J M S K
Jurnal Matematika, Statistika & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
e-ISSN: 2614-8811
Vol .15, No. 2, 95-103, January 2019

Pengembangan Modul PreprocessingTeks untuk Kasus Formalisasi dan Pengecekan Ejaan Bahasa Indonesia pada Aplikasi Web Mining Simple Solution (WMSS)

Umi Chuzaimah Zulkifli[1], Lya Hulliyyatus Suadaa

Abstract

Data of social media currently has been much used to analyze both sentiment analysis and another analysis. In fact, data that is obtained from the social media in generally has some mistakes which can influence the spelling in writing of words. The solution offered is word formalization and spelling check. Based on the problem, it will be built a preprocessing model to overcome two the mistakes. The method that will be used in formalization is to change the words to be formal form based on KBBI, while the method used for spelling check is spelling correction. Spelling correction method consists of *distance edit, bigram* and *distance edit rule.* In this study, in addition the application of both methods, also it will be analyzed comparing the result of spelling correction. From the result of analysis shows that distance edit rule has higher accuracy, namely 83.39% than using both edit distance and bigram method. In addition, edit distance rule method also has faster performance than another both methods. Overall, method to change word to formal word were based on KBBI and spelling correction has been able to overcome the problem of two cases, such that it can increase accuracy of the result of the analysis.

Keywords: preprocessing, spelling correction, edit distance, bigram

Abstrak

Data media sosial saat ini telah banyak digunakan untuk melakukan analisis baik analisis sentimen maupun analisis terkait lainnya. Nyatanya, data yang diperoleh dari media sosial tersebut pada umumnya memiliki kesalahan yang akan mempengaruhi hasil analisis. Kesalahan tersebut berupa penggunaan kata yang tidak baku dan adanya kesalahan ejaan dalam penulisan kata. Solusi yang ditawarkan berupa formalisasi kata dan pengecekan ejaan. Berdasarkan masalah tersebut, akan dibangun modul preprocessing untuk mengatasi dua kesalahan di atas. Metode yang digunakan pada formalisasi adalah mengubah kata ke bentuk formal berdasarkan KBBI sedangkan metode yang digunakan pada pengecekan ejaan adalah spelling correction. Metode spelling correction tersebut terdiri dari tiga yaitu edit distance, bigram dan edit distance + rule. Pada penelitian ini, selain penerapan kedua metode juga akan dilakukan analisis untuk melihat perbandingan hasil pada metode spelling correction. Dari hasil analisis tersebut, diketahui bahwa metode edit distance + rule memiliki akurasi yang lebih tinggi yaitu sebesar 83,39% dibandingkan dengan kedua metode lainnya yaitu edit distance dan bigram. Selain itu, metode edit distance + rule juga memiliki performa tercepat dibandingkan kedua metode lainnya. Secara keseluruhan, metode mengubah kata ke bentuk formal berdasarkan KBBI dan spelling correction telah mampu mengatasi masalah pada dua kasus di atas sehingga dapat meningkatkan akurasi hasil analisis.

Kata Kunci:*preprocessing, spelling correction, edit distance, bigram*

Sekolah Tinggi Ilmu Stastistik, Jakarta
[1]rwiedyanusantara@gmail.com,

1. Pendahuluan

Media sosial saat ini menjadi salah satu kebutuhan masyarakat. Menurut Jan H.Kietzmann (Liliweri,2015:291-292), salah satu fungsi media sosial yaitu *sharing*, dimana media sosial dapat membantu para pengguna melakukan distribusi pesan dan melakukan *sharing* atas pesan. Pesan tersebut sebagai informasi bagi pengguna baik berupa artikel, berita ataupun laporan tertentu. Pengguna media sosial tidak hanya menyebarkan informasi, pengguna lain juga dapat memberikan opini baik berupa pendapat atau komentar. Fungsi *sharing* tersebut akhirnya menjadikan media sosial sebagai salah satu wadah bagi masyarakat untuk beropini.Ini dikarenakan masyarakat berharap opini tersebut dapat dibaca dan dijangkau oleh pengguna media sosial lainnya sehingga pengguna bisa setuju ataupun tidak setuju dengan opini tersebut.

