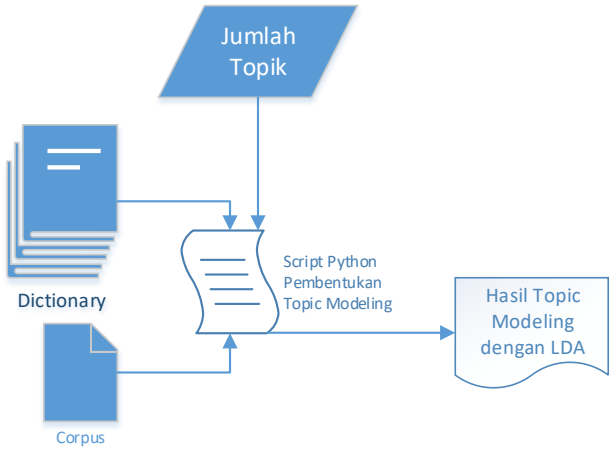
Source Code yang diimplementasikan untuk pembentukan topik dengan LDA ditampilkan pada kode 10 berikut:

```
model = LdaModel(corpus=corpus_tfidf, id2word=dictionary,
num_topics=3)
for idx, topic in model.print_topics(-1):
    print('Topic: {} Word: {}'.format(idx, topic))

import pandas as pd
top_words_per_topic = []
for t in range(model.num_topics):
    top_words_per_topic.extend([(t, ) + x for x in
model.show_topic(t, topn = 10)])
df = pd.DataFrame(top_words_per_topic, columns=['Topic',
'Word','P']).to_csv("top_words.csv")
import gensim
```

```
import pyLDAvis.gensim;pyLDAvis.enable_notebook()
data = pyLDAvis.gensim.prepare(model, corpus_tfidf, dictionary)
print(data)
pyLDAvis.save_html(data, 'lda-gensim.html')
```

Kode 10. Source Code Pembentukan Topik dengan LDA

Alur proses pembentukan topik dengan LDA ditampilkan pada gambar 16 berikut:



Gambar 16. Alur Topic Modeling dengan LDA

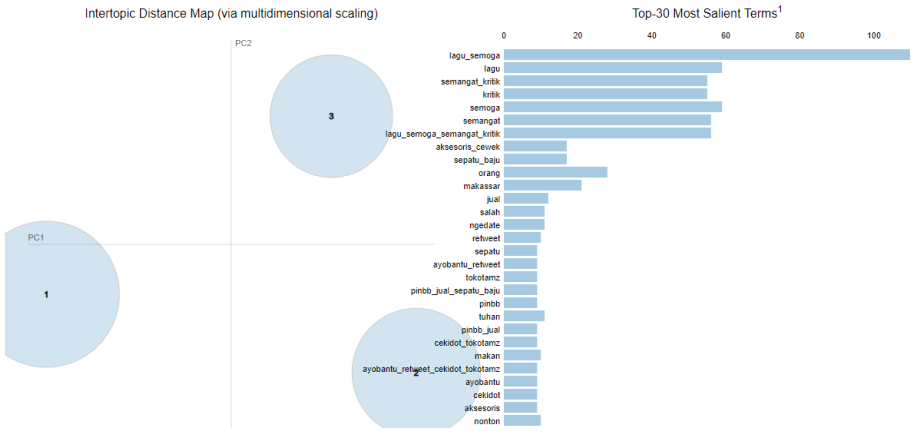
4. Visualisasi Hubungan antara Topik

Dari hasil penentuan jumlah topik didapatkan 3 topik utama yang dibicarakan dengan nilai akurasi yang tinggi dan saling berelasi seperti pada tabel 10 dibawah ini:

Tabel 10. Bobot Kata terhadap Tiga Topik Teratas

Topik 1	Topik 2	Topik 3
0.035*"orang" + 0.016*"salah" + 0.016*"tuhan" + 0.015*"makan" + 0.015*"selamat" + 0.015*"nonton" + 0.014*"kalo" + 0.013*"gitu" + 0.013*"main" + 0.012*"happy"	0.129*"lagu_semoga" + 0.069*"lagu" + 0.065*"semoga" + 0.064*"semangat_kritik" + 0.064*"kritik" + 0.063*"semangat" + 0.062*"lagu_semoga_semangat_kritik" + 0.015*"membelah_laut" + 0.013*"udah" + 0.009*"foto"	0.028*"makassar" + 0.026*"aksesoris_cewek" + 0.026*"sepatu_baju" + 0.019*"jual" + 0.016*"ngedate" + 0.015*"retweet" + 0.014*"suka" + 0.014*"sepatu" + 0.014*"cewek" + 0.013*"ayobantu_retweet"

Hubungan atau korelasi antara topik satu dengan yang lain diperlihatkan dalam visualisasi grafik pada gambar 17.



Gambar 17. Visualisasi Hubungan antar Topik

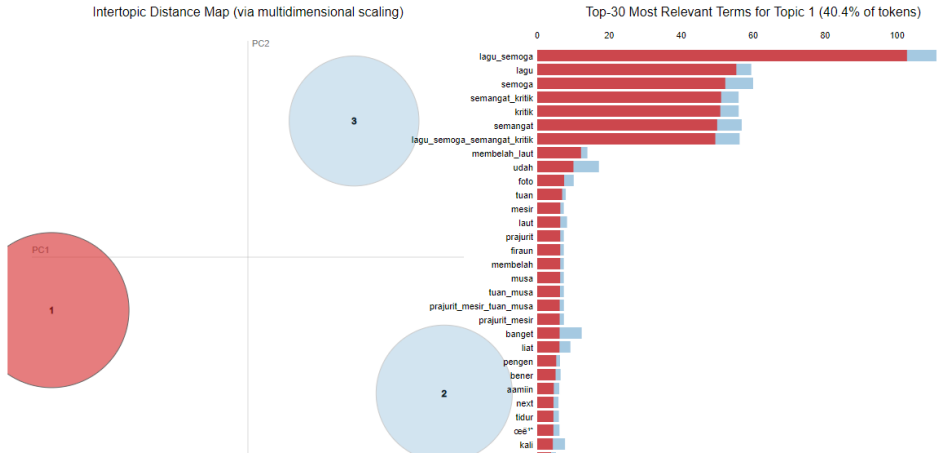
Gambar 17 menampilkan 30 kata atau istilah paling penting yang saling terkait antara topik yang satu dengan topik lainnya. 30 kata tersebut dapat dilihat pada tabel 11 berikut:

