Jurnal Pendidikan Teknologi dan Kejuruan
Vol. 21, No. 1, Januari 2024
P-ISSN: 0216-3241 E-ISSN: 2541-0652

IDENTIFIKASI DAN NORMALISASI TEKS *SLANG* DENGAN *FASTTEXT* PADA *TWITTER DALAM BAHASA INDONESIA*

P. M. S. Ardinata[1)], A. A. J. Permana[2)], I. N. S. W Wijaya[3)]

[1,2,3] Fakultas Teknik dan Kejuruan, Universitas Pendidikan Ganesha
Email: pande.sindu@undiksha.ac.id, agus.aan@undiksha.ac.id, wahyu.wijaya@undiksha.ac.id.

ABSTRAK

Salah satu dampak yang signifikan dari popularitas media sosial adalah munculnya istilah *slang* yang semakin banyak. Istilah *slang* adalah bahasa yang digunakan oleh kelompok-kelompok tertentu untuk berkomunikasi secara informal. *Slang* juga dapat muncul melalui singkatan, penggunaan kata-kata yang berbeda dari arti aslinya, atau penggabungan kata-kata yang tidak konvensional. Dalam pengolahan bahasa alami *(Natural Language Processing) Slang* sering kali memiliki makna yang tidak jelas atau ambigu, dan kata-kata *slang* dapat memiliki konotasi yang berbeda tergantung pada konteks dan subkultur tertentu. Ini dapat menyebabkan kesalahan dalam pemrosesan bahasa alami dan menghasilkan hasil yang tidak akurat atau salah dalam tugas seperti klasifikasi teks atau analisis sentimen. Dari permasalahan tersebut dalam penelitian ini dikembangkan suatu metode untuk mengidentifikasi dan melakukan normalisasi *slang* pada kalimat yang akan diproses oleh NLP. Proses normalisasi *slang* ke bahasa yang lebih standar dilakukan dengan memanfaatkan pretrain model dari *fasttext* untuk mencari kata – kata yang memiliki kedekatan dengan *slang.* Data yang digunakan pada penelitian ini didapatkan dari sosial media *twitter.* Sebelum dinormalisasi data melewati beberapa proses seperti *preprocessing* data yang meliputi proses *cleaning, case folding,* dan *stopword removal* kemudian dilanjutkan dengan proses identifikasi *slang* pada kalimat dan terakhir dilakukan proses normalisasi slang yang didapatkan. Penelitian ini menemukan bahwa metode fasttext masih belum cukup baik melakukan normalisasi slang dikarenakan masih ada sekitar 1329 data dari 3239 data yang tidak berhasil dinormalisasi dengan baik yaitu sekitar 41%. Penelitian ini memberikan kontribusi dalam membantu proses pengolahan kata yang lebih baik untuk NLP.

Kata kunci: *slang*, normalisasi, *fasttext*, NLP

ABSTRACT

One significant impact of the popularity of social media is the proliferation of slang terms. Slang refers to language used by specific groups for informal communication. Slang can emerge through abbreviations, the usage of words in different contexts from their original meanings, or the combination of unconventional words. In Natural Language Processing (NLP), slang often carries ambiguous or unclear meanings, and slang words can have different connotations depending on the context and specific subculture. This can lead to errors in natural language processing and result in inaccurate or erroneous outcomes in tasks such as text classification or sentiment analysis. To address these challenges, this research develops a method for identifying and normalizing slang in sentences to be processed by NLP. The process of normalizing slang into a more standard language is achieved by utilizing pre-trained models from FastText to search for words that have semantic proximity to slang. The data used in this study is obtained from Twitter social media. Before normalization, the data undergo several preprocessing steps, including data cleaning, case folding, and stopword removal. Then, the process continues with the identification of slang in sentences, followed by the normalization of the identified slang. The findings of this research indicate that the FastText method is not entirely effective in normalizing slang, as approximately 41% of the data (1329 out of 3239 data) did not undergo successful normalization. Despite this limitation, this research contributes to improving the word processing process for NLP tasks.