Opini pengguna umumnya berkaitan dengan kejadian yang sedang hangat ataupun peristiwa penting.Opini ini sangat banyak dan terkadang mengarah pada suatu kata kunci tertentu.Oleh karena itu, opini dari pengguna media sosial ini dapat dikumpulkan oleh pengembang sistem untuk selanjutnya dianalisis sehingga menghasilkan informasi yang berguna.Informasi ini dapat menjadi masukan baik untuk masyarakat maupun pemerintah dalam menentukan kebijakan publik.

Penelitian ini akan berfokus pada tahap *preprocessing text*untuk aplikasi WMSS. WMSS sendiri merupakan suatu *text mining tools* yang saat ini dikembangkan, dimana tahap – tahap *text mining* tersebut terdiri dari *text crawling, preprocessing* dan analisis. *Preprocessing text* merupakan salah satu tahap penting dalam proses pengolahan data dengan menggunakan *text mining tools*. Ini dikarenakan pada tahap ini data mentah akan mengalami proses *cleaning data;* dimana data mentah tersebut akan diubah menjadi data yang berkualitas.

Alasan perlunya *preprocessing* untuk contoh kasus analisis sentimen adalah mengurangi *noise* atau gangguan sehingga dapat meningkatkan performa dan juga mempercepat proses klasifikasi pada kasus tersebut. Alasan tersebut tentu berlaku pada kasus analisis lainnya sehingga tahap ini diperlukan pada aplikasi *text mining* sehingga hasil yang diperoleh lebih berkualitas.

Masalah yang umumnya dihadapi pada tahap *preprocessing text* yaitu penggunaan bahasa yang tidak formal dan adanya kesalahan pengetikan pada teks tersebut. Masyarakatterkadang menggunakan kata yang tidak baku ataupun bahasa gaul dalam beropini. Selain itu, penulisan kata yang salah ejaan atau umumnya disebut *typo* juga sering terjadi pada suatu kalimat.Jika kita tetap memproses kata – kata tersebut, analisis yang dihasilkan akan tidak akurat.

Berdasarkan masalah di atas, akan dilakukan penerapan metode yang sesuai guna meningkatkan akurasi dari analisis data teks. Metode yang digunakan dalam penelitian ini yaitu *mengubah kata ke bentuk formal berdasarkan KBBI* untuk menyelesaikan permasalahan mengenai formalisasi kata dan metode *Spelling Correction* untuk menyelesaikan permasalahan mengenai ejaan yang kurang tepat. Metode *Spelling Correction* adalah metode untuk memperbaiki kesalahan ejaan pada kata. Beberapa metode yang termasuk *Spelling Correction* yaitu *n-gram, semantic relatadness* dan *biased correction.*

Untuk penelitian terkait dengan kasus masalah pengecekan ejaan, terdapat salah satu sistem yang bernama “typoonline”.Sistem ini bertujuan untuk membantu mengecek *typo* atau kesalahan ejaan pada dokumen, kalimat dan situs web.Sistem “typoonline” tentu terkait dengan penelitian ini dalam hal penemuan kesalahan ejaan.Namun perbedaan sistem tersebut dengan penelitian ini adalah adanya penambahan fitur untuk memperbaiki ejaan pada kalimat sehingga modul yang dibangun bukan hanya dapat digunakan untuk menemukan kesalahan ejaan tetapi juga memperbaiki kesalahan ejaan tersebut.

Selain sistem "typoonline" tersebut, terdapat penelitian lain terkait kasus masalah pengecekan ejaan yang berjudul *"Text normalization in social media: progress, problems and applications for a pre-processing system of casual English"*. Dalam penelitian tersebut, terdapat 8 kesalahan yang umumnya terdapat pada bahasa Inggris di sosial media. Salah satunya adalah *"typing error/misspelling"* atau kesalahan ejaan pada bahasa Inggris, contohnya *"wouls"* menjadi *"would"*. Kesalahan ejaan tersebut sama dengan penelitian yang saat ini dilakukan namun algoritma yang digunakan berbeda. Algoritma yang digunakan dalam penelitian ini yaitu *Edit Distance, Bigram* serta gabungan penggunaan *Rule* dan *Edit Distance*.

