Aplikasi Pre-Processing Pembentuk Data Latih Terhadap Review Novel Indonesia Pada Situs Goodreads

Deki Irawan, Diana Ikasari, Widiastuti
Jurusan Sistem Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi
Universitas Gunadarma
Jl. Margonda Raya No. 100, Jawa Barat
deki@student.gunadarma.ac.id

***Abstrak—Goodreads* merupakan situs jaringan sosial yang mengkhususkan pada katalogisasi buku. *Goodreads* mempunyai konten *friend, group,* maupun *discussion*. *Goodreads* memungkinkan anggota untuk menampilkan daftar buku sudah dibaca (*read*), buku yang sedang dibaca (*currently reading*), dan akan dibaca (*to read*). Dalam situs ini, pengguna dapat saling berbagi rekomendasi buku bacaan dengan memberikan *review* maupun komentar. *Review* tersebut dapat membantu penulis dalam menganalisis riset atas opini publik, serta untuk membantu pengguna atau pembaca sebelum membeli buku dengan melihat sentimen *review* yang ada, apakah buku tersebut layak dibeli atau tidak. Membaca komentar *review* secara keseluruhan dapat memakan waktu, namun jika hanya sedikit komentar *review* yang dibaca maka evaluasi akan menjadi bias. Untuk mengatasi masalah tersebut, digunakan klasifikasi analisis sentimen yang digunakan untuk mengelompokkan *review* menjadi opini positif atau opini negatif secara otomatis. Tahap yang akan dilakukan sebelum klasifikasi adalah *text pre-processing* sebagai langkah awal dari *text mining*, proses ini bertujuan untuk mengubah dan menyiapkan data/dokumen ke bentuk yang lebih cocok dan tepat dalam merepresentasikan data untuk digunakan pada proses-proses selanjutnya. Dalam penelitian ini digunakan untuk membentuk data latih pada tahap klasifikasi. Tahapan pada *pre-processing* terdiri dari *case folding, tokenizing, stopword removal,* dan *stemming* menggunakan Algoritma Nazief Adriani. Data *review* yang digunakan dalam sistem ini terdiri dari 330 data latih.**

Kata kunci - Novel; Pre-Processing; Review; Stemming

I. PENDAHULUAN

Menurut Hamzah, opini adalah suatu bagian terpenting dalam pengambilan keputusan untuk suatu kebijakan. Keputusan yang tepat juga dipengaruhi oleh analisis opini dari berbagai sumber yang terkait dengan pengambilan keputusan [2]. Umumnya opini muncul sebagai respon dari suatu kejadian, misalnya *review* yang diberikan *customer* setelah membeli barang dan menggunakan produk tersebut, pengalaman dan kesan yang dirasakan juga menjadi masukkan yang dapat digunakan sebagai *review*. Dengan berkembangnya ketersediaan dan popularitas akan sumber yang kaya opini seperti website review online dan blog pribadi, kesempatan baru dan tantangan muncul semenjak orang-orang sekarang bisa dan menggunakan secara aktif informasi dan teknologi untuk mencari dan memahami opini orang lain. Menurut Yessenov (Yessenov, 2009) ada beberapa contoh website yang bisa mereview produk, seperti Amazon, atau situs review film seperti Rotten Tomatoes yang memungkinkan untuk memberikan rating pada produk, biasanya beberapa skala ditentukan sama dengan review personal yang dibuat [6]. Opini juga diperoleh dari beragam cara yaitu permintaan saran dalam aktivitas penelitian menggunakan angket, atau pesan yang ditinggalkan pada kolom *komentar*, baik pada blog, media sosial, *website* atau forum *online*. Tidak jarang opini sebagai suatu informasi yang begitu penting tidak tersentuh, karena jumlahnya yang sangat banyak dan kurangnya pengetahuan bagaimana cara mengelolanya. Salah satu cabang riset yang kemudian berkembang dari situasi ini adalah analisis sentimen. Cabang ini menjadi riset yang menantang karena di dalamnya terdapat akumulasi dari berbagai tantangan riset, yaitu antara lain *information extraction*, *information summarization*, *document classification* [3].

Selama ini analisis sentimen umumnya mengunakan *rating* yang dinyatakan dalam jumlah bintang. Penggunaan *rating* tidak relevan jika digunakan untuk menilai apakah isi dalam novel bisa diterima dengan baik oleh konsumen atau tidak, hal tersebut didasarkan pada penelitian Santika, Arifin dan Purwitasari (2015), yang menyatakan bahwa *rating* atau penilaian tingkat kepuasan pengguna terhadap suatu produk harus diberikan secara manual oleh penulis opini [5].

