

**PENERAPAN SUPPORT VECTOR MACHINE UNTUK
DETEKSI SARKASME PADA ANALISIS SENTIMEN MEDIA
SOSIAL INDONESIA**

TUGAS AKHIR

Diajukan sebagai syarat untuk menyelesaikan
Program Studi Strata-1 Departemen Informatika

Disusun Oleh:
Candra Ricky Susanto
1114046



**INSTITUT
TEKNOLOGI
HARAPAN
BANGSA**
School of Telematics

**DEPARTEMEN INFORMATIKA
INSTITUT TEKNOLOGI HARAPAN BANGSA
BANDUNG
2017**



LEMBAR PENGESAHAN

**PENERAPAN SUPPORT VECTOR MACHINE UNTUK DETEKSI
SARKASME PADA ANALISIS SENTIMEN MEDIA SOSIAL INDONESIA**



Disusun oleh:

Nama: Candra Ricky Susanto

NIM : 1114046

Telah Disetujui dan Disahkan

Sebagai laporan Tugas Akhir Departemen Informatika

Institut Teknologi Harapan Bangsa

Bandung, Desember 2017

Disetujui,

Pembimbing

Ria Chaniago, S.T., M.T.

NIK.



PERNYATAAN HASIL KARYA PRIBADI

Saya yang bertanda tangan di bawah ini:

Nama : Candra Ricky Susanto

NIM : 1114046

Dengan ini menyatakan bahwa laporan Tugas Akhir dengan Judul : ”
PENERAPAN SUPPORT VECTOR MACHINE UNTUK DETEKSI
SARKASME PADA ANALISIS SENTIMEN MEDIA SOSIAL INDONESIA”
adalah hasil pekerjaan saya dan seluruh ide, pendapat atau materi dari sumber lain
telah dikutip dengan cara penulisan referensi yang sesuai.

Pernyataan ini saya buat dengan sebenar-benarnya dan jika pernyataan ini tidak
sesuai dengan kenyataan maka saya bersedia menanggung sanksi yang akan
dikenakan pada saya.

Bandung, Desember 2017

Yang membuat pernyataan,

Candra Ricky Susanto

ABSTRAK

Analisis sentimen pada media sosial telah menjadi salah satu topik penelitian yang paling ditargetkan pada *Natural Language Processing* (NLP) [2]. Analisis sentimen ini bertujuan untuk menentukan nilai polaritas dari sebuah dokumen secara otomatis. Salah satu tantangan pada analisis sentimen adalah melakukan klasifikasi terhadap teks sarkasme [3]. Dalam penelitian ini, dikembangkan sistem analisis sentimen yang dapat melakukan klasifikasi teks positif, teks negatif, teks netral, dan teks sarkasme. Metode klasifikasi yang digunakan adalah *Support Vector Machine* (SVM). Beberapa fitur yang digunakan untuk memberikan informasi dari dokumen adalah *number of interjection word*, *question word* [5], *unigram*, *sentiment score*, *capitalization*, *topic* [4], *part of speech* dan *punctuation based* [3]. Pengujian dilakukan dengan 2 teknik klasifikasi, yaitu *levelled method* dan *direct method* [5]. Berdasarkan pengujian yang dilakukan, hasil akurasi mencapai 72% yang diperoleh menggunakan metode SVM dengan teknik klasifikasi *direct method*.

Kata Kunci: *Natural Language Processing*, *Classification*, *Support Vector Machine*, Analisis Sentimen

ABSTRACT

Sentiment analysis on social media has become one of the most targeted research topics in Natural Languange Processing (NLP) [2]. This sentiment analysis aims to determine the polarity value of a document automatically. One of the challenges in the sentiment analysis is to classify sarcasm text [3]. In this study, developed a system of sentiment analysis that can classify positive text, negative text, neutral text, and sarcasm text. The classification method used is Support Vector Machine (SVM). Some of the features used to provide information from documents are number of interjection word, question word [5], unigram, sentiment score, capitalization, topic [4], part of speech and punctuation based [3]. Testing is done by 2 classification techniques, namely levelled method and direct method [5]. Based on the tests performed, the accuracy result was 72% obtained using the SVM method with the classification technique direct method.

Keyword: Natural Language Processing, Classification, Support Vector Machine, Sentiment Analysis

PEDOMAN PENGGUNAAN TUGAS AKHIR

Laporan tugas akhir yang tidak dipublikasikan terdaftar dan tersedia di Perpustakaan Institut Teknologi Harapan Bangsa, dan terbuka untuk umum dengan ketentuan bahwa hak cipta ada pada pengarang dan pembimbing Tugas Akhir. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau peringkasan hanya dapat dilakukan dengan seizin pengarang atau pembimbing Tugas Akhir dan harus disertai dengan ketentuan penulisan ilmiah untuk menyebutkan sumbernya.

Tidak diperkenankan untuk memperbanyak atau menerbitkan sebagian atau seluruh laporan tugas akhir tanpa seizin dari pengarang atau pembimbing Tugas Akhir yang bersangkutan.

KATA PENGANTAR

Terima kasih kepada Tuhan yang Maha Esa karena dengan bimbingan-Nya dan karunia-Nya penulis dapat melaksanakan Tugas Akhir yang berjudul **"PENERAPAN SUPPORT VECTOR MACHINE UNTUK DETEKSI SARKASME PADA ANALISIS SENTIMEN MEDIA SOSIAL INDONESIA TENTANG PENELITIAN TUGAS AKHIR DEPARTEMEN INFORMATIKA"**. Laporan ini disusun sebagai salah satu syarat kelulusan di Institut Teknologi Harapan Bangsa. Pada kesempatan ini penulis menyampaikan terima kasih yang sebesar-besarnya kepada:

1. Tuhan Yang Maha Esa, karena oleh bimbingan-Nya penulis selalu mendapat pengharapan untuk menyelesaikan tugas akhir ini.
2. Ibu Ria Chaniago, S.T., M.T., selaku pembimbing I Tugas Akhir yang senantiasa memberi dukungan, semangat, ilmu-ilmu, saran dan dukungan kepada penulis selama tugas akhir berlangsung dan selama pembuatan laporan tugas akhir ini.
3. Bapak Firhat Hidayat, S.T., M.T., selaku penguji I Tugas Akhir. Terima kasih atas dukungan, semangat, ilmu-ilmu, dan masukan yang telah diberikan kepada penulis dalam menyelesaikan Laporan Tugas Akhir ini
4. Ibu Ir. Inge Martina, M.T., selaku penguji II dalam Tugas Akhir Terima kasih atas dukungan, semangat, ilmu-ilmu, dan masukan yang telah diberikan kepada penulis dalam menyelesaikan Laporan Tugas Akhir ini.
5. Seluruh dosen dan staff Departemen Teknik Informatika ITHB yang telah membantu dalam menyelesaikan Laporan Tugas Akhir ini.

6. Segenap jajaran staf dan karyawan ITHB yang turut membantu kelancaran dalam menyelesaikan Laporan Tugas Akhir ini.
7. Kedua orang tua tercinta yang selalu menyediakan waktu untuk memberikan doa, semangat dan dukungan yang tak habis-habisnya kepada penulis untuk menyelesaikan Laporan Tugas Akhir ini. Terima kasih untuk nasihat, masukan, perhatian, teguran dan kasih sayang yang diberikan hingga saat ini.

Penulis menyadari bahwa laporan ini masih jauh dari sempurna karena keterbatasan waktu dan pengetahuan yang dimiliki oleh penulis. Oleh karena itu, kritik dan saran untuk membangun kesempurnaan tugas akhir ini sangat diharapkan. Semoga tugas akhir ini dapat membantu pihak-pihak yang membutuhkannya.

Bandung, Agustus 2016

Hormat Saya,
Candra Ricky Susanto.

DAFTAR ISI

LEMBAR PENGESAHAN	i
LEMBAR PERNYATAAN HASIL KARYA PRIBADI	ii
ABSTRAK	iii
ABSTRACT	iv
PEDOMAN PENGGUNAAN TUGAS AKHIR	v
KATA PENGANTAR	vi
DAFTAR ISI	x
DAFTAR TABEL	xiv
DAFTAR GAMBAR	xv
I PENDAHULUAN	1-1
1.1 Latar Belakang Masalah	1-1
1.2 Rumusan Masalah	1-2
1.3 Batasan Masalah	1-2
1.4 Tujuan Penelitian	1-3
1.5 Kontribusi Penelitian	1-3
1.6 Metode Penelitian	1-3
1.7 Sistematika Penulisan	1-4
II LANDASAN TEORI	2-1
2.1 Tinjauan Pustaka	2-1
2.1.1 Pemrosesan Bahasa Alami	2-1

2.1.2	Analisis Sentimen	2-2
2.1.3	Pembelajaran Mesin	2-2
2.1.4	Text Preprocessing	2-3
2.1.5	<i>Feature Extraction</i>	2-8
2.1.6	Definisi Support Vector Machine (SVM)	2-13
2.1.7	Simplified Sequential Minimal Optimization (SMO)	2-17
2.1.8	Twitter-scraper	2-19
2.1.9	Natural Language Toolkit (NLTK)	2-19
2.1.10	Tweet-preprocessor	2-19
2.1.11	Sastrawi	2-20
2.1.12	IPosTagger	2-21
2.1.13	Gensim	2-21
2.1.14	NumPy	2-22
2.1.15	Scikit-learn	2-23
2.1.16	Bottle	2-24
2.2	Tinjauan Studi	2-25
2.3	Objek Penelitian	2-27
2.3.1	Twitter	2-27
2.3.2	Sarkasme	2-27

III ANALISIS DAN PERANCANGAN SISTEM	3-1	
3.1	Analisis Masalah	3-1
3.2	Kerangka Pemikiran	3-4
3.3	<i>Flowchart</i> Sistem Analisis Sentimen	3-4
3.3.1	Analisis Data	3-5
3.3.2	<i>Text Preprocessing</i>	3-6
3.3.3	Feature Extraction	3-12
3.3.4	Perhitungan <i>Support Vector Machine</i>	3-20
3.3.5	<i>Class Diagram</i>	3-25

IV IMPLEMENTASI DAN PENGUJIAN	4-1
4.1 Lingkungan Aplikasi	4-1
4.2 Daftar <i>Class</i> dan <i>Method</i>	4-1
4.2.1 <i>Class Loader</i>	4-1
4.2.2 <i>Class Preprocessing</i>	4-2
4.2.3 <i>Class Features</i>	4-5
4.2.4 <i>Class Feature Extraction</i>	4-7
4.2.5 <i>Class Sentiment Extraction</i>	4-11
4.2.6 <i>Class Negation</i>	4-12
4.2.7 <i>Class Topic</i>	4-12
4.2.8 <i>Class SVM</i>	4-13
4.2.9 <i>Class FMeasure</i>	4-16
4.2.10 <i>Class Learning</i>	4-17
4.2.11 <i>Class Main</i>	4-18
4.3 Implementasi Perangkat Lunak	4-19
4.3.1 Implementasi Pengambilan Data	4-19
4.3.2 Implementasi <i>Text Preprocessing</i>	4-20
4.3.3 Implementasi <i>Feature Extraction</i>	4-21
4.3.4 Implementasi SVM dengan SMO	4-23
4.4 Pengujian	4-24
4.4.1 Pengujian Kombinasi Fitur	4-24
4.4.2 Pengujian Parameter pada SMO	4-35
4.4.3 Pengujian Klasifikasi 4 kelas, 3 kelas dan 1 kelas	4-36
V PENUTUP	5-1
5.1 Kesimpulan	5-1
5.2 Saran	5-2

DAFTAR TABEL

2.1	Tabel <i>Derivation Prefix Rule</i>	2-7
2.1	Tabel <i>Derivation Prefix Rule</i>	2-8
2.2	Tabel <i>Tag Set</i> pada Bahasa Indonesia	2-9
2.2	Tabel <i>Tag Set</i> pada Bahasa Indonesia	2-10
2.3	Tabel persamaan kernel pada SVM	2-15
2.4	Tabel <i>One versus Rest</i>	2-16
2.5	Tabel <i>Pseudocode SMO</i>	2-18
2.6	Tabel metode pada <i>library</i> Twitter-scraper	2-19
2.7	Tabel metode pada <i>library</i> NLTK	2-19
2.8	Tabel metode pada <i>library</i> Tweet-preprocessor	2-20
2.9	Tabel metode pada <i>library</i> Sastrawi	2-20
2.9	Tabel metode pada <i>library</i> Sastrawi	2-21
2.10	Tabel metode pada <i>library</i> IPosTagger	2-21
2.11	Tabel metode pada <i>library</i> Gensim	2-22
2.12	Tabel metode pada <i>library</i> NumPy	2-23
2.13	Tabel metode pada <i>library</i> Scikit-learn	2-23
2.13	Tabel metode pada <i>library</i> Scikit-learn	2-24
2.14	Tabel metode pada <i>library</i> Bottle	2-24
2.15	Tabel Tinjauan Studi	2-25
2.15	Tabel Tinjauan Studi	2-26
3.1	Tabel Hasil <i>Case Folding</i>	3-7
3.2	Tabel Hasil <i>Remove Hashtag, URL, Mention</i>	3-8
3.3	Tabel Hasil <i>Remove Punctuation</i>	3-8
3.4	Tabel Hasil <i>Tokenization</i>	3-8
3.4	Tabel Hasil <i>Tokenization</i>	3-9

3.5	Tabel Hasil <i>Misuse of Word</i>	3-9
3.5	Tabel Hasil <i>Misuse of Word</i>	3-10
3.6	Tabel Hasil <i>Abbreviation Word</i>	3-10
3.6	Tabel Hasil <i>Abbreviation Word</i>	3-11
3.7	Tabel Hasil <i>Stopword Removal</i>	3-11
3.8	Tabel Hasil <i>Stemming</i>	3-12
3.9	Contoh Perhitungan Fitur <i>Unigram</i>	3-13
3.10	Contoh Perhitungan Fitur <i>Part of Speech</i>	3-13
3.11	Contoh Perhitungan Fitur <i>Sentiment Score</i>	3-14
3.12	Contoh Perhitungan Fitur <i>Punctuation Based</i>	3-15
3.13	Contoh Perhitungan Fitur <i>Capitalization</i>	3-15
3.14	Tabel Data <i>Training</i>	3-15
3.15	<i>Word Index</i>	3-16
3.16	<i>Token-topic</i>	3-16
3.17	<i>Word-topic 1</i>	3-16
3.18	<i>Word-topic 2</i>	3-16
3.19	<i>Word-topic 3</i>	3-17
3.20	<i>Document-topic</i>	3-17
3.21	Probabilitas <i>Word-topic</i>	3-17
3.22	Probabilitas <i>Document-topic</i>	3-17
3.23	Contoh Hasil Perhitungan Fitur <i>Topic</i>	3-18
3.24	Contoh Hasil Perhitungan Fitur <i>Interjection</i>	3-18
3.25	Contoh Hasil Perhitungan Fitur <i>Question Word</i>	3-18
3.26	Contoh Data untuk Perhitungan TF-IDF	3-19
3.27	Contoh Hasil Perhitungan TF	3-19
3.28	Contoh Hasil Perhitungan IDF	3-19
3.29	Contoh Hasil Perhitungan Fitur TF-IDF	3-20
3.30	Data <i>Training</i>	3-21

3.31 Data <i>Testing</i>	3-22
3.32 Hasil Perhitungan <i>Kernel Linear</i> pada Data <i>Training</i>	3-22
3.33 Nilai Alpha dan Bias pada perhitungan SMO	3-23
3.34 Hasil Perhitungan <i>Kernel Linear</i> pada Data <i>Testing</i>	3-24
4.1 Daftar <i>Method</i> pada <i>Class Loader</i>	4-2
4.2 Daftar <i>Method</i> pada <i>Class Preprocessing</i>	4-2
4.3 Daftar <i>Method</i> pada <i>Class Features</i>	4-5
4.4 Daftar <i>Method</i> pada <i>Class Feature Extraction</i>	4-8
4.5 Daftar <i>Method</i> pada <i>Class Sentiment Extraction</i>	4-11
4.6 Daftar <i>Method</i> pada <i>Class Negation</i>	4-12
4.7 Daftar <i>Method</i> pada <i>Class Topic</i>	4-12
4.8 Daftar <i>Method</i> pada <i>Class SVM</i>	4-13
4.9 Daftar <i>Method</i> pada <i>Class FMeasure</i>	4-16
4.10 Daftar <i>Method</i> pada <i>Class FMeasure</i>	4-17
4.11 Daftar <i>Method</i> pada <i>Class FMeasure</i>	4-18
4.12 Tabel Pengujian Kombinasi Fitur 1	4-25
4.12 Tabel Pengujian Kombinasi Fitur 1	4-26
4.13 Tabel Pengujian Kombinasi Fitur 2	4-27
4.14 Analisis <i>Error</i> Fitur <i>Unigram</i> pada Klasifikasi Non-Sarkasme	4-28
4.15 Analisis <i>Error</i> Fitur <i>Sentiment Score</i> pada Klasifikasi Non-Sarkasme	4-29
4.16 Analisis <i>Error</i> Fitur <i>Punctuation Based</i> pada Klasifikasi Non-Sarkasme	4-29
4.17 Analisis <i>Error</i> Fitur <i>Unigram</i> pada Klasifikasi Sarkasme	4-30
4.18 Analisis Fitur <i>Topic</i>	4-31
4.18 Analisis Fitur <i>Topic</i>	4-32
4.18 Analisis Fitur <i>Topic</i>	4-33
4.19 Analisis <i>Error</i> Fitur <i>Topic</i> pada Klasifikasi Sarkasme	4-34
4.20 Analisis <i>Error</i> Fitur <i>Capitalization</i> pada Klasifikasi Sarkasme	4-35

4.21 Pengujian Parameter SMO	4-36
4.22 Pengujian Klasifikasi 4 Kelas (Positif, Negatif, Netral, Sarkasme) . .	4-37
4.23 Pengujian Klasifikasi 3 Kelas (Positif, Negatif, Netral)	4-37
4.24 Pengujian Klasifikasi 1 Kelas (Sarkasme/Non-Sarkasme)	4-37

DAFTAR GAMBAR

2.1	<i>Graphical Model</i> pada <i>Smoothed LDA</i>	2-11
2.2	<i>Hyperplane</i> pada <i>SVM</i>	2-14
3.1	Klasifikasi dengan <i>Levelled Method</i>	3-2
3.2	Klasifikasi dengan <i>Levelled Method</i>	3-2
3.3	Kerangka kerja klasifikasi teks	3-4
3.4	<i>Flowchart</i> Sistem Analisis Sentimen	3-5
3.5	<i>Flowchart Text Preprocessing</i>	3-7
3.6	<i>Flowchart</i> Klasifikasi <i>Support Vector Machine</i> (SVM) dengan <i>SMO</i>	3-21
3.7	<i>Class Diagram</i> Sistem Analisis Sentimen	3-25

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Facebook adalah media sosial berbasis teks yang memiliki sejumlah konten buatan pengguna. Pada tahun 2011 tercatat 200 juta *tweet* yang dibuat setiap harinya dan mencakup berbagai topik [1]. Dengan jumlah data yang signifikan ini, data tersebut memiliki potensi penelitian terkait dengan *text mining*. Salah satu yang dapat dieksplorasi pada data media sosial adalah analisis opini dan sentimen [1]. Analisis sentimen pada media sosial telah menjadi salah satu topik penelitian yang paling ditargetkan pada *Natural Language Processing* (NLP) dalam dekade terakhir, seperti yang ditunjukkan dalam beberapa survei baru-baru ini [2]. Tujuan analisis sentimen adalah untuk mendeteksi polaritas secara otomatis pada sebuah dokumen, salah satu tantangan besar dalam analisis sentimen adalah ironi dan sarkasme [3]. Pada 5 Juni 2014, BBC melaporkan bahwa U.S. *Secret Service* sedang mencari sistem perangkat lunak yang dapat mendeteksi sarkasme pada data media sosial (BBC, 2014) dengan tujuan otomatisasi pengawasan media sosial dan analisis perangkat data media sosial dalam ukuran besar [4].

Sarkasme adalah teks ironi untuk mengejek atau menyampaikan penghinaan. Sarkasme sendiri mengubah nilai dari sebuah teks ke nilai yang berlawanan. Berdasarkan penelitian yang dilakukan mengenai deteksi sarkasme, 2 dari 100 teks mengandung sarkasme pada *microblogging* teks dengan topik pembicaraan seperti makanan, dan kesehatan [5]. *Microblogging* adalah sebuah *blog* yang menyediakan layanan untuk menulis pesan seperti Twitter dengan jumlah karakter kurang dari 200 [6]. Dan pada topik yang sensitif seperti pemerintahan, *brand*, atau politik ditemukan 18 dari 100 teks mengandung sarkasme [5].

Pada umumnya fitur-fitur yang digunakan pada analisis sentimen adalah *n-gram* (*unigram*, *bigram*). Fitur seperti *negativity* dan *number of interjection word* merupakan sebuah fitur yang digunakan untuk memberikan informasi terkait teks sarkasme [5]. Tujuan dari fitur *negativity* ini adalah untuk mendapatkan informasi global berupa *negativity* dari sebuah topik, dan fitur *number of interjection word* untuk menentukan kecenderungan sebuah teks dianggap

sarkasme berdasarkan kemunculan *interjection word* [5]. Fitur *negativity* ini memakan banyak waktu, Karena untuk memberi topik dan nilai *negativity* terhadap teks diperlukan pengetahuan yang cukup mengenai isi dari teks. Oleh karena itu, pada penelitian ini akan digunakan fitur *topic modelling* (LDA) untuk mendapatkan informasi global dari sebuah teks.

Beberapa klasifikasi yang sering digunakan untuk analisis sentimen adalah *Naive Bayes*, *SVM*, dan *Maximum Entropy*. Kelemahan dari metode *Naive Bayes* mengasumsikan antar variabel sebagai variabel bebas atau *independent*. Kelebihan dari *Naive Bayes* diantaranya mudah diimplementasikan, dan dapat memberikan hasil yang baik untuk banyak kasus [5]. Kelebihan dari *SVM* adalah dapat menggeneralisasi sampel dengan baik, jika parameter C dipilih dengan baik. Hal ini membuat *SVM* dapat menghindari *overfitting* jika memilih parameter yang sesuai. Kekurangan dari *SVM* adalah kurangnya transparansi hasil. *SVM* tidak dapat mewakili semua nilai sebagai fungsi parametrik sederhana, karena dimensinya sangat tinggi [7]. Kelebihan dari metode klasifikasi *Maximum Entropy* adalah fleksibel, karena *stochastic rule system* yang diperkuat dengan fitur *syntactic*, *semantic* dan *pragmatic* [8].

Berdasarkan penelitian yang telah dilakukan [3][4][5], secara keseluruhan *SVM* dapat memberikan hasil akurasi yang baik. Oleh Karena itu, pada penelitian ini akan digunakan metode *Support Vector Machine*. Fitur yang digunakan pada penelitian ini adalah kombinasi fitur *number of interjection word*, *question word* [5], *unigram*, *sentiment score*, *capitalization*, *topic* [4], *part of speech* dan *punctuation based* [3].

1.2 Rumusan Masalah

Berikut ini adalah rumusan masalah yang dibuat berdasarkan latar belakang di atas:

1. Bagaimana kombinasi fitur yang dapat memberikan hasil yang terbaik?
2. Bagaimana pengaruh fitur *topic* dengan LDA terhadap klasifikasi sarkasme?
3. Bagaimana akurasi dari klasifikasi jika menangani sarkasme, tidak menangani sarkasme, dan hanya menangani sarkasme?

1.3 Batasan Masalah

Berikut ini adalah batasan masalah dalam pembahasan dan pengembangan yang dilakukan:

1. Data yang digunakan untuk penelitian ini adalah data Twitter Bahasa Indonesia.

2. Semua teks sarkasme dianggap sebagai positif sarkasme.
3. Sistem yang dikembangkan tidak menangani *emoticon* pada teks.
4. Setiap teks yang mengandung kata tanya dan tanda tanya akan dianggap sebagai teks netral.

1.4 Tujuan Penelitian

Berdasarkan batasan masalah di atas, berikut ini adalah tujuan penelitian dari tugas akhir ini:

1. Mengimplementasikan SVM pada sistem untuk klasifikasi analisis sentimen.
2. Menganalisis dan menentukan sebuah opini merupakan positif, netral, negatif atau sarkasme.
3. Menentukan teknik klasifikasi yang terbaik untuk analisis sentimen.

1.5 Kontribusi Penelitian

Berikut ini adalah kontribusi penelitian yang diberikan pada pengembangan sistem analisis sentimen ini:

1. Menggunakan fitur *unigram*, *part of speech*, *punctuation based*, *capitalization*, *topic* dan *interjection* untuk *feature set* pada klasifikasi sarkasme.
2. Melakukan klasifikasi 4 kelas, yaitu positif, negatif, netral dan sarkasme.