Sedangkan penelitian terkait dengan kasus masalah formalisasisalah satunya yaitu "*Mining and Modeling Relations between Formal and Informal Chinese Phrases from Web Corpora"*. Dalam penelitian ini, kata informal dalam bahasa Cina akan diubah menjadi kalimat formal dengan menggunakan model Log-linear. Penelitian ini sama dengan penelitian yang saat ini dilakukan yaitu mengubah kata informal menjadi formal, namun perbedaan dengan penelitian ini adalah metode yang digunakan.

2. Metodologi

2.1. Metode Pengumpulan Data

Penelitian ini bertujuan untuk melakukan *preprocessing* terhadap teks berita dan media sosial.Oleh karena itu, data yang digunakan berasal dari portal berita dan media sosial. Selain itu, *preprocessing* dapat dilakukan pada teks dan fail yang dimasukkan oleh pengguna.

Data yang digunakan pada penelitian ini yaitu data sekunder.Data tersebut berasal dari hasil *crawling* pada portal berita dan media sosial *Twitter* dan *Facebook*.*Crawling* merupakan proses pengambilan data dari suatu halaman web. Data tersebut berupa teks berita, twit masyarakat dimana twit tersebut umumnya berupa status, komentar ataupun informasi, serta status pengguna. Untuk data twit dan status pengguna umumnya telah dikategorikan berdasarkan kata pencarian tertentu sesuai kebutuhan.

Selain data *crawling*, pengguna juga dapat melakukan masukan berupa kalimat atau teks serta fail sehingga pengguna dapat melihat hasilnya secara langsung.

2.2 Metode Analisis

Metode analisis yang digunakan dalam penelitian ini terdiri dari beberapa tahap.Tahap tersebut terdiri dari *parsing*, tokenisasi dan penggunaan algoritma untuk setiap kasus. Berikut akan dijelaskan satu persatu tahap di atas.

1. *Parsing*
   Tahap pertama yaitu *parsing* yaitu pemecahan dokumen menjadi komponen – komponen terpisah. Misalnya, untuk data kumpulan twit akan dipecah menjadi twit yang terpisah.
2. Tokenisasi
   Pada tahap ini, dilakukan penghilangan angka, tanda baca ataupun karakter. Penghilangan ini dilakukan karena dianggap tidak memiliki pengaruh pada *preprocess* ini.
3. Penggunaan metode yang akan digunakan pada setiap kasus
   Untuk kasus pengecekan ejaan, metode yang digunakan yaitu *Edit Distance, Bigram*, dan penggabungan *Rule & Edit Distance.*Untuk kasus formalisasi, metode yang digunakan yaitu dengan mengubah kata ke bentuk formal berdasarkan KBBI.
4. Analisis hasil metode
   Khusus untuk kasus pengecekan ejaan, selain dilakukan penerapan dari ketiga metode juga akan dilakukan perbandingan untuk melihat metode mana yang menghasilkan hasil terbaik.

Untuk mengetahui metode terbaik, akan dilihat persentase akurasi kata yang berhasil diperbaiki dari masing – masing metode. Selain itu juga akan dilihat performa dari setiap metode.

2.3 Metode Penelitian

Metode penelitian yang digunakan dalam pembangunan aplikasi ini adalah *Design Research.*