Situs *goodreads* adalah salah satu situs jaringan sosial yang mengkhususkan pada katalogisasi buku. *Goodreads* memiliki konten *friend*, *group*, maupun *discussio*n. *Goodreads* memiliki 85 juta pembaca dengan 2,2 miliar buku yang terdata di sana, dan ada 87 juta *review* buku [8]. *Goodreads* memungkinkan anggota untuk menampilkan daftar buku yang sudah dibaca (*read*), buku yang sedang dibaca (*currently reading*), dan buku yang akan dibaca (*to read*). Situs ini merupakan tempat yang tepat bagi pengguna untuk berdiskusi dan memungkinkan pengguna saling memberikan rekomendasi buku favorit [8].

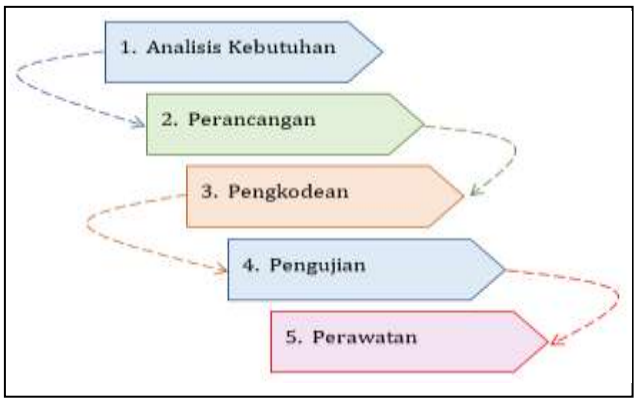
Membaca *review* secara keseluruhan dapat memakan waktu, namun jika hanya sedikit komentar *review* yang dibaca evaluasi akan menjadi bias [7]. Salah satu cara untuk mengatasi

masalah tersebut adalah dengan menggunakan klasifikasi analisis sentimen yang digunakan untuk mengelompokkan review menjadi opini positif atau opini negatif secara otomatis. Tahap yang akan dilakukan sebelum klasifikasi adalah melakuan *input review* dari pengguna dilanjutkan dengan proses teks *pre-processing*. Tahapan pada *pre-processing* terdiri dari *case folding*, *tokenizing*, *stopword removal*, dan *stemming* menggunakan Algoritma Nazief Adriani. Kata dasar digunakan untuk mengetahui makna dari kata, untuk mendapatkan kata dasar dilakukan proses *stemming* yang menjadi bagian dari proses *text pre-processing*. Hal ini yang menjadi tujuan dari penelitian ini, yaitu membentuk data latih hasil dari proses *pre-processing* sebagai sebuah proses untuk bisa dilanjutkan ke proses klasifikasi sentimen analisis yang mendefinisikan kecenderungan *review* bermakna positif atau bermakna negatif [1].

II. METODE

A. System Development Life Cycle

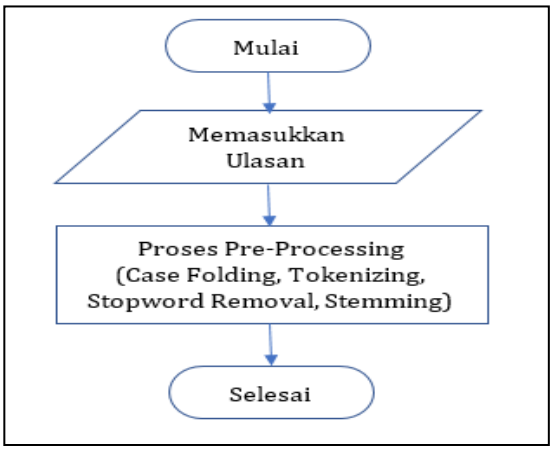
Metode *System Development Life Cycle* (SDLC) dengan model *Waterfall* menurut Roger S. Presman (2001) digunakan sebagai metode dalam penelitian ini [4]. Model ini merupakan pendekatan yang sistematis dan terurut, dimulai dari kegiatan analisis kebutuhan, perancangan, pengkodean, pengujian dan perawatan seperti terlihat pada gambar 1.