Tabel 11. 30 Istilah Paling Penting pada 3 Topik

No	Kata atau Istilah	No	Kata atau Istilah
1	lagu_semoga	**16**	sepatu
2	lagu	**17**	ayobantu_retweet
3	semangat_kritik	**18**	tokotamz
4	kritik	**19**	pinbb_jual_sepatu_baju
5	semoga	**20**	pinbb
6	semangat	**21**	tuhan
7	lagu_semoga_semangat_kritik	**22**	pinbb_jual
8	aksesoris_cewek	**23**	cekidot_tokotamz
9	sepatu_baju	**24**	makan
10	orang	**25**	ayobantu_retweet_cekidot_tokotamz
11	makassar	**26**	ayobantu
12	jual	**27**	cekidot
13	salah	**28**	aksesoris

No	Kata atau Istilah	No	Kata atau Istilah
14	ngedate	**29**	nonton
15	retweet	**30**	selamat

Frekuensi kata atau istilah dalam topik 1 dan hubungannya dengan topik lainnya diperlihatkan pada gambar 18 berikut:



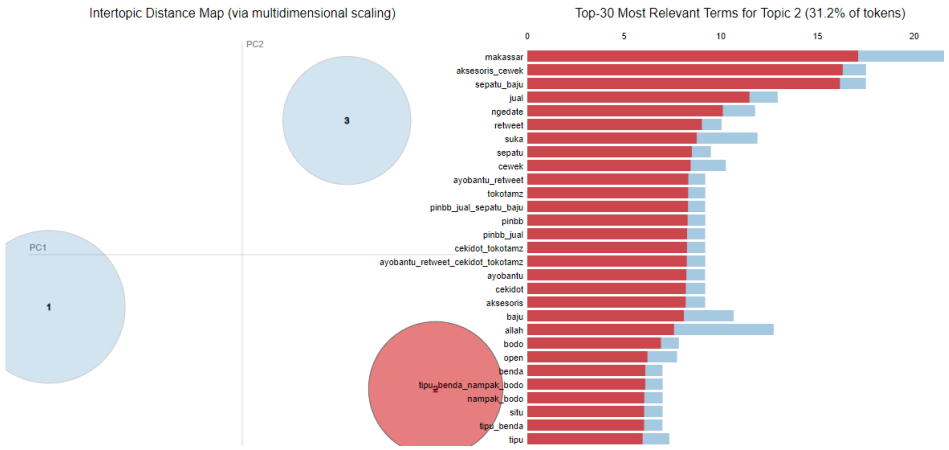
Gambar 18. Visualisasi Frekuensi Kata dan Hubungan pada Topik 1

Berdasarkan visualisasi grafik frekuensi kata pada gambar 18, Tabel 12 menampilkan 7 kata atau istilah pada topik 1 dengan frekuensi tertinggi.

Tabel 12. 7 Istilah Paling Penting pada Topik 1

No	Kata atau Istilah
1	lagu_semoga
2	lagu
3	semoga
4	semangat_kritik
5	kritik
6	semangat
7	membelah laut

Frekuensi kata atau istilah dalam topik 2 dan hubungannya dengan topik lainnya diperlihatkan pada gambar 19 berikut:



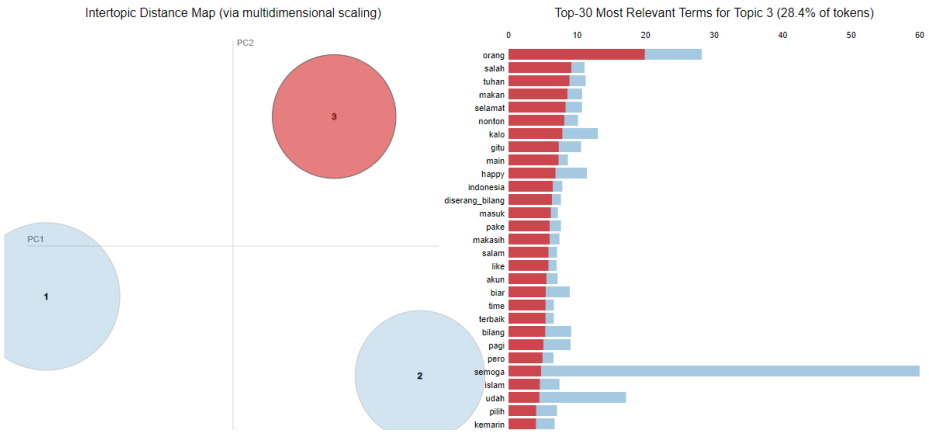
Gambar 19. Visualisasi Frekuensi Kata dan Hubungan pada Topik 2

Berdasarkan visualisasi grafik frekuensi kata pada gambar 19, Tabel 13 menampilkan 20 kata atau istilah pada topik 2 dengan frekuensi tertinggi.