1.6 Metode Penelitian

Tahapan-tahapan yang akan dilakukan dalam pelaksanaan penelitian ini adalah sebagai berikut:

1. Studi kepustakaan

Tahap pertama penulisan ini adalah studi kepustakaan, yaitu mengumpulkan bahan referensi dari berbagai sumber seperti buku, jurnal, laporan penelitian ataupun situs-situs internet. Materi yang dicari dan dipelajari adalah mengenai pembelajaran mesin, pemrosesan bahasa alami, klasifikasi analisis sentimen pada media sosial Indonesia.

2. Analisis dan perancangan

Dalam tahap ini, penulis melakukan perancangan sistem. Dimulai dari alur bisnis proses sistem, merancang sistem dengan menerapkan metode-metode yang ada.

3. Data sampling

Data teks Twitter yang digunakan dalam penelitian ini dikumpulkan menggunakan *library* Python Twitter-scraping.

4. Klasifikasi

I. PENDAHULUAN

Dalam tahap ini, data teks akan didahului dengan melakukan *text preprocessing*, *feature extraction* kemudian klasifikasi. Klasifikasi akan menggunakan metode klasifikasi *Support Vector Machine* (SVM).

1.7 Sistematika Penulisan

Pada penelitian ini peneliti menyusun berdasarkan sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi Latar Belakang Masalah, Rumusan Masalah, Batasan Masalah, Tujuan Penelitian, Kontribusi Penelitian, Metodologi Penelitian, dan Sistematika Penulisan.

BAB II LANDASAN TEORI

Bab ini berisi teori dasar yang digunakan pada penyusunan laporan ini yang meliputi penjelasan mengenai *Support Vector Machine* dan LDA.

BAB III ANALISIS DAN PERANCANGAN

Bab ini berisi perancangan sistem yang meliputi perancangan aplikasi “Penerapan *Support Vector Machine* untuk Deteksi Sarkasme pada Analisis Sentimen Media Sosial Indonesia”.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi implementasi dan analisis hasil penelitian terhadap sistem yang dibangun, apakah sesuai dengan tujuan yang diharapkan atau belum.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran dari seluruh kegiatan yang bisa digunakan sebagai masukan untuk pengembangan sistem informasi lebih lanjut yang nantinya akan dikembangkan.

BAB II

LANDASAN TEORI

Pada bab ini akan dijelaskan mengenai teori dasar dan metode yang digunakan pada penelitian ini.

2.1 Tinjauan Pustaka

Pada bagian ini akan dijelaskan teori-teori terkait yang akan digunakan dalam aplikasi klasifikasi teks

2.1.1 Pemrosesan Bahasa Alami

Pemrosesan Bahasa Alami atau *Natural Language Processing* (NLP) adalah bidang penelitian dan aplikasi yang mengeksplorasi bagaimana komputer dapat digunakan untuk memahami dan memanipulasi teks atau pidato bahasa alami. Penelitian NLP bertujuan untuk mengumpulkan pengetahuan tentang bagaimana manusia memahami dan menggunakan bahasa sehingga alat dan teknik yang tepat dapat dikembangkan untuk membuat sistem komputer memahami dan memanipulasi bahasa alami untuk melakukan tugas yang diinginkan. Dasar-dasar NLP terletak pada sejumlah ilmu, seperti ilmu komputer dan informasi, linguistik, matematika, teknik elektro dan elektronika, kecerdasan buatan dan robotika, dan psikologi. Aplikasi NLP mencakup sejumlah bidang studi, seperti terjemahan mesin, pemrosesan teks bahasa alami dan *summarization*, antarmuka pengguna, pencarian informasi multibahasa dan lintas bahasa (CLIR), pengenalan ucapan, kecerdasan buatan, dan sistem pakar. Di bawah ini terdapat 7 tingkatan *interdependent* yang digunakan untuk mengekstrak makna dari teks atau bahasa lisan [9]:

1. *Phonetic* atau *phonological level* adalah tingkatan yang berurusan dengan pengucapan.
2. *Morphological level* adalah tingkatan yang berhubungan dengan makna kata, awalan kata dan akhiran kata.
3. *Lexical Level* adalah tingkatan yang berhubungan dengan leksikal kata dan *Part Of Speech*
4. *Syntactic level* adalah tingkatan yang berhubungan dengan tata Bahasa dan

struktur kalimat.

5. *Semantic Level* adalah tingkatan yang berhubungan dengan makna kata dan kalimat.
6. *Discourse Level* adalah tingkatan yang berhubungan dengan struktur berbagai jenis teks menggunakan struktur dokumen.
7. *Pragmatic Level* adalah tingkatan yang berhubungan dengan pengetahuan dari luar pada isi sebuah dokumen.

2.1.2 Analisis Sentimen

Analisis sentimen atau biasa disebut *opinion mining* adalah bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap, dan emosi orang-orang terhadap entitas seperti produk, layanan, organisasi, individu, isu, peristiwa, topik, dan atribut mereka. Ada juga istilah yang lain dengan tugas yang tidak jauh berbeda, misalnya analisis sentimen, *opinion mining*, *opinion extraction*, *sentiment mining*, *subjectivity analysis*, *affect analysis*, *emoticon analysis*, *review mining*, dan sebagainya. Istilah tersebut merupakan bagian dari analisis sentimen atau *opinion mining*. Istilah analisis sentimen pertama kali muncul di Nasukawa dan Yi (2003), dan istilah *opinion mining* pertama kali muncul di Dave et al. (2003). Namun, penelitian tentang sentimen dan pendapat muncul lebih awal (Das and Chen, 2001; Morinaga et al., 2002; Pang et al., 2002; Tong, 2001; Turney, 2002; Wiebe, 2000). Analisis sentimen atau *opinion mining* berfokus pada opini yang mengungkapkan atau menyiratkan sentimen positif atau negatif [10].

2.1.3 Pembelajaran Mesin

Tujuan utama Pembelajaran Mesin adalah untuk memodelkan hubungan antara masukan dan keluaran. Pembelajaran mesin menyediakan teknik yang secara otomatis dapat membangun model komputasi dari hubungan kompleks dengan memproses data yang ada dan memaksimalkan kriteria kinerja yang bergantung pada masalah. Proses otomatis pembuatan model disebut *training* dan data yang digunakan untuk tujuan pelatihan disebut data *training*. Model yang sudah dilatih dapat memberikan informasi baru tentang bagaimana variabel masukan dipetakan ke keluaran dan dapat digunakan untuk membuat prediksi untuk nilai masukan baru yang bukan merupakan bagian dari data *training*. Untuk dapat mempelajari model yang akurat, algoritme pembelajaran mesin membutuhkan sejumlah besar data *training*. Oleh karena itu, langkah pertama yang penting dalam menggunakan teknik pembelajaran mesin adalah

mengumpulkan sekumpulan contoh data *training* dan menyimpannya dalam bentuk yang sesuai untuk keperluan komputasi.

Pembelajaran mesin dapat dilakukan di banyak domain seperti diagnosis medis, bioinformatika, informatika kimia, analisis jaringan sosial, analisis pasar saham, dan robotika. Kinerja model pembelajaran mesin bergantung pada banyak faktor seperti jumlah dan kualitas data pelatihan, kompleksitas dan bentuk hubungan antara variabel masukan dan keluaran, dan kendala komputasi seperti waktu pelatihan dan memori yang tersedia [11].

2.1.4 Text Preprocessing

Text Preprocessing bertujuan untuk meminimalisir kata yang akan digunakan dalam teks. Pada media sosial Indonesia seperti Twitter atau Facebook, pengguna sering menggunakan bahasa non-formal dibanding formal seperti mengganti kata dengan angka, mengulangi huruf yang sama, dan menggunakan kata non-formal untuk menggantikan kata formal [5]. Untuk memproses kata-kata seperti itu, maka akan dilakukan beberapa teknik *text preprocessing* sebagai berikut:

2.1.4.1 Case Folding

Case Folding adalah tahap mengubah teks menjadi huruf kecil. Huruf-huruf yang ditangani dimulai dari a-z.

2.1.4.2 Remove Hashtag, URL, Mention

Remove Hashtag, URL, Mention adalah tahap menghapus semua *hashtag*, *URL*, dan *mention* yang terdapat pada teks. Contoh ”@andi filmnya bagus https://... #bagus”, maka akan menjadi ”filmnya bagus”.

2.1.4.3 Remove Punctuation

Remove Punctuation adalah tahap menghapus tanda baca yang terdapat pada teks. Tanda petik (”), tanda petik tunggal (’), tanda seru (!), tanda tanya (?) dan tanda pemisah (-) akan menjadi pengecualian pada *preprocessing* ini, karena akan digunakan sebagai fitur *punctuation based*. Contoh tanda baca yang akan dihapus adalah ”%”, ”&”, ”*”, ”{}”, ”()”, ”[]”, ”：“, dan lain-lain.

2.1.4.4 Tokenization

Tokenization adalah tahap memecahkan kalimat menjadi token-token. Tiap token yang dihasilkan tidak harus terdiri dari satu kata. Satu token bisa saja

menghasilkan dua kata.

2.1.4.5 *Misuse of Word*

Setelah tahap *tokenization*, tahap selanjutnya adalah mengubah penyalahgunaan kata, sebagai contoh: "ASIIKKK", terdapat penyalahgunaan huruf yaitu "I" dan "K". Sehingga akan diubah menjadi "ASIK". Setiap kata yang terdapat huruf sama dan bersebelahan akan dijadikan menjadi 1 huruf, meskipun kata tersebut menjadi kata yang tidak memiliki makna, contohnya kata "tanggung", kata tersebut akan diubah menjadi "tangung" sehingga menjadi tidak memiliki makna. Hal ini tidak masalah untuk dilakukan, karena pada penelitian ini tidak dilakukan pada *semantic level* atau tidak mementingkan makna dari sebuah kata.

2.1.4.6 *Abbreviation Word*

Setelah tahap *misuse of word*, tahap selanjutnya adalah mengubah kata-kata yang menggunakan singkatan. Sebagai contoh: "km" menjadi kamu.

2.1.4.7 *Stopword Removal*

Setelah tahap *abbreviation word*, maka teks akan siap memasuki tahap berikutnya yaitu *stopword removal*. Pada tahap ini kata-kata yang terdapat pada daftar *stopword* akan dihilangkan dari teks. *Stopword* adalah kata-kata yang dianggap tidak memberi pengaruh terhadap teks. Sebagai contoh, kata "yang", "dan", "di", "ke", dan "dari". Masih terdapat banyak lagi kata-kata yang termasuk dalam *stopword* ini.

2.1.4.8 *Stemming*

Setelah tahap *stopword removal*, maka teks akan siap memasuki tahap berikutnya yaitu *stemming*. Pada tahap ini token-token pada teks akan diubah menjadi kata dasar. Fitur ini akan menggunakan *library python* yaitu *stemmer Sastrawi*. Pada tahap ini penulis akan menggunakan library yang terdapat pada *python* untuk melakukan *stemming*. *Stemming* yang digunakan adalah algoritme Nazief dan Adriani. *Stemming* ini biasanya bertujuan untuk menghapus *suffixes*, dan sering digunakan pada *text search*, *machine translation*, *document summarization*, *text classification*. *Affixes* dapat dibagi menjadi dua, yaitu *inflectional* atau *derivational*. Contoh pada kata kerja *English*, "teach" dapat ditambah dengan *inflectional suffixes* "-es" sehingga menjadi kata kerja tunggal "teaches". Selain itu kata "teach" juga dapat ditambah dengan *derivational suffixes* "-er" sehingga menjadi kata benda "teacher". Dalam Indonesia, *inflectional*

II. LANDASAN TEORI

suffixes hanya terdapat pada *suffixes*, sedangkan *derivational suffixes* terdapat pada *prefix*, *suffixes* atau gabungan keduanya [12].

Dalam Indonesia, terdapat dua macam *inflectional suffixes* [Moeliono dan Dardjowidjojo 1988]:

1. *Particle* (P) “-kah”, ”-lah”, ”-tah”, “-pun”

Particle di atas adalah imbuhan yang tidak akan mengubah kata menjadi kata kerja atau kata benda. Contoh, *suffix* ”-lah” ketika ditambahkan pada kata ”duduk” akan menjadi ”duduklah”.

2. *Possessive Pronouns* (PP) “-ku”, “-mu”, “-nya”

Possessive Pronouns akan mengubah kata menjadi hubungan kepemilikan. Contoh, *suffix* ”-nya” ketika ditambahkan pada kata ”buku” akan menjadi ”bukunya”.

Derivational Prefixes (DP) {”be-,” ”di-,” ”ke-,” ”me-,” ”pe-,” ”se-,” and ”te-”}. *Prefix* ”me-” dapat muncul sebagai ”me-”, ”mem-”, ”meng-”, ”men-”, dan ”meny-” bergantung dengan kata dasarnya. Sebagai contoh, *derivational prefixes* ”se-”, ”peng-”, ”ke-” ditambahkan pada kata ”tahu” dan dengan *suffixes* ”-an”, ”-ku” akan menghasilkan kata ”sepenteahuaku” [12].

Derivational Suffixes (DS) {”-i”, ”-kan”, ”-an”} hanya boleh digunakan satu *suffixes* saja pada kata dasar. Sebagai contoh, kata ”lapor” ditambahkan dengan *suffix* ”-kan” menghasilkan ”laporkan” [12].

Derivational Config (DC) {”be-an”, ”me-i”, ”me-kan”, ”di-i”, ”dikan”} adalah gabungan dari *prefix* dan *suffix*. Sebagai contoh, *prefix* ”ke-” dan *suffix* ”-an” ditambahkan pada kata dasar ”dalam” menghasilkan ”kedalaman”. Berikut adalah contoh *affix* model yang mungkin terbentuk pada kata dasar:

[[DP +][DP +] DP +] kata dasar [[+DS][+PP][+P]]

Berikut merupakan langkah dan proses yang dilakukan dalam *stemming* [12]:

1. Menghapus semua *inflectional suffixes* pada kata masukan. Proses *stemming* berhenti, jika kata masukan terdapat pada kamus kata dasar. Jika tidak, maka berdasarkan *affix* model, yang tersisa adalah *derivational suffixes*. Berikut adalah model yang mungkin terbentuk:

[[DP +][DP +] DP +] kata dasar [+DS]
--

II. LANDASAN TEORI

2. Menghapus semua *derivational suffixes*. Berikut model yang terbentuk:

[[DP +][DP +] DP +] kata dasar

3. Menghapus semua *derivational prefixes*.

- 1 Syarat proses stemming berhenti:

- Saat kombinasi *prefix* dan *suffix* yang dihapus pada langkah 3 tidak *valid*
- Saat *prefix* yang dihapus sama dengan *prefix* yang telah dihapus sebelumnya
- Saat penghapusan *prefix* telah dilakukan 3 kali.

- 2 Identifikasi tipe *prefix*:

- *Plain*, seperti "di-", "ke-", "se-" dapat langsung dihapus.
- *Complex*, seperti "be-", "te-", "me-", atau "pe" memiliki banyak variasi dan *prefix disambiguation*. Contoh, *prefix* "me-", bisa menjadi "mem-", "men-", "meny-", atau "meng-" bergantung dengan huruf depan dari kata dasar. Hal ini dapat diatasi dengan tabel 2.1. Jika kata masukan dapat ditemukan pada kamus kata dasar, maka proses berhenti.
- Jika tidak ditemukan, maka akan dilakukan pengulangan secara rekursif pada langkah 4.

4. Jika kata masukan masih belum dapat ditemukan maka akan dilakukan pengecekan pada tabel 2.1. Tabel tersebut berisi variasi *prefix* dengan melakukan pengecekan *prefix* dan huruf tertentu dari kata dasar. Contoh, kata "menangkap" memenuhi *rule* 15, dengan *prefix* "me-"(inisial *prefix* "men-" dan diikuti huruf vokal "a-"). Pada *rule* 15, ada 2 kemungkinan huruf yang akan didapatkan, yaitu "n" ("men-nV") dan "a" ("men-tV"). Jika menggunakan huruf "n", maka menghasilkan kata "nangkap" (tidak ada pada kamus), dan jika menggunakan huruf "t" maka menghasilkan "tangkap" (ada pada kamus).
5. Jika semua langkah tidak berhasil, maka akan mengembalikan kata masukan seperti kondisi awalnya.

Huruf "V" pada tabel merupakan *vowel*. Huruf "C" pada tabel merupakan *consonant*. Huruf "A" pada tabel merupakan *any letter*. Huruf "P" merupakan *fragment* kata pendek "er". Berikut adalah tabel *rule* pada *prefix disambiguation*:

II. LANDASAN TEORI

Tabel 2.1 Tabel Derivation Prefix Rule

Rule	Construct	Return
1	berV...	ber-V... be-rV
2	berCAP...	ber-CAP... dimana C! = 'r' dan P!+'er'
3	berCAerV...	ber-CAerV... dimana C! = 'r'
4	belajar...	bel-ajar...
5	beC ₁ erC...	be-C ₁ erC... dimana C ₁ ! { 'r' 'l' }
6	terV...	ter-V... te-rV...
7	terCP...	ter-CP... dimana C! = 'r' dan P! = 'er'
8	terCer...	ter-Cer... dimana C! = 'r'
9	teC ₁ erC ₂ ...	teC ₁ erC ₂ ... dimana C! = 'r'
10	me{l r w y}V...	me-{l r w y}V...
11	mem{b f v}...	mem-{b f v}...
12	mempe...	mem-pe...
13	mem{rV V}...	me-m{rV V}... me-p{rV V}...
14	men{c d j z}...	men-{c d j z}...
15	menV...	me-nV... me-tV...
16	meng{g h q k}	meng-{g h q k}
17	mengV...	meng-V... meng-kV...
18	menyV...	meny-sV...
19	mempV	mem-pV... dimana V! = 'e'
20	pe{w y}V...	pe-{w y}V...
21	perV...	per-V... pe-rV...
22	perCAP...	per-CAP... dimana C! = 'r' dan p!= 'er'
23	perCAerV...	per-CAerV... dimana C! = 'r'
24	pem{b f v}...	pem-{b f v}...

II. LANDASAN TEORI

Tabel 2.1 Tabel Derivation Prefix Rule

Rule	Construct	Return
25	pem{rV V}...	pe-m{rV V}... pe-p{rV V}...
26	pen{c d j z}...	pen-{c d j z}...
27	penV...	pe-nV... pe-tV...
28	peng{g h q}...	peng-{g h q}...
29	pengV...	peng-V... peng-kV...
30	penyV...	peny-sV...
31	pelV...	pe-lV... kecuali kata "pelajar", return "ajar"
32	peCP...	pe-CP... dimana C! = {r w y l m n} dan P!='er'
33	peCerV...	pe-CerV... dimana C! = {r w y l m n}

2.1.5 Feature Extraction

Setelah melakukan pemrosesan teks menjadi lebih terstruktur, setiap kata dalam dokumen teks diekstraksi agar setiap teks memperoleh dan mengkalkulasi berdasarkan kata yang penting untuk diolah lebih lanjut saat mengklasifikasi. Pada penelitian ini, digunakan proses *unigram*, *part of speech*, *sentiment score*, *punctuation based*, capitalization, topic, *interjection*, *question word* dan TF-IDF.

2.1.5.1 Unigram

Fitur *Unigram* adalah fitur untuk memisahkan sebuah kalimat menjadi kumpulan token yang terdiri dari 1 kata pada setiap tokennya.

2.1.5.2 Part of Speech

Fitur *Part of Speech* adalah fitur untuk menghitung jumlah kemunculan kata benda, kata sifat, kata kerja, kata keterangan dan kata negasi. Fitur ini akan menggunakan *library* IPosTagger python. Pada umumnya Bahasa Indonesia memiliki 5 *part of speech* yaitu, kata benda (*noun*), kata sifat (*adjective*), kata kerja (*verb*), kata keterangan (*adverb*) dan kata tugas (*function word*). Kata benda dapat dibagi menjadi beberapa *subcategories* yaitu *uncountable common nouns*, *genitive common nouns*, *proper nouns*, dan *various pronouns*. Berikut daftar *part of speech* Bahasa Indonesia [13]:

II. LANDASAN TEORI

Tabel 2.2 Tabel Tag Set pada Bahasa Indonesia

No	Tag	Description	Example	No	Tag	Description	Example
1	(<i>Opening parenthesis</i>	({ [19	VBT	<i>Transitive Verbs</i>	Makan, tidur
2)	<i>Closing parenthesis</i>) }]	20	VBI	<i>Intransitive Verbs</i>	Bерmain, terdiam
3	,	<i>Comma</i>	,	21	MD	<i>Modal or auxiliaries verbs</i>	Sudah, boleh, harus
4	.	<i>Sentence terminator</i>	. ? !	22	JJ	<i>Adjectives</i>	Mahal, kaya, malas
5	:	<i>Colon or ellipsis</i>	: ;	23	CDP	<i>Primary cardinal numeral</i>	Satu, juta, milyar
6	-	<i>Dash</i>	-	24	CDO	<i>Ordinal cardinal numerals</i>	Pertama, kedua
7	"	<i>quotation</i>	' "	25	CDI	<i>Irregular cardinal numerals</i>	Beberapa, segala, semua
8	WP	<i>WH-pronouns</i>	Apa, Siapa, mengapa	26	CDC	<i>Collective cardinal numerals</i>	Bertiga, bertujuh
9	\$	<i>Dollar</i>	\$	27	NEG	<i>Negations</i>	Bukan, tidak
10	Rp	<i>Rupiah</i>	Rp	28	IN	<i>Prepositions</i>	Di, Ke, Dari
11	SYM	<i>Symbols</i>	% \$ ' ")) * + , . < = > @	29	CC	<i>Coordinate conjunction</i>	Dan, atau

II. LANDASAN TEORI

Tabel 2.2 Tabel Tag Set pada Bahasa Indonesia

No	Tag	Description	Example	No	Tag	Description	Example
12	NNC	<i>Countable common nouns</i>	Buku, rumah, karyawan	30	SC	<i>Subordinate conjunction</i>	Yang, ketika
13	NNU	<i>Uncountable common nouns</i>	Air, gula, nasi, hujan	31	RB	<i>Adverbs</i>	Sekarang, nanti, sementara
14	NNG	<i>Genitive common nouns</i>	Idealnya	32	UH	<i>Interjection</i>	Wah, wow, aduh, oh
15	NNP	<i>Proper nouns</i>	Jakarta, BCA	33	DT	<i>Determiners</i>	Para, ini, itu
16	PRP	<i>Personal pronouns</i>	Saya, aku	34	WDT	<i>WH-determiners</i>	Apa, siapa
17	PRN	<i>Number pronouns</i>	Kedua-duanya	35	RP	<i>Particles</i>	Kan, kah, lah
18	PRL	<i>Locative pronouns</i>	Sini, situ, sana	36	FW	<i>Foreign Word</i>	Absurd, list

2.1.5.3 Sentiment Score

Fitur *Sentiment Score* adalah fitur untuk menghitung nilai sentimen dari sebuah teks. Fitur ini dilakukan dengan melakukan pengecekan tiap kata pada teks terhadap data SentiWordNet. SentiWordNet adalah kumpulan kata yang sudah diberi *tag* seperti *noun*, *adjective*, *verb*, *adverb* dan memiliki nilai sentimen. Data SentiWordNet merupakan hasil yang diubah dari Barasa WordNet. Barasa WordNet adalah kamus Bahasa Melayu dan Bahasa Indonesia yang dibuat berdasarkan Princeton WordNet. Barasa WordNet memiliki 49.668 *synsets (sets of synonyms)*, 145.696 *senses* dan 64.431 *unique words*. Untuk menggunakan SentiWordNet, harus mendapatkan POS *Tag* dari setiap kata. Sebagai contoh: "rumah/noun" memiliki nilai sentimen positif dan negatif 0. Pada fitur ini akan mengatasi kata negasi. Negasi adalah kata seperti "tidak" atau "bukan" yang dapat mengubah nilai sentimen dari sebuah kata. Sebagai Contoh, "Sekolah tidak ada

masalah sm guru tidak asyik wkwk”, kata ”asyik” pada teks tersebut akan berubah menjadi negatif, karena didahului kata negasi ”tidak”. Penulis menangani hal ini dengan menambah nilai sentimen sebaliknya dan dikalikan 2 [5]. Sebagai contoh, kata ”asyik” memiliki sentimen positif 0.35 dan sentimen negatif 0.0. maka jika didahului kata negasi, kata ”asyik” akan memiliki nilai sentimen positif 0.35 dan sentimen negatif 0.70.