Gambar 1. System Development Life Cycle Model Waterfall

B. Analisis Sistem

Analisis sistem pada penelitian ini dimulai dari memasukkan ulasan yang telah diklasifikasikan secara manual ke dalam 2 kelas sentimen yaitu kelas sentiman positif dan kelas sentimen negatif. Tahapan selanjutnya adalah melakukan proses *pre-processing* yang memiliki beberapa kegiatan pemrosesan yaitu *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Data latih yang dihasilkan dari proses Pre-Processing ini dapat digunakan untuk proses sentimen analisis. Kegiatan ini terlihat pada gambar 2.



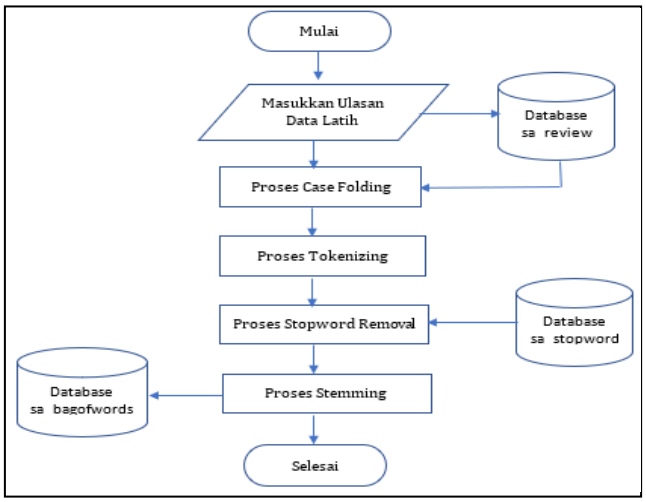
Gambar 2. Analisis Sistem

Kebutuhan data dalam penelitian yang digunakan sebagai data *review* dari novel Indonesia diambil dari situs https://www.goodreads.com. Goodreads memberikan fasillitas kepada *reviewer* untuk memberikan komentar terhadap novel berbahasa Indonesia. Data yang digunakan dalam penelitian ini sebanyak 330 *review*.

Kebutuhan fungsional yang diberikan kepada pengguna adalah, pertama pengguna dapat melakukan proses menambah merubah dan menghapus *review*. Kedua adalah penggunna dapat menambah, merubah dan menghapus kata dasar. Ketiga, bahwa pengguna dapat menambah, merubah dan menghapus kata *stopwords* dan terutama adalah bahwa sistem akan menampilkan hasil dari proses *tokenizing*, *stopword removal*, dan *stemming*.

C. Perancangan Flowchart

Langkah dari proses pembentukan data latih terlihat pada gambar 3. Pemrosesan dimulai dengan memasukkan data latih yang telah diberi sentimen positif dan negatif yang disimpan dalam database sa_review, selanjutnya data ini akan masuk ke dalam proses *case folding*, dan hasilnya akan masuk ke dalam proses *tokenizing*. Hasil dari *tokenizing* masuk ke dalam proses *stopword removal* dengan memeriksa apakah ada kesamaan kata yang terdapat pada database sa_stopword atau tidak. Hasil data *stopword removal* yang tidak sama dengan database sa_stopword akan masuk ke tahap proses *stemming* menggunakan Algoritma Nazief dan Adriani. Pada tahap ini dilakukan pencarian kata dasar dengan menghilangkan imbuhan yang melekat pada kata tersebut untuk disimpan dalam database sa_bagofwords. Proses *stemming* adalah langkah akhir dalam melakukan *pre-processing* untuk mendapatkan data latih.



Gambar 3. Flowchart Pembentukan Data Latih

D. Perancangan Basisdata

Penelitian ini menggunakan 1 database dengan 5 tabel. Struktur dari masing-masing tabel tersebut terlihat sebagai berikut:

1) *Struktur Tabel sa_user*

Tabel sa_user berfungsi untuk menyimpan data *user*, yang terdiri dari id, username, nama dan password. Struktur tabel ini dapat dilihat pada tabel 1.

TABEL 1. STRUKTUR TABEL SA_USER

Kolom	*Tipe*	*Panjang*	*Keterangan*
Id	Int	11	Kunci
Username	Varchar	20	
Nama	Varchar	40	
PAssword	Varchar	40	

2) *Struktur Tabel sa_review*

Tabel sa_review untuk menyimpan data *review* yang sudah diberi sentimen yang terdiri dari id_review, judul_review, sentimen_review, kategori_review, isi_review, dan user_review. Struktur tabel ini dapat dilihat pada tabel 2.