1. *Awareness of Problem*
   Pada tahap ini akan dilakukan pencarian masalah. Setelah melakukan pengamatan, diketahui bahwa umumnya data pada sosial media memiliki struktur kata yang tidak baku dan umumnya banyak terdapat kesalahan ejaan atau "typo".
2. *Suggestion*
   Setelah diketahui masalah di atas, pada tahap ini akan dicari solusi untuk menyelesaikan masalah tersebut. Pada penelitian ini, solusi yang ditawarkan adalah dengan menerapkan metode pengubahan kata ke bentuk formal berdasarkan KBBIuntuk penyelesaian kata yang tidak baku dan metode *Spelling Correction* untuk penyelesaian kesalahan penulisan pada modul *preprocessing.*
3. *Development*
   Pada tahap ini, solusi yang ditawarkan tersebut akan diimplementasi dengan melakukan pembangunan modul *preprocessing.* Modul tersebut akan dibangun dengan menerapkan kedua metode yang telah disebutkan pada tahap sebelumnya.
4. *Evaluation*
   Pada tahap ini, akan dilakukan tahap uji coba untuk melihat apakah modul yang dibangun telah berhasil menghasilkan *output* untuk menyelesaikan masalah yang telah disebutkan pada tahap awal. Selain itu, akan dilakukan analisis lanjutan untuk kasus penyelesaian kesalahan ejaan.
5. *Conclusion*
   Tahap ini dilakukan dengan menyimpulkan hasil implementasi dari model *preprocessing* dengan menerapkan kedua metode tersebut. Dari kesimpulan tersebut, peneliti dapat memberikan saran sehingga dapat dilakukan pengembangan modul *preprocessing* pada penelitian selanjutnya.

3. Hasil dan Pembahasan

3.1 Analisis Perancangan Sistem

Rancangan aplikasi dibutuhkan agar aplikasi yang dihasilkan sesuai dengan tujuan yang telah ditetapkan pengembang aplikasi. Dalam pembuatan modul ini, terdapat tiga rancangan aplikasi yaitu : rancangan program, rancangan antarmuka dan rancangan metode. Rancangan antarmuka akan ditampilkan pada Lampiran 1, sedangkan rancangan metode akan dilampirkan pada Lampiram 2 dan Lampiran 3.

3.2 Implementasi Aplikasi

Pada pengembangan sistem ini, bahasa pemograman yang digunakan adalah Python versi 3 dengan IDE Jupyter. Untuk pengaturan tampilan menggunakan platform Django versi 10. Selain itu *corpus* yang digunakan untuk metode *Spelling Correction* menggunakan data dari *Leipzig Corpora Collection.*

3.3 Implementasi Metode

Implementasi sistem terdiri dari tahap – tahap implementasi yang terdapat pada metode analisis dan rancangan algortime sebelumnya.Penjelasan untuk tiap metode adalah sebagai berikut.

1. *Parsing*

   Untuk pengembangan modul ini, *parsing* dilakukan dengan memecah kumpulan twit yang telah di *crawling* sebelumnya menjadi twit – twit yang terpisah. Selain twit, kalimat atau berita baik yang di *crawling* ataupun diinput manual akan dilakukan *parsing* dengan memecah kalimat menjadi kata – kata atau kumpulan kalimat.
   Contoh tahap *parsing* pada modul ini adalah sebagai berikut :
   - Monitoring evaluasi Dana Desa, bersama Dinas PMP, TA, PD, PLD menuju Lampung Barat Hebat. @jokowi @wapres_ri @EkoSandjojo @taufikmadjid71

   Hasil *parsing* kalimat di atas yaitu :
   “Monitoring”, “evaluasi”, “Dana”, “Desa”, “bersama”, “Dinas”, “PMP”, “TA”, “PD”, “PLD”, “menuju” ,”Lampung”, “Barat”, “Hebat”, “@jokowi”, “@wapres_ri”, “@EkoSandjojo”, “@taufikmadjid71”

2. Tokenisasi

   Pada pengembangan modul ini, tokenisasi dilakukan dengan menghilangkan tanda baca selain @ dan # dengan menggunakan *package Regex*.Selain itu semua kata diatur menjadi *lower case*. Berdasarkan hasil *parsing* di atas, maka hasil tokenisasi dari kalimat tersebut adalah :
   “monitoring”, “evaluasi”, “dana”, “desa”, “bersama”, “dinas”, “pmp”, “ta”, “pd”, “pld”, “menuju” ,”lampung”, “barat”, “hebat”, “@jokowi”, “@wapres_ri”, “@ekosandjojo”, “@taufikmadjid71”

3. Penggunaan metode
   - Mengubah kata ke bentuk formal berdasarkan KBBI

     Untuk menerapkan metode ini, peneliti menggunakan data yang berasal dari Kateglo.Data tersebut didapatkan dengan menggunakan API yang berasal dari website Kateglo tersebut.
     Contoh penerapan :
     Kata “gue” akan diubah menjadi kata “saya”
   - *Spelling Correction*

     Pada metode ini, akan diterapkan tiga metode untuk mengatasi masalah kesalahan ejaan. Metode tersebut yaitu *Edit Distance,* Bigramdan penggabungan *Rule & Edit Distance.*
     Contoh penerapan :
     “mentri” diubah menjadi “menteri”.