2.1.5.4 Punctuation Based

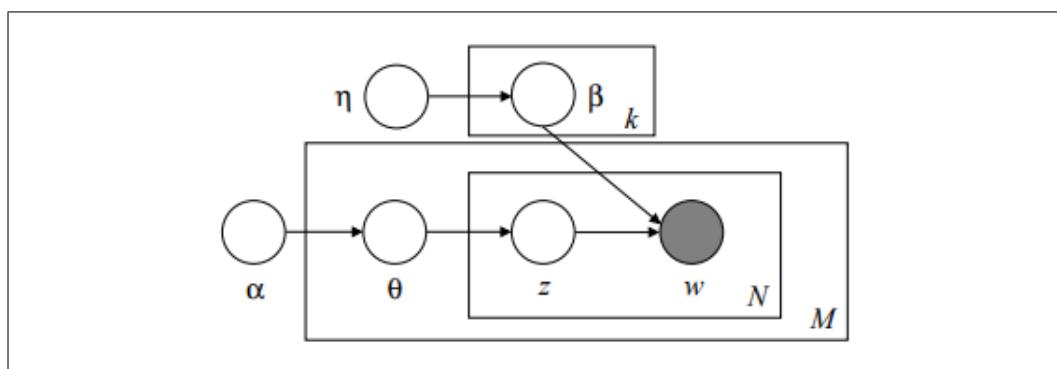
Fitur *Punctuation Based* adalah fitur untuk menghitung jumlah kemunculan dari tanda baca yang terdapat pada teks. Contoh tanda baca yang akan dihitung kemunculannya seperti tanda seru(!), tanda tanya (?), tanda petik (”) dan tanda petik tunggal (’). Kemunculan setiap tanda baca pada teks akan dibagi dengan kemunculan maksimal tanda baca pada keseluruhan data *training*.

2.1.5.5 Capitalization

Fitur *Capitalization* adalah fitur untuk menghitung jumlah kemunculan kata kapital dari sebuah teks. Kemunculan setiap kata kapital pada teks akan dibagi dengan kemunculan maksimal kata kapital pada keseluruhan data *training*.

2.1.5.6 Topic

Fitur *Topic* adalah fitur yang menggunakan *topic modelling Latent Dirichlet Allocation* (LDA) untuk mempelajari topik yang terdapat pada sebuah teks. Tujuan dari penggunaan fitur ini adalah untuk mendapatkan informasi global dari sebuah teks.



Gambar 2.1 Graphical Model pada *Smoothed LDA*

Gambar di atas merupakan model *smoothed LDA*. Dengan kotak pada bagian luar merepresentasikan dokumen, dan kotak pada bagian dalam merepresentasikan pengulangan topik dan kata dalam dokumen. Berikut penjelasan simbol di atas :

II. LANDASAN TEORI

1. α : parameter *dirichlet prior* distribusi topik terhadap dokumen.
2. η : parameter *dirichlet prior* distribusi kata terhadap topik.
3. θ : distribusi topik terhadap dokumen.
4. β : distribusi kata terhadap topik.
5. Z : identitas topik untuk seluruh kata dalam seluruh dokumen
6. W : identitas seluruh kata dalam seluruh dokumen
7. K : jumlah topik
8. M : jumlah dokumen
9. N : jumlah kata pada dokumen

Berikut ini adalah proses dalam melakukan training LDA topic modelling:

1. Menentukan jumlah topik yang diinginkan (K), parameter α , parameter η .
2. Memberikan topik terhadap setiap kata secara acak untuk sementara.
3. Menghitung nilai probabilitas *topic-word* dan *document-topic*
4. Mengalikan hasil *topic-word* dan *document-topic* (Z).
5. Melakukan random antara 0-1, dan membandingkannya dengan hasil pada tahap 4. Hasil tersebut akan digunakan untuk memberi nilai topik kembali pada setiap kata. Proses ini diulang sampai memenuhi jumlah iterasi.

2.1.5.7 *Interjection*

Fitur *Interjection* adalah fitur untuk menghitung jumlah kata *interjection* yang terdapat pada sebuah teks. Contoh kata *interjection* adalah "aha", "bah", "wew", "wow", "yay", "nah", "uh", dan lain-lain. Fitur ini digunakan untuk membantu dalam klasifikasi teks sarkasme, Karena dari 100 data sarkasme ditemukan 20 teks mengandung kata interjeksi. Ketika sebuah teks mengandung kata interjeksi, teks tersebut cenderung dianggap teks sarkasme [5].

2.1.5.8 *Question Word*

Fitur *Question Word* adalah fitur untuk memberikan nilai *true* atau *false*, jika pada teks terdapat kata tanya seperti apa, dimana, mengapa, siapa, dan bagaimana. Kata tanya ini biasa terdapat pada teks netral, sehingga dapat membantu dalam pemberian informasi pada klasifikasi teks netral.

II. LANDASAN TEORI

2.1.5.9 TF-IDF

TF-IDF merupakan gabungan dari frekuensi kata (TF) dan frekuensi kata pada dokumen (IDF). *Term Frequency* (TF) merupakan metode penghitungan jumlah kemunculan kata pada dokumen dibagi jumlah kata pada dokumen. TF memiliki persamaan sebagai berikut [15]:

$$TF = \frac{\text{frekuensi kemunculan kata pada dokumen}}{\text{jumlah kata pada dokumen}} \quad (2 . 1)$$

IDF memiliki persamaan sebagai berikut:

$$IDF = \log \frac{\text{jumlah dokumen}}{\text{jumlah dokumen yang mengandung kata tertentu}} \quad (2 . 2)$$

TF-IDF memiliki persamaan sebagai berikut:

$$TF-IDF = TF * IDF \quad (2 . 3)$$

Ciri-ciri hasil TF-IDF sebagai berikut:

1. Bobot akan tinggi jika intensitas kemunculan kata rendah pada banyak dokumen dan intensitas kemunculan kata tinggi pada sebuah dokumen.
2. Bobot akan rendah jika intensitas kemunculan kata tinggi pada banyak dokumen dan intensitas kemunculan kata rendah pada sebuah dokumen
3. Bobot akan sangat rendah jika kemunculan suatu kata terdapat pada setiap dokumen.

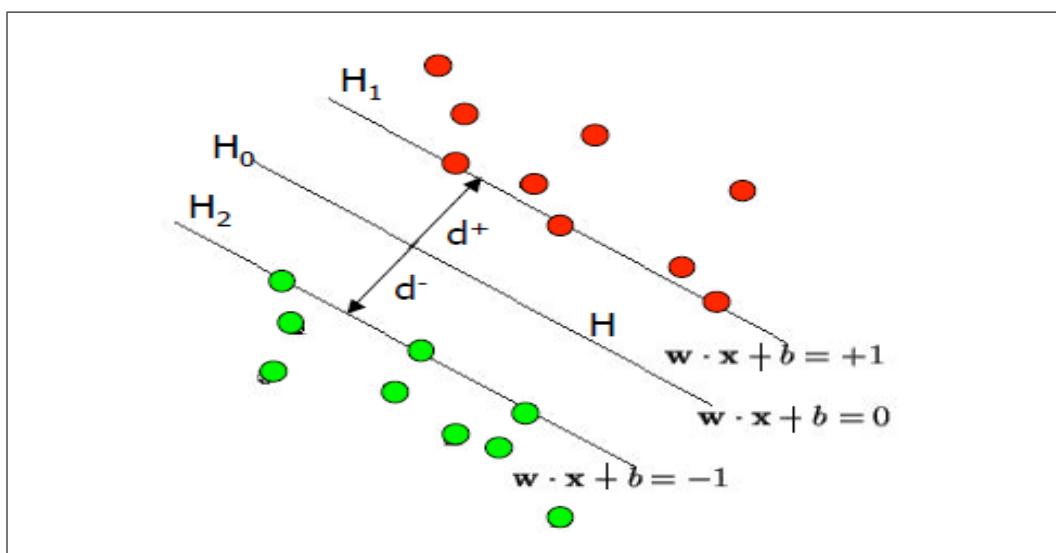
2.1.6 Definisi Support Vector Machine (SVM)

Ide di balik SVM adalah membangun *hyperplane* yang optimal agar dapat digunakan dalam klasifikasi pola yang dapat dipisahkan secara linear. *Hyperplane* yang optimal adalah *hyperplane* yang dipilih dari satu himpunan *hyperplane* yang memaksimalkan *margin* pada *hyperplane*. *Margin* adalah jarak dari bidang *hyper*

ke titik terdekat dari pola. SVM dibuat khusus untuk memaksimalkan *margin*, yang akan menjamin bahwa pola masukan akan diklasifikasikan dengan benar. SVM biasanya digunakan untuk mengklasifikasikan pola. Pola tersebut bisa linear atau non-linear [7].

2.1.6.1 Support Vector Machine (SVM)

Konsep dasar dari *Support Vector Machine* (SVM) adalah membangun *hyperplane* untuk memisahkan 2 kelas, yaitu kelas positif dan kelas negatif dan memaksimalkan *margin*. *Margin* adalah jarak antara *hyperplane* dan titik data terdekat. Titik data yang terdekat yang menyentuh *hyperplane* disebut *Support Vector*. Pada gambar 2.2 terdapat 6 *support vector*, 3 *support vector* dari kelas +1, dan 3 *support vector* dari kelas -1 [16].



Gambar 2.2 *Hyperplane* pada SVM

Garis lurus di tengah pada gambar 2.2 adalah *hyperplane*. Untuk memisahkan 2 kelas pada *Support Vector Machine* dibutuhkan *hyperplane* yang optimal. Optimasi bidang *hyperplane* dapat diselesaikan dengan teknik optimasi, yaitu *Lagrangian Multiplier*. Persamaan pada *hyperplane* dapat didefinisikan sebagai berikut [16]:

$$w^T \cdot x + b = 0 \quad (2.4)$$

$$w^T \cdot x + b \geq +1, \text{ dimana } y_i = +1 \quad (2.5)$$

$$w^T \cdot x + b \leq -1, \text{ dimana } y_i = -1 \quad (2.6)$$

II. LANDASAN TEORI

Dimana :

X_i : Vektor masukan

y_i : Kelas

w^T : Vektor *weight*

b : Bias

Persamaan kernel pada Support Vector Machine sebagai berikut:

Tabel 2.3 Tabel persamaan kernel pada SVM

Kernel	Persamaan
Linear	$K(X_i, X_j) = X_i^T X_j \quad (2 . 7)$
Polynomial	$K(X_i, X_j) = (\gamma X_i^T X_j + \gamma)^4, \gamma > 0 \quad (2 . 8)$
Radial Basis Function (RBF)	$K(X_i, X_j) = \exp\left(-\frac{ X_i - X_j ^2}{2\sigma^2}\right) \quad (2 . 9)$
Sigmoid	$K(X_i, X_j) = \tanh(\gamma X_i^T X_j + \gamma) \quad (2 . 10)$

Perhitungan w dapat dilakukan dengan persamaan sebagai berikut:

$$w = \sum_{i=1}^m \alpha_i y_i x_i \quad (2 . 11)$$

Proses klasifikasi kernel linear dapat dilakukan dengan persamaan:

$$f(x) = \sum_{i=1}^m \alpha_i y_i K(X_i, X) + b, \text{ m=jumlah data training} \quad (2 . 12)$$

Jika nilai $f(x_i) \leq +1$ maka tweet masuk ke kelas positif dan sebaliknya jika $f(x_i) \leq -1$ maka tweet masuk ke kelas negatif.

2.1.6.2 *Multiclass Support Vector Machine*

Support Vector Machine (SVM) pada awalnya hanya dibuat untuk mengklasifikasikan dua kelas dengan memaksimalkan margin. Pada kenyataanya klasifikasi selalu memiliki lebih dari 2 kelas, sehingga dikembangkan Multiclass SVM agar dapat menangani masalah kelas yang lebih dari 2. Salah satu algoritme untuk klasifikasi multiclass adalah algoritme One versus Rest [17].

Prinsip dari One versus Rest adalah membangun kelas sebanyak jumlah klasifikasi kelas. Jika terdapat tiga kelas, maka akan dilakukan perbandingan sebanyak tiga kali. Dan mengambil hasilnya dari nilai keluaran tertinggi. Masalah utama pada klasifikasi ini adalah jumlah data training yang tidak merata [17].

Tabel 2.4 Tabel *One versus Rest*

Y=+1	Y=-1	Hipotesis
Kelas 1	Bukan kelas 1	$f^1(x) = w^{T1} \cdot x + b^1$
Kelas 2	Bukan kelas 2	$f^2(x) = w^{T2} \cdot x + b^2$
Kelas 3	Bukan kelas 3	$f^3(x) = w^{T3} \cdot x + b^3$

II. LANDASAN TEORI

2.1.7 Simplified Sequential Minimal Optimization (SMO)

Simplified Sequential Minimal Optimization (SMO) algoritme adalah teknik optimasi untuk mengatasi dual problem (problem optimization). Algoritme SMO akan dilakukan pengulangan sampai konvergen [18]. Perhitungan nilai α (alpha) dapat dilakukan dengan persamaan sebagai berikut:

$$\text{if } Y_i \neq Y_j, L = \max(0, \alpha_j - \alpha_i), H = \min(C, C + \alpha_j - \alpha_i) \quad (2.13)$$

$$\text{if } Y_i = Y_j, L = \max(0, \alpha_j + \alpha_i - C), H = \min(C, \alpha_j + \alpha_i) \quad (2.14)$$

$$\alpha_j = \alpha_j - \frac{Y_j(E_i - E_j)}{\eta} \quad (2.15)$$

$$\alpha_i = \alpha_i + Y_i Y_j (\alpha_j(\text{old}) - \alpha_i), \alpha_j = \begin{cases} H, & \text{if } \alpha_j > H \\ \alpha_j, & \text{if } L \leq \alpha_j \leq H \\ L, & \text{if } \alpha_j < L \end{cases} \quad (2.16)$$

$$E_k = F(X_k) - Y_k \quad (2.17)$$

$$\eta = 2 * K(X_i, X_j) - K(X_i, X_i) - K(X_j, X_j) \quad (2.18)$$

Perhitungan nilai bias (b) dilakukan dengan persamaan sebagai berikut:

$$b_1 = b - E_i - Y_i(\alpha_i - \alpha_i(\text{old}))K(X_i, X_i) - Y_j(\alpha_j - \alpha_j(\text{old}))K(X_i, X_j) \quad (2.19)$$

$$b_2 = b - E_j - Y_i(\alpha_i - \alpha_i(\text{old}))K(X_i, X_j) - Y_j(\alpha_j - \alpha_j(\text{old}))K(X_j, X_j) \quad (2.20)$$

$$b = \begin{cases} b_1, & \text{if } 0 < \alpha_i < C \\ b_2, & \text{if } 0 < \alpha_j < C \\ (b_1 + b_2)/2, & \text{lainnya} \end{cases} \quad (2.21)$$

II. LANDASAN TEORI

Berikut ini adalah pseudocode Simplified Sequential Minimal Optimization (SMO):

Tabel 2.5 Tabel *Pseudocode* SMO

```
Input:  
C: regularization parameter  
tol: numerical tolerance  
max passes: max # of times to iterate over's without changing  
(x(1), y(1)), . . . ,(x(m), y(m)): training data  
Output:  
 $\alpha$  : Lagrange multipliers for solution  
b : threshold for solution  
Initialize  $\alpha_i = 0$ ,  $b = 0$ .  
Initialize passes = 0.  
while (passes < max passes)  
    num changed alphas = 0.  
    for i = 1, . . . ,m,  
        Calculate  $E_i = f(x(i)) - y(i)$  using (2.21).  
        if ((y(i)Ei <-tol &&  $\alpha_i < C$ ) || (y(i)Ei > tol &&  $\alpha_i > 0$ ))  
            Select  $j \neq i$  randomly  
            Calculate  $E_j = f(x(j)) - y(j)$  using (2.16).  
            Save old  $\alpha$ 's:  $\alpha_i(\text{old}) = \alpha_i$ ,  $\alpha_j(\text{old}) = \alpha_j$ .  
            Compute L and H by (2.12) or (2.13).  
            if (L == H)  
                continue to next i.  
            Compute  $\eta$  by (2.17).  
            if ( $\eta \geq 0$ )  
                continue to next i.  
            Compute and clip new value for  $\alpha_j$  by (2.14).  
            if ( $|\alpha_j - \alpha_j(\text{old})| < 10^{-5}$ )  
                continue to next i.  
            Determine value for  $\alpha_i$  by (2.15).  
            Compute  $b_1$  and  $b_2$  by (2.18) and 2.19).  
            Compute b by (2.20).  
            num changed alphas := num changed alphas + 1.  
        end if  
    end for  
    if (num changed alphas == 0)  
        passes := passes + 1  
    else  
        passes := 0  
    end while
```

2.1.8 Twitter-scaper

Twitter-scaper adalah sebuah *script* yang digunakan untuk mengambil data *tweet* menggunakan *library* BeautifulSoup4. BeautifulSoup4 adalah *library* untuk mengambil data dari sebuah HTML atau XML. Berikut ini adalah *script* cara penggunaan Twitter-scaper pada *command prompt* :

Tabel 2.6 Tabel metode pada *library* Twitter-scaper

No	Script	Keterangan
1	Twitterscraper dpr%20lang%3Aid -l 500 -o dpr-500.json	<i>Script</i> ini digunakan untuk megambil data dengan <i>keyword</i> pencarian DPR, jumlah <i>tweet</i> 500, Bahasa Indonesia, dengan keluaran file berupa json. -l adalah <i>length</i> atau jumlah dari <i>tweet</i> yang ingin diambil. -o adalah <i>output</i> atau keluaran berupa nama file yang diinginkan, dengan <i>extension</i> json.

2.1.9 Natural Language Toolkit (NLTK)

NLTK adalah *library* python yang digunakan untuk bekerja dengan Bahasa Manusia. Hal-hal yang dapat ditangani NLTK, yaitu *tokenize*, *tagging*, *parsing*, dan *semantic reasoning*. Berikut ini adalah tabel metode pada *library* NLTK:

Tabel 2.7 Tabel metode pada *library* NLTK

No	Metode	Keterangan
1	word_tokenize (String)	Metode ini digunakan untuk melakukan tokenisasi pada sebuah teks, sehingga menghasilkan token-token kata. Keluaran dari metode ini adalah array yang berisi token-token kata.

2.1.10 Tweet-preprocessor

Library Tweet-preprocessor adalah *library* yang digunakan untuk melakukan *preprocessing* pada data *tweet*. *Library* ini menyediakan fitur untuk melakukan proses *cleaning*, *tokenizing*, dan *parsing* pada *tweet*. Berikut ini adalah tabel metode pada *library* Tweet-preprocessor:

II. LANDASAN TEORI

Tabel 2.8 Tabel metode pada *library* Tweet-preprocessor

No	Metode	Keterangan
1	set_options(OPT)	Metode ini digunakan untuk melakukan konfigurasi pengaturan teks yang akan dihapus. Pengaturan yang tersedia adalah URL, mention, hashtag, reserved words, emoji, smiley, number.
2	clean (String)	Metode ini digunakan untuk menhapus hashtag, mentions, emojis, reserved words (RT, FAV), smileys dan URLs yang terdapat pada teks. Keluaran dari metode ini adalah String.
3	parse(String)	Metode ini digunakan untuk mengambil teks tertentu pada tweet, seperti URL, mention, hashtag, reserved words, emoji, smiley, number.
4	tokenization(String)	Metode ini digunakan untuk mengubah tweet yang mengandung kata seperti hashtag menjadi \$hashtag\$. Sebagai contoh, hashtag "#DPR", akan menjadi "\$HASHTAG\$"

2.1.11 Sastrawi

Library Sastrawi adalah *library stemmer* yang digunakan untuk mengubah kata-kata pada Bahasa Indonesia menjadi kata dasar. Berikut ini adalah tabel metode pada *library* Sastrawi:

Tabel 2.9 Tabel metode pada *library* Sastrawi

No	Metode	Keterangan
1	StemmerFactory ()	Metode ini merupakan <i>constructor</i> untuk membuat <i>class stemmer</i> .
2	create_stemmer ()	Metode ini untuk menginisialisasi <i>stemmer</i> yang akan digunakan untuk melakukan <i>stemming</i> .

II. LANDASAN TEORI

Tabel 2.9 Tabel metode pada *library* Sastrawi

No	Metode	Keterangan
3	stem (String)	Metode ini digunakan untuk melakukan pengubahan kata-kata pada sebuah string berupa teks menjadi kata dasar. Keluaran dari metode ini adalah String, yaitu kalimat yang sudah diubah menjadi kata dasar.

2.1.12 IPosTagger

Library IPosTagger adalah *library* yang digunakan untuk memberikan *tagging* terhadap kata-kata pada sebuah kalimat. Berikut ini adalah metode pada *library* IPosTagger:

Tabel 2.10 Tabel metode pada *library* IPosTagger

No	Metode	Keterangan
1	MainTagger (String, String, int, int, int, int, boolean, float, int, float, int)	Metode ini merupakan <i>constructor</i> untuk membuat <i>class</i> <i>tagger</i> . Parameter masukan yang diperlukan adalah <i>file lexicon</i> , <i>file n-gram</i> , tipe <i>n-gram</i> , <i>maxAffixLength</i> , <i>threshold</i> , <i>minWordFreq</i> , <i>modeAffixTree</i> , <i>lambdaBigram</i> , <i>TwoPhaseType</i> , <i>BeamFactor</i> , <i>UseLexicon</i> .
2	taggingStr (String)	Metode ini digunakan untuk memberi <i>tagging</i> pada setiap kata dalam teks. Keluaran dari metode ini adalah array yang berisi kata-kata yang sudah diberi <i>tag</i> .

2.1.13 Gensim

Gensim adalah *library open-source* yang menyediakan *toolkit* pemodelan topik yang diimplementasikan dengan Python. Berikut ini adalah tabel metode pada *library* Gensim:

II. LANDASAN TEORI

Tabel 2.11 Tabel metode pada *library* Gensim

No	Metode	Keterangan
1	LdaModel (Array [] [], Array [] [], int, String)	Metode ini merupakan <i>constructor</i> yang akan menginisialisasi array[][] <i>corpus</i> yang berisi <i>id</i> kata dan kemunculan kata, array[][] kamus yang berisi kata dan <i>id</i> kata, jumlah topik, dan nilai alpha.
2	dictionary (Array [])	Metode ini akan membentuk sebuah kata yang <i>unique</i> dengan <i>id</i> kata tersebut berdasarkan kata-kata dari data dokumen <i>training</i> .
3	doc2bow (String)	Metode ini akan menghitung kemunculan kata dengan <i>id</i> katanya.
4	Models.Save (String)	Metode ini untuk menyimpan hasil dari <i>model training</i> yang ke dalam <i>file</i> .
5	Corpora.save (String)	Metode ini untuk menyimpan hasil dari kata <i>unique</i> dan <i>id</i> kata dari data <i>training</i> ke dalam <i>file</i> .
6	Models.load (String)	Metode ini untuk mengambil kembali <i>model</i> yang telah disimpan pada sebuah <i>file</i> dengan menerima masukan nama <i>file</i> .
7	Corpora.load (String)	Metode ini untuk mengambil kembali data yang telah disimpan pada sebuah <i>file</i> dengan menerima masukan nama <i>file</i> .

2.1.14 NumPy

NumPy adalah *library* yang digunakan untuk perhitungan *scientific* pada Python. NumPy memiliki fitur seperti, pembuatan array object. Berikut ini adalah tabel metode pada *library* NumPy:

II. LANDASAN TEORI

Tabel 2.12 Tabel metode pada *library NumPy*

No	Metode	Keterangan
1	array (Array)	Metode ini mengubah array <i>default</i> dari Python masukan menjadi array numpy atau disebut <i>ndarray</i> .
2	dot (ndarray, ndarray)	Metode ini digunakan untuk mengalikan nilai dari 2 <i>ndarray</i> .
3	mean (ndarray)	Metode ini digunakan untuk mendapatkan nilai rata-rata dari sebuah array yang berisi nilai angka.
4	save (String)	Menyimpan nilai dari <i>ndarray</i> ke dalam sebuah <i>file</i> . Metode ini menerima masukan berupa nama <i>file</i> .
5	load (String)	Mengambil kembali data yang telah disimpan di dalam sebuah <i>file</i> . Metode ini menerima masukan berupa nama <i>file</i> yang ingin diambil datanya.

2.1.15 Scikit-learn

Library Scikit-learn merupakan *tools* yang digunakan untuk *data mining* dan data *analysis*. Berikut ini adalah tabel metode pada *library Scikit-learn*:

Tabel 2.13 Tabel metode pada *library Scikit-learn*

No	Metode	Keterangan
1	shuffle (array [], array [])	Metode ini digunakan untuk melakukan <i>random</i> data yang akan digunakan untuk data <i>training</i> dan <i>testing</i> .
2	DictVectorizer ()	Metode ini merupakan <i>constructor</i> yang akan digunakan untuk memetakan nilai fitur menjadi vector.
3	Fit_transform (array [], array [])	Metode ini digunakan untuk memetakan array dari nilai fitur menjadi vektor berdasarkan fitur pada data <i>training</i> .