TABEL 2. STRUKTUR TABEL SA_REVIEW

Kolom	*Tipe*	*Panjang*	*Keterangan*
Id_review	Int	11	Kunci
Judul_review	Varchar	100	
Sentimen_review	Enum('POSITIF','NEGATIF')		
Kategori_review	Enum('DATA LATIH','DATA UJI')		
Isi_review	Text		
User_review	Varchar	50	

3) *Struktur Tabel sa_bagofwords*

Tabel sa_bagofwords berfungsi untuk menyimpan data hasil *pre-processing*, yang terdiri dari id_bagofwords, id_review, term_tokenized, term_filtered, term_stemmed. Struktur tabel ini dapat dilihat pada tabel 3.

TABEL 3. STRUKTUR TABEL SA_BAGOFWORDS

Kolom	*Tipe*	*Panjang*	*Keterangan*
Id_bagofwords	Int	11	Kunci
Id_review	Int	11	
Term_tokenized	Text		
Term_filtered	Text		
Term_stemmed	Text		

4) *Struktur Tabel sa_katadasar*

Tabel sa_katadasar berfungsi untuk menyimpan data kata dasar yang terdiri dari id_katadasar, kata_katadasar. Struktur tabel ini dapat dilihat pada tabel 4.

TABEL 4. STRUKTUR TABEL SA_KATADASAR

Kolom	*Tipe*	*Panjang*	*Keterangan*
Id_katadasar	Int	10	Kunci
Kata_katadasar	Varchar	30	

5) *Struktur Tabel sa_stopwords*

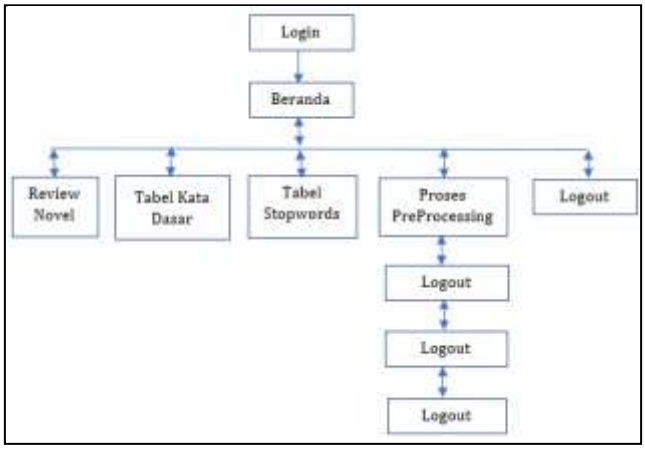
Tabel sa_stopwords berfungsi untuk menyimpan data *stopwords* yang terdiri dari id_stopwords, kata_stopwords. Struktur tabel ini dapat dilihat pada tabel 5.

TABEL 5. STRUKTUR TABEL SA_USER

Kolom	*Tipe*	*Panjang*	*Keterangan*
Id_stopwords	Int	10	Kunci
Kata_stopwords	Varchar	30	

E. Struktur Navigasi Aplikasi

Gambar 4 merupakan struktur navigasi dari aplikasi dalam penelitian ini yang menceritakan mengenai keterhubungan antara satu halaman dengan halaman lain yang berfungsi menghasilkan data keluaran sebagai *pre-processing* dari sentimen analisis novel berbahasa Indonesia. Aplikasi dimulai dari proses *login* yang dilanjutkan masuk ke halaman Beranda di mana di dalamnya terdapat halaman Tabel Review Novel, halaman Tabel Kata Dasar, Tabel Stopwords dan halaman Proses PreProcessing. Halaman Stopword berisi tabel dari kumpulan kata *stopword removal* dari proses *pre-processing* yang sudah tersimpan dan dapat ditampilkan kembali jika diperlukan.



Gambar 4. Struktur Navigasi Aplikasi

Pembuatan aplikasi dari penelitian ini menggunakan seperangkat komputer dengan spesifikasi *processor* Core i5. RAM 4GB, *harddisk* 500GB, dan *memory* NVIDIA 920M 2GB. Perangkat lunak yang dibutuhkan adalah sistem operasi Windows 10 Home (64-bit), XAMPP, Sublime Text, Adobe Dreamweaver CS4, dan Mozilla Firefox.

III. HASIL DAN DISKUSI

Hasil dari penelitian ini adalah data latih sebagai hasil dari proses *pre-processing* yang kelanjutannya dapat digunakan untuk melakukan proses sentimen analisis.