Keywords : *slang, normalization, fasttext, NLP*

1. PENDAHULUAN

Media sosial memberikan ruang yang luas bagi pengguna untuk mengekspresikan diri, terlibat dalam komunitas online, dan mengikuti tren yang sedang populer. Salah satu dampak yang signifikan dari popularitas media sosial adalah munculnya istilah *slang* yang semakin banyak. Istilah *slang* adalah bahasa yang digunakan oleh kelompok-kelompok tertentu untuk berkomunikasi secara informal [1]. Dalam konteks media sosial, pengguna sering kali menciptakan kosakata dan frasa baru yang menggambarkan tren, *meme*, dan referensi populer [2]. *Slang* juga dapat muncul melalui singkatan, penggunaan kata-kata yang berbeda dari arti aslinya, atau penggabungan kata-kata yang tidak konvensional [3]. Penggunaan *slang* dalam komunikasi juga menimbulkan tantangan tersendiri. *Slang* seringkali sulit dipahami oleh mereka yang tidak akrab dengan kosakata dan konvensi bahasa yang digunakan [4]. Hal ini dapat menyebabkan kesalahpahaman atau bahkan hambatan dalam komunikasi antara individu yang berasal dari latar belakang yang berbeda.

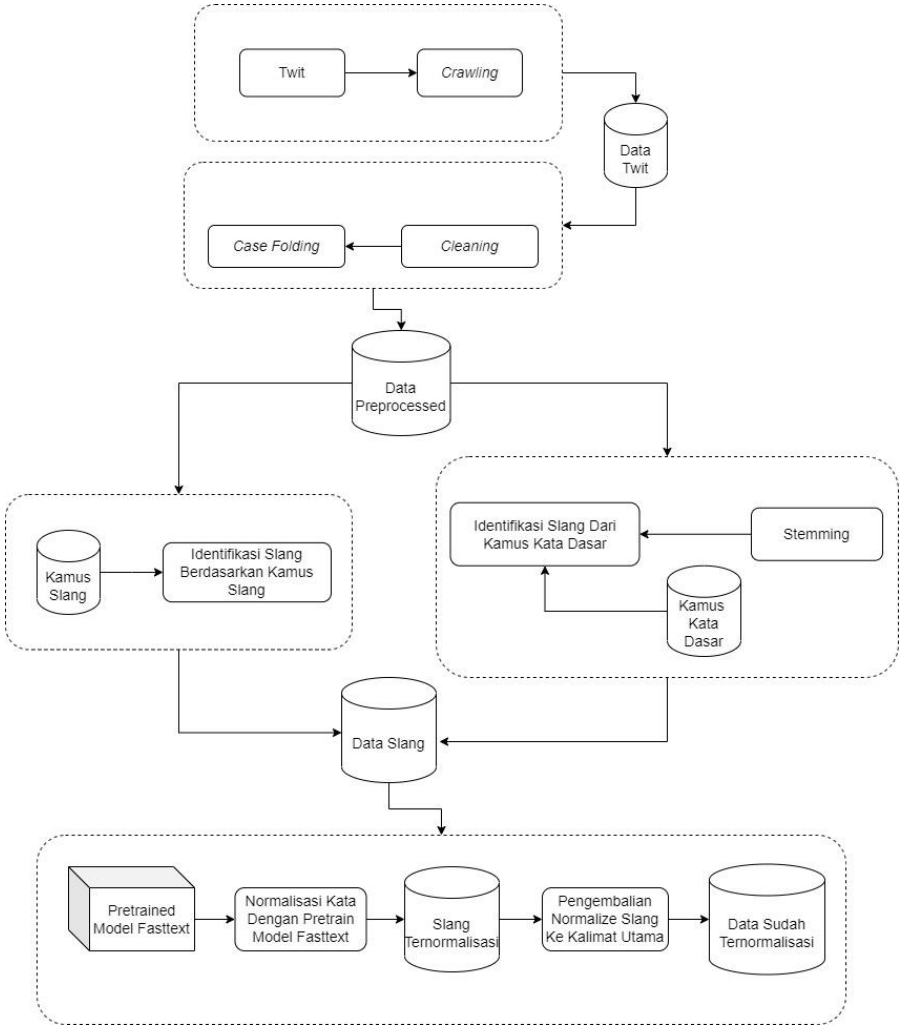
Dalam pengolahan bahasa alami *(Natural Language Processing) Slang* sering kali memiliki makna yang tidak jelas atau ambigu, dan kata-kata *slang* dapat memiliki konotasi yang berbeda tergantung pada konteks dan subkultur tertentu. Ini dapat menyebabkan kesalahan dalam pemrosesan bahasa alami dan menghasilkan hasil yang tidak akurat atau salah dalam tugas seperti klasifikasi teks atau analisis sentimen [5]. Dalam teks klasifikasi, slang dapat mempengaruhi kemampuan model untuk mengenali dan mengkategorikan teks dengan benar [6]. Model klasifikasi yang dilatih menggunakan data formal dan standar mungkin mengalami kesulitan dalam mengenali dan memahami teks yang mengandung *slang* [7]. Hal ini dapat mengakibatkan klasifikasi yang salah atau tidak akurat, karena makna dasar dari teks slang mungkin hilang atau disalahartikan.