II. LANDASAN TEORI

Tabel 2.13 Tabel metode pada *library* Scikit-learn

No	Metode	Keterangan
4	Transform (array [], array [])	Metode ini digunakan untuk memetakan array dari nilai fitur menjadi vektor berdasarkan fitur pada data <i>testing</i> , dengan fitur yang tidak terdapat saat melakukan <i>fit_transform()</i> akan diabaikan.
5	load (String)	Mengambil kembali data yang telah disimpan di dalam sebuah <i>file</i> . Metode ini menerima masukan berupa nama <i>file</i> yang ingin diambil datanya.

2.1.16 Bottle

Bottle adalah sebuah *mini web-framework* pada Python. Bottle menyediakan fitur-fitur standar seperti *routing*, *template*, *utilities*, dan *server*. Berikut ini adalah metode yang digunakan pada sistem analisis sentimen ini:

Tabel 2.14 Tabel metode pada *library* Bottle

No	Metode	Keterangan
1	Run()	Metode ini digunakan untuk menjalankan localhost.
2	@route(String)	Metode ini digunakan untuk <i>routing</i> web atau mengakses css dan <i>javascript</i>
3	@view(String)	Metode ini digunakan untuk menggunakan template view html.

II. LANDASAN TEORI

2.2 Tinjauan Studi

Tabel 2.15 Tabel Tinjauan Studi

No	Peneliti	Judul	Tahun	Masalah	Metode/Accuracy
1	Edwin Lunando, Ayu Purwarianti	<i>Indonesian Social Media Analysis sentimen with Sarcasm Detection</i>	2013	Deteksi teks sarkasme pada analisis sentimen yang masih sulit dilakukan	Pada klasifikasi positif, netral, dan negatif: 1. <i>Naive Bayes</i> - 77.40% 2. <i>Maximum Entropy</i> - 78.40% 3. <i>SVM</i> - 77.80% Pada klasifikasi sarkasme dan non-sarkasme: 1. <i>Naive Bayes</i> - 53.10% 2. <i>Maximum Entropy</i> - 53.80% 3. <i>SVM</i> - 54.10%
2	Tomas Ptacek, Ivan Habernal, Jun Hong	<i>Sarcasm Detection on Czech and English Twitter</i>	2014	Deteksi teks sarkasme pada analisis sentimen yang masih sulit dilakukan pada Bahasa Czech Karena banyaknya jenis syntax dan morfologi	Pada klasifikasi sarkasme dan non-sarkasme 1. <i>Maximum Entropy</i> - 94.66% 2. <i>SVM</i> - 93.13%

II. LANDASAN TEORI

Tabel 2.15 Tabel Tinjauan Studi

No	Peneliti	Judul	Tahun	Masalah	Metode/Accuracy
3	Chun-Che Peng, Mohammad Lakis, Jan Wei Pan	<i>Detecting Sarcasm in Text: An Obvious Solution to a Trivial Problem</i>	2015	Deteksi teks sarkasme pada analisis sentimen yang masih sulit dilakukan	Pada klasifikasi sarkasme dan non-sarkasme: 1. <i>Naive Bayes</i> - 62.02% 2. <i>One Class SVM</i> - 50.00% 3. <i>Gaussian Kernel</i> - 82.20%

Pada penelitian satu, Edwin Lunando dan Ayu Purwarianti menggunakan fitur *unigram*, *negativity*, *number of interjection word*, *sentiment score* dan *question word*. Pada penelitian ini Edwin Lunando dan Ayu Purwarianti membuktikan fitur *negativity* dapat meningkatkan akurasi dalam menentukan teks sarkasme. Untuk menggunakan fitur *negativity* akan memakan banyak waktu. Hal ini disebabkan oleh perlunya pengetahuan yang cukup mengenai isi dari teks. Hasil klasifikasi dengan *direct method* menghasilkan akurasi yang lebih baik dibanding *levelled method*. Penelitian ini menggunakan klasifikasi *Naive Bayes*, *Maximum Entropy* dan *SVM*, dengan akurasi tertinggi dihasilkan oleh *Maximum Entropy* [5].

Pada penelitian dua, Tomas Ptacek, Ivan Habernal dan Jun Hong menggunakan berbagai macam fitur seperti *n-gram*, *pattern*, *part of speech*, *emoticons*, *punctuation based*, *pointedness*, *word-case*. Fitur *pointedness* adalah fitur yang menangani tanda seperti tanda baca tanya (?), tanda baca seru(!) dan *emoticon*. Fitur tersebut tidak digunakan karena penelitian ini tidak menangani *emoticon* pada *tweet*. Fitur *pattern* adalah fitur yang digunakan untuk menghitung kemunculan kata jika memenuhi *threshold* minimal kemunculan kata. Fitur tersebut tidak digunakan, karena penelitian ini sudah menggunakan fitur *unigram* untuk menghitung jumlah kemunculan kata. Pada penelitiannya didapatkan fitur terbaik seperti *n-gram*, *pointedness*, *part of speech*, *emoticon* dan *word-case*. Penelitian ini menggunakan klasifikasi *Maximum Entropy*, dan *SVM*, dengan

akurasi tertinggi dihasilkan oleh *Maximum Entropy* [3].

Pada penelitian tiga, Chun-Che Peng, Mohammad Lakis, Jan Wei Pan menggunakan fitur *n-gram*, *sentiment score*, *part of speech*, *capitalization* dan *topic*. Pada penelitian tiga, penggunaan fitur *topic* akurasi meningkat secara signifikan. Penelitian ini menggunakan klasifikasi *Naive Bayes*, *One Class SVM*, dan *Gaussian Kernel*, dengan akurasi tertinggi dihasilkan oleh klasifikasi *Gaussian Kernel* [4].

2.3 Objek Penelitian

Pada bagian ini akan dibahas objek-objek yang terkait dengan analisis sentimen.

2.3.1 Twitter

Twitter adalah sebuah *microblogging service* dengan jumlah pengguna 41 juta pada juli 2009. *Tweet* pada Twitter memiliki beragam topik dengan batas teks mencapai 140 karakter. Berdasarkan penelitian yang dilakukan pada juli 2009, Twitter memiliki 41.7 juta pengguna, 1.47 juta relasi social, 4262 *trending topic* dan 106 juta *tweet*. Pada Twitter pengguna dapat melakukan *follow* ke pengguna lain, dan pengguna yang di *follow* tidak perlu melakukan *follow* kembali. Menjadi seorang *follower*, pengguna akan menerima semua pesan yang dibuat oleh pengguna yang diikuti. Twitter memiliki fitur *re tweet* yang digunakan untuk menyebarkan informasi kepada *follower* pengguna. Istilah pada Twitter terus berkembang, contoh RT adalah singkatan untuk *retweet*, "@" diikuti oleh nama pengguna atau biasa disebut *mention*, dan "#" diikuti oleh sebuah kata yang disebut *hashtag* [19].

2.3.2 Sarkasme

Ironi adalah sebuah bentuk penyampaian secara tidak langsung untuk menyampaikan perasaan. Karakteristik dari ironi adalah pertentangan antara makna dari sebuah kalimat dan maksud dari pembicara yang sebenarnya. Salah satu bentuk ironi adalah sarkasme. Sarkasme digunakan untuk menyampaikan kritik kepada pendengar atau situasi yang disertai dengan penghinaan. Sebagai contoh, ada seorang pegawai yang sedang tidur siang saat bekerja, kemudian atasannya berkomentar "jangan bekerja terlalu keras!". Jika dilihat dari kalimat, kalimat tersebut tidak memiliki makna yang negatif, tetapi kalimat tersebut sebenarnya digunakan untuk mengkritik pegawainya yang malas [20].

BAB III

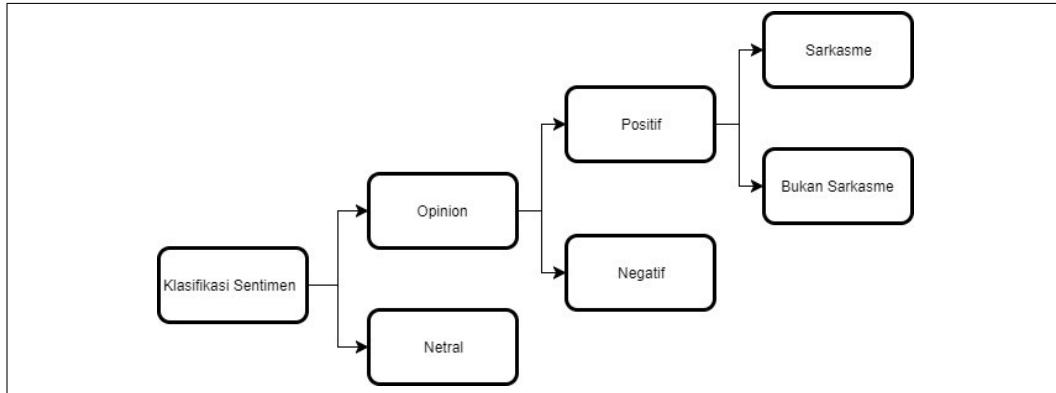
ANALISIS DAN PERANCANGAN SISTEM

Bab ini menjelaskan analisis masalah beserta pendekatan dan alur kerja dari aplikasi yang akan dikembangkan, dimulai dari *preprocessing*, implementasi metode dan hasil yang ditampilkan.

3.1 Analisis Masalah

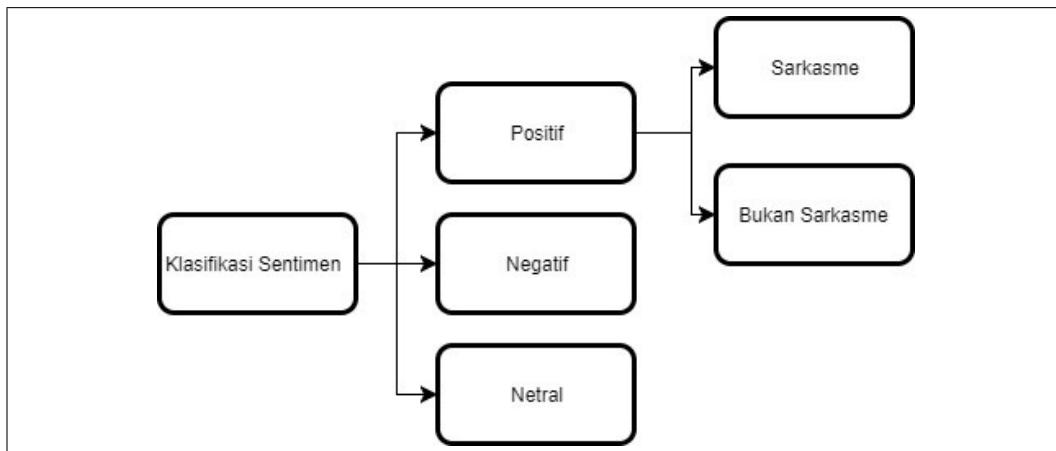
Dalam penelitian ini, penulis akan menggunakan data berupa *tweet* dari media sosial Twitter untuk data *training* dan data *testing*. Penulis memilih media sosial Twitter sebagai objek penelitian, karena Twitter memiliki fitur *hashtag* yang dapat digunakan untuk mendapatkan data sarkasme, yang kemunculannya sedikit. Selain itu, pada tahun 2011 tercatat ada 200 juta *tweet* setiap harinya. Penulis akan menggunakan klasifikasi *Support Vector Machine* (SVM) pada pengembangan sistem analisis sentimen ini. *Kernel SVM* yang dipilih adalah linear, karena jumlah fitur lebih banyak dibanding jumlah data. Data diambil dengan *scraping* html pada halaman Twitter menggunakan *library* Python, yaitu Twitter-scraper. Data diambil berdasarkan pencarian pada Twitter dengan berbagai *keyword*. *Keyword* yang digunakan adalah "DPR", "film", "sekolah" dan "internet" untuk mendapatkan data dengan label positif, negatif dan netral. Sedangkan data dengan label sarkasme diambil dengan menggunakan *keyword* seperti "DPR #sarkasme", "film #sarkasme", "sekolah #sarkasme", dan "internet #sarkasme". Digunakannya *keyword* tersebut karena *keyword* tersebut memiliki cukup banyak data sarkasme, dengan 13 teks sarkasme pada *keyword* DPR, 3 teks sarkasme pada *keyword* film, 12 teks sarkasme pada *keyword* sekolah dan 7 teks sarkasme pada *keyword* internet. Dalam penelitian ini akan dilakukan klasifikasi dengan 2 teknik klasifikasi, yaitu *levelled method* dan *direct method* [5]. Pengklasifikasian akan menganggap teks sarkasme sebagai positif sarkasme, karena teks sarkasme cenderung terlihat seperti teks positif, namun bernilai negatif [5]. Berikut ini adalah *flowchart* proses klasifikasi dengan *levelled method*:

III. ANALISIS DAN PERANCANGAN SISTEM



Gambar 3.1 Klasifikasi dengan *Levelled Method*

Berikut ini adalah *flowchart* proses klasifikasi dengan *direct method*:



Gambar 3.2 Klasifikasi dengan *Levelled Method*

Klasifikasi *levelled method* dan *direct method* akan membangun sebanyak 4 model. Berikut ini adalah model yang diperlukan untuk klasifikasi *levelled method* dan *direct method*:

1. Model 1 : kelas netral atau bukan netral.
2. Model 2 : kelas positif atau bukan positif.
3. Model 3 : kelas negatif atau bukan negatif.
4. Model 4 : kelas sarkasme atau bukan sarkasme.

Pada klasifikasi *levelled method* akan dilakukan klasifikasi teks termasuk sebagai kelas netral atau kelas opini. Jika teks merupakan kelas opini, maka akan diklasifikasikan menjadi teks positif atau negatif. Jika hasilnya kelas positif, maka akan diklasifikasikan menjadi teks sarkasme atau non-sarkasme. Sedangkan pada klasifikasi *direct method* teks akan langsung diklasifikasikan sebagai 3 kelas, yaitu positif, negatif atau netral. Jika teks merupakan positif, maka akan diklasifikasikan

III. ANALISIS DAN PERANCANGAN SISTEM

sebagai sarkasme atau non-sarkasme. Penelitian ini akan membandingkan akurasi klasifikasi *levelled method* dan *direct method*.

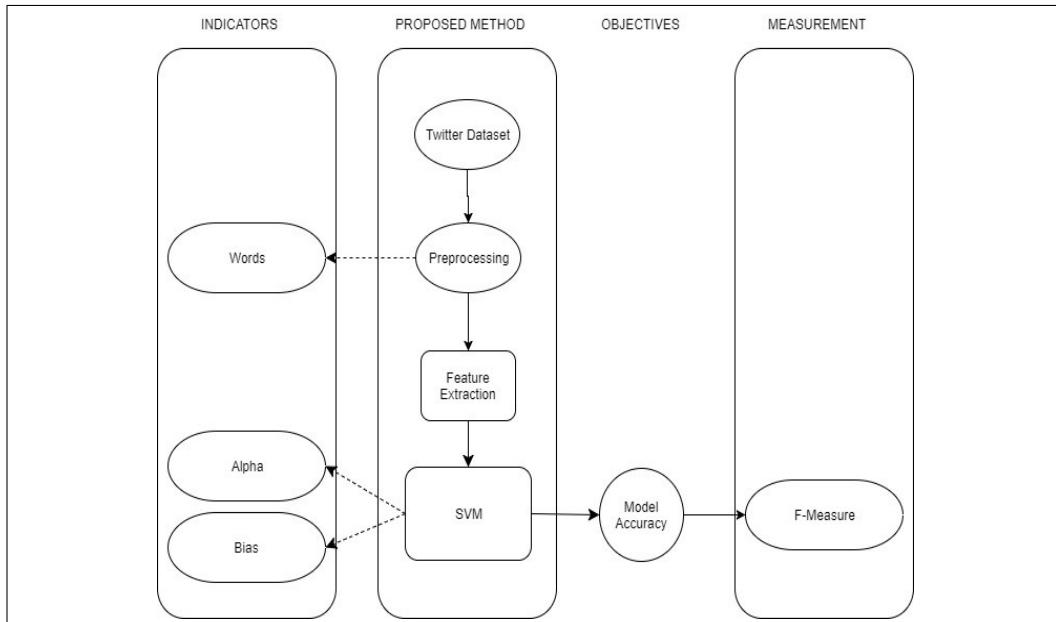
Data *tweet* yang dikumpulkan akan diberi label secara manual. Data *tweet* yang diambil tersebut diberi label sebagai positif, negatif, netral, sarkasme. Salah satu fitur yang dianggap membantu dalam penentuan teks sarkasme adalah menggunakan fitur seperti *emoticon*, kemunculan kata *adjective* dan *adverb*, kemunculan *interjection* dan penggunaan *punctuation* [5]. Fitur-fitur yang akan digunakan pada penelitian ini adalah *unigram*, *interjection*, *question word*, *sentiment score*, *capitalization*, *topic*, *part of speech* dan *punctuation-based*. Fitur *unigram* lebih dipilih dibanding *bigram*, karena kata-kata pada media social terlalu beragam, sehingga sulit untuk menemukan kata yang sama jika menggunakan *bigram*. Berikut ini adalah contoh teks yang akan menganggap sebuah kata beda jika menggunakan *bigram*:

1. Se7...kalau boleh setiap hari kemerdekaan WAJIB TAYANG **DI SEKOLAH**
2. Ceritanya menarik. Karakternya juga unik. Salah satu karakter favorit, Endong. Aktingnya cukup keren. Sangat cocok di tonton **anak sekolah**.
3. Tiap sore **pulang sekolah** ngarepin paket dating

Berdasarkan data di atas, jika menggunakan *bigram*, akan menghasilkan token [”tayang sekolah”], [”anak sekolah”], dan [”pulang sekolah”]. Hal tersebut menyebabkan setiap token tersebut akan dianggap berbeda dan nilai fitur tidak bagus. Kata ”di” pada teks 1 dihapus, karena kata ”di” tidak memiliki makna, oleh karena itu fitur kata *bigram* menjadi [”tayang sekolah”]. Sedangkan jika menggunakan *unigram*, akan menghasilkan token [”tayang”, ”sekolah”], [”anak”, ”sekolah”], dan [”pulang”, ”sekolah”]. Sehingga kata ”sekolah” dapat memberikan nilai fitur yang lebih baik karena kemunculannya adalah 3.

3.2 Kerangka Pemikiran

Berikut ini adalah kerangka pemikiran dari metode yang diusulkan untuk melakukan klasifikasi:



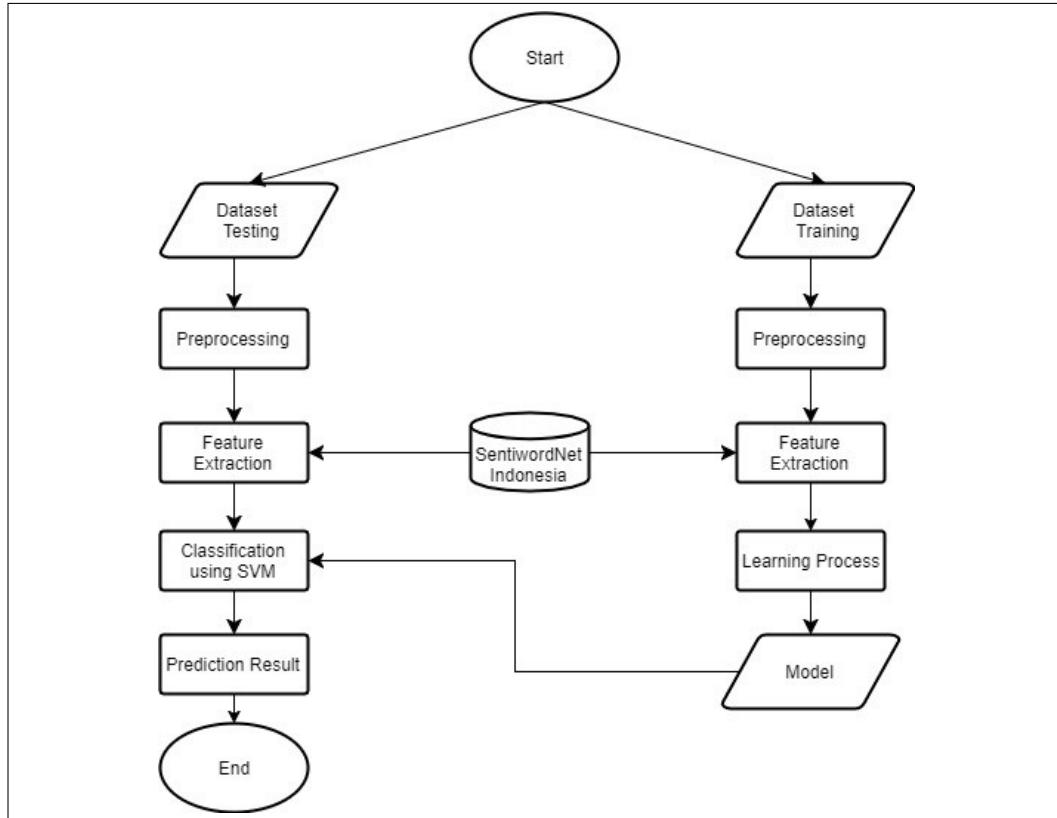
Gambar 3.3 Kerangka kerja klasifikasi teks

Sistem akan dimulai dari masukan data Twitter. Kemudian melakukan proses *text preprocessing* dengan menggunakan data Twitter yang sudah diberi label. Tahap selanjutnya setelah *preprocessing* adalah melakukan *feature extraction*. Hasil *feature extraction* akan digunakan untuk pemodelan klasifikasi SVM. Setelah mendapatkan modelnya, maka dapat dihitung akurasinya dengan *f-measure*.

3.3 Flowchart Sistem Analisis Sentimen

Berikut adalah flowchart untuk sistem analisis sentimen dalam penelitian ini:

III. ANALISIS DAN PERANCANGAN SISTEM



Gambar 3.4 Flowchart Sistem Analisis Sentimen

Alur proses sistem dimulai dari data *training*. Kemudian dilanjutkan dengan *preprocessing*. Setelah itu, data *training* yang sudah di *preprocessing* akan masuk ke tahap *feature extraction* untuk mendapatkan informasi dari setiap teks. *Feature extraction* akan menggunakan data SentiWordNet untuk mendapatkan fitur nilai sentimen dari sebuah teks. Kemudian melakukan *learning process* untuk menghasilkan model klasifikasi SVM. Setelah model didapatkan, data *testing* dapat melakukan tahap-tahap seperti pada data *training*, yaitu *preprocessing* dan *feature extraction*. Setelah mendapat fitur dari data *testing*, selanjutnya dapat melakukan klasifikasi teks dengan model yang sudah melalui *learning process*.

3.3.1 Analisis Data

Dalam penelitian ini, akan digunakan 236 data *tweet* yang sudah diberi kelas pada setiap *tweet*. Kelas tersebut dibagi menjadi 4 kelas yaitu, kelas positif, kelas negatif, kelas netral dan kelas sarkasme. Selain itu, akan digunakan 177 data *tweet* sebagai data *training*, dan 59 data *tweet* sebagai data *testing*. Berikut contoh dari data *tweet* yang akan digunakan:

1. Kelas Positif

Sukses untuk pak @aniesbaswedan, stop reklamasi !!

III. ANALISIS DAN PERANCANGAN SISTEM

2. Kelas Negatif

Anggota DPR skrg tidak mementingkan rakyat.

3. Kelas Netral

Menurut Netizen, Apakah DPR RI/DPRD sudah mewakili aspirasi rakyat?