4. Analisis hasil metode

Analisis ini dilakukan untuk melihat perbandingan hasil dari perbaikan ejaan dengan menggunakan tiga metode di atas. Analisis ini dilakukan dengan menggunakan data twit dengan keyword "Jokowi" . Setelah dilakukan tiga tahap di atas yaitu *parsing,* tokenisasi dan penggunaan metode maka diketahui hasil akurasi dari ketiga metode untuk kasus pengecekan ejaan. Hasil akurasi itu dilakukan dengan membandingkan data twit yang diharapkan dengan data twit yang dihasilkan. Dari perbandingan tersebut akan dibuat Confusion Matrix sehingga dapat dihitung nilai akurasinya.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Berikut hasil dari ketiga algoritma tersebut:

Tabel 1. Hasil akurasi dengan metode pada metode *Spelling Correction.*

No.	Metode	Akurasi
(1)	(2)	(3)
1.	*Edit Distance*	79,46%
2.	*Edit Distance + Rule*	83,39%
3.	Bigram	39,84%

Berdasarkan tabel di atas, metode *Edit Distance + Rule* menghasilkan akurasi yang paling tinggi dibandingkan dengan kedua metode lainnya yaitu *Edit Distance* dan Bigram yakni sebesar 83,39% . Metode Bigram menghasilkan akurasi yang paling kecil karena metode ini bergantung pada ukuran *corpus.*

Selain akurasi, kita juga dapat melihat bagaimana performa setiap metode. Kita akan menghitung waktu dari modul ini dengan menggunakan data twit yang sama dengan sebelumnya yaitu data twit dengan *keyword* "Jokowi". Data twit ini terdiri dari 50 twit yang akan dilihat waktu untuk melakukan *preprocessing* dari setiap metode pada *spelling correction.* Berikut hasil performa dari metode tersebut yang dihitung dalam detik :

Tabel 2.Performa dengan metode pada *Spelling Correction.*

No.	Metode	Waktu (detik)
(1)	(2)	(3)
1.	*Edit Distance*	24.42
2.	*Edit Distance + Rule*	17.21
3.	Bigram	45.6

4. Kesimpulan

Pada penelitian ini dapat ditarik kesimpulan :

1. Telah dibangun modul *preprocessing* untuk membantu menyelesaikan masalah pada data twit, kalimat ataupun berita.
2. Metode pengubahan kata ke bentuk formal berdasarkan KBBI dan *Spelling Correction* dapat membantu menyelesaikan masalah yang dihadapi yaitu formalisasi kata dan perbaikan ejaan.

3. Untuk metode *Spelling Correction,* metode *Edit Distance + Rule* menghasilkan akurasi yang lebih baik dibandingkan kedua metode lainnya yaitu *Edit Distance* dan Bigram.
4. Telah dirancang desain antarmuka sehingga memudahkan pengguna dalam menggunakan modul ini.
5. Modul *preprocessing* telah dapat diintegrasikan dengan modul lainnya yaitu *web crawler* dan analisis pada aplikasi WMSS.

5. Saran

Saran yang dapat diberikan untuk penelitian ini yaitu :

1. Perlu pencarian data *corpus* yang lebih baik lagi, baik dalam hal jumlah ataupun kualitas sehingga dapat meningkatkan akurasi dari metode perbaikan ejaan.
2. Pengembangan modul *preprocessing* untuk kasus lainnya seperti kasus bahasa alay atau penggunaan bahasa asing dan bahasa daerah.

Daftar Pustaka

[1]. Badan Pengembangan dan Pembinaan Bahasa, 2016. *Pedoman Umum Ejaan Bahasa Indonesia*, Kementrian Pendidikan dan Kebudayaan, Jakarta.