Untuk mengatasi tantangan ini penting untuk mengidentifikasi dan menormalisasi *slang* sebelum memasukkan teks ke dalam proses NLP [8]. Dalam konteks teks klasifikasi dan analisis sentimen langkah-langkah normalisasi *slang* dapat melibatkan penggantian kata-kata *slang* dengan bentuk formalnya atau dengan kata-kata yang lebih umum dan dapat dipahami oleh model NLP. Dengan melakukan normalisasi *slang*, teks yang mengandung *slang* dapat diubah menjadi bentuk yang lebih baku dan mudah diinterpretasikan oleh model NLP [9].

Dalam penelitian ini akan dilakukan upaya untuk menemukan *slang* pada suatu kalimat dan melakukan normalisasi pada *slang* tersebut menggunakan *FastText* untuk mencari kata formal yang sesuai untuk menggantikan kata slang pada satu kalimat. *FastText* adalah model *word embedding* yang telah dilatih sebelumnya menggunakan jumlah besar teks dari berbagai sumber [10]. Model ini memiliki pemahaman yang luas tentang kata-kata dan maknanya dalam konteks yang beragam[11]. *Fasttext* dapat menangkap ragam bahasa, idiom, dan penggunaan kata dalam berbagai konteks. Ini memungkinkan model untuk mengenali hubungan semantik dan asosiasi antara kata-kata berdasarkan bagaimana kata-kata tersebut muncul dalam teks yang berbeda-beda [11]. *Fasttext* yang memperhatikan informasi subkata juga berkontribusi pada pemahaman yang lebih baik tentang kata-kata dalam bahasa yang memiliki afiksasi, reduplikasi, dan kata-kata yang jarang ditemukan [12]. Diharapkan dengan mengatasi tantangan *slang* dapat membantu proses pengolahan bahasa alami terutama pada data yang berasal dari sosial media. Sumber data dalam penelitian ini adalah dari data *twitter* karena pengguna platform ini cenderung menggunakan bahasa yang lebih informal dan *slang* dalam cuitan mereka. Teks-teks yang ditemukan di Twitter mencerminkan percakapan sehari-hari, tren, dan topik terkini yang sedang dibicarakan oleh pengguna. *Pretrain* model *fasttext* akan digunakan dalam penelitian ini untuk membantu menormalisasi *slang* dengan mencari alternatif kata-kata yang lebih formal atau umum yang dapat menggantikan kata-kata *slang* dalam teks. Model *FastText* memiliki pengetahuan tentang hubungan antara kata-kata dalam konteks yang beragam, sehingga dapat membantu dalam mencari pengganti yang cocok untuk kata-kata *slang* yang ditemukan.