4. Kelas Sarkasme

Semoga FH dan FZ ada di DPR selamanya.. krn sepi dunia kalau gak ada mereka.. :)

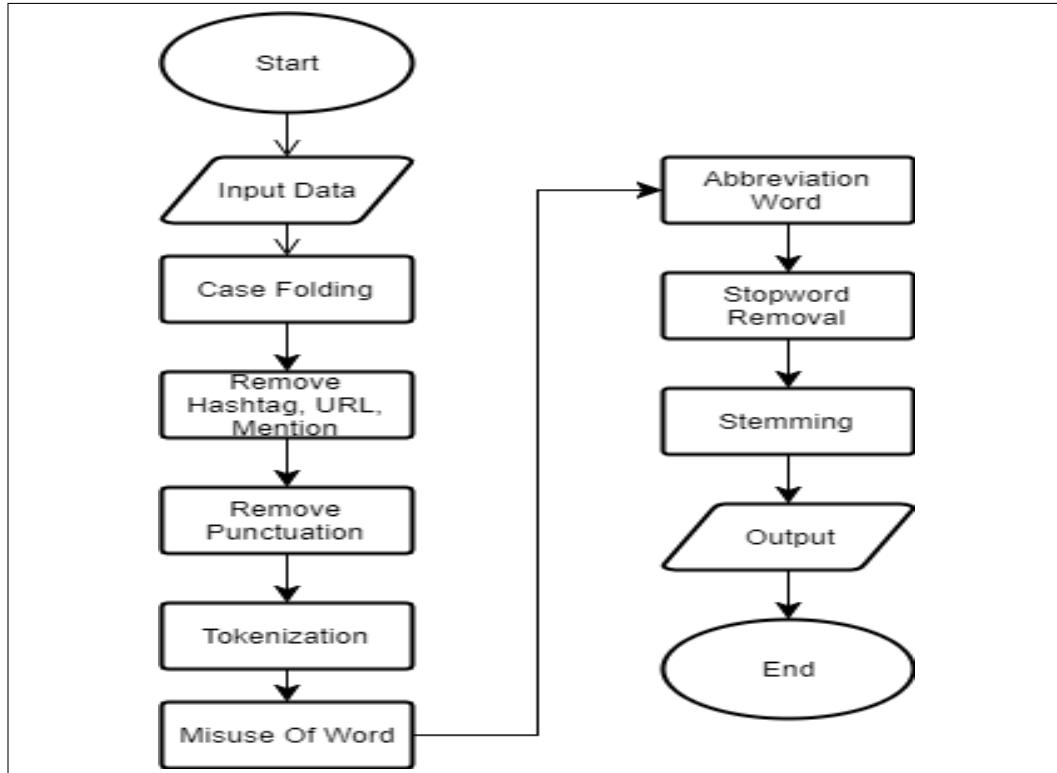
Karakteristik data yang digunakan:

1. Semua sarkasme akan dianggap sebagai sarkasme positif.
2. Jika ada perbandingan pada sebuah kalimat, pemberian label berdasarkan kalimat yang lebih memberi nilai positif maupun negatif, sebagai contoh *tweet* "Tapi, penilaian sy kinerja pemerintahan dan kpk lebih baik daripada yang di dpr", teks tersebut akan diberi label positif.

3.3.2 *Text Preprocessing*

Sebelum melakukan ekstraksi fitur, akan dilakukan *text preprocessing* pada data *training* dan data *testing*. *Text preprocessing* dilakukan untuk mengurangi *noise*, dan mengurangi kata-kata non-formal. Beberapa metode untuk menangani *noise* atau kata-kata non-formal pada *text preprocessing* seperti *case folding*, *remove hashtag*, URL dan *mention*, *remove punctuation*, *tokenization*, *misuse of word*, *abbreviation word*, *stopword removal*, dan *stemming*. Berikut ini adalah *flowchart text preprocessing* yang akan dilakukan:

III. ANALISIS DAN PERANCANGAN SISTEM



Gambar 3.5 Flowchart Text Preprocessing

3.3.2.1 Case Folding

Pada tahap *Case Folding*, kalimat pada teks akan diubah menjadi huruf kecil. *Case folding* dilakukan untuk mengatasi masalah kata yang sama, hanya berbeda huruf kapital.

Tabel 3.1 Tabel Hasil Case Folding

Teks	Hasil
BPK terbaik. BI terbaik. DPR ter? @Fahrihamzah ter? Mandi dlu Ooom!	bpk terbaik. bi terbaik. dpr ter? @fahrihamzah ter? mandi dlu ooom!

3.3.2.2 Remove Hashtag, URL, Mention

Pada tahap ini, semua *hashtag*, *URL*, *mention* akan dihapus dari teks. Penghapusan ini dilakukan karena dalam teks sarkasme sering terdapat *hashtag* #sarkasme, hal ini dapat membantu dalam sebuah fitur jika tidak menghapusnya. Namun setiap teks yang memiliki *hashtag* #sarkasme, tidak menentukan bahwa sebuah teks adalah sarkasme, sehingga *hashtag* akan dihapus. Berikut ini adalah contoh hasil *remove hashtag*, *URL*, dan *mention*:

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.2 Tabel Hasil Remove Hashtag, URL, Mention

Teks	Hasil
bpk terbaik. bi terbaik. dpr ter? @fahrihamzah ter? mandi dlu ooom!	bpk terbaik. bi terbaik. dpr ter? ter? mandi dlu ooom!

3.3.2.3 Remove Punctuation

Pada tahap ini, semua tanda baca akan dihapus dari teks, kecuali tanda petik ('), tanda petik tunggal ('), tanda seru (!), tanda tanya (?) dan tanda pemisah (-). Tanda baca tertentu tidak dihapus, karena akan digunakan sebagai salah satu fitur yang ada. Contoh tanda baca yang akan dihapus adalah "%", "&", "*", "{}", "()", "[]", ":";, dan lain-lain. Penghapusan tanda baca dilakukan untuk memperkecil fitur. Berikut ini adalah contoh hasil *remove punctuation*:

Tabel 3.3 Tabel Hasil Remove Punctuation

Teks	Hasil
bpk terbaik. bi terbaik. dpr ter? ter? mandi dlu ooom!	bpk terbaik bi terbaik dpr ter? ter? mandi dlu ooom!

3.3.2.4 Tokenization

Tokenization adalah tahap memecahkan kalimat menjadi token-token. *Tokenization* dilakukan untuk mengubah kalimat menjadi token kata, dan kata-kata yang diikuti tanda baca tanpa diberi jarak spasi akan ikut dipisahkan. Sebagai contoh kalimat "asyik!!!", kata asyik akan terpisah dari tanda baca seru (!), sehingga menjadi token kata "asyik", "!", "!", "!". Pada tahap ini akan digunakan *library* NLTK, sebagai *tokenization*. Penulis lebih memilih NLTK dibanding Tweet-preprocessor, karena Tweet-preprocessor tidak dapat mengatasi kata yang diikuti tanda baca. Berikut ini adalah *tweet* yang akan dilakukan *tokenization*:

bpk terbaik bi terbaik dpr ter? ter? mandi dlu ooom!

Hasil dari *tokenization* dari teks tersebut adalah:

Tabel 3.4 Tabel Hasil Tokenization

Token
bpk

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.4 Tabel Hasil *Tokenization*

Token
terbaik
bi
terbaik
dpr
ter
?
ter
?
mandi
dlu
oom

3.3.2.5 *Misuse of Word*

Setelah tahap *tokenization*, tahap selanjutnya adalah mengubah penyalahgunaan kata atau huruf sama yang saling bersebelahan. Tahap ini diperlukan supaya mengurangi kesalahan penggunaan kata. *Method Python* yang digunakan untuk menangani masalah ini adalah “`itertools.groupby(string)`”. Berikut hasil dari *misuse of word*:

Tabel 3.5 Tabel Hasil *Misuse of Word*

Token	Hasil
bpk	bpk
terbaik	terbaik
bi	bi
terbaik	terbaik
dpr	dpr

Tabel 3.5 Tabel Hasil *Misuse of Word*

Token	Hasil
ter	ter
?	?
ter	ter
?	?
mandi	mandi
dlu	dlu
ooom	om

3.3.2.6 Abbreviation Word

Setelah tahap *misuse of word*, tahap selanjutnya adalah mengubah kata-kata yang menggunakan singkatan. Pada tahap ini akan dilakukan pencarian token kata ke dalam kamus kata *abbreviation* yang sudah dibuat secara manual dan menggantikan token kata tersebut dengan persamaan katanya. Berikut hasil dari *abbreviation word* :

Tabel 3.6 Tabel Hasil *Abbreviation Word*

Token	Hasil
bpk	bpk
terbaik	terbaik
bi	bi
terbaik	terbaik
dpr	dpr
ter	ter
?	?
ter	ter
?	?

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.6 Tabel Hasil *Abbreviation Word*

Token	Hasil
mandi	mandi
dlu	dulu
om	om

3.3.2.7 *Stopword Removal*

Setelah tahap *abbreviation word*, maka teks akan siap memasuki tahap berikutnya yaitu *Stopword Removal*. Pada tahap ini, token-token kata yang terdapat pada daftar *stopword* akan dihilangkan, karena dianggap sebagai kata-kata yang tidak penting. Berikut ini adalah hasil dari *stopword removal*:

Tabel 3.7 Tabel Hasil *Stopword Removal*

Token	Hasil
bpk	bpk
terbaik	terbaik
bi	bi
terbaik	terbaik
dpr	dpr
ter	ter
?	?
ter	ter
?	?
mandi	mandi
dulu	-
om	om

3.3.2.8 *Stemming*

Setelah tahap *stopword removal*, maka teks akan siap memasuki tahap berikutnya yaitu *stemming*. Pada tahap *stemming* ini, token yang ada akan diubah menjadi kata dasar. Pada tahap ini penulis akan menggunakan *library* sastrawi yang terdapat pada python untuk melakukan *stemming*. Berikut ini adalah hasil dari tahap *stemming*:

Tabel 3.8 Tabel Hasil *Stemming*

Token	Hasil
bpk	bpk
terbaik	baik
bi	bi
terbaik	baik
dpr	dpr
ter	ter
?	?
ter	ter
?	?
mandi	mandi
om	om

3.3.3 Feature Extraction

Setelah melakukan pemrosesan teks menjadi lebih terstruktur, setiap kata dalam dokumen teks diekstraksi agar setiap teks memperoleh dan mengkalkulasi berdasarkan kata yang penting untuk diolah lebih lanjut saat mengklasifikasi. Pada penelitian ini, digunakan proses *unigram*, *part of speech*, *sentiment Score*, *punctuation based*, *capitalization*, *topic*, *interjection* dan *question word*.

3.3.3.1 *Unigram*

Unigram adalah fitur untuk mengambil kata dari sebuah teks. *Unigram* merupakan fitur yang paling sesuai untuk media sosial Indonesia, Karena struktur

III. ANALISIS DAN PERANCANGAN SISTEM

kata yang digunakan pada media sosial Indonesia sangat beragam dan tidak formal [5]. Fitur ini akan digunakan untuk menghitung kemunculan kata pada sebuah teks.

Tabel 3.9 Contoh Perhitungan Fitur *Unigram*

Kata	D1	D2
DPR	1	1
Sukses	2	0

Tabel di atas menunjukkan kata "DPR" muncul sebanyak 1 kali pada dokumen D1 dan D2, sedangkan kata "Sukses" muncul sebanyak 2 kali pada dokumen D1 dan tidak muncul pada dokumen D2.

3.3.3.2 *Part of Speech*

Fitur ini digunakan untuk menghitung jumlah kata benda, kata sifat, kata kerja dan kata keterangan yang terdapat pada teks. Sebelum dapat menghitung kemunculannya, diperlukan *tagging* terlebih dahulu untuk tiap teks. *Tagging* adalah proses menandai sebuah kata pada teks sesuai dengan *tag* yang ada. Berikut ini hasil dari proses *tagging* pada *tweet* "Anggota DPR sekarang tidak mementingkan rakyat":

Anggota / NN	DPR / IN	Sekarang / NN	Tidak / NEG	Mementingkan / VBT	Rakyat / NN
--------------	----------	---------------	-------------	--------------------	-------------

Berikut perhitungan kemunculan POS *tag* pada *tweet* "Anggota DPR sekarang tidak mementingkan rakyat":

Tabel 3.10 Contoh Perhitungan Fitur *Part of Speech*

POS Tag	Jumlah Kemunculan
Jumlah kata benda	3 (Anggota, Sekarang, Rakyat)
Jumlah kata keterangan	1 (DPR)
Jumlah kata negasi	1 (Tidak)
Jumlah kata kerja	1 (Mementingkan)
Jumlah kata sifat	0

III. ANALISIS DAN PERANCANGAN SISTEM

3.3.3.3 *Sentiment Score*

Tweet yang sudah di preprocessing kemudian dicari nilai sentimennya pada SentiWordNet yang disediakan. Untuk mencari nilai sentimen teks, sebelumnya harus melakukan pos tagging terlebih dahulu untuk setiap kata. Kemudian hasilnya digunakan untuk mencari kata yang terdapat pada SentiWordNet. Masukan yang diperlukan untuk menggunakan SentiWordNet adalah kata dan *tag* dari kata tersebut. Sebagai Contoh, "makan/v". Berikut adalah daftar POS Tagging yang diperlukan untuk menggunakan SentiWordNet: *Noun* (n), *Verb* (v), *Adverb* (r), dan *Adjective* (a).

Berikut perhitungan nilai sentimen yang didapatkan dari SentiWordNet Indonesia:

Tabel 3.11 Contoh Perhitungan Fitur *Sentiment Score*

Kata/POS Tag	Nilai Sentimen (Positif, Negatif)
Anggota / NN (n)	(0.01209677, 0.01612903)
DPR / IN (-)	-
Sekarang / NN (n)	(0.04166667, 0.01388889)
Tidak / NEG (-)	-
Mementingkan / VBT (v)	(0.01785714, 0.03571428)
Rakyat / NN (n)	(0.01785714, 0.00892857)
Total Nilai Sentimen (Positif, Negatif)	(0.08947773, 0.05680363)

3.3.3.4 *Punctuation Based*

Punctuation Based adalah fitur yang digunakan untuk menghitung tanda seru (!), tanda tanya (?), dan tanda petik (‘, ’) pada teks [3]. Berikut perhitungan kemunculan *punctuation* atau tanda baca pada *tweet* "Angota DPR sekarang tidak mementingkan rakyat", dengan masing-masing maksimal kemunculan tanda baca seru (!), tanda tanya (?), tanda petik (‘, ’) pada data *training* adalah 5, 8, dan 2. Berikut ini adalah hasil dari perhitungan fitur *punctuation based*:

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.12 Contoh Perhitungan Fitur *Punctuation Based*

Tanda Baca	Jumlah Kemunculan
<i>Exclamation Mark (!)</i>	$0/5=0$
<i>Question Mark (?)</i>	$0/8=0$
<i>Quotation Mark (", ')</i>	$0/2=0$

3.3.3.5 *Capitalization*

Capitalization adalah fitur yang digunakan untuk menghitung jumlah kata yang memiliki keseluruhan hurufnya merupakan huruf kapital pada sebuah *tweet*. Berikut perhitungan kemunculan *capitalization* pada *tweet* "Anggota DPR sekarang tidak mementingkan rakyat", dengan kemunculan maksimal kata kapital pada data *training* adalah 5:

Tabel 3.13 Contoh Perhitungan Fitur *Capitalization*

Kata kapital	Jumlah kemunculan
Jumlah kemunculan kata kapital	$1/5 = 0.2$ (DPR)

3.3.3.6 *Topic*

Fitur ini digunakan untuk mencari topik dari sebuah teks menggunakan *library* LDA pada python yaitu gensim. Contoh perhitungan topik *modelling* pada *tweet* "Anggota DPR sekarang tidak mementingkan rakyat", dengan 4 data *training* dan banyak topik adalah 2:

Tabel 3.14 Tabel Data *Training*

Dokumen	Kalimat
D1	Sudah pak mundur dari DPR saja, ngeluh mulu kapan kerjanya
D2	Titip salam buat bu Prabowo, semoga sukses.
D3	DPR = Dewan Perwakilan Rakyat
D4	Tapi , penilaian saya kinerja pemerintahan dan kpk lebih baik daripada yang di dpr

Mengganti kata-kata *unique* pada teks dengan indeks kata.

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.15 Word Index

Dokumen	Kalimat
D1	1 2 3 4 5 6 7 8 9 10
D2	11 12 13 14 15 16 17
D3	5 18 19 20 21
D4	22 23 24 25 26 27 28 29 30 31 32 33 5

Selanjutnya memberikan topik terhadap tiap token pada dokumen secara random.

Tabel 3.16 Token-topic

Dokumen	Indeks Kata/Topik													
	1	2	3	4	5	6	7	8	9	10				
D1	1	2	1	1	2	1	1	2	2	1				
	1	2	1	1	1	2	2							
D2	11	12	13	14	15	16	17							
	1	2	1	1	1	2	2							
D3	5	18	19	20	21									
	2	1	1	2	2									
D4	22	23	24	25	26	27	28	29	30	31	32	33	5	
	2	2	1	2	2	1	1	1	2	2	1	2	1	

Kemudian menghitung kemunculan *topic* pada setiap *word index*.

Tabel 3.17 Word-topic 1

Topik/Indeks Kata	1	2	3	4	5	6	7	8	9	10	11
1	1	0	1	1	1	1	1	0	0	1	1
2	0	1	0	0	2	0	0	1	1	0	0

Tabel 3.18 Word-topic 2

Topik/Indeks Kata	12	13	14	15	16	17	18	19	20	21	22
1	0	1	1	1	0	0	1	1	0	0	0
2	1	0	0	0	1	1	0	0	1	1	1

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.19 Word-topic 3

Topik/Indeks Kata	23	24	25	26	27	28	29	30	31	32	33
1	0	1	0	0	1	1	1	0	0	1	0
2	1	0	1	1	0	0	0	1	1	0	1

Setelah itu menghitung kemunculan *topic* pada setiap *document*.

Tabel 3.20 Document-topic

Dokumen	Topik 1	Topik 2
D1	6	4
D2	4	3
D3	2	3
D4	6	7

Kemudian dilakukan perhitungan dengan persamaan LDA. Pengulangan dilakukan sampai dokumen terakhir selesai dihitung. Berikut merupakan contoh hasil probabilitas dari setiap kata terhadap topik:

Tabel 3.21 Probabilitas Word-topic

Topic / Indeks Kata	1	2	3	4	5	6	7	8
1	0.0578	0.0134	0.0621	0.0278	0.0244	0.0198	0.0172	0.0273
2	0,037	0.0246	0.012	0.012	0.0319	0.0287	0.652	1

Berikut merupakan contoh hasil probabilitas dari dokumen terhadap topik:

Tabel 3.22 Probabilitas Document-topic

Dokumen	Topik 1	Topik 2
D1	0.5	0.5
D2	0.45	0.55
D3	0.25	0.75
D4	0.65	0.35

III. ANALISIS DAN PERANCANGAN SISTEM

Berikut adalah contoh hasil probabilitas topik pada *tweet* "Anggota DPR sekarang tidak mementingkan rakyat":

Tabel 3.23 Contoh Hasil Perhitungan Fitur *Topic*

Topik	Probabilitas
Topik 1	0.3567
Topik 2	0.456

3.3.3.7 *Interjection*

Fitur ini untuk menghitung jumlah kemunculan kata interjeksi yang terdapat pada sebuah teks. Contoh dari kata interjeksi adalah "aha", "bah", "wew", "wow", "yay", "nah", "uh", dan lain-lain. Fitur ini digunakan untuk pengklasifikasian kelas sarkasme. Berikut adalah perhitungan kemunculan kata *interjection* pada *tweet* "Angota DPR sekarang tidak mementingkan rakyat":

Tabel 3.24 Contoh Hasil Perhitungan Fitur *Interjection*

Teks	Jumlah kemunculan kata interjeksi
Angota DPR sekarang tidak mementingkan rakyat	0

3.3.3.8 *Question Word*

Fitur ini digunakan untuk mengklasifikasikan teks netral. Dengan mendeteksi kata tanya seperti "siapa", "apa", "kapan", "mengapa", "dimana", dan "bagaimana", kata-kata tersebut akan memberikan nilai sentimen netral pada sebuah teks. Fitur ini akan memberikan nilai *true* (1) jika sebuah teks mengandung kata tanya dan sebaliknya akan memberikan nilai *false* (0) jika sebuah teks tidak mengandung kata tanya. Berikut ini adalah contoh hasil fitur ekstraksi *question word*:

Tabel 3.25 Contoh Hasil Perhitungan Fitur *Question Word*

Teks	Nilai (<i>True/False</i>)
Angota DPR sekarang tidak mementingkan rakyat	<i>False</i> , karena tidak terdapat kata tanya pada teks

III. ANALISIS DAN PERANCANGAN SISTEM

3.3.3.9 TF-IDF

Setelah mendapatkan nilai tiap fitur, dilakukan fitur ekstraksi TF-IDF. Berikut ini adalah contoh untuk perhitungan TF-IDF pada fitur *unigram*:

Tabel 3.26 Contoh Data untuk Perhitungan TF-IDF

Kata	D1	D2
DPR	5	2
Sukses	2	0
Semoga	2	8
Total Kata	9	10

Dengan menggunakan rumus TF, diperoleh nilai sebagai berikut:

Tabel 3.27 Contoh Hasil Perhitungan TF

TF	D1	D2
TF (DPR)	$5/9 = 0.56$	$2/10 = 0.2$
TF (Sukses)	$2/9 = 0.22$	$0/10 = 0$
TF (Semoga)	$2/9 = 0.22$	$8/10 = 0.8$

Dengan menggunakan IDF, diperoleh nilai sebagai berikut :

Tabel 3.28 Contoh Hasil Perhitungan IDF

IDF	Nilai IDF
IDF (DPR)	$\text{Log}(2/2) = 0$
IDF (Sukses)	$\text{Log}(2/1) = 0.301$
IDF (Semoga)	$\text{Log}(2/2) = 0$

Setelah mendapatkan nilai TF dan nilai IDF, maka TF-IDF dapat dihitung, berikut ini adalah hasil dari TF-IDF:

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.29 Contoh Hasil Perhitungan Fitur TF-IDF

Kata	D1	D2
DPR	$0.56 * 0 = 0$	$0.2 * 0 = 0$
Sukses	$0.22 * 0.301 = 0.0661$	$0 * 0.301 = 0$
Semoga	$0.22 * 0 = 0$	$0.8 * 0 = 0$

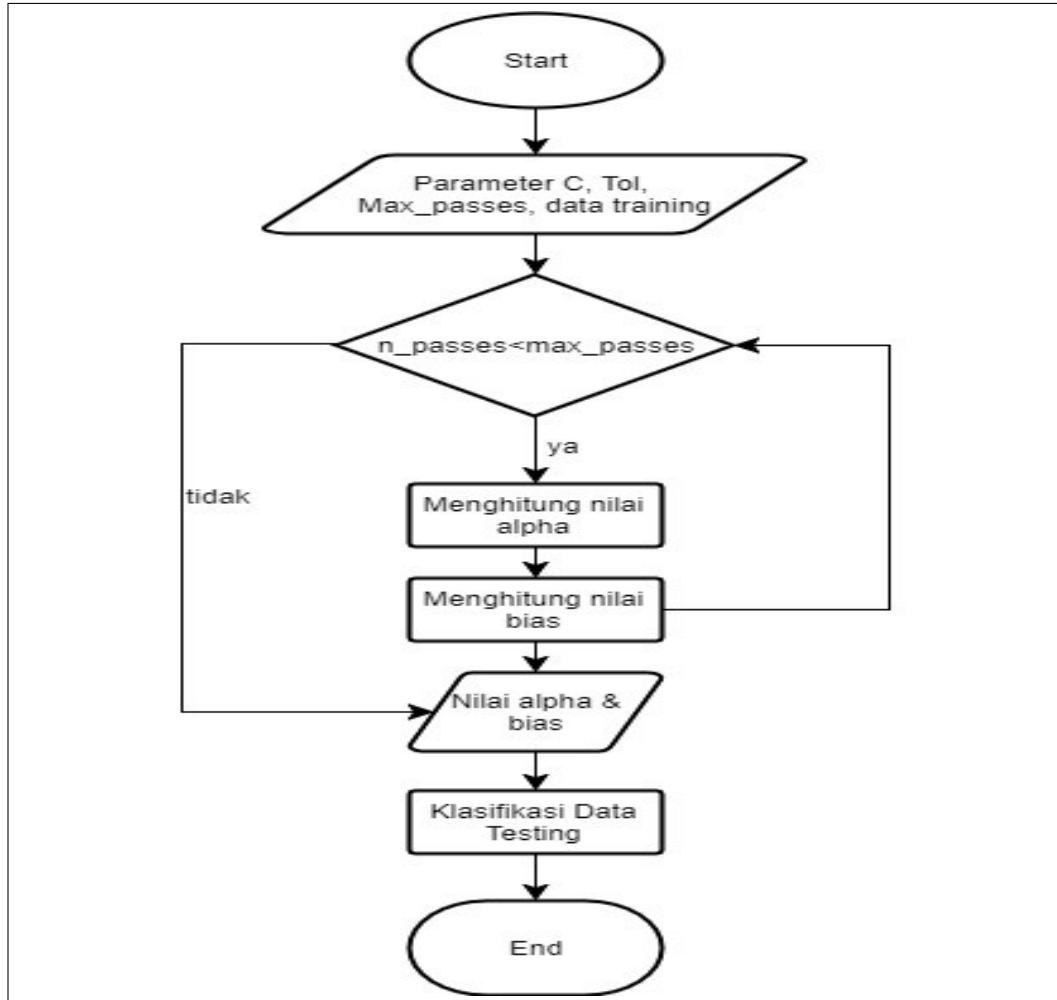
Langkah di atas dilakukan untuk setiap fitur yang ada, hingga dapat nilai TF-IDFnnya.