[2]. Clark, Eleanor dan Kenji A., 2011. *Text normalization in social media: progress, problems and applications for a pre-processing system of casual English*,Pacific Association for Computational Linguistics (PACLING 2011), pp.2-11.

[3]. D. Goldhahn, T. Eckart & U. Quasthoff, 2012. *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages*,Proceedings of the 8th International Language Ressources and Evaluation (LREC'12)

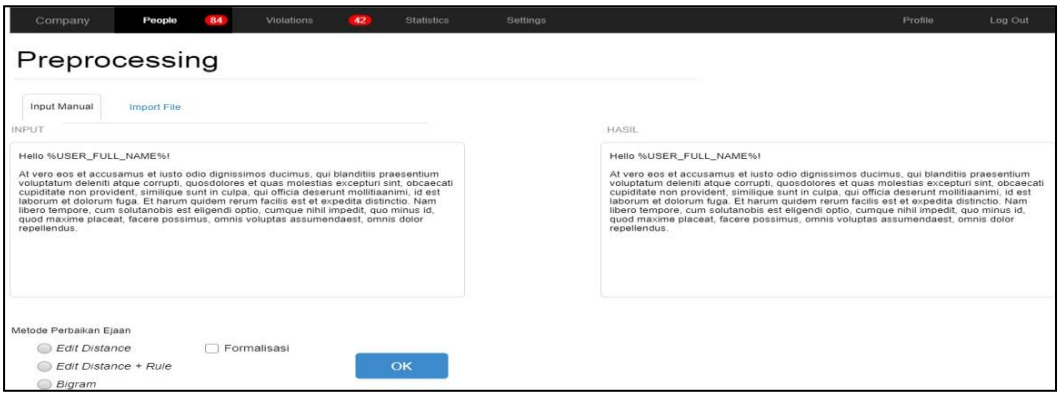
[4]. Flor, Michael, 2012. *Four types of context for automatic spelling correction,* TAL. Volume 53 – n° 3/2012, pp. 61-99.

[5]. Haddi, Emma dkk, 2013.*The Role of Text Pre-processing in Sentiment Analysis*, Procedia Computer Science 17, pp. 26 – 32.

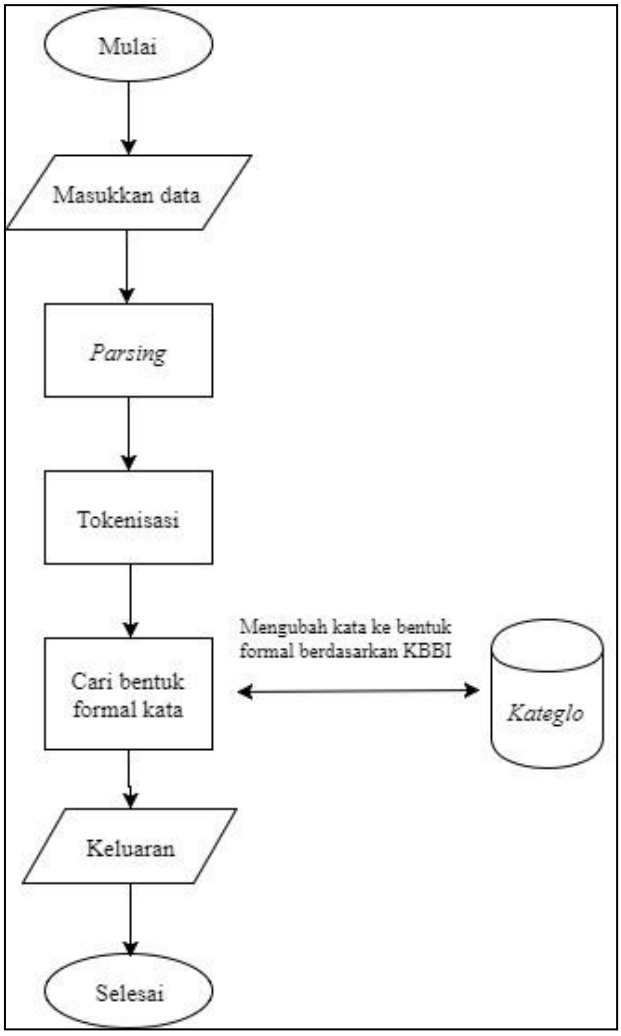
[6]. Jody, dkk, 2015. *Analisis dan Implementasi Algoritma Winnowing dengan Synonym Recognition pada Deteksi Plagiarisme untuk Dokumen Teks Berbahasa Indonesia,* Telkom University, Bandung.