2. METODE

Dalam penelitian identifikasi dan normalisasi *slang* bahasa indonesia ini akan dilakukan melalui beberapa proses seperti pada Gambar 1 mulai dari proses pengumpulan data, pengumpulan kamus *slang*, proses *preprocessing* data yang terdiri dari proses *cleaning, casefolding* dan *stopword removal,* kemudaian dilanjutkan dengan proses pencarian *slang* dari kamus slang dan pencarian *slang* dari proses *stemming* yang dibandingkan dengan kamus kata dasar dan normalisasi *slang.*



Gambar 1. Gambar Umum Penelitian

A. Pengumpulan Data

Data dalam penelitian ini menggunakan data yang bersumber dari sosial media *twitter*. Dilipilihnya twitter untuk menjadi sumber data pada penelitian ini disebabkan pada *twitter* lebih banyak penggunanya menggunakan kalimat – kalimat yang informal, singkatan dan istilah asing. Data pada *twitter* diperoleh dengan menggunakan metode *crawling* dengan memanfaatkan *library snscrapper* untuk menelusuri data – data *twit* berdasarkan kata kunci yang dimasukan. Data hasil *crawling* akan disimpan dalam format .csv untuk diproses selanjutnya. Data hasil dari proses *crawling* selanjutnya akan dipilih untuk menghilangkan data yang tidak relevan seperti *link website* dan iklan produk.

B. Pengumpulan Kamus *Slang*

Pengumpulan kamus *slang* merupakan tahap pembuatan kamus dari kata – kata *slang* yang sering digunakan. Kamus *slang* ini dibuat berdasarkan kamus yang sudah disediakan sebelumnya oleh *library* NLP yang kemudian diperbaharui dengan data–data kata *slang* terbaru. Data *slang* baru diperoleh dengan pedoman dari ahli bahasa dan keyword dari resiliensi yang katanya sudah disepadankan berdasarkan kamus bahasa indonesia. Hasil penambangan data juga memiliki keyword tersebut dan belum masuk di library NLP sehingga perlu ditambahkan secara manual. Kamus kumpulan *slang* ini dibuat dengan tujuan untuk mencari kata *slang* pada kalimat yang sama pada kamus.

C. *Preprocessing* data

Setelah data berhasil dikumpulkan data masih dalam bentuk data teks yang belum terstruktur. Data masih memiliki banyak noise atau gangguan pada data yang terdiri dari emoji, *link,* dan karakter karakter lain sehingga data perlu melewati proses *preprocessing* terlebih dahulu. Pada tahapan ini

maka dilakukan proses *preprocessing* pada data untuk mendapatkan data yang hanya berupa data dalam bentuk teks saja. Tahap – tahap pada proses *preprocessing* yang dilakukan terdiri dari proses berikut :

1) Data Cleaning
   Data *cleaning* berfungsi untuk menghapus karakter – karakter yang tidak diperlukan atau *noise* pada data dengan menghilangkan tanda baca, emoji, angka dan *link* [13]. Proses ini dilakukan karena keberadaan elemen-elemen tersebut dapat membuat data menjadi tidak efektif. Proses *cleaning* akan dilakukan dengan memanfaatkan *library regular expression.*
2) *Case Folding*
   Case folding merupakan tahapan untuk menyamakan bentuk dari setiap kata agar tidak ada perbedaan pada kata yang serupa. *Case folding* merubah keseluruhan huruf pada kata menjadi bentuk *lowercase* [13]. Untuk melakukan proses *case folding* akan dimanfaatkan *library* yang disediakan oleh *python.*
3) *Stopword Removal*
   Stopword removal merupakan tahapan atau proses dalam pemrosesan teks yang bertujuan untuk menghapus kata-kata yang dianggap tidak memiliki makna atau kontribusi yang signifikan dalam analisis teks. Kata-kata ini seringkali merupakan kata-kata umum seperti “dan,” “atau,” “di,” “dari,” “ke,” dan lain-lain [13]. Dalam banyak kasus, kata-kata ini tidak menyampaikan informasi khusus tentang konten teks dan cenderung muncul secara berulang dalam berbagai dokumen.