Tabel 13. 20 Istilah Paling Penting pada Topik 2

No	Kata atau Istilah	No	Kata atau Istilah
1	makassar	**11**	tokotamz
2	aksesoris_cewek	**12**	pinbb_jual_sepatu_baju
3	sepatu_baju	**13**	pinbb
4	jual	**14**	pinbb_jual
5	ngedate	**15**	cekidot_tokotamz
6	retweet	**16**	ayobantu_retweet_cekidot_tokotamz
7	suka	**17**	ayobantu
8	sepatu	**18**	cekidot
9	cewek	**19**	aksesoris
10	ayobantu_retweet	**20**	baju

Frekuensi kata atau istilah dalam topik 3 dan hubungannya dengan topik lainnya diperlihatkan pada gambar 20 berikut:



Gambar 20. Visualisasi Frekuensi Kata dan Hubungan pada Topik 3

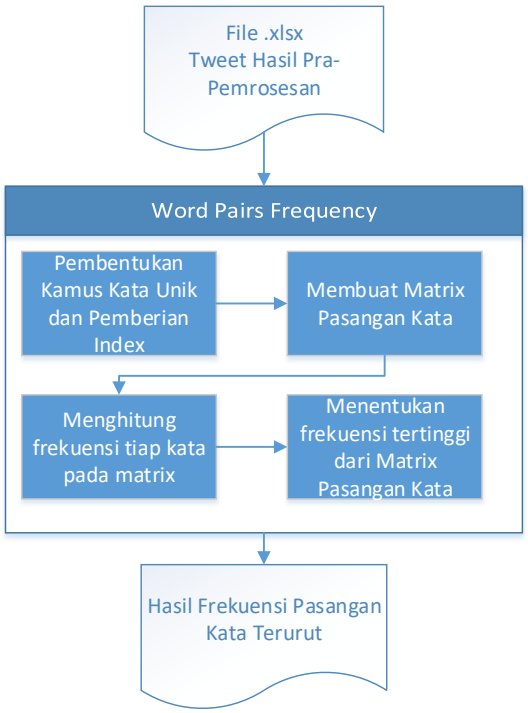
Berdasarkan visualisasi grafik frekuensi kata pada gambar 20, Tabel 13 menampilkan 6 kata atau istilah pada topik 3 dengan frekuensi tertinggi.

Tabel 14. 6 Istilah Paling Penting pada Topik 3

No	Kata atau Istilah
1	Orang
2	Salah
3	Tuhan
4	Makan
5	Selamat
6	nonton

D. WORD PAIRS FREQUENCY

Tahap pembentukan Word Pairs Frequency dapat dilihat pada gambar 21 berikut:



Gambar 21. Tahapan Pembentukan Word Pairs Frequency

1. Pembentukan Kamus Kata

Tahap proses pembentukan kamus kata pada metode word pairs frequency memiliki kemiripan konsep dengan pembentukan Dictionary dan Corpus pada Topic Modeling dengan menggunakan LDA. Tiap kata yang diproses diberi indeks dan disusun dalam dokumen berbentuk matrix.

Source Code yang diimplementasikan untuk pembentukan kamus kata dengan metode word pairs frequency ditampilkan pada kode 11 berikut:

```
xrange = range
columns = word_list
ncols = word_list_len + 1

term_doc = pd.DataFrame(columns = columns)
term_doc.insert(0, "Tweet", " ")
term_doc["Tweet"] = tweetDF["text"]
term_doc.fillna(0, inplace=True)

i_row = 0
```

```
for line in tweet_clean_fin:
    for word in line:
        for col in xrange(1, ncols-1):
            if word == term_doc.columns[col]:
term_doc.iloc[i_row, col] += 1
    i_row += 1

statDF = copy.deepcopy(term_doc)
columns_cl = ["Tweet", "Sim"]
tweet_sim = pd.DataFrame(columns = columns_cl)
tweet_sim = tweetDF["text"]
tweet_sim.fillna(0.0, inplace=True)

row_sum = statDF.sum(axis=1)
statDF["Total"] = row_sum
print('Row Max Value = ', row_sum.max())
print("Max Value DF = ", statDF["Total"].max(axis=0))

col_list = list(statDF)
col_list.remove('Tweet')

rsum = {col: statDF[col].sum() for col in col_list}

sum_df = pd.DataFrame(rsum, index=["Total"])
statDF = statDF.append(sum_df)
```

Kode 11. Source Code Pembentukan Kamus Kata

2. Pembuatan Matrix Pasangan Kata

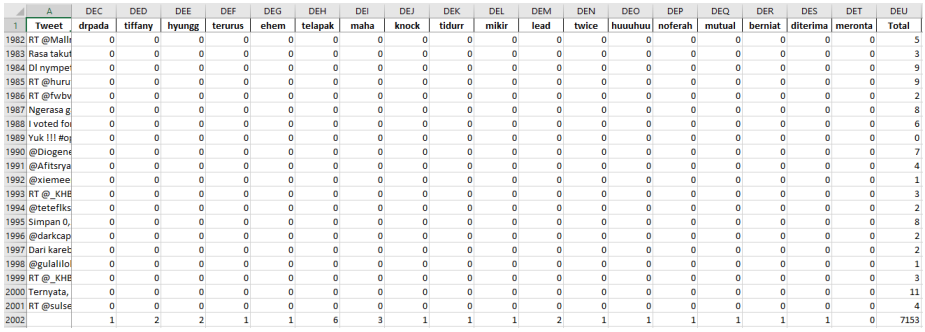
	A	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	Tweet	duit	everything	nyariin	peca	pusat	golok	minhyun	anfaal	lewati	villain	guay	gila	salon	bellaku	engkau	rumit	enyehatk	innovator	sampul
2	RT @ladaz	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	RT @hanb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	RT @fucel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	@bbyidio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	@ayanger	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	RT @ladaz	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	RT @ladaz	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Nungguin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	@tengom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	@mochaa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	@bayjing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	@tengom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	@szadgyr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	@sanca07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	@Reginaf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	RT @ladaz	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	@ayayawi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	@Virgoni	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	S a m p a h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	RT @chae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	New post	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 22. Hasil Pembuatan Matrix Pasangan Kata