3.3.4 Perhitungan *Support Vector Machine*

Hasil perhitungan TF-IDF di atas akan digunakan sebagai nilai masukan dalam SVM. Pada penelitian ini, proses klasifikasi teks menggunakan SVM linear dengan metode *One versus Rest*. Digunakannya SVM linear Karena fitur yang dimiliki lebih banyak dibanding dataset yang ada. Klasifikasi akan dilakukan menggunakan *levelled method* [5] dan *direct method* [5].

Untuk dapat melakukan klasifikasi diperlukan *training* model terlebih dahulu. Langkah awal dalam *training* model adalah memberi nilai masukan seperti C, tol, *max_passes*, fitur-fitur data *training*, dan label dari setiap data. Pengulangan perhitungan pada nilai alpha dan bias akan dilakukan sampai memenuhi *max_passes* yang sudah ditentukan. Ketika pengulangan selesai, hasil perhitungan nilai alpha dan bias akan disimpan. Berikut adalah *flowchart* sistem klasifikasi pada *Support Vector Machine* (SVM) dengan *Simplified Sequential Minimal Optimization* (SMO):

III. ANALISIS DAN PERANCANGAN SISTEM



Gambar 3.6 Flowchart Klasifikasi Support Vector Machine (SVM) dengan SMO

Berikut ini adalah contoh perhitungan klasifikasi teks pada Twitter kelas netral dengan data *training* sebagai berikut:

Tabel 3.30 Data *Training*

Dokumen	Kata					Kelas
	DPR	Sukses	Semoga	Bagus	Jelek	
D1	0.01	0.02	0.04	0.02	0.03	1
D2	0	0.003	0.02	0.01	0	1
D3	0.02	0.01	0	0.01	0	-1
D4	0.02	0	0.03	0.04	0.05	-1
D5	0	0.03	0.02	0.01	0.02	1

Berikut ini adalah data *testing* yang akan digunakan untuk klasifikasi:

III. ANALISIS DAN PERANCANGAN SISTEM

Tabel 3.31 Data Testing

Dokumen	Kata					Kelas
	DPR	Sukses	Semoga	Bagus	Jelek	
T1	0.03	0	0.03	0.02	0	?

Pada tabel 3.30, kelas 1 menunjukkan kelas opini positif, sedangkan kelas -1 menunjukkan kelas opini lainnya. Proses dimulai dengan menghitung *kernel* linear dengan menggunakan rumus tabel 2.7. Berikut ini contoh perhitungan nilai *kernel* data *training* pada tabel 3.30:

$$K(D1,D2) = (0.01*0.01) + (0.02*0.02) + (0.04*0.04) + (0.02*0.02) + (0.03*0.03) = 0.0034$$

$$K(D1,D2) = (0.01*0) + (0.02*0.003) + (0.04*0.02) + (0.02*0.01) + (0.03*0) = 0.00106$$

$$K(D1,D3) = (0.01*0.02) + (0.02*0.01) + (0.04*0) + (0.02*0.01) + (0.03*0) = 0.0006$$

$$K(D1,D4) = (0.01*0.02) + (0.02*0) + (0.04*0.03) + (0.02*0.04) + (0.03*0.05) = 0.0037$$

$$K(D1,D5) = (0.01*0) + (0.02*0.03) + (0.04*0.02) + (0.02*0.01) + (0.03*0) = 0.0022$$

Berikut ini adalah tabel hasil dari perhitungan *kernel* pada setiap dokumen:

Tabel 3.32 Hasil Perhitungan *Kernel* Linear pada Data *Training*

	D1	D2	D3	D4	D5
D1	0.0034	0.00106	0.0006	0.0037	0.0022
D2	0.00106	0.000509	0.00013	0.001	0.00059
D3	0.0006	0.00013	0.0006	0.0008	0.0004
D4	0.0037	0.001	0.0008	0.0054	0.002
D5	0.0022	0.00059	0.0004	0.002	0.0018

Setelah mendapatkan perhitungan *kernel* linear tahap selanjutnya mencari nilai alpha dan bias dengan algoritme *Simplified SMO* sebagai berikut:

III. ANALISIS DAN PERANCANGAN SISTEM

```

C=0.05 (Parameter Regularisasi)
Tol=0.0001 (Toleransi Numerik)
maxIter= 2 (maksimal pengulangan yang dilakukan ketika alpha tidak berubah)
 $\alpha_i = 0$ 
 $\alpha_i(\text{old}) = 0$ 
 $b=0$ 
iter=0
iter < maxIter //kondisi while
    jum_perubahan_alpha = 0
    i = 1//loop sebanyak jumlah dokumen training
         $E_i = f(x_1) - y_1$ 
         $= [(0 * 1 * 0.0034) + (0 * 1 * 0.000106) + (0 * -1 * 0.0006) + (0 * -1 * 0.0037) + (0 * 1 * 0.0022)) + 0] - 1$ 
        //kondisi if terpenuhi
        j = 3 //pilih j random, j != i
         $E_3 = f(x_3) - y_3$ 
         $= [(0 * 1 * 0.0006) + (0 * 1 * 0.00013) + (0 * -1 * 0.0006) + (0 * -1 * 0.0008) + (0 * 1 * 0.0004)) + 0] - -1$ 
         $= 1$ 
         $\alpha_1(\text{old}) = 0$ 
         $\alpha_3(\text{old}) = 0$ 
        L = max(0,(0-0))=0
        H = min(0.05,(0.05 + 0 - 0)) = 0.05
        //kondisi L!=H //kondisi L!=H terpenuhi
         $\eta = 2 * 0.06 - 0.16 - 0.09 = (-0.13)$ 
        //kondisi eta < 0 terpenuhi
         $\alpha_3 = 0 - [ ((-1 * (-1 - 1))) / (-0.13) ] = 15.385$ 
        kondisi  $\alpha_3 > H$  terpenuhi
         $\alpha_3 = 0.05$ 
        // kondisi abs( $\alpha_3 - \alpha_3(\text{old})$ ) >  $10^{-5}$  terpenuhi
         $\alpha_1 = 0 + (1 * -1 * (0 - 0.05)) = 0.05$ 
         $b_1 = 0 - (-1) - (1 * (0.05 - 0) * 0.16) - (-1 * (0.05 - 0) * 0.06) = 0.995$ 
         $b_2 = 0 - (1) - (1 * (0.05 - 0) * 0.06) - (-1 * (0.05 - 0) * 0.09) = -0.9985$ 
         $b = (0.995 + (-0.9985))/2 = -0.0035$ 
        jum_perubahan_alpha = 1
    //pengulangan dilakukan terus sampai data ke i=5
    iter = 0 // kondisi jum_perubahan_alpha !=0
//pengulangan dilakukan terus sampai kondisi while tidak terpenuhi

```

Berikut ini adalah nilai alpha dan bias dari perhitungan *simplified SMO*:

Tabel 3.33 Nilai Alpha dan Bias pada perhitungan SMO

α_1	α_2	α_3	α_4	α_5	bias
0.78	0.5	0.3	0.21	0.25	0.135

III. ANALISIS DAN PERANCANGAN SISTEM

Setelah mendapatkan nilai alpha dan bias, maka dapat dilakukan klasifikasi teks pada tabel 3.31. Perhitungan klasifikasi *testing* juga dilakukan dengan perhitungan *kernel* linear sebagai berikut:

$$K(T1,D1) = (0.03*0.01) + (0*0.02) + (0.03*0.04) + (0.02*0.02) + (0*0.03) = 0.0019$$

$$K(T1,D2) = (0.03*0) + (0*0.03) + (0.03*0.02) + (0.02*0.01) + (0*0) = 0.0008$$

$$K(T1,D3) = (0.03*0.02) + (0*0.01) + (0.03*0) + (0.02*0.01) + (0*0) = 0.0008$$

$$K(T1,D4) = (0.03*0.02) + (0*0) + (0.03*0.03) + (0.02*0.04) + (0*0.05) = 0.0023$$

$$K(T1,D5) = (0.03*0) + (0*0.03) + (0.03*0) + (0.02*0.03) + (0*0.02) = 0.0008$$

Berikut ini adalah hasil perhitungan *kernel* linear pada data *testing* :

Tabel 3.34 Hasil Perhitungan *Kernel* Linear pada Data *Testing*

	D1	D2	D3	D4	D5
T1	0.0019	0.0008	0.0008	0.0023	0.0008

Setelah perhitungan *kernel* linear dilakukan maka dapat dilakukan perhitungan klasifikasi sebagai berikut:

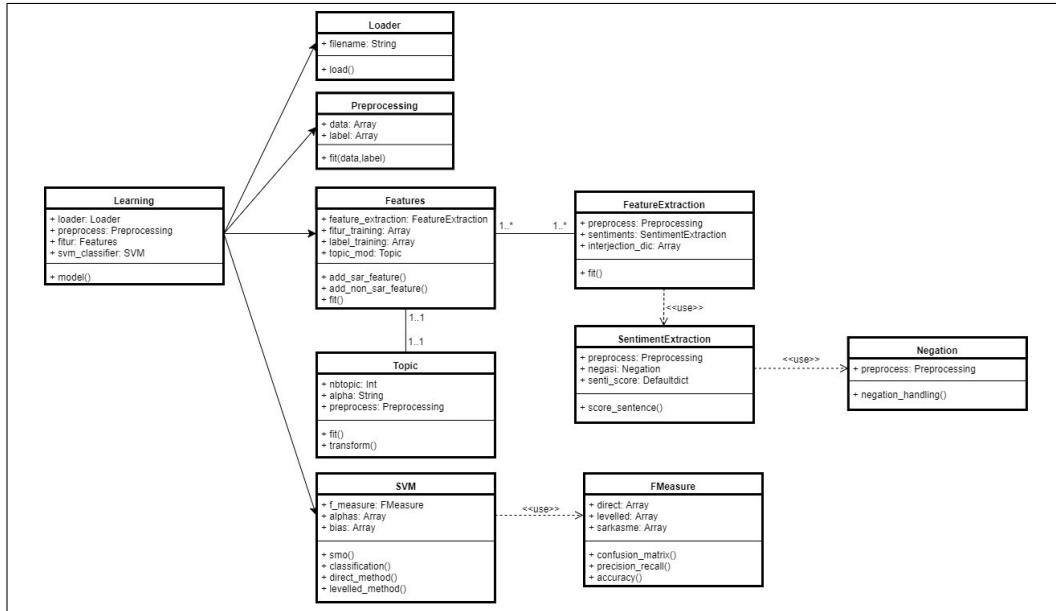
$$F(x) = (0.78 * 1 * 0.0019) + (0.5 * 1 * 0.0008) + (0.3 * -1 * 0.0008) + (0.21 * -1 * 0.0023) + (0.25 * 1 * 0.0008) + 0.135 = -1.86 = -1$$

Berdasarkan perhitungan di atas, dokumen T1 termasuk dalam kelas 1 yang merupakan kelas bukan netral.

III. ANALISIS DAN PERANCANGAN SISTEM

3.3.5 Class Diagram

Berikut ini adalah *design class diagram* pada sistem analisis sentimen yang akan dibuat:



Gambar 3.7 Class Diagram Sistem Analisis Sentimen

BAB IV

IMPLEMENTASI DAN PENGUJIAN

Pada bab ini akan menjelaskan tentang pengimplementasian dan pengujian terhadap analisis sentimen yang telah dibangun berdasarkan bab-bab sebelumnya.

4.1 Lingkungan Aplikasi

Dalam aplikasi terbagi menjadi dua bagian, yaitu lingkungan implementasi perangkat keras dan perangkat lunak. Di dalam penelitian ini, perangkat keras yang digunakan adalah:

1. Lenovo Ideapad 310
2. Processor Intel R Core T i5-6200U CPU 2.30GHz-2.40GHz
3. RAM 8 GB.

Spesifikasi perangkat lunak yang digunakan untuk pengembangan sistem adalah:

1. Sistem Operasi : Windows 10 Home Single Language 64-bit.
2. Tool Pengembangan : Python 2.7.14 64-bit, Pycharm (Python IDE) 2017.2.4, Bottle 0.12.13.
3. *Library* : Scikit-learn, Gensim, IPosTagger, Sastrawi, NLTK, NumPy.

4.2 Daftar *Class* dan *Method*

Pada bagian ini akan dijelaskan mengenai *class* dan *method* yang digunakan dalam pengembangan sistem analisis sentimen

4.2.1 *Class Loader*

Class loader merupakan *class* yang digunakan untuk mengambil data dan label dari sebuah *file* dengan *extension* .txt. Berikut ini adalah daftar *method* pada *class* loader:

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.1 Daftar *Method* pada *Class Loader*

Variabel:	
String	Filename

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	Load	-	-	Array, Array	<i>Method</i> ini digunakan untuk mengambil data teks, beserta label pada file .txt.

4.2.2 *Class Preprocessing*

Class preprocessing merupakan *class* yang digunakan untuk meminimalisir kata, mengurangi bahasa yang non-formal, dan menghapus huruf ataupun tanda baca yang tidak digunakan. Berikut ini adalah daftar *method* pada *class preprocessing*:

Tabel 4.2 Daftar *Method* pada *Class Preprocessing*

Variabel:		Variabel:	
Stemmer (Class Lib)	stemmer	Array	data
Array (String)	question_dic	Array	label
Array (String)	negasi	Int	size
Array (String)	stopword	Array(String)	abbreviation_dic

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	case_folding	- String	- tweet	String	<i>Method</i> ini digunakan untuk mengecilkan setiap huruf pada tweet.

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
2	remove_hashtag_mention_url	- String	- tweet	String	<i>Method</i> ini digunakan untuk menghapus semua textithashtag, <i>mention</i> , URL pada <i>tweet</i> .
3	remove_punctuation	- String	- tweet	String	<i>Method</i> ini digunakan untuk menghapus semua tanda baca yang ada pada <i>tweet</i> kecuali tanda seru (!), tanda tanya (?), tanda petik (”), tanda petik tunggal (‘), dan tanda pemisah (-).
4	tokenize	- String	- tweet	Array	<i>Method</i> ini digunakan untuk memisahkan teks menjadi token-token yang terdiri dari satu kata.
5	misuse_of_word	- String	- tweet	String	<i>Method</i> ini digunakan untuk menghilangkan huruf sama yang saling bersebelahan.
6	abbreviation_word	- String	- tweet	String	<i>Method</i> ini berguna untuk mengubah kata-kata singkatan menjadi kata persamaanya.
7	stopword_removal	- String	- tweet	String	<i>Method</i> ini digunakan untuk menghilangkan kata-kata yang dianggap tidak memiliki makna.

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
8	stemming	- String	- tweet	String	<i>Method</i> ini digunakan untuk mengubah kata menjadi kata dasar.
9	get_count_max	- String	- tweet	Int, Int, Int, Int	<i>Method</i> ini digunakan untuk mendapatkan kemunculan terbanyak tanda baca tanya (?), tanda seru (!), tanda quotation (", '), dan kata kapital.
10	write_file	- Array - Array	- data - label	Array, Array	<i>Method</i> ini digunakan untuk melakukan random terhadap data dan label, kemudian menyimpan data dan label kedalam file .txt.
11	get_size	- Array	- label	Int	<i>Method</i> ini digunakan untuk menghitung jumlah data yang akan digunakan sebagai data <i>training</i> , dengan mengalikan jumlah keseluruhan data dengan 75%.
12	Fit	- Array - Array	- data - label	Int, Int, Int, Int, Array	Fitur ini digunakan untuk melakukan <i>preprocessing</i> secara keseluruhan terhadap data <i>tweet</i> , serta mengembalikan hasil kemunculan maksimal tanda baca, dan kata kapital.

IV. IMPLEMENTASI DAN PENGUJIAN

4.2.3 Class Features

Class features merupakan *class* yang digunakan untuk menangani hal terkait *feature*, dimulai dari penambahan *feature set*, penyimpanan nilai fitur ke dalam *file*, pengambilan *feature* berdasarkan jenis klasifikasi. Jenis klasifikasi yang ada pada sistem ini adalah klasifikasi 4 kelas (positif, negatif, netral, sarkasme), klasifikasi 3 kelas (positif, negatif, netral), 1 kelas (sarkasme/non-sarkasme). Berikut ini adalah *method* pada *class features*:

Tabel 4.3 Daftar *Method* pada *Class Features*

Variabel:			Variabel:
FeatureExtraction (Class)	feature_extraction	Array	IDF
Array	feature_training	Array	label_training
Topic (Class)	topic_mod		

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	get_sar_feature	<ul style="list-style-type: none"> - Array - Array 	<ul style="list-style-type: none"> - data - label 	Array, Array	<i>Method</i> ini digunakan untuk menghapus semua data kecuali data positif dan sarkasme.
2	get_net_feature	<ul style="list-style-type: none"> - Array - Array 	<ul style="list-style-type: none"> - data - label 	Array, Array	<i>Method</i> ini digunakan untuk menghapus semua data sarkasme.
3	add_non_sar_feature	<ul style="list-style-type: none"> - String 	<ul style="list-style-type: none"> - tweet 	Array	<i>Method</i> ini digunakan untuk mendapatkan semua nilai fitur untuk kelas positif, negatif dan netral.
4	add_sar_feature	<ul style="list-style-type: none"> - String 	<ul style="list-style-type: none"> - tweet 	Array	<i>Method</i> ini digunakan untuk mendapatkan semua nilai fitur untuk kelas sarkasme.

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
5	Fit	<ul style="list-style-type: none"> - Array - String 	<ul style="list-style-type: none"> - data - type 	Array	<p><i>Method</i> ini digunakan untuk memanggil <i>method</i> add_sar_feature dan add_non_sar_feature untuk menambahkan fitur dari data <i>training</i> berdasarkan <i>type</i>, jika <i>type</i> sama dengan sarkasme, maka akan menambahkan fitur sarkasme, dan sebaliknya.</p>
6	save_feature	<ul style="list-style-type: none"> - Int - Array - Array 	<ul style="list-style-type: none"> - n_classify - temp_train - temp_test 	-	<p><i>Method</i> ini digunakan untuk menyimpan hasil fitur ekstraksi data <i>training</i> dan labelnya ke dalam file .txt.</p>

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
7	get_train_feature	<ul style="list-style-type: none"> - Int - Array - Array - Array - Array 	<ul style="list-style-type: none"> - n_classify - data_train - label_train - data_test - label_test 	<ul style="list-style-type: none"> Array, Array, Array, Array 	<i>Method</i> ini digunakan untuk memisahkan dan mengekstraksi fitur dari data yang akan digunakan menjadi dua, yaitu data untuk fitur ekstraksi non-sarkasme, dan data untuk fitur ekstraksi sarkasme.
8	get_test_feature	<ul style="list-style-type: none"> - Int - Array - Array 	<ul style="list-style-type: none"> - n_classify - data_test - label_test 	<ul style="list-style-type: none"> Array, Array 	<i>Method</i> ini digunakan untuk menghapus data sarkasme pada klasifikasi tiga kelas, dan menghapus data negatif dan netral pada klasifikasi 1 kelas.

4.2.4 *Class FeatureExtraction*

Class FeatureExtraction merupakan *class* yang digunakan untuk mengekstraksi atau mengambil nilai dari sebuah *tweet* yang akan digunakan sebagai fitur. Berikut ini adalah *method* pada *class FeatureExtraction*:

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.4 Daftar *Method* pada *Class FeatureExtraction*

Variabel:		Variabel:	
Preprocessing (Class Lib)	preprocess	Int	max_qout
SentimentExtraction (Class)	sentiments	Int	max_cap
Tagger (Class Lib)	tagger	Int	max_excl
Array (String)	interjection_dic	Int	max_quest

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	unigram	- Array - String - Array	- features - tweet - IDF	Array	<i>Method</i> ini digunakan untuk menghitung kemunculan kata pada teks, dan mengembalikan nilai fitur kata yang sudah dihitung menggunakan TF-IDF
2	tagging	- String	- tweet	String	<i>Method</i> ini digunakan untuk melakukan <i>tagging</i> terhadap setiap kata, sebagai contoh "mementingkan/VBT"

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
3	part_of_speech	- Array - String	- features - tweet	Array	Method ini digunakan untuk menerima <i>tweet</i> yang sudah diberi <i>tag</i> , dan akan menghitung kemunculan <i>tag</i> seperti kemunculan kata benda, kata sifat, kata kerja, kata keterangan dan kata negasi.
4	sentiment_score	- Array - String	- features - tweet	Array	<i>Method</i> ini digunakan untuk menerima parameter <i>tweet</i> yang sudah diberi <i>tag</i> , kemudian melakukan <i>scoring</i> sentimen terhadap <i>tweet</i> .
5	punctuation_based	- Array - String - String	- features - tweet - type	Array	<i>Method</i> ini digunakan untuk menghitung kemunculan tanda tanya (?), tanda seru (!), tanda petik (") dan tanda petik tunggal (''). Setiap kemunculan akan dibagi jumlah kemunculan terbanyak pada data <i>training</i> .

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
6	capitalization	- Array - String	- features - tweet	Array	<i>Method</i> ini digunakan untuk menghitung kemunculan kata kapital kemudian jumlah kemunculan dibagi jumlah kemunculan terbanyak pada data <i>training</i> .
7	stopword_removal	- String	- tweet	String	<i>Method</i> ini digunakan untuk menghilangkan kata-kata yang dianggap tidak memiliki makna.
8	Topic	- Array - String - Class	- features - tweet - topic_mod	Array	<i>Method</i> ini digunakan untuk menerima <i>class topic</i> yang sudah dilatih pada data <i>training</i> , dan digunakan untuk mendapatkan topik dari <i>tweet</i> .
9	interjection_word	- Array - String	- features - tweet	Array	<i>Method</i> ini digunakan untuk menghitung jumlah kemunculan kata <i>interjection</i> seperti, "wow", "wahh".
10	question_word	- Array - String	- features - tweet	Array	<i>Method</i> ini digunakan untuk memberikan nilai fitur kata tanya sebagai 1 (<i>true</i>) atau 0 (<i>false</i>), jika terdapat kata tanya pada <i>tweet</i> .

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
11	Idf	- Array - Array	- data - label	Array	<i>Method</i> ini digunakan untuk menghitung nilai IDF dari masukan data <i>training</i> .

4.2.5 Class SentimentExtraction

Class SentimentExtraction merupakan *class* yang digunakan untuk mendapatkan nilai sentimen positif dan negatif dari sebuah *tweet*. Berikut ini adalah *method* pada *class* SentimentExtraction:

Tabel 4.5 Daftar *Method* pada *Class* SentimentExtraction

Variabel:	Variabel:		
Defaultdict (Collection)	senti_score	Negation (class)	negasi
Preprocessing (Class)	preprocess		

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	score_sentence	- Array	- tweet	Array [][]	<i>Method</i> ini akan menerima masukan array <i>tweet</i> yang sudah diberi <i>tag</i> , kemudian melakukan iterasi setiap kata, dan memanggil <i>method</i> score_word untuk mendapatkan nilai sentimen positif dan negatif.
2	score_word	- String - String	- word - tag	Array [][]	<i>Method</i> ini akan memberi nilai sentimen positif dan negatif berdasarkan kata dan <i>tag</i> dari kata.

IV. IMPLEMENTASI DAN PENGUJIAN

4.2.6 Class Negation

Class Negation merupakan *class* yang digunakan untuk mengatasi kata negasi yang terdapat pada *tweet*. Berikut ini adalah *method* pada *class Negation*:

Tabel 4.6 Daftar *Method* pada *Class Negation*

Variabel:	
Preprocessing (Class)	preprocess

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	negation_handling	- String - String - Array [][], - Bool	- prev - pprev - score - change	Array [], Bool	<i>Method</i> ini akan digunakan pada <i>sentiment score</i> , untuk melakukan pengecekan kata negasi, jika terdapat kata negasi maka nilai dari kata setelah kata negasi akan ditambah dan dikali dua.

4.2.7 Class Topic

Class Topic merupakan *class* yang digunakan untuk melakukan *topic modelling* yang akan dilatih dari data *training*. Berikut ini adalah *method* pada *class Topic*:

Tabel 4.7 Daftar *Method* pada *Class Topic*

Variabel:		Variabel:	
Int	nbtopic	String	alpha
Preprocessing (Class)	preprocess	LdaModel (Class Lib)	lda
Corpora (Class Lib)	dictionary		

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	Fit	- Array	- tweet	Array	<i>Method</i> ini akan digunakan untuk melakukan <i>topic modelling</i> dengan <i>library</i> LDA dari gensim.
2	transform	- String	- tweet	Array	<i>Method</i> ini akan digunakan untuk menadapatkan topik dari sebuah <i>tweet</i> .

4.2.8 Class SVM

Class SVM merupakan *class* yang digunakan untuk melakukan *training* pada data *training* dan mendapatkan hasil dari klasifikasi pada data *testing*. Berikut ini adalah *method* pada *class SVM*:

Tabel 4.8 Daftar *Method* pada *Class SVM*

Variabel:		Variabel:	
Float	c	Array	alphas
Float	tol	Array	bias
Int	max_passes	FMeasure (class)	f_measure

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	smo	- Array - Array	- features - label	Array, Float	<i>Method</i> ini akan digunakan untuk mendapatkan nilai alpha dan bias, dari data <i>training</i> .