[7]. Li, Zhifei, dan David Yarowsky, 2008. *Mining and Modeling Relations between Formal and Informal Chinese Phrases from Web Corpora*, Proceedings of the 2008 Conference on Empirical Methods in Natural Languange Processing, pp. 1031-1040.

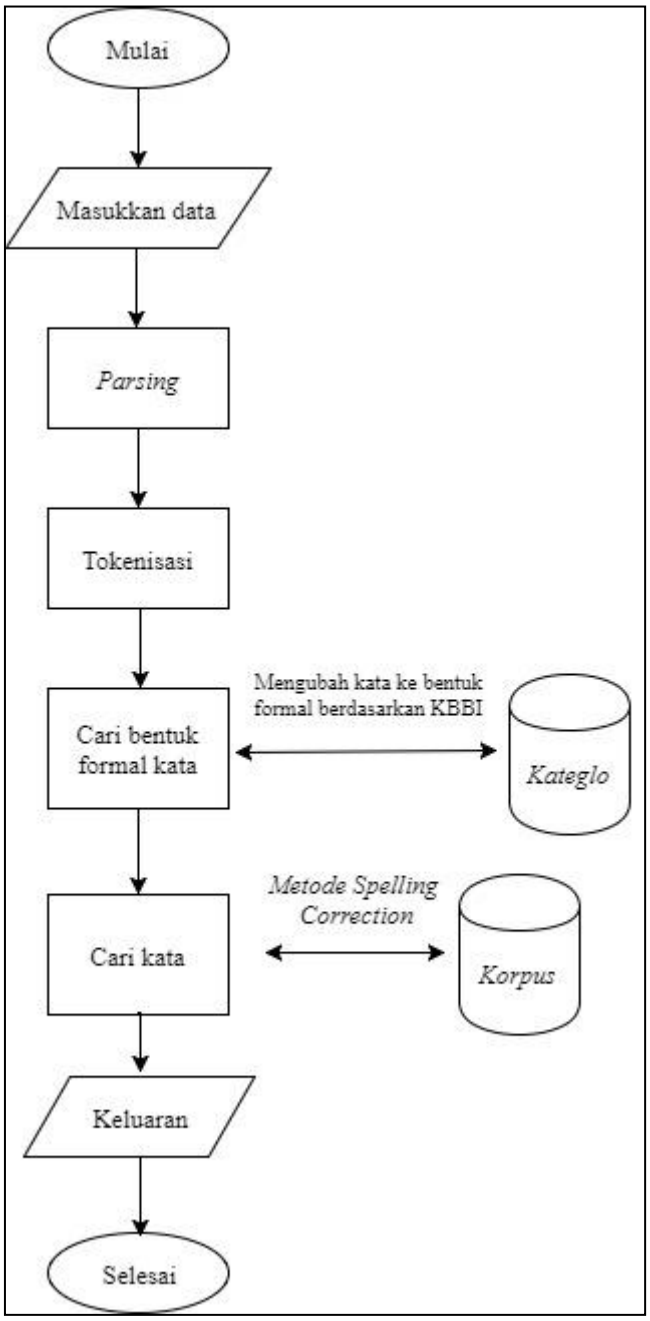
Lampiran 1. Rancangan Antarmuka



Lampiran 2. Rancangan Metode untuk Formalisasi



Lampiran 3. Rancangan Metode untuk Formalisasi + Perbaikan Ejaan


Mulai
Masukkan data
Parsing
Tokenisasi
Cari bentuk formal kata
Mengubah kata ke bentuk formal berdasarkan KBBI
Kateglo
Cari kata
Metode Spelling Correction
Korpus
Keluaran
Selesai