Gambar 22 menampilkan hasil pembuatan matrix untuk pasangan kata yang akan dihitung frekuensi kemunculannya. Setiap tweet yang telah melalui tahap pra-pemrosesan data dibentuk kedalam list baris, dan setiap kata unik dari tweet tersebut dibentuk kedalam list kolom. Setelah melalui

tahap pra-pemrosesan dan pembentukan kamus kata, diperoleh sejumlah 2.853 kata unik dari 2000 data tweet yang uji.

3. Menghitung Frekuensi Kata pada Matrix



	A	DEC	DED	DEE	DEF	DEG	DEH	DEI	DEJ	DEK	DEL	DEM	DEN	DEO	DEP	DEQ	DER	DES	DET	DEU
1	Tweet	drpada	tiffany	hyungg	terurus	ehem	telapak	maha	knock	tidurr	mikir	lead	twice	huuuhuu	noferah	mutual	berniat	diterima	meronta	Total
1982	RT @Mall	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
1983	Rasa takut	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
1984	DI nympe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
1985	RT @huru	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
1986	RT @fvbv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1987	Ngerasa g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
1988	I voted for	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
1989	Yuk !!! #o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1990	@Diogene	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
1991	@Afitsrya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
1992	@xiemee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1993	RT @_KHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
1994	@tetefiks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1995	Simpan 0,	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
1996	@darkcap	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1997	Dari kareb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1998	@gulalilol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1999	RT @_KHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
2000	Ternyata,	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
2001	RT @sulse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
2002		1	2	2	1	1	6	3	1	1	1	2	1	1	1	1	1	1	0	7153

Gambar 23. Matrix Nilai Hasil Frekuensi Kata

Untuk memperoleh nilai hasil frekuensi kata, setiap kata unik pada matrix dipasangkan dengan setiap tweet yang diproses. Jika kata unik memiliki kesamaan dengan tweet, maka pada titik pertemuannya diberi nilai penambahan 1.

Nilai dari setiap titik adalah total jumlah dari pertemuan antara kata unik dan tweet. Pada tabel 15 diperlihatkan 10 total nilai dari tweet dan kata unik.

Tabel 15. Total Nilai Tweet pada Matrix

No	Tweet	Nilai
1	bunsay foto keluarga azula family	5
2	deket udah akrab banget kalo sokap gimana temen xixi	9
3	nungguin admin kirim link toefl bikin diriku ngantuk beraaatt nungguin	10
4	yukk minat produk young living langsung	6
5	prajurit mesir tuan musa membelah laut firaun	7
6	dahlah cape	2
7	ahahaha emang orang jawa	4

No	Tweet	Nilai
8	new post rapat bamus dprd makassar membahas jadwal reses sosper berubah been published baca pesan new	14
9	terurus kebersihannya urus makassar	4
10	nomor pelaporan maret kali mati lampu bintuni papua baratâ	9

Source Code yang diimplementasikan untuk membentuk matrix dan menghitung frekuensi kata ditampilkan pada kode 13 berikut:

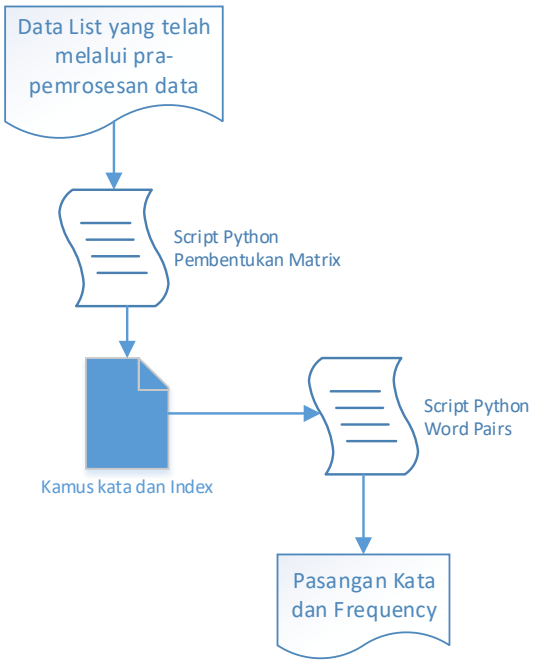
```
tup_word = []
sim_word = np.zeros((ncols, ncols))

for i in xrange(ncols-1):
    v1 = [0.0]*ncols
    v1 = term_doc.iloc[:, i+1]

    for k in xrange(ncols-1):
        v2 = [0.0]*ncols
        if i >= k: pass
        else:
            v2 = term_doc.iloc[:, k+1]
            similar = cosine_sim(v1, v2)
            tup_w = (similar, list(columns)[i], list(columns)[k])

            tup_word.append(tup_w)
            sim_word[i,k] = similar
            sim_word[k,i] = similar

    sim_word[i,i] = 1.0

sim_word[ncols-1,ncols-1] = 1.0

print('Similarity for Words: Words = ', word_list_len)
print(sim_word)

tu_tweet = []
sim_tweet = np.zeros((num_tweets, num_tweets))

for i in xrange(num_tweets):
    v1 = [0.0]*num_tweets
    v1 = term_doc.iloc[i, 1:]

    for k in xrange(num_tweets):
        v2 = [0.0]*num_tweets
        if i >= k: pass
        else:
            v2 = term_doc.iloc[k, 1:]
            similar = cosine_sim(v1, v2)
```

```
            tup_twe = (similar, term_doc['Tweet'][i],
term_doc['Tweet'][k])
            tu_tweet.append(tup_twe)

            sim_tweet[i, k] = similar
            sim_tweet[k, i] = similar
    sim_tweet[i,i] = 1.0

print('
')
print("Similarity for Tweets: Tweets = ", num_tweets)
print(sim_tweet)

statDF.tail()

statDF.to_excel('sim_twett.xlsx', index=False, encoding='utf-8')
```