IV. IMPLEMENTASI DAN PENGUJIAN

No	Method	Input		Output	Keterangan
		Tipe	Variabel		
2	error	- Array - Array - Array - Float - Int	- feature - label - alpha - bias - i	Float	<i>Method</i> ini akan digunakan untuk mendapatkan nilai <i>error</i> yang akan digunakan pada method smo.
3	classification	- Array - Float - Array - Array - Array - Array - Array	- alpha - b - feature_train - feature_test - label_train - label_test - model_type	Array, Array, Array, Array	<i>Method</i> ini untuk menghitung hasil prediksi menggunakan persamaan SVM, mengembalikan nilai <i>f-measure</i> , akurasi dari setiap metode klasifikasi yang digunakan, dan hasil prediksi.
4	f_measure_sarkasme	- Array - Array	- value_prediction - label_actual	Float	<i>Method</i> ini digunakan untuk mengubah nilai value_prediction menjadi nilai 1 atau -1, khusus klasifikasi 1 kelas.

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
5	direct_method	- Array - Array	- value_prediction - label_actual	Array, Array, Array	<i>Method</i> ini digunakan untuk melakukan klasifikasi dengan <i>direct method</i> . Keluaran dari <i>method</i> ini adalah <i>f-measure</i> dan prediksi.
6	levelled_method	- Array - Array	- value_prediction - label_actual	Array, Array, Array	<i>Method</i> ini digunakan untuk melakukan klasifikasi dengan <i>levelled method</i> . Keluaran dari <i>method</i> ini adalah <i>f-measure</i> dan prediksi.
7	convert_label	- Array - String	- label - type	Array	<i>Method</i> ini digunakan untuk mengubah label dari data sesuai dengan kelas yang akan melalui proses <i>training</i> .
8	get_data_label	- String	- prediksi	String	<i>Method</i> ini digunakan untuk mendapatkan nilai label seperti semula. Sebagai contoh, jika prediksi "sarkasme" maka akan diubah kembali menjadi 2, untuk melalui proses perhitungan akurasi.

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
9	save_alpha_bias	- Int - Array - Float	- n_classify - alpha - bias	-	<i>Method</i> ini untuk menyimpan nilai alpha dan bias ke dalam file.

4.2.9 Class FMeasure

Class FMeasure merupakan *class* yang digunakan untuk mendapatkan akurasi dari klasifikasi data *testing*. Berikut ini adalah *method* pada *class* FMeasure:

Tabel 4.9 Daftar *Method* pada *Class* FMeasure

Variabel:		Variabel:	
Float	c	Array	alphas
Float	tol	Array	bias
Int	max_passes	FMeasure (class)	f_measure

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	confusion_matrix	- Array - Array	- label_actual - label_prediction	Array, Array, Array	<i>Method</i> ini digunakan untuk mendapatkan nilai <i>True Positive</i> , <i>False Positive</i> dan <i>False Negative</i> dari label prediksi dan label sebenarnya.

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
2	f_measure_all			Array	<i>Method</i> ini digunakan untuk menghitung f_measure dari setiap kelas yang ada.
3	precision_recall			Float	<i>Method</i> ini untuk mendapatkan nilai precision dan recall berdasarkan true positive, false positive dan false negative yang sudah didapatkan.
4	accuracy			Float	<i>Method</i> ini digunakan untuk mendapatkan f-measure berdasarkan nilai precision dan recall.

4.2.10 Class Learning

Class Learning merupakan *class* yang digunakan untuk melakukan pemodelan pada data *training* dan klasifikasi pada data *testing* dimulai dari pengambil data menggunakan *class* Loader, melakukan *preprocessing* menggunakan *class* Preprocessing, mendapatkan *feature set* menggunakan *class* Feature hingga melakukan klasifikasi menggunakan *class* SVM. Berikut ini adalah *method* pada *class* Learning:

Tabel 4.10 Daftar *Method* pada *Class FMeasure*

Variabel:	Variabel:		
Loader (Class)	load	Preprocessing (Class)	preprocess
Features (Class)	fitur	SVM (Class)	svm_classifier

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	get_model_type	- Int	- n_classify	Array	<i>Method</i> ini akan digunakan untuk mengembalikan jenis klasifikasi yang akan dilakukan, jika n_classify=0, maka akan dilakukan klasifikasi 4 kelas yaitu positif, negatif, netral dan sarkasme. Jika n_classify=1, akan melakukan klasifikasi tanpa sarkasme. Jika n_classify=2, akan melakukan klasifikasi sarkasme atau non-sarkasme.
2	model	- String	- file name	-	<i>Method</i> ini berguna untuk melakukan <i>learning</i> terhadap data <i>training</i> dan melakukan klasifikasi terhadap data <i>testing</i> .

4.2.11 Class Main

Class Main merupakan *class* yang digunakan untuk menjalankan sistem berupa *website* pada *localhost*. Berikut ini adalah *method* pada *class Main*:

Tabel 4.11 Daftar *Method* pada *Class FMeasure*

Variabel:	
Learning (Class)	learn

IV. IMPLEMENTASI DAN PENGUJIAN

No	<i>Method</i>	<i>Input</i>		<i>Output</i>	Keterangan
		Tipe	Variabel		
1	classify	- String - teks		String	<i>Method</i> ini untuk mengembalikan hasil klasifikasi dari teks masukan kepada UI.
2	upload	-	-	Array, Array, Array, Array, Array, Array, Array	<i>Method</i> ini digunakan untuk menerima <i>file</i> yang diisi pada <i>form</i> masukan <i>file</i> pada website, kemudian melakukan <i>training</i> pada <i>file</i> tersebut. Hasil akurasi dan data yang digunakan untuk <i>training</i> dan <i>testing</i> akan ditampilkan pada tampilan.
3	home	-	-	-	<i>Method</i> ini untuk melakukan <i>redirect</i> ke halaman awal website.

4.3 Implementasi Perangkat Lunak

Pada bagian ini akan dijelaskan mengenai implementasi aplikasi analisis sentimen dimulai dari pengambilan data, *text preprocessing*, *feature extraction*, hingga klasifikasi SVM dengan SMO.

4.3.1 Implementasi Pengambilan Data

Pada bagian ini akan dilakukan proses pengambilan data dari *file* ekstensi .txt. Berikut ini adalah proses yang akan dilakukan dalam pengambilan data:

1. Memilih data *file* dengan ekstensi .txt yang berisi data dan labelnya yang dipisahkan dengan *tab* (\t)
2. Membuka *file* dengan *method* pada Python, yaitu `open(filename)`, kemudian memisahkan data dan label berdasarkan *tab* (\t) dengan *method* pada Python `reader(file, delimiter = "\t")`.
3. Melakukan *looping* atau pengulangan pada data yang sudah dipisahkan berdasarkan *tab*, dan menyimpan data beserta labelnya ke dalam variabel dengan tipe array.

4.3.2 Implementasi *Text Preprocessing*

Sebelum melakukan pengambilan data beserta labelnya, setiap data akan melalui proses *text preprocessing* dengan melakukan iterasi atau pengulangan pada data yang sudah disimpan pada array. *Text preprocessing* yang dilakukan secara umum adalah *remove hashtag*, *URL*, *mention*, *remove punctuation*, *tokenize*, *stopword removal*, yang lainnya akan dijalankan saat penambahan fitur. Sebagai contoh fitur *part of speech* hanya akan melakukan *case folding*. Berikut ini merupakan proses yang akan dilakukan dalam *text preprocessing*:

1. *Case Folding*

Mengubah setiap huruf pada *tweet* menjadi huruf kecil dengan *method* variabel.lower().

2. *Remove Hashtag, URL, Mention*

Menghapus semua *hashtag*, *URL*, *mention* pada *tweet* dengan *method* pada *library* Tweet-preprocessor, yaitu clean(*tweet*).

3. *Remove Punctuation*

- 1 Menginisialisasikan tanda baca yang ada ke sebuah variabel dengan tipe array dengan *method* pada Python, yaitu set(string.punctuation)
- 2 Menghapus semua tanda baca kecuali tanda baca tanya (?), tanda seru (!), tanda petik(“, ’) dan tanda pemisah (-) dengan melakukan iterasi setiap *character* pada *tweet*, dan menghapus kemunculan tanda baca yang terdapat pada variabel tanda baca pada tahap 1.

4. *Tokenization*

Melakukan *tokenize* pada *tweet* dengan *method* pada *library* NLTK, yaitu word_tokenize(*tweet*).

5. *Misuse of Word*

Menggabungkan setiap huruf yang sama dan bersebelahan, dengan melakukan iterasi setiap token yang dihasilkan pada tahap *tokenization*, dengan *method* itertools.groupby(token).

6. *Abbreviation Word*

- 1 Menginisialisasi kamus kata singkatan yang dibuat secara manual pada *file* ekstensi .txt ke dalam variabel *abbreviation.dic* tipe array.
- 2 Melakukan iterasi setiap token dan menggantinya dengan persamaannya jika terdapat pada variabel kamus kata singkatan.

IV. IMPLEMENTASI DAN PENGUJIAN

7. Stopword Removal

- 1 Menginisialisasi kamus kata *stopword* pada *file* ekstensi .txt ke dalam variabel *stopword* tipe array.
- 2 Melakukan iterasi setiap token dan menghapus token tersebut jika terdapat pada variabel *stopword*.

8. Stemming

Mengubah kata menjadi bentuk kata dasarnya dengan *method* pada *library* sastrawi, yaitu *stem(tweet)*.

4.3.3 Implementasi *Feature Extraction*

Setelah *text preprocessing*, akan dilakukan *feature extraction* untuk mendapatkan nilai fitur dari setiap teks. Berikut ini merupakan proses yang akan dilakukan dalam *feature extraction*:

1. Unigram

- 1 Melakukan *case folding* terhadap *tweet*, *stemming*, *misuse of word*, *tokenize*.
- 2 Menghitung jumlah kemunculan kata pada *tweet* dan disimpan ke dalam variabel tipe object ({}).
- 3 Menghitung nilai TF setiap kata dan dikalikan dengan IDF katanya.
- 4 Hasil fitur akan disimpan dalam variabel *features* tipe object ({}).

2. Part of Speech

- 1 Melakukan *case folding* terhadap *tweet*.
- 2 Melakukan *tagging* terhadap *tweet* dengan *method* pada *library* IPosTagger, yaitu *taggingStr(tweet)*
- 3 Hasil *tweet* yang sudah diberi *tag* akan dilakukan perhitungan kemunculan setiap *tag*.
- 4 Hasil fitur akan disimpan dalam variabel *features* tipe object ({}).

3. Sentiment Score

- 1 Menginisialisasi variabel *senti_score* dengan tipe *defaultdict* untuk menyimpan kata, *tag* beserta nilai sentimen positif dan negatifnya.
- 2 Melakukan *case folding* terhadap *tweet*.
- 3 Melakukan *tagging* terhadap *tweet* dengan *method* pada *library* IPosTagger, yaitu *taggingStr(tweet)*
- 4 Hasil *tweet* yang sudah diberi *tag* akan dicek ke dalam variabel *senti_score* untuk mendapatkan nilai sentimennya.
- 5 Hasil fitur akan disimpan dalam variabel *features* tipe object ({}).

IV. IMPLEMENTASI DAN PENGUJIAN

4. *Punctuation Based*

- 1 Menghitung kemunculan tanda baca tanya (?), tanda seru (!) dan tanda petik (', '') dengan *method* pada Python, yaitu `count(tanda_baca)`.
- 2 Hasil fitur akan disimpan dalam variabel features tipe object ({}).

5. *Capitalization*

- 1 Menghitung kemunculan kata kapital, dan membagi jumlah kemunculan kapital dengan kemunculan maksimal kata kapital pada data *training*.
- 2 Hasil fitur akan disimpan dalam variabel features tipe object ({}).

6. *Topic*

- 1 Melakukan *training topic modelling* LDA dengan *method* pada *library gensim*, yaitu `LdaModel(corpus, dictionary, jumlah_topik, alpha)`. Menyimpan model dan *dictionary* kedalam *file pickle* dengan *method save(filename)*. Model dan *dictionary* disimpan ke dalam variabel `lda` dan `dictionary`.
- 2 Setelah mendapatkan model *topic*, lakukan *case_folding, stemming, misuse of word, tokenize* pada *tweet*.
- 3 Melakukan perhitungan kemunculan kata dengan *method* pada *gensim*, yaitu `doc2bow(token)` dan menyimpan hasilnya ke dalam variabel `corpus_sentence`.
- 4 Mendapatkan probabilitas *tweet* terhadap topik yang ada dengan memanggil model LDA, yaitu `lda[corpus_sentence]`.
- 5 Hasil fitur akan disimpan dalam variabel features tipe object ({}).

7. *Interjection*

- 1 Melakukan *case folding, misuse of word, tokenize*.
- 2 Melakukan iterasi dan menghitung kemunculan kata interjeksi
- 3 Hasil fitur akan disimpan dalam variabel features tipe object ({}).

8. *Question Word*

- 1 Melakukan *case folding, misuse of word, tokenize*.
- 2 Melakukan iterasi dan menghitung kemunculan kata tanya
- 3 Hasil fitur akan disimpan dalam variabel features tipe object ({}).

9. *TF-IDF*

- 1 Melakukan *case_folding*, menghapus semua kata negasi pada variabel array, dan kata tanya pada variabel array. Setelah itu *stemming, misuse of word, tokenize*.
- 2 Menghitung kemunculan kata pada sejumlah data *training*.
- 3 Menghitung nilai IDF yang disimpan pada variabel tipe `defaultdict`.

Semua hasil fitur ekstraksi pada data *training* akan disimpan ke dalam file *fitur_training.pickle* dengan *method* pada *library numpy*, yaitu *numpy.save(filename)*.

4.3.4 Implementasi SVM dengan SMO

Setelah melalui proses *feature extraction*, selanjutnya melakukan proses *learning* dengan *Simplified Sequential Minimal Optimization* (SMO) dan klasifikasi menggunakan SVM linear. Berikut ini merupakan proses yang dilakukan dalam *learning* dan klasifikasi:

1. Menggunakan hasil fitur ekstraksi data *training* untuk mencari nilai alpha dan bias.
2. Mengubah setiap label menjadi +1 atau -1 sesuai dengan model yang ingin dibuat, yaitu *label_train['all']* untuk label model positif, negatif, netral, dan *label_train['sar']* untuk label model sarkasme.
3. Menghitung nilai alpha dan bias untuk setiap model menggunakan SMO. Hasil dari perhitungan alpha dan bias akan disimpan kedalam *file pickle* dengan metode *numpy.save(filename)*.
4. Pada tahap testing akan melakukan tahap 1 sampai dengan 2, kemudian menghitung nilai $f(x)$ menggunakan alpha dan bias pada tahap 3 berdasarkan masing-masing model.
5. Pengklasifikasian pada data *testing* akan dilakukan dengan 2 macam cara klasifikasi, yaitu *direct method* dan *levelled method*.
6. Pengecekan pada *direct method* akan melakukan perhitungan $f(x)$ pada model positif, negatif, netral dan sarkasme. Kemudian hasil dari model positif, negatif, netral akan dicari nilai tertingginya. Jika nilai yang didapatkan adalah model positif, maka akan melakukan pengecekan model sarkasme. Jika nilai $\text{sign}(f(x)) \geq 1$ maka *tweet* akan diklasifikasikan sebagai sarkasme, jika bukan maka *tweet* akan diklasifikasikan sebagai positif.
7. Pengecekan pada *levelled method* akan melakukan perhitungan $f(x)$ pada model netral. Jika $\text{sign}(f(x)) \geq 1$ maka akan diklasifikasikan *tweet* sebagai netral, sebaliknya termasuk kelas opini. Jika termasuk kelas opini, maka akan melakukan pengecekan pada model positif dan negatif. Perhitungan nilai $\text{sign}(f(x))$ pada model positif dan negatif akan diambil nilai tertinggi. Jika nilai tertinggi adalah pada model positif, maka akan dilakukan pengecekan pada

model sarkasme. Jika $\text{sign}(f(x)) \geq 1$ maka akan termasuk kelas sarkasme, sebaliknya kelas positif.

4.4 Pengujian

Pada bab ini, akan dilakukan berbagai pengujian untuk menentukan kombinasi fitur terbaik, perbandingan tahap *preprocessing*, menentukan parameter regularisasi SMO terbaik, dan menentukan metode klasifikasi terbaik untuk analisis sentimen.

4.4.1 Pengujian Kombinasi Fitur

Pada bagian ini akan dilakukan pengujian kombinasi fitur untuk mendapatkan kombinasi fitur terbaik. Kombinasi fitur yang akan dilakukan pada pengujian ini ada dua, yaitu KF1 dan KF 2. KF1 adalah kombinasi fitur 1 pada kelas positif, negatif dan netral yang terdiri dari fitur *unigram*, *sentiment score*, *question word*, dan *punctuation based*. KF 2 adalah kombinasi fitur 2 pada kelas sarkasme yang terdiri dari fitur *unigram*, *sentiment score*, *topic*, *part of speech*, *punctuation based*, *interjection word* dan *capitalization*.

Berikut ini adalah singkatan fitur yang akan digunakan pada tabel pengujian:

1. U adalah *Unigram*
2. SS adalah *Sentiment Score*
3. QW adalah *Question Word*
4. PB adalah *Punctuation Based*
5. T adalah *Topic*
6. POS adalah *Part of Speech*
7. IW adalah *Interjection Word*
8. C adalah *Capitalization*.

Berikut ini adalah singkatan metode klasifikasi pada tabel pengujian:

1. DM adalah *Direct Method*
2. LM adalah *Levelled Method*

Berikut ini adalah singkatan dari setiap *f-measure* pada tabel pengujian:

IV. IMPLEMENTASI DAN PENGUJIAN

1. FP_L/FP_D adalah *f-measure* positif *Levelled Method* dan *f-measure* positif *Direct Method*.
2. FG_L/FG_D adalah *f-measure* negatif *Levelled Method* dan *f-measure* negatif *Direct Method*.
3. FT_L/FT_D adalah *f-measure* netral *Levelled Method* dan *f-measure* netral *Direct Method*.
4. FS_L/FS_D adalah *f-measure* sarkasme *Levelled Method* dan *f-measure* sarkasme *Direct Method*.

Pengujian dilakukan dengan data *training* 177, data *testing* 59, dengan parameter C=5, tol=0.001, dan max_passes=5. Pengujian ini akan mencari kombinasi fitur terbaik untuk klasifikasi 4 kelas (positif, negatif, netral, sarkasme) dan 3 kelas (positif, negatif, netral). Berikut ini adalah tabel pengujian pada kombinasi fitur 1 pada klasifikasi 4 kelas dan 3 kelas:

Tabel 4.12 Tabel Pengujian Kombinasi Fitur 1

No / Kelas	Kombinasi fitur non-sarkasme	Kombinasi fitur sarkasme	Akurasi <i>F-Measure</i>				
			FP_L/ FP_D	FN_L/ FN_D	FT_L/ FT_D	FS_L/ FS_D	LM/ DM
1/4	KF1	KF2	0.72 / 0.71	0.71 / 0.69	0.82 / 0.83	0.24 / 0.25	0.62 / 0.62
1/3	KF1	-	0.88 / 0.81	0.75 / 0.75	0.88 / 0.83	-	0.84 / 0.80
2/4	U + PB + QW	KF2	0.61 / 0.63	0.55 / 0.53	0.73 / 0.79	0.13 / 0.13	0.51 / 0.52
2/3	U + PB + QW	-	0.63 / 0.62	0.63 / 0.59	0.91 / 0.81	-	0.72 / 0.67
3/4	U + SS + QW	KF2	0.71 / 0.71	0.65 / 0.65	0.74 / 0.74	0.27 / 0.27	0.59 / 0.59
3/3	U + SS + QW	-	0.79 / 0.71	0.71 / 0.71	0.78 / 0.74	-	0.76 / 0.72
4/4	U + SS + PB	KF2	0.70 / 0.74	0.73 / 0.72	0.75 / 0.83	0.25 / 0.29	0.61 / 0.65

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.12 Tabel Pengujian Kombinasi Fitur 1

No / Kelas	Kombinasi fitur non-sarkasme	Kombinasi fitur sarkasme	Akurasi F-Measure				
			FP_L/ FP_D	FN_L/ FN_D	FT_L/ FT_D	FS_L/ FS_D	LM/ DM
4/3	U + SS + PB	-	0.82 / 0.79	0.81 / 0.77	0.88 / 0.83	-	0.84 / 0.80
5/4	U + SS	KF2	0.63 / 0.63	0.74 / 0.79	0.65 / 0.72	0.29 / 0.29	0.58 / 0.61
5/3	U + SS	-	0.67 / 0.67	0.74 / 0.76	0.69 / 0.72	-	0.70 / 0.72
6/4	U + PB	KF2	0.69 / 0.69	0.63 / 0.65	0.76 / 0.78	0.24 / 0.13	0.58 / 0.56
6/3	U + PB	-	0.58 / 0.6	0.67 / 0.67	0.85 / 0.88	-	0.70 / 0.72
7/4	U + QW	KF2	0.72 / 0.72	0.53 / 0.57	0.74 / 0.70	0.14 / 0.14	0.53 / 0.53
7/3	U + QW	-	0.65 / 0.60	0.59 / 0.61	0.80 / 0.76	-	0.68 / 0.66
8/4	SS + PB + QW	KF2	0.63 / 0.63	0.44 / 0.44	0.81 / 0.81	0.19 / 0.19	0.52 / 0.52
8/3	SS + PB + QW	-	0.63 / 0.63	0.63 / 0.59	0.87 / 0.84	-	0.71 / 0.69

Kolom "No/Kelas" pada pengujian tabel di atas, menunjukkan nomor urut pengujian dan klasifikasi kelasnya. Sebagai contoh, 1/4 adalah nomor urut pengujian ke-1 dengan klasifikasi 4 kelas (positif, negatif, netral, sarkasme), dan 1/3 adalah nomor urut pengujian ke-1 dengan klasifikasi 3 kelas (positif, negatif, netral). Berdasarkan pengujian kombinasi fitur di atas, kombinasi fitur terbaik yang dihasilkan pada fitur non-sarkasme klasifikasi 4 kelas dan 3 kelas adalah *unigram*, *punctuation based* dan *sentiment score*. Pengujian kombinasi fitur 1 untuk klasifikasi 3 kelas tidak akan dilakukan lagi, karena kombinasi fitur 2 tidak akan digunakan untuk klasifikasi 3 kelas. Setelah mendapatkan kombinasi fitur

IV. IMPLEMENTASI DAN PENGUJIAN

terbaik pada pengujian tabel 4.11, maka selanjutnya mencari kombinasi fitur terbaik pada kombinasi fitur 2 dengan percobaan menggunakan hasil terbaik dari kombinasi fitur 1. Berikut ini adalah tabel pengujian pada kombinasi fitur 2 pada klasifikasi 4 kelas:

Tabel 4.13 Tabel Pengujian Kombinasi Fitur 2

No	Kombinasi fitur non-sarkasme	Kombinasi fitur sarkasme	Akurasi F-Measure				
			FP_L/ FP_D	FN_L/ FN_D	FT_L/ FT_D	FS_L/ FS_D	LM/ DM
1	U + SS + PB	KF2	0.70 / 0.74	0.73 / 0.72	0.75 / 0.83	0.25 / 0.29	0.61 / 0.65
2	U + SS + PB	T + IW + C + POS + SS + PB	0.69 / 0.71	0.73 / 0.75	0.75 / 0.83	0.22 / 0.25	0.60 / 0.64
3	U + SS + PB	U + IW + C + POS + SS + PB	0.76 / 0.78	0.73 / 0.75	0.75 / 0.83	0.25 / 0.29	0.62 / 0.66
4	U + SS + PB	U + T + IW + C + SS + PB	0.70 / 0.74	0.73 / 0.75	0.75 / 0.83	0.38 / 0.40	0.64 / 0.68
5	U + SS + PB	U + T + IW + C + PB	0.72 / 0.73	0.73 / 0.75	0.75 / 0.83	0.47 / 0.47	0.67 / 70
6	U + SS + PB	U + T + IW + C	0.75 / 0.76	0.71 / 0.71	0.79 / 0.83	0.57 / 0.57	0.70 / 0.72
7	U + SS + PB	U + T + C	0.75 / 0.76	0.71 / 0.71	0.79 / 0.83	0.57 / 0.57	0.70 / 0.72
8	U + SS + PB	U + T	0.78 / 0.77	0.71 / 0.71	0.79 / 0.83	0.46 / 0.5	0.69 / 0.70
9	U + SS + PB	U + C	0.75 / 0.76	0.71 / 0.71	0.79 / 0.83	0.57 / 0.57	0.70 / 0.72
10	U + SS + PB	U	0.74 / 0.75	0.71 / 0.71	0.79 / 0.83	0.36 / 0.36	0.65 / 0.66

Berdasarkan pengujian di atas, kombinasi fitur 2 yang terbaik adalah *unigram, topic, dan capitalization*, dengan metode klasifikasi terbaik adalah *direct method*.