D. Identifikasi *Slang* Pada Kalimat

Pada tahap ini akan dilakukan proses identifikasi dengan menelusuri data yang terkumpul dan mencocokkan kata-kata dengan kamus *slang* yang telah disusun sebelumnya. Kata-kata dalam data akan diperiksa secara terpisah, kemudian dibandingkan dengan kata-kata yang terdapat dalam kamus *slang* yang telah disusun sebelumnya. Dalam proses ini, jika sebuah kata dalam data ditemukan sesuai dengan kata-kata yang ada dalam kamus *slang* maka kata tersebut akan diidentifikasi sebagai *slang* dan disimpan pada satu file baru. Dengan demikian, metode ini memungkinkan peneliti untuk mengenali dan memisahkan kata-kata *slang* yang telah terdefinisi sebelumnya dalam kamus *slang* yang disiapkan. Selain itu cara lain yang dapat dilakukan adalah dengan menggunakan metode *stemming*. Kata-kata dalam data akan diproses menggunakan algoritma *stemming* untuk mendapatkan bentuk dasarnya. Setelah proses *stemming*, kata-kata yang tidak berhasil diubah menjadi bentuk dasar akan dicocokkan dengan kamus kata dasar Bahasa Indonesia. Jika kata tersebut tidak ditemukan dalam kamus kata dasar maka kata tersebut akan diidentifikasi sebagai *slang* dan disimpan pada sebuah file. Dengan menggunakan kedua cara tersebut dapat secara efektif mengidentifikasi kata-kata *slang* dalam data baik dengan mencocokkan dengan kamus *slang* maupun dengan mencari kata-kata yang tidak berhasil diubah melalui proses *stemming* berdasarkan kamus kata dasar Bahasa Indonesia.

E. Normalisasi *Slang*

Setelah *slang* berhasil teridentifikasi maka proses selanjutnya adalah melakukan normalisasi pada data *slang*. Proses normalisasi akan menggunakan *pretrain* model *fasttext* untuk mengubah *slang* menjadi bentuk yang lebih sesuai. *Pretrain* model *fasttext* yang telah dilatih sebelumnya pada dataset yang luas. *Pretrain* model ini memiliki pengetahuan tentang penggunaan kata-kata dalam berbagai konteks bahasa yang berbeda termasuk *slang*. Model *fasttext* akan menghasilkan kemungkinan pengganti untuk setiap kata *slang* dalam kalimat berdasarkan pola dan konteks yang dipelajari selama proses pelatihan. *Fasttext* menggunakan representasi vektor kata untuk mencari alternatif baku yang cocok untuk setiap kata *slang*. Alternatif baku ini bisa berupa kata-kata yang lebih umum, standar, atau *non-slang* yang memiliki representasi vektor yang mirip dengan kata *slang* yang akan digantikan. *Fasttext* akan menghitung kemiripan antara vektor representasi kata *slang* dan vektor representasi kata alternatif yang ada dalam kamus baku. Alternatif yang memiliki kemiripan terdekat dengan kata *slang* akan dipilih sebagai pengganti *slang*.

3. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan dengan mengikuti beberapa langkah – langkah mulai pengumpulan data, pembuatan kamus *slang, preprocessing* data, identifikasi *slang,* dan normalisasi *slang.*

A. Pengumpulan Data

Hasil pengumpulan data yang dilakukan pada sosial media *twitter* dengan teknik *crawling* dimana data yang diambil pada panelitian ini adalah data *twit* yang berkaitan dengan ungkapan

perasaan mahasiswa. Komponen dari data *twitter* yang diambil adalah data *username* dan *twit*. Hasil dari proses *crawling* dan pemilihan data relevan didapatkan data sejumlah 1550 data. Contoh data yang berhasil dikumpulkan dapat diperhatikan pada Tabel 1.

Tabel 1. Contoh Data Twit

No	User	Twit
1	ArchieButet	G kaget. Praktik lama. Beberapa mahasiswa skripsi jg sengaja dibuat penelitiannya berat, hasil penelitian dipake kenaikan pangkat ybs. Atau diajak join nulis jurnal, nanti nama pertama y dosennya, kadang mahasiswa penulis malah g dicantumkan. Haha
2	whoseeindahh	Aku yg mahasiswa semester 4 kek berat bgt ,berngkt pagi pulang mlm nangis nangis kerjain tugas.
3	iqbalkanzo_	Sikap dan ucapan dospem tipe sulit kadang-kadang memang tidak jelas, yang mengharuskan mahasiswa bimbingannya melakukan tugas ganda. Tugas kesatunya ialah melakukan penelitian dan menulis karya ilmiah, sedangkan tugas keduanya, yang lebih berat, ialah memahami dospemnya.
4	iqbalkanzo_	Hanya, beban mental saat berhadapan dengan dospem tipe elitis membuat penyelesaian tugas akhir terasa berat bagi mahasiswa.
5	pinkvelt	Susah banget hidup sebagai mahasiswa di Kalimantan. Strugglenya berat, mau cari kerjaan atau part time salah satu syaratnya tidak sedang kuliah. Hampir semua lowongan nih begini. Huaaa, gimana bisa bertahan hidup disiniðŸ˜