Kode 12. Source Code Pembentukan Matrix dan Menghitung Frekuensi Kata

4. Penentuan Frekuensi Pasangan Kata

Setelah melalui tahap perhitungan frekuensi kata pada matrix, dilakukan proses penentuan frekuensi pasangan kata. Alur proses penentuan frekuensi pasangan kata dapat dilihat pada gambar 24 berikut:



Gambar 24. Alur Penentuan Frekuensi Pasangan Kata

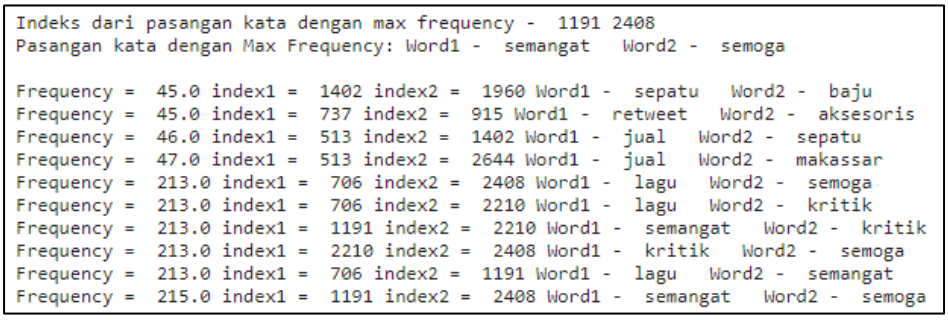
Dalam prosesnya, langkah-langkah perhitungan frekuensi pasangan kata yaitu sebagai berikut:

1) Mencari kata dengan frekuensi tertinggi pertama,
2) Mencari kata dengan frekuensi tertinggi kedua dengan mengacu pada kata dengan frekuensi tertinggi yang pertama,
3) Membentuk pasangan dari kedua kata tersebut,
4) Menghitung frekuensi kemunculan dari setiap pasang kata,
5) Mengurutkan frekuensi pasangan kata dari yang terbesar ke yang terkecil,
6) Menampilkan 10 pasang kata dengan frekuensi tertinggi.

Source Code yang diimplementasikan untuk penentuan frekuensi pasangan kata ditampilkan pada kode 14 berikut:

```
def tweet_prep(df):
    tweet_list = df['Tweet'].tolist()
    tweet_list_clean = df['Clean_Tweet'].tolist()
    word_list_cl = [[word for word in str(line).split()] for line
in tweet_list_clean]
    word_list_tot = list(chain.from_iterable(word_list_cl))

    set_word = set(word_list_tot)

    return Pair_words(set_word, tweet_list_clean, n_top)

print("Top ", n_top, " pairs of words counter")
most_comm = Pair_words(word_list, tweet_clean_fin, n_top)
most_comm
```

Kode 13. Source Code Penentuan Frekuensi Pasangan Kata



```
Indeks dari pasangan kata dengan max frequency -  1191 2408
Pasangan kata dengan Max Frequency: Word1 -  semangat   Word2 -  semoga

Frequency =  45.0 index1 =  1402 index2 =  1960 Word1 -  sepatu   Word2 -  baju
Frequency =  45.0 index1 =  737 index2 =  915 Word1 -  retweet   Word2 -  aksesoris
Frequency =  46.0 index1 =  513 index2 =  1402 Word1 -  jual   Word2 -  sepatu
Frequency =  47.0 index1 =  513 index2 =  2644 Word1 -  jual   Word2 -  makassar
Frequency =  213.0 index1 =  706 index2 =  2408 Word1 -  lagu   Word2 -  semoga
Frequency =  213.0 index1 =  706 index2 =  2210 Word1 -  lagu   Word2 -  kritik
Frequency =  213.0 index1 =  1191 index2 =  2210 Word1 -  semangat   Word2 -  kritik
Frequency =  213.0 index1 =  2210 index2 =  2408 Word1 -  kritik   Word2 -  semoga
Frequency =  213.0 index1 =  706 index2 =  1191 Word1 -  lagu   Word2 -  semangat
Frequency =  215.0 index1 =  1191 index2 =  2408 Word1 -  semangat   Word2 -  semoga
```

Gambar 25. Frekuensi Pasangan Kata

Hasil frekuensi pasangan kata pada gambar 25 memperlihatkan 10 pasangan kata dengan bobot dan indeks masing-masing kata. Frekuensi tertinggi memiliki bobot 215.0 pada pasangan kata “semangat”, indeks 1191 dan kata “semoga”, indeks 2408. Berikut ditampilkan pada tabel 16 untuk frekuensi setiap pasangan kata.