4.4.1.1 Analisis *Error* Fitur Non-Sarkasme

Pada bagian ini akan dilakukan analisis *error* terhadap setiap fitur yang ada pada fitur non-sarkasme, yaitu *unigram*, *punctuation based* dan *sentiment score*.

1. Fitur *Unigram*

Pada bagian ini akan dilakukan analisis *error* ketika hanya menggunakan fitur *sentiment score* dan *punctuation based*, tanpa fitur *unigram*. Berikut ini adalah analisis *error* pada fitur *unigram*:

Tabel 4.14 Analisis *Error* Fitur *Unigram* pada Klasifikasi Non-Sarkasme

KF	Teks	Fitur	Label	Prediksi
U + PB + SS	sarankan tema mading sekolah	positive_sentiment: 0.031	Netral	Netral
		negative_sentiment: 0.0		
		questionM: 0.0		
		sekolah: 0.13		
		saran: 0.48		
		tema: 0.48		
SS + PB	sarankan tema mading sekolah	positive_sentiment: 0.031	Netral	Positif
		negative_sentiment: 0.0		
		questionM: 0.0		

Berdasarkan hasil percobaan klasifikasi di atas, tanpa fitur *unigram*, teks akan selalu diklasifikasikan berdasarkan nilai sentimen. Hal ini terjadi karena tidak ada fitur kata seperti "saran" dan "tema" yang termasuk kata netral pada data *training*, sehingga klasifikasi tanpa *unigram* bergantung dengan nilai sentimen dari sebuah kalimat. Sebagai contoh teks netral pada data *training* yang menggunakan kata "saran": "aku minta saran dong enaknya jualan apa ya modalnya gak terlalu banyak bisa online+tawarin temen sekolah. makasih sarannya<3".

2. Fitur *Sentiment Score*

Pada bagian ini akan dilakukan analisis *error* ketika hanya menggunakan fitur *unigram* dan *punctuation based*, tanpa fitur *sentiment score*. Berikut ini adalah analisis *error* pada fitur *sentiment score*:

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.15 Analisis Error Fitur *Sentiment Score* pada Klasifikasi Non-Sarkasme

KF	Teks	Fitur	Label	Prediksi
U + PB + SS	ruang sekolah	questionM: 0.0	Netral	Netral
		sekolah: 0.27		
		positive_sentiment: 0.034		
		negative_sentiment: 0.017		
U + PB	ruang sekolah	questionM: 0.0	Netral	Positif
		sekolah: 0.27		

Berdasarkan hasil percobaan klasifikasi di atas, tanpa fitur *sentiment score* teks netral yang pendek dan tidak memiliki tanda baca tanya (?) tidak akan bisa diklasifikasikan ke kelas yang seharusnya, yaitu netral. Hal ini disebabkan oleh klasifikasi yang menjadi bergantung terhadap fitur kemunculan kata atau *unigram*. Pada percobaan di atas, teks netral diprediksi sebagai teks positif, karena cukup banyak kemunculan kata "sekolah" pada data *training* yang termasuk pada teks positif.

3. Fitur *Punctuation Based*

Pada bagian ini akan dilakukan analisis *error* ketika hanya menggunakan fitur *unigram* dan *sentiment score*, tanpa fitur *punctuation based*. Berikut ini adalah analisis *error* pada fitur *punctuation based*:

Tabel 4.16 Analisis Error Fitur *Punctuation Based* pada Klasifikasi Non-Sarkasme

KF	Teks	Fitur	Label	Prediksi
U + PB + SS	Sekolah mu berani terima tantangan ? keberanian mention ig	positive_sentiment: 0.69	Netral	Netral
		negative_sentiment: 0.49		
		questionM: 0.2		
		sekolah: 0.069		
		terima: 0.16		
U + SS	sekolah mu berani terima tantangan ? keberanian mention ig	positive_sentiment: 0.69	Netral	Positif
		negative_sentiment: 0.49		
		sekolah: 0.069		
		terima: 0.16		

Berdasarkan hasil percobaan klasifikasi di atas, tanpa fitur *punctuation based*, teks netral yang memiliki nilai sentimen tidak akan bisa diklasifikasikan ke

IV. IMPLEMENTASI DAN PENGUJIAN

kelas yang seharusnya, yaitu netral. Hal ini terjadi karena klasifikasi menjadi sangat bergantung terhadap fitur *sentiment score*. Pada percobaan di atas, nilai sentimen positif lebih tinggi dibanding negatif, oleh karena itu teks diklasifikasikan sebagai teks positif.

4.4.1.2 Analisis Error Fitur Sarkasme

Pada bagian ini akan dilakukan analisis *error* terhadap setiap fitur yang ada pada fitur sarkasme, yaitu *unigram*, *topic*, dan *capitalization*.

1. Fitur *Unigram*

Pada bagian ini akan dilakukan analisis *error* ketika hanya menggunakan fitur *topic* dan *capitalization*, tanpa fitur *unigram*. Berikut ini adalah analisis *error* pada fitur *unigram*:

Tabel 4.17 Analisis Error Fitur *Unigram* pada Klasifikasi Sarkasme

KF	Teks	Fitur	Label	Prediksi
U + T + C	@Budhiheriawan @richard_errik @whytas @riri_hesria @Ria1Kartolo @AdieRinaldi @fuadpuad makin pinter aja anggota DPR #sarkasme	topic 0: 0.055	Sarkasme	Sarkasme
		topic 1: 0.059		
		topic 2: 0.036		
		topic 3: 0.84		
		capitalization: 0.14		
		dpr: 0.26		
		pintar: 0.74		
		angota: 0.44		
T + C	@Budhiheriawan @richard_errik @whytas @riri_hesria @Ria1Kartolo @AdieRinaldi @fuadpuad makin pinter aja anggota DPR #sarkasme	topic 0: 0.055	Sarkasme	Positif
		topic 1: 0.059		
		topic 2: 0.036		
		topic 3: 0.84		
		capitalization: 0.14		

Berdasarkan hasil percobaan klasifikasi di atas, tanpa fitur *unigram*, klasifikasi akan salah, karena fitur yang digunakan tidak memberikan informasi yang cukup untuk mengklasifikasikan sarkasme.

2. Fitur *Topic*

Pada bagian ini akan dilakukan analisis *error* ketika hanya menggunakan fitur *unigram*, tanpa fitur *topic*. Fitur *capitalization* tidak digunakan untuk analisis

IV. IMPLEMENTASI DAN PENGUJIAN

error karena pada pengujian fitur tabel 4.13 menunjukkan tidak ada perubahan akurasi saat menggunakan semua fitur terbaik sarkasme, yaitu *unigram*, *capitalization*, *topic* dan tanpa fitur *topic*. Berikut ini adalah analisis fitur *topic* pada data *training* dengan label sarkasme:

Tabel 4.18 Analisis Fitur *Topic*

No	Teks	Topic 0	Topic 1	Topic 2	Topic 3
1	Dan besok sekolah, hell yeah ! :D #sarkasme	0.88	0.046	0.028	0.038
2	BPK terbaik. BI terbaik. DPR.....ter..? @Fahrihamzah ... ter...? Mandi dulu Ooom!	0.93	0.025	0.015	0.020
3	astaga, besar sekali hati internet provider satu ini, isinya minta maaf trus #sarkasme	0.92	0.032	0.019	0.027
4	Wow, abis ditinggal sesiangan, jam 8 malam internet @innovateIND @HOMElinksBSD udah nyala lagi #sarkasme	0.93	0.024	0.015	0.020
5	Pantesan namanya dia & papa bs hilang dari list anggota DPR kasus e-KTP Ternyata mahluk paling suci sebagai gedung DPR	0.016	0.95	0.010	0.014
6	Oh @fadlizon betapa beruntungnya DPR RI punya sampeyan. #sarkasme #ILC	0.036	0.038	0.023	0.90
7	Nilai-nilai gue angkanya gede-gede bgt ya wow #sarkasme #edisidepresi	0.036	0.040	0.023	0.89
8	hebat fadlizon ketemu sahabat lama sepaham dan seidiologi dari rusia	0.027	0.028	0.017	0.92

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.18 Analisis Fitur *Topic*

No	Teks	Topic 0	Topic 1	Topic 2	Topic 3
9	jangan salahkan film horor lokal, mereka hanya menuruti KEMAUAN dan KEMALUAN pasar #sarkasme	0.023	0.93	0.015	0.020
10	Owh internet connection-nya keren sekali, saat dibutuhkan sangat bisa diandalkan #sarkasme #sinis #speechless	0.90	0.038	0.023	0.031
11	Hari ini ga ketemu @gistaanindy ckck sekolah segede apa sih? #sarkasme	0.91	0.033	0.020	0.027
12	TERIMA KASIH INTERNET	0.85	0.058	0.036	0.049
13	Salam buat Ketua Dewan-nya ya. Dah sehatkah...?	0.043	0.88	0.028	0.038
14	Semoga FH dan FZ ada di DPR selamanya.. krn sepi dunia kalau gak ada mereka.. :)	0.021	0.94	0.013	0.018
15	film berkualitas "Mr Bean Kesurupan DEPE" jadi bangga dengan negeri ini,, #sarkasme	0.93	0.025	0.015	0.020
16	Sepertinya sudah masuk jam2nya nih, internet di sini luar biasa!!!!!!#sarkasme	0.05	0.85	0.05	0.049
17	Lg istirahat nih di rumah habis sekolah dari pagi sampe sore...!!!	0.92	0.032	0.019	0.027
18	Anggota DPR Pancasilais. Mereka amalkan sila ke-4. Musyawarah dan mufakat untuk membagi2 jarak atas negeri ini. #Sarkasme	0.014	0.018	0.94	0.015
19	Yah ga bisa liat sekolah!!!	0.88	0.047	0.028	0.039

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.18 Analisis Fitur *Topic*

No	Teks	Topic 0	Topic 1	Topic 2	Topic 3
20	besok sekolah. Hatiku gembiraaa! #sarkasme	0.88	0.046	0.028	0.038
21	masuk sekolah/.	0.077	0.081	0.049	0.79
22	RT @fnabila: Asiiik bentar lagi sekolah!!!!!! Ga sabar!!!! #sarkasme	0.036	0.90	0.023	0.032
23	Super sekali internet smartfren utk streaming...muter terusss...sampe ga keliatan... Waakakakkakak #sarkasme	0.024	0.93	0.015	0.020
24	Anda lebih hebat lagi pak @Fahrihamzah tidak punya partai bisa jadi anggota dpr.	0.036	0.038	0.023	0.90
25	Minggu....les : besok...sekolah:— oke,gw suka belajar. Sukaa bgt. #sarkasme	0.021	0.94	0.018	0.013
26	Entar sist kalo anggarannya UDAH NAIK.	0.80	0.080	0.049	0.066
Total		12	8	1	5

Berdasarkan analisis fitur *topic* pada 26 data *training* dengan label sarkasme, dapat disimpulkan setiap teks sarkasme cenderung termasuk pada topik 0, 1 dan 3, dari jumlah topik yang ada adalah 4. Hasil analisis fitur *topic* di atas akan membantu dalam analisis *error* pada fitur *topic* yang akan dilakukan. Berikut ini analisis *error* fitur *topic*:

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.19 Analisis *Error* Fitur *Topic* pada Klasifikasi Sarkasme

KF	Teks	Fitur	Label	Prediksi
U + T	Berkaitan akses situs porno anggota dewan Badan Kehormatan DPR adakan investigasi siapa ditonton	topic 0: 0.044	Sarkasme	Sarkasme
		topic 1: 0.88		
		topic 2: 0.02		
		topic 3: 0.03		
		dewan: 0.16		
		angota: 0.11		
		tonton: 0.18		
U	Berkaitan akses situs porno anggota dewan Badan Kehormatan DPR adakan investigasi siapa ditonton	dpr: 0.06	Sarkasme	Positif
		dewan: 0.16		
		angota: 0.11		
		tonton: 0.18		

Berdasarkan hasil percobaan klasifikasi di atas, tanpa fitur *topic*, klasifikasi akan salah, jika terdapat kata yang kemunculannya sering terdapat pada kelas selain kelas sebenarnya. Pada kasus penggunaan di atas kata "tonton" menjadi fitur yang paling menentukan kelas dari sebuah teks, sehingga ketika fitur *topic* tidak digunakan klasifikasi sarkasme menjadi salah. Sebagai contoh teks positif pada data *training* yang menggunakan kata "tonton", yaitu "Jadi support film Indonesia, tapi juga harus jeli apa yang kita tonton ... hmm, aulion bener nih."

3. Fitur *Capitalization*

Pada bagian ini akan dilakukan analisis *error* ketika hanya menggunakan fitur *unigram*, *topic*, dan *sentiment score*, tanpa fitur *capitalization*. Berikut ini adalah analisis *error* fitur *capitalization*:

IV. IMPLEMENTASI DAN PENGUJIAN

Tabel 4.20 Analisis *Error* Fitur *Capitalization* pada Klasifikasi Sarkasme

KF	Teks	Fitur	Label	Prediksi
U + T + C	Hampir 11 jam dekat sekolah I LOVE MY SENIOR LIFE PEEPS!	topic 0: 0.46	Sarkasme	Sarkasme
		topic 1: 0.060		
		topic 2: 0.036		
		topic 3: 0.043		
		capitalization: 4.0		
		sekolah: 0.079		
		life: 0.32		
		jam: 0.23		
		cinta: 0.27		
U + T	Hampir 11 jam dekat sekolah I LOVE MY SENIOR LIFE PEEPS!	topic 0: 0.46	Sarkasme	Positif
		topic 1: 0.060		
		topic 2: 0.036		
		topic 3: 0.043		
		sekolah: 0.079		
		life: 0.32		
		jam: 0.23		
		cinta: 0.27		

Berdasarkan hasil percobaan klasifikasi di atas, tanpa fitur *capitalization*, klasifikasi akan salah, jika terdapat sebuah teks yang memiliki nilai topik yang kecil. Hal ini terjadi karena nilai topik dari teks hanya 0.46, sedangkan kata "cinta" yang cenderung positif pada data *training*, hal ini yang menyebabkan klasifikasi salah ketika tidak ada fitur *capitalization*.

4.4.2 Pengujian Parameter pada SMO

Pada bagian ini, akan dilakukan pengujian parameter SMO yaitu, C, *tolerance*, dan *max_passes*. Nilai parameter C yang akan digunakan untuk pengujian adalah 1, 5 dan 10. Nilai parameter *max_passes* yang akan digunakan untuk pengujian adalah 5 dan 10. Nilai parameter tol yang akan digunakan untuk pengujian adalah 0.001 dan 0.0001. Berikut ini adalah tabel pengujian parameter SMO:

Tabel 4.21 Pengujian Parameter SMO

Parameter			Akurasi <i>F-Measure</i>				
C	Tol	MaxPasses	FP.D	FN.D	FT.D	FS.D	DM
1	0.001	5	0.65	0.52	0.89	0.18	0.56
1	0.001	10	0.65	0.52	0.89	0.18	0.56
1	0.0001	5	0.65	0.52	0.89	0.18	0.56
1	0.0001	10	0.65	0.52	0.89	0.18	0.56
5	0.001	5	0.76	0.71	0.83	0.57	0.72
5	0.001	10	0.76	0.71	0.83	0.57	0.72
5	0.0001	5	0.76	0.71	0.83	0.57	0.72
5	0.0001	10	0.76	0.71	0.83	0.57	0.72
10	0.001	5	0.74	0.75	0.81	0.57	0.72
10	0.001	10	0.74	0.75	0.81	0.57	0.72
10	0.0001	5	0.74	0.75	0.81	0.57	0.72
10	0.0001	10	0.74	0.75	0.81	0.57	0.72

Berdasarkan pengujian parameter di atas, parameter yang menghasilkan akurasi terbaik adalah parameter C=5 dan C=10. Parameter C=10 memberikan hasil akurasi yang lebih merata pada setiap kelasnya, sedangkan pada parameter C=5, memberikan akurasi kelas positif yang lebih tinggi, namun akurasi kelas negatif menjadi rendah. Oleh karena itu akan dipilih parameter C=10 dengan *tolerance* 0.001 dan *max passes* 5. Alasan pemilihan parameter tersebut adalah semakin tinggi parameter *tolerance* dan *max passes* tidak terjadi adanya perubahan pada akurasi, selain itu alasan pemilihan parameter *tolerance* dan *max passes* yang lebih rendah adalah supaya *training* yang dilakukan menjadi lebih cepat.

4.4.3 Pengujian Klasifikasi 4 kelas, 3 kelas dan 1 kelas

Pada bagian ini akan dilakukan pengujian untuk membandingkan akurasi *f-measure* jika melakukan klasifikasi 4 kelas, yaitu positif, negatif, netral, sarkasme, kemudian mengklasifikasikan 3 kelas, yaitu positif, negatif dan netral, dan yang terakhir mengklasifikasikan 1 kelas sarkasme atau non-sarkasme. Klasifikasi kelas akan menggunakan fitur yang dihasilkan pada pengujian kombinasi fitur yang telah dilakukan, yaitu *unigram*, *punctuation based* dan *sentiment score* untuk klasifikasi non-sarkasme, dan fitur *unigram*, *capitalization*, *topic* dan *sentiment score* untuk

IV. IMPLEMENTASI DAN PENGUJIAN

klasifikasi sarkasme. Berikut ini adalah tabel pengujian klasifikasi 4 kelas:

Tabel 4.22 Pengujian Klasifikasi 4 Kelas (Positif, Negatif, Netral, Sarkasme)

Klasifikasi	<i>F-Measure</i>				
	FP.D	FN.L	FT.D	FS.D	DM
Klasifikasi 4 kelas	0.74	0.75	0.81	0.57	0.72

Tabel 4.23 Pengujian Klasifikasi 3 Kelas (Positif, Negatif, Netral)

Klasifikasi	<i>F-Measure</i>			
	FP.D	FN.L	FT.D	DM
Klasifikasi 3 kelas	0.79	0.81	0.86	0.82

Tabel 4.24 Pengujian Klasifikasi 1 Kelas (Sarkasme/Non-Sarkasme)

Klasifikasi	<i>F-Measure</i>
Klasifikasi 1 kelas (sarkasme/non-sarkasme)	0.75

Berdasarkan hasil pengujian di atas, hasil akurasi meningkat sebanyak 10% ketika hanya mengklasifikasikan 3 kelas. Dan akurasi deteksi sarkasme meningkat sebanyak 18% ketika hanya melakukan deteksi sarkasme atau non-sarkasme.

BAB V

PENUTUP

Bab ini berisi kesimpulan dan saran dari sistem deteksi sarkasme pada analisis sentimen media sosial.

5.1 Kesimpulan

Kesimpulan dari pembuatan sistem analisis sentimen dan pengujian-pengujian yang telah dilakukan adalah sebagai berikut:

1. Kombinasi fitur terbaik untuk model positif, negatif dan netral adalah *unigram*, *punctuation based* dan *sentiment score*. Dan kombinasi fitur terbaik untuk model sarkasme adalah *unigram*, *topic*, dan *capitalization*.
2. Fitur *unigram* sangat berpengaruh dalam klasifikasi teks baik positif, negatif, netral, ataupun sarkasme. Fitur *punctuation based* dapat meningkatkan akurasi klasifikasi teks netral hingga 9%. Fitur *sentiment score* dapat meningkatkan akurasi hingga 7% pada teks positif, dan 10% pada teks negatif.
3. Fitur *topic* dapat meningkatkan akurasi klasifikasi sarkasme hingga 10% saat digabungkan dengan fitur *unigram*. Fitur *capitalization* dapat meningkatkan akurasi klasifikasi sarkasme hingga 7%.
4. Fitur *interjection word* tidak berpengaruh karena kemunculan *interjection word* yang jarang pada data sarkasme yang dilatih. Fitur *part of speech* menurunkan akurasi klasifikasi sarkasme, karena kemunculan kata benda, kata sifat, kata kerja, kata keterangan, dan kata negasi juga banyak terdapat pada teks-teks selain sarkasme. Fitur *sentiment score* menurunkan akurasi klasifikasi sarkasme karena teks sarkasme memiliki nilai sentimen yang selalu berlawanan dari yang terlihat dari teks. Fitur *punctuation based* menurunkan klasifikasi karena tanda baca juga terdapat banyak pada teks selain sarkasme.
5. Fitur *question word* menurunkan akurasi, karena saat digabungkan dengan fitur *punctuation based*, setiap teks positif, negatif atau sarkasme yang memiliki kata tanya dan tanda baca tanya (?) akan otomatis terkласifikasi sebagai teks netral.

6. Kombinasi parameter SMO yang terbaik adalah C=10, tol=0.001, dan max_passes=5. Akurasi klasifikasi *direct method* lebih tinggi dibanding *levelled method*.
7. Kombinasi klasifikasi pada 4 kelas (positif, negatif, netral, sarkasme) menghasilkan akurasi sebesar 72% dan pada kombinasi klasifikasi 3 kelas (positif, negatif, netral) menghasilkan akurasi sebesar 82%. Akurasi pada klasifikasi 3 kelas tanpa data sarkasme meningkat cukup tinggi. Dan pada klasifikasi 1 kelas, yaitu sarkasme dan non-sarkasme menghasilkan 75%.
8. Klasifikasi pada 4 kelas menghasilkan akurasi yang kecil, karena teks sarkasme pada data yang digunakan terdapat teks yang tidak terlihat positif, sehingga terjadi kesalahan klasifikasi. Sebagai contoh teks sarkasme yang lebih terlihat seperti netral, yaitu: "Apa bedanya anggota DPR dengan sebuah baterai? Baterai punya sisi positif. #Sarkasme".

5.2 Saran

Saran dari penulis untuk pengembangan yang dilakukan untuk sistem deteksi sarkasme pada analisis sentimen adalah:

1. Menggunakan data yang lebih banyak dan kategori data yang lebih banyak, sehingga fitur *topic* dapat digunakan untuk memberikan informasi global dari sebuah teks sarkasme dengan lebih baik serta menggunakan jenis kernel RBF untuk klasifikasi.
2. Menggunakan fitur *emoticon* untuk memberikan informasi teks sarkasme.

DAFTAR PUSTAKA

- [1] Ramadhani, R. A. A., Indriani, F., & Nugrahadi, D. T. *Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis. Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis.*
- [2] Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3), 478-514
- [3] Ptáček, T., Habernal, I., & Hong, J. (2014). Sarcasm detection on czech and english Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 213-223).
- [4] Peng, C. C., Lakis, M., & Pan, J. W. Detecting Sarcasm in Text.
- [5] Lunando, E., & Purwarianti, A. (2013, September). *Indonesian social media sentiment analysis with sarcasm detection*. In *Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on* (pp. 195-198). IEEE.
- [6] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*(pp. 56-65). ACM.
- [7] Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & javad Rajabi, M. (2014, September). Advantage and drawback of *Support Vector Machine* functionality. In *Computer, Communications, and Control Technology (I4CT), 2014 International Conference on* (pp. 63-65). IEEE.
- [8] Malouf, R. (2002, August). A comparison of algorithms for *Maximum Entropy* parameter estimation. In *proceedings of the 6th conference on Natural language learning-Volume 20* (pp. 1-7). Association for Computational Linguistics.
- [9] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1), 51-89.
- [10] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.

DAFTAR PUSTAKA

- [11] Baştanlar, Y., & Özysal, M. (2014). Introduction to machine learning. *miRNomics: MicroRNA Biology and Computational Analysis*, 105-128.
- [12] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4), 1-33.
- [13] Wicaksono, A. F., & Purwarianti, A. (2010, August). HMM based part-of-speech tagger for Bahasa Indonesia. In *Fourth International MALINDO Workshop, Jakarta*.
- [14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- [15] Shabtai, A., Moskovitch, R., Elovici, Y., & Glezer, C. (2009). Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey. *information security technical report*, 14(1), 16-29.
- [16] Berwick, R. (2003). An Idiot's guide to Support Vector Machines (SVMs). *Retrieved on October*, 21, 2011.
- [17] Wang, Z., & Xue, X. (2014). Multi-class *Support Vector Machine*. In *Support Vector Machines Applications* (pp. 23-48). Springer International Publishing.
- [18] CS 229, Autumn 2009 “The Simplified SMO Algorithm”.
- [19] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.
- [20] Shamay-Tsoory, S. G., Tomer, R., & Aharon-Peretz, J. (2005). The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology*, 19(3), 288.

LAMPIRAN