A. Hasil Tahapan Pembentukan Data Latih

Pembentukan data latih melewati tahapan proses *case folding, tokenizing, stopword removal* dan *stemming*. Berikut adalah hasil dari proses-proses tersebut yang datanya diambil dari tiga novel berbahasa Indonesia yang ada di situs Good Reads.

Tabel 6. memperlihatkan hasil proses *tokenizing* yang dihasilkan dari aplikasi dengan logika memotong *string* dari penyusunnya. Pengecekan *review* dari karakter pertama sampai karakter terakhir. Apabila karakter ke-i bukan tanda pemisah kata seperti titik (.), koma (,), spasi dan tanda pemisah lainnya, maka akan digabungkan dengan karakter selanjutnya.

TABEL 6. KUMPULAN HASIL TOKENIZING

No	*Judul Novel*	*Review Asli*	*Hasil Tokenizing*
1	Anak Semua Bangsa	Buku kedua semakin menarik. Banyak tokoh yang menjadi karakter dalam tulisan Minke. Anak Semua Bangsa menekankan kompleksitas latar, tokoh & penokohan, serta perluasan alur dari buku pertama.	buku kedua semakin menarik banyak tokoh yang menjadi karakter dalam tulisan minke anak semua bangsa menekankan kompleksitas latar tokoh penokohan serta perluasan alur dari buku pertama
2	Burung-Burung Manyar	Buku roman percintaan yang sangat indah, romantis, dewasa, toleran dan penuh suasana humor yang "njawani". Cinta dalam hal ini dikemas menjadi sebuah perjuangan untuk membahagiakan orang yang dikasihi.	buku roman percintaan yang sangat indah romantis dewasa toleran dan penuh suasana humor yang njawani cinta dalam hal ini dikemas menjadi sebuah perjuangan untuk membahagiakan orang yang dikasihi
3	Di Bawah Lindungan Ka'bah	Plot yang agak mendatar. Tiada apa yang istimewa bagi buku ini.	plot yang agak mendatar tiada apa yang istimewa bagi buku ini

Tabel 7 memperlihatkan hasil proses *stopwords removal* yang dihasilkan dari aplikasi dengan logika Setiap kata pada *review* akan diperiksa. Jika terdapat kata sambung, kata depan, kata ganti atau kata yang tidak ada hubungannya dalam analisis sentimen, maka kata tersebut akan dihilangkan.

Kumpulan Hasil Stopwords Removal

No	*Judul Novel*	*Hasil Tokenizing*	*Hasil Stopwords Removal*
1	Anak Semua Bangsa	buku kedua semakin menarik banyak tokoh yang menjadi karakter dalam tulisan minke anak semua bangsa menekankan kompleksitas latar tokoh penokohan serta perluasan alur dari buku pertama	buku menarik tokoh karakter tulisan minke anak bangsa menekankan kompleksitas latar tokoh penokohan serta perluasan alur buku
2	Burung-Burung Manyar	buku roman percintaan yang sangat indah romantis dewasa toleran dan penuh suasana humor yang njawani cinta dalam hal ini dikemas menjadi sebuah perjuangan untuk membahagiakan orang yang dikasihi	buku roman percintaan indah romantis dewasa toleran penuh suasana humor njawani cinta dikemas sebuah perjuangan membahagiakan orang dikasihi
3	Di Bawah Lindungan Ka'bah	plot yang agak mendatar tiada apa yang istimewa bagi buku ini	plot mendatar tiada istimewa buku

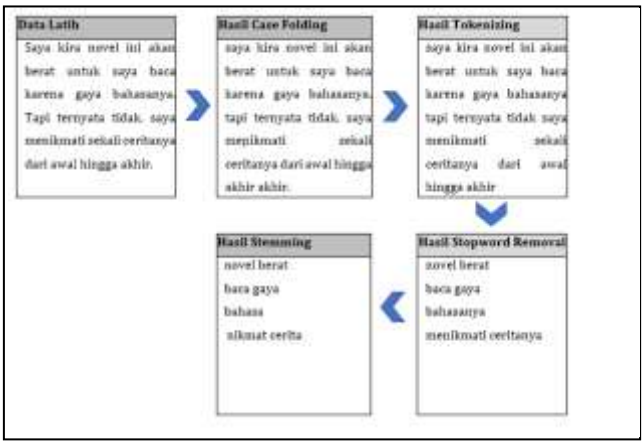
Tabel 8 memperlihatkan hasil proses *stemming* yang dihasilkan dari aplikasi dengan logika mengubah kata ke bentuk dasarnya dengan cara menghilangkan imbuhan-imbuhan pada kata dalam dokumen. Algoritma *stemming* yang digunakan dalam penelitian ini adalah Algoritma Nazief Adriani.