Tabel 16. Frekuensi Pasangan Kata

No	Frekuensi	Indeks1	Kata1	Indeks2	Kata2
1	45.0	1402	sepatu	1960	baju
2	45.0	737	retweet	915	aksesoris
3	46.0	513	jual	1402	sepatu
4	47.0	513	jual	2644	makassar
5	213.0	706	lagu	2408	semoga
6	213.0	706	lagu	2210	kritik
7	213.0	1191	semangat	2210	kritik
8	213.0	2210	kritik	2408	semoga
9	213.0	706	lagu	1191	semangat
10	215.0	1191	semangat	2408	semoga

E. ANALISIS HASIL

Proses pembentukan frekuensi kata yang dilakukan pada penelitian ini menggunakan 2 metode untuk memperoleh 2 hasil model yang dibutuhkan

untuk dianalisa. Yang pertama menggunakan metode Topic Modeling dengan LDA untuk menentukan frekuensi kata berdasarkan topik yang memiliki skor koherensi terbaik. Yang kedua menggunakan metode Word Pairs Frequency untuk mendapatkan hasil frekuensi dari sejumlah pasangan kata. Kedua metode tersebut membutuhkan data berupa teks yang telah melalui tahap pengambilan data dan tahap pra-pemrosesan data.

Kualitas hasil yang diperoleh dari kedua metode tersebut sangat bergantung pada data yang diambil dan pada tahap pra-pemrosesan data. Jika data yang diambil dan proses pada tahap pra-pemrosesan data masih terdapat kekurangan, maka bentuk kekurangan tersebut akan tetap diproses dan akan mempengaruhi kualitas dari hasil yang diharapkan.

1. Analisis Hasil Topic Modeling dengan LDA

Topic Modeling dengan LDA pada penelitian ini digunakan untuk mengukur frekuensi kemunculan kata dan membaginya menjadi bentuk topik-topik yang berbeda. Setiap topik diwakilkan dengan 30 kata atau istilah dengan frekuensi kemunculan terbanyak.

Jumlah topik yang dihasilkan dari penelitian ini sebanyak 3 topik berdasarkan skor korensi tertinggi pada 20 topik yang telah diuji. 3 dari 20 topik dengan skor koherensi tertinggi ditampilkan pada tabel 17 berikut:

Tabel 17. 3 Topik dengan Skor Koherensi Tertinggi

No	Nomor Topik	Skor Koherensi
1	13	0.665045
2	4	0.649874
3	9	0.642377

Jumlah topik yang diperoleh kemudian divisualisasikan dalam bentuk grafik Topic Modelling untuk melihat kata atau istilah yang memiliki frekuensi tertinggi dari masing-masing topik. Frekuensi kemunculan tiap kata atau istilah dari masing-masing topik dapat dilihat pada tabel 18 berikut:

Tabel 18. Frekuensi Kemunculan Kata berdasarkan Topic Modeling

Topik 1	**Skor**	**Topik 2**	**Skor**	**Topik 3**	**Skor**
lagu_semoga	0.129	makassar	0.028	orang	0.035
lagu	0.069	aksesoris_cewek	0.026	salah	0.016
semoga	0.065	sepatu_baju	0.026	tuhan	0.016
semangat_kritik	0.064	jual	0.019	makan	0.015
kritik	0.064	ngedate	0.016	selamat	0.015
semangat	0.063	retweet	0.015	nonton	0.015
membelah laut	0.015	suka	0.014	kalo	0.014
udah	0.013	sepatu	0.014	gitu	0.013
tuan	0.011	cewek	0.014	main	0.013
foto	0.009	ayobantu_retweet	0.013	happy	0.012

Masing-masing topik memuat kata atau istilah yang dapat dianggap sebagai sebuah topik. Topik 1 dapat diasumsikan bahwa topik yang dibicarakan yaitu tentang sebuah lagu yang dapat dijadikan penyemangat bagi para pengguna twitter saat itu. Topik 2 membicarakan tentang aktivitas jual beli *online* yang berlangsung di kota Makassar dan topik 3 tentang aktivitas orang-orang di kehidupan sehari-hari.

Dengan menggunakan pendekatan Topic Modeling, frekuensi kemunculan setiap kata atau istilah tidak ditampilkan berdasarkan jumlah kemunculan melainkan berdasarkan bobot skor yang telah melalui proses

menggunakan algoritma LDA. Sehingga pada tabel 18 dapat dilihat secara keseluruhan bahwa frekuensi kata dengan skor tertinggi dan terendah dihasilkan oleh topik 1 pada kata atau istilah “lagu_semoga” dengan bobot nilai 0.129 dan kata “foto” dengan bobot nilai 0.009.

Pembentukan topik dengan pendekatan Topic Modeling menunjukkan frekuensi kata atau istilah berdasarkan urutan dari topik yang terbentuk. Topik 1 dengan frekuensi tertinggi, topik 2 dengan frekuensi sedang dan topik 3 dengan frekuensi rendah.

2. Analisis Hasil Word Pairs Frequency

Metode Word Pairs Frequency pada penelitian ini digunakan untuk menentukan frekuensi kemunculan pasangan kata yang saling terhubung.

```
Top  10  pairs of words counter
[(('semangat', 'semoga'), 215),
 (('lagu', 'semoga'), 213),
 (('lagu', 'kritik'), 213),
 (('semangat', 'kritik'), 213),
 (('kritik', 'semoga'), 213),
 (('lagu', 'semangat'), 213),
 (('jual', 'makassar'), 47),
 (('jual', 'sepatu'), 46),
 (('sepatu', 'cekidot'), 45),
 (('ayobantu', 'jual'), 45)]
```

Gambar 26. 10 Urutan Tertinggi Hasil Frekuensi Kata

Dari hasil penentuan frekuensi pasangan kata, dilakukan pengurutan pada pasangan kata berdasarkan frekuensi tertinggi. Gambar 25 menampilkan 10 urutan teratas dari tiap pasangan kata dan ditampilkan secara rinci pada tabel 19 berikut:

Tabel 19. Urutan Pasangan Kata 10 Teratas

No	Pasangan Kata	Frekuensi
1	“semangat”, “semoga”	215
2	“lagu”, “semoga”	213
3	“lagu”, “kritik”	213
4	“semangat”, “kritik”	213
5	“kritik”, “semoga”	213
6	“lagu”, “semangat”	213
7	“jual”, “makassar”	47
8	“jual”, “sepatu”	46
9	“sepatu”, “cekidot”	45
10	“ayobantu”, “jual”	45

Dapat dilihat pada Tabel 19, nomor urut 1 hingga nomor urut 6 menunjukkan sebuah topik yang sama yaitu tentang sebuah lagu yang dapat dijadikan penyemangat bagi para pengguna twitter saat itu. Dan nomor urut 7 hingga nomor urut 10 membicarakan tentang aktivitas jual beli *online* yang berlangsung di kota Makassar.