B. Pengumpulan Kamus *Slang*

Pengumpulan kamus *slang* yang telah dilakukan berhasil mendapatkan sejumlah 2742 kata yang termasuk sebagai kamus *slang.* Data *slang* tersebutdiperoleh dari keyword padanan kata yang diberikan oleh pakar yang belum masuk pada library sehingga perlu ditambahkan lagi supaya dapat dikenali oleh mesin saat proses pembelajaran. Contoh dari kamus *slang* yang berhasil dibangun dapat diperhatikan pada Tabel 2.

Tabel 2. Contoh Data Slang

No	*Slang*
1	aj
2	ajep ajep
3	ak
4	akika
5	akkoh
6	akuwh
7	alay
8	alow
9	ambilin
10	ancur

C. Preprocessing Data

Data yang berhasil dikumpulkan pada proses pengumpulan data masih merupakan data masih harus melewati proses *cleaning* dan proses *case folding* agar nantinya data siap untuk digunakan. Hasil dari proses *cleaning, case folding,* dan *stopword removal* dapat diperhatikan pada Tabel 3.

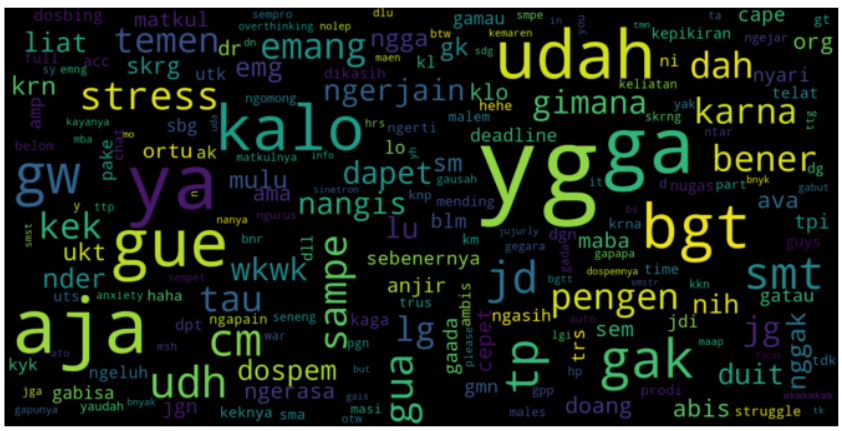
Tabel 3. Preprocessing Data

Twit Hasil Crawling	Hasil *Cleaning*	Hasil *Case Folding*	Hasil *Stopword Removal*
G kaget. Praktik lama. Beberapa mahasiswa skripsi jg sengaja dibuat	G kaget Praktik lama Beberapa mahasiswa skripsi jg sengaja dibuat	g kaget praktik lama beberapa mahasiswa skripsi jg sengaja dibuat	g kaget praktik mahasiswa skripsi jg sengaja