TABEL 7. KUMPULAN HASIL STEMMING

No	*Judul Novel*	*Hasil Stopwords Removal*	*Hasil Stemming*
1	Anak Semua Bangsa	buku menarik tokoh karakter tulisan minke anak bangsa menekankan kompleksitas latar tokoh penokohan serta perluasan alur buku	buku tarik tokoh karakter tulis minke anak bangsa tekan kompleksitas latar tokoh tokoh serta luas alur buku
2	Burung-Burung Manyar	buku roman percintaan indah romantis dewasa toleran penuh suasana humor njawani cinta dikemas sebuah perjuangan membahagiakan orang dikasihi	buku roman cinta indah romantis dewasa toleran penuh suasana humor njawani cinta kemas sebuah juang bahagia orang kasih
	Di Bawah Lindungan Ka'bah	plot mendatar tiada istimewa buku	plot datar tiada istimewa buku

B. Ilustrasi Hasil Tahapan

Ilustrasi dari tahapan pembentukan data latih terlihat seperti gambar 5. Data latih diambil dari salah satu *review* terhadap novel berbahasa Indonesia, kemudian masuk ke dalam tahap *case folding*, dilanjutkan ke tahap *tokenizing*, kemudian masuk ke tahap *stopword removal* dan tahap akhir dari *pre-processing* adalah masuk ke dalam *stemming*.



Gambar 5. Ilustrasi Hasil Proses

C. Halaman Aplikasi

Hasil jadi aplikasi terlihat seperti gambar 6, di mana tampilan halaman dan keterhubungannya disesuaikan dengan struktur navigasi. Pada halaman Beranda berisi informasi jumlah data latih, perbandingan jumlah data latih positif dan jumlah data latih negatif dalam bentuk diagram pie. Pada halaman Tabel Review Novel berisi dataset *review* novel dengan aktivitas yang dapat dilakukan adalah menambah, merubah atau menghapus *review* serta mencari *review* yang ingin ditampilkan.

Gambar 6. Tampilan Halaman Aplikasi

IV. KESIMPULAN

Aplikasi *pre-processing* untuk analisis sentimen telah berhasil dibuat hingga menghasilkan data latih. Pada proses *pre-processing* terdapat beberapa tahapan pemrosesan yang harus dilakukan terdiri dari *case folding, tokenizing, stopword removal*, dan *stemming*. *Stemming* yang digunakan dalam proses ini yaitu *stemming* Nazief dan Adriani. *Term* (kata) hasil *stemming* dari proses *pre-processing* dijadikan sebagai model data latih yang nantinya akan dapat digunakan sebagai data uji pada proses sentimen analisis terhadap *review* novel berbahasa Indonesia khususnya pada situs Goodreads.

DAFTAR PUSTAKA

[1] Adriani, M., Asian, J., Nazief, B. Tahaghoghi, S.M.M., Williams, H.E., "Stemming Indonesian: A Confix-Stripping Approach. Transaction on Asian Langeage Information Processing", 2007.

[2] Hamzah, A., "Sentiment Analysis Untuk Memanfaatkan Saran Kuesioner Dalam Evaluasi Pembelajaran Dengan Menggunakan Naive Bayes Classifier (Nbc)", in Proc. SNAST 2014, diakses tanggal 25 Maret 2019.

[3] Pang B, Lee L, Vaithyanathan S, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02), USA, pp. 79 – 86, 2002.

[4] Pressman, Roger, S., "Software Engineering: A Practitioner's Approach", Fifth Ed. New York, McGraw-Hill Book Company, 2001.

[5] Santika P.P, Arifin A.A, Purwitasari D, "Pembentukan Thesaurus yang Sensitif Terhadap Tingkat Polaritas Review pada Cross-Domain Sentiment Classification", 2015, http://jurnal.akba.ac.id/index.php/inspiration/article/view/69, diakses tanggal 28 Maret 2019.

[6] Yessenov, Kuat and Misailovic, Sasa, "Sentiment Analysis of Movie Review Comments". 6863 Spring final project. 2009.

[7] Z. Zhang, Q. Ye, Z. Zhang and Y. & Li. "Sentiment Classification of Internet Restaurant Reviews Written in Cantonese," in Expert Systems with Applications. 2011. pp. 7674-7682.

[8] https://www.goodreads.com/about/us, diakses tanggal 28 Maret 2019.