Dengan pendekatan Word Pairs Frequency, terdapat pasangan kata atau istilah yang berhubungan maupun tidak berhubungan. Pada tabel 19, nomor urut pasangan kata 9 pada kata “cekidot” merupakan sebuah kata yang tidak baku dan dianggap tidak berhubungan dengan kata yang dimaksud pada penelitian ini karena merupakan bentuk bahasa yang tidak baku. Berdasarkan penelitian tentang penggunaan bahasa gaul oleh Nurul Wijiasih (2016), kata atau istilah “cekidot” sebenarnya adalah bentuk akronim dari kata berbahasa inggris yaitu *check it out* yang bermakna “silahkan dicek” atau “silahkan dilihat” (Wijiasih, 2016). Kata “cekidot” tersebut muncul akibat dari tidak terstrukturnya data teks tweet yang digunakan sebagai objek pada penelitian ini dan tidak termasuk dalam kata yang harus dihilangkan pada tahap pra-pemrosesan data.

BAB V
KESIMPULAN DAN SARAN

A. KESIMPULAN

Berdasarkan hasil analisis diatas, dapat disimpulkan:

1. Dengan pendekatan Topic Modeling, terdapat frekuensi kata atau istilah dengan frekuensi tertinggi, sedang dan rendah berdasarkan urutan topik yang terbentuk.

   Topik 1 dengan frekuensi tertinggi pada kata atau istilah “lagu_semoga”, “lagu”, “semoga”, “semangat_kritik”, “kritik”, dan “semangat”. Topik 2 dengan frekuensi sedang pada kata atau istilah “makassar”, “aksesoris_cewek”, “sepatu_baju”, dan “jual” dan topik 3 dengan frekuensi yang rendah pada kata atau istilah “orang”, “salah”, “tuhan”, “makan”, “selamat”, dan “nonton”.
2. Dengan menggunakan Topic Modeling, tidak dapat diketahui frekuensi setiap kata atau istilah berdasarkan jumlah kemunculan. Frekuensi kemunculan setiap kata atau istilah ditampilkan berdasarkan bobot skor yang telah melalui proses menggunakan algoritma LDA. Sehingga kata atau istilah dengan frekuensi yang muncul hanya sekali saja tidak dapat ditampilkan.
3. Terdapat pasangan kata yang saling berhubungan dengan menggunakan pendekatan Word Pairs Frequency. Pasangan-pasangan kata “semangat semoga”, “lagu semoga”, “lagu kritik”, “semangat kritik”, “kritik semoga”, dan “lagu semangat” menunjukkan sebuah topik yang sama yaitu tentang lagu yang dapat dijadikan penyemangat bagi para pengguna twitter saat itu. Dan pasangan-pasangan kata “jual makassar”, “jual sepatu”, “sepatu

cekidot", dan "ayobantu jual" membicarakan tentang aktivitas jual beli online yang berlangsung di kota Makassar.

4. Terdapat kata atau istilah pada pasangan kata yang tidak baku yaitu pada kata "cekidot". Kata "cekidot" tersebut muncul akibat dari tidak terstrukturnya data teks tweet yang digunakan sebagai objek pada penelitian ini.

B. SARAN

Sumber data pada penelitian ini berasal dari media sosial Twitter dengan pengambilan data tweet secara umum. Hal ini mengakibatkan adanya sejumlah kata yang tidak sesuai kaidah penulisan ikut serta dalam tahap proses oleh sistem, sehingga akurasi perhitungan frekuensi kata menjadi rendah. Dari permasalahan diatas, maka saran untuk pengembangan penelitian ini yaitu penentuan sumber data untuk perhitungan frekuensi kata menggunakan metode pada penelitian ini lebih dikhususkan pada sumber-sumber informasi yang menggunakan bahasa yang baku dan terstruktur, sehingga dalam pemrosesannya setiap kata dari sumber data memiliki nilai informasi yang akurat dan frekuensi kata dan pasangan kata yang terbentuk memiliki informasi dengan akurasi yang tepat.

Pada penelitian ini, tahap proses dengan Topic Modeling dengan LDA dan tahap proses dengan metode Word Pairs Frequency dilakukan secara terpisah. Maka saran untuk pengembangan penelitian ini yaitu menggabungkan kedua metode tersebut dan memperoleh hasil Topic Modeling pada pasangan kata.

DAFTAR PUSTAKA

Agustina, L. (2018). Pemanfaatan Media Sosial untuk Implementasi e-Government. *Mediakom*, *13*(November 2015), 0–6.

Ardi, Z., & Sukmawati, I. (2017). Social Media and the Quality of Subjective Well-Being ; Counseling Perspective in Digital Era. *International Conseling and Education Seminar : The Responsibility of Counselor and Educator in Millennium Era*, 28–35.