Twit Hasil Crawling	Hasil *Cleaning*	Hasil *Case Folding*	Hasil *Stopword Removal*
penelitiannya berat, hasil penelitian dipake kenaikan pangkat ybs. Atau diajak join nulis jurnal, nanti nama pertama y dosennya, kadang mahasiswa penulis malah g dicantumkan. Haha	penelitiannya berat hasil penelitian dipake kenaikan pangkat ybs Atau diajak join nulis jurnal nanti nama pertama y dosennya kadang mahasiswa penulis malah g dicantumkan Haha	penelitiannya berat hasil penelitian dipake kenaikan pangkat ybs atau diajak join nulis jurnal nanti nama pertama y dosennya kadang mahasiswa penulis malah g dicantumkan haha	penelitiannya berat hasil penelitian dipake kenaikan pangkat ybs diajak join nulis jurnal nama y dosennya kadang mahasiswa penulis g dicantumkan haha
Aku yg mahasiswa semester 4 kek berat bgt ,berngkt pagi pulang mlm nangis nangis kerjain tugas.	Aku yg mahasiswa semester kek berat bgt berngkt pagi pulang mlm nangis nangis kerjain tugas	aku yg mahasiswa semester kek berat bgt berngkt pagi pulang mlm nangis nangis kerjain tugas	yg mahasiswa semester kek berat bgt berngkt pagi pulang mlm nangis nangis kerjain tugas
Sikap dan ucapan dospem tipe sulit kadang-kadang memang tidak jelas, yang mengharuskan mahasiswa bimbingannya melakukan tugas ganda. Tugas kesatunya ialah melakukan penelitian dan menulis karya ilmiah, sedangkan tugas keduanya, yang lebih berat, ialah memahami dospemnya.	Sikap dan ucapan dospem tipe sulit kadangkadang memang tidak jelas yang mengharuskan mahasiswa bimbingannya melakukan tugas ganda Tugas kesatunya ialah melakukan penelitian dan menulis karya ilmiah sedangkan tugas keduanya yang lebih berat ialah memahami dospemnya	sikap dan ucapan dospem tipe sulit kadangkadang memang tidak jelas yang mengharuskan mahasiswa bimbingannya melakukan tugas ganda tugas kesatunya ialah melakukan penelitian dan menulis karya ilmiah sedangkan tugas keduanya yang lebih berat ialah memahami dospemnya	sikap ucapan dospem tipe sulit kadangkadang mengharuskan mahasiswa bimbingannya tugas ganda tugas kesatunya penelitian menulis karya ilmiah tugas berat memahami dospemnya

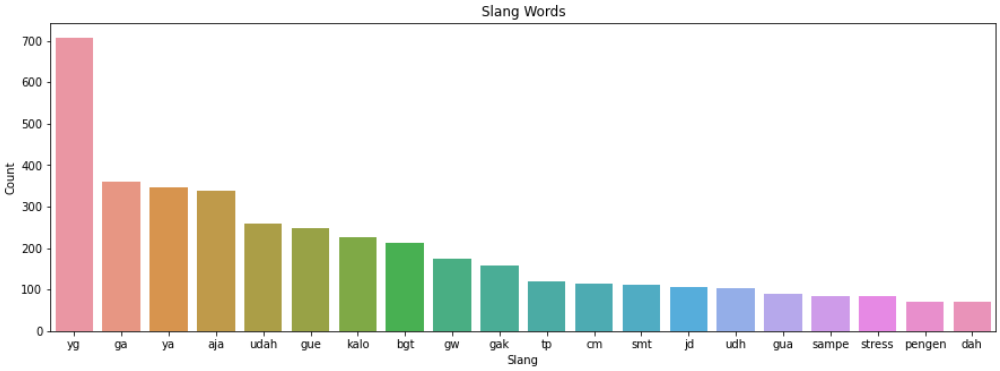
D. Identifikasi *Slang* Pada Kalimat

Data setelah melewati proses *preprocessing* selanjutnya untuk melakukan identifikasi *slang* dapat dilakukan dengan mengidentifikasi slang yang sudah terdaftar dalam kamus *slang* yang dibentuk sebelumnya. Proses Identifikasi *slang* yang terdaftar pada kata dilakukan dengan menelusuri setiap kata pada kalimat dan jika ditemukan kata yang terdaftar pada kamus slang pada kalimat maka kata tersebut akan disimpan pada file baru. Hasil dari proses identifikasi *slang* menggunakan kamus *slang* yang didapatkan dari 1550 kalimat yaitu sebanyak 8813 slang. *Slang* yang berhasil ditemukan dapat diperhatikan pada Gambar 2.



Gambar 2. Contoh *Slang Word*

Jika diperhatikan pada *wordcloud* yang ada pada Gambar 3, dapat diamati bahwa terdapat dominasi penggunaan *slang* yang berbentuk singkatan kata dalam kalimat. *Slang* dengan bentuk singkatan ini secara konsisten muncul lebih sering daripada jenis *slang* lainnya dalam teks yang dianalisis. Fenomena ini menunjukkan bahwa penggunaan singkatan dalam *slang* memiliki prevalensi yang signifikan dalam data yang diamati. *Slang* berupa singkatan seringkali digunakan untuk menyampaikan pesan dengan singkat dan cepat, sekaligus mencerminkan dinamika dan perubahan bahasa yang terus berkembang. Jumlah kemunculan dari *slang* dapat diperhatikan pada Gambar. Tujuan dari analisis ini adalah ternyata kondisi dalam kamus data menggunakan kata bukan singkatan, sehingga berdasarkan analisis ini ini perlu diarahkan bahwa singkatan tersebut memiliki mengacu ke makna kata tertentu.

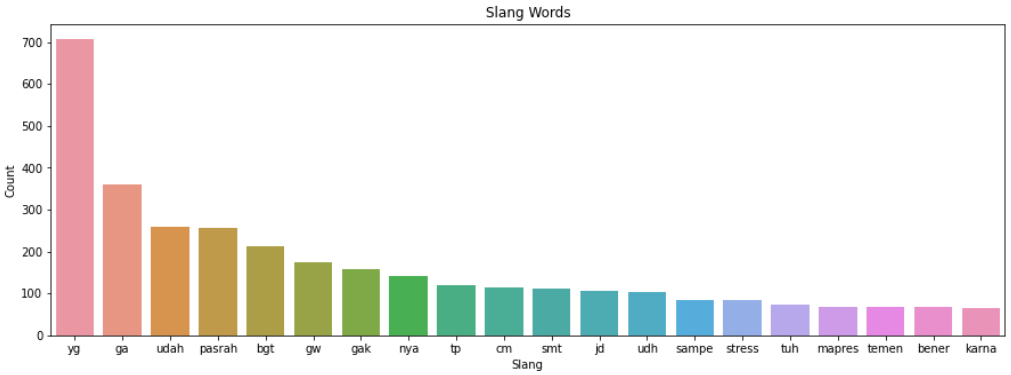


Gambar 3. Grafik *Wordcloud*

Pada tahap selanjutnya dari penelitian ini, identifikasi *slang* dilakukan menggunakan pendekatan yang lebih formal dan ilmiah yaitu dengan menerapkan metode *stemming* untuk menentukan kata dasar dari setiap kata dalam kalimat yang dianalisis. Pada penelitian ini memanfaatkan pretrain model bahasa indonesia yang telah disediakan oleh fasttext. Pretrain model merupakan model yang telah dilatih sebelumnya dengan jutaan corpus bahasa indonesia sehingga menghasilkan model dengan representasi vektor yang lebih akurat serta mampu mencari kemiripan suatu kata dengan lebih akurat. Proses *stemming* bertujuan untuk menyederhanakan kata-kata dalam bentuk yang lebih standar dan memperoleh akar kata yang sesuai dengan kamus kata dasar yang telah ditetapkan. Hasil *stemming* ini kemudian dibandingkan dengan kumpulan kata-kata dalam kamus kata dasar. Dalam prosedur ini apabila kata hasil *stemming* tidak terdapat dalam kamus kata dasar, maka diasumsikan bahwa kata tersebut termasuk ke dalam kategori *slang*. Dalam konteks ini kata *slang* akan dianggap sebagai kata-kata yang tidak dapat direduksi menjadi bentuk dasar yang tercatat dalam kamus kata. Oleh karena itu, kata-kata yang teridentifikasi sebagai *slang* akan dipisahkan dan disimpan dalam file baru sebagai hasil dari proses identifikasi slang yang berlangsung. Jumlah data *slang* yang ditemukan dari 1550 data *twit* adalah sebanyak 11479 *slang*. Hasil dari identifikasi kata–kata *slang* yang ditemukan dengan metode perbandingan hasil *stemming* dan kamus kata dasar dapat diperhatikan pada