Badan Pusat Statistik. (2010). *Sensus Penduduk 2010 - Kota Makassar*. https://sp2010.bps.go.id/index.php/site?id=7371000000&wilayah=Kota-Makassar

Buchanan, E., & Padfield, W. (2019). Using Word Frequencies to Analyze Political Language and Moral Focus. *Using Word Frequencies to Analyze Political Language and Moral Focus*. https://doi.org/10.4135/9781526491398

Cakranegara, P. A., & Susilowati, E. (2017). Analisis strategi implementasi media sosial (studi kasus ukm “xyz”). *Perusahaan Studi Manajemen*, *2*(2), 1–16.

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. *Jurnal Teknologi Informasi*, *3*(2), 179. https://doi.org/10.36294/jurti.v3i2.1077

Dicle, M. F., & Dicle, B. (2018). Content Analysis: Frequency Distribution of Words. *SSRN Electronic Journal*, *October*. https://doi.org/10.2139/ssrn.2997101

Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts and Humanities*, *3*(1). https://doi.org/10.1080/23311983.2016.1171458

Hu, W. (2013). Real-Time Twitter Sentiment toward Thanksgiving and Christmas Holidays. *Social Networking*, *02*(02), 77–86. https://doi.org/10.4236/sn.2013.22009

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., Kolodziej, J., Wang, L., Chen, D., Rayes, A., Tziritas, N., Xu, C. Z., Zomaya, A. Y., Alzahrani, A. S., & Li, H. (2015). A survey on text mining in social networks. *Knowledge Engineering Review*, *30*(2), 157–170. https://doi.org/10.1017/S0269888914000277

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. In *Multimedia Tools and Applications* (Vol. 78, Issue 11). https://doi.org/10.1007/s11042-018-6894-4

Juditha, C. (2017). Memahami Struktur Jaringan Media Sosial sebagai Cara Strategis Periklanan di Era Ekonomi Digital Understanding Social Media Network Structure as a Strategic Way of Advertising in Digital Economy Era. *Jurnal Pekommas, Vol. 2 No. 1, April 2017: 99-114*, *2*(1), 99–114.

Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line trend analysis with topic models: Twitter trends detection topic model online. *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, *2*(December), 1519–1534.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331. https://doi.org/10.1017/S0003055403000698

Lijffijt, J., Papapetrou, P., Puolamäki, K., & Mannila, H. (2011). Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6912 LNAI*(PART 2), 341–357. https://doi.org/10.1007/978-3-642-23783-6_22

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Lrec 2012*, *June*, 15–22. http://www.lrec-conf.org/proceedings/lrec2012/workshops/21.LREC2012 NLP4UGC Proceedings.pdf#page=20%5Cnhttp://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf

Negara, E. S., Andryani, R., & Saksono, P. H. (2016). Analisis Data Twitter: Ekstraksi dan Analisis Data G eospasial. *Jurnal INKOM*, *10*(1), 27. https://doi.org/10.14203/j.inkom.433

Pennebaker, J. W., & Chung, C. K. (2013). Counting little words in big data: The psychology of individuals, communities, culture, and history. *Social Cognition and Communication*, 25–42. https://doi.org/10.4324/9780203744628

Piepenbrink, A., & Gaur, A. S. (2017). Topic models as a novel approach to identify themes in content analysis: The example of organizational research

methods. *2017 Annual Meeting of the Academy of Management, AOM 2017*, *2017-Augus*(August). https://doi.org/10.5465/AMBPP.2017.141

Rana, S. M. R. (2015). Location Based Popularity Analysis of Twitter Data. *Tesis*, *151*, 10–17. https://doi.org/10.1145/3132847.3132886

Setiawan, H., & Santoso, P. (2013). Model Optimalisasi Peluang Pemanfaatan Media Jejaring Sosial dalam Implementasi E-Governance di Indonesia. *Jurnal Informatika. UPN "Veteran" Yogyakarta.*, *2013*(semnasIF), 147–154.

Siddharth, S., Darsini, R., & Sujithra, M. (2018). Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python. *ISSN (Online) 2394-2320 International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, *5*(2), 285–291.

Situmorang, S. H., Mulyono, H., & Berampu, L. T. (2018). Peran dan Manfaat Sosial Media Marketing bagi Usaha Kecil. *AJEFB - Asian Journal of Entrepreneurship and Family Business*, *1*(2), 77–84.

Sri Arini, N. W., Putu Widja, I. B., & Yasa Negara, I. K. R. (2019). Analisis Frekuensi Kata untuk Mengekstrak Kata Kunci dari Artikel Ilmiah Berbahasa Indonesia. *Eksplora Informatika*, *8*(2), 80–84. https://doi.org/10.30864/eksplora.v8i2.162

Tala, F. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tong, Z., & Zhang, H. (2016). *A Text Mining Research Based on LDA Topic Modelling*. 201–210. https://doi.org/10.5121/csit.2016.60616

Twitter. (2020). *Tentang API Twitter*. https://help.twitter.com/id/rules-and-policies/twitter-api

We Are Social. (2020). Digital 2020: Indonesia. *Global Digital Insights*, 17. https://datareportal.com/reports/digital-2020-indonesia

Wijiasih, N. (2016). Penggunaan Kata Gaul pada Mahasiswa Pendidikan Bahasa dan Sastra Jawa Unnes. In *Universitas Negeri Semarang*. Universitas Negeri Semarang.

Yeh, J.-F., Tan, Y.-S., & Lee, C.-H. (2016). Topic detection and tracking for

conversational content by using conceptual dynamic latent Dirichlet allocation. *Neurocomputing*, *216*, 310–318. https://doi.org/10.1016/j.neucom.2016.08.017

Yuyun, Nuzir, F. A., & Dewancker, B. J. (2017). Dynamic land-use map based on twitter data. *Sustainability (Switzerland)*, *9*(12), 1–20. https://doi.org/10.3390/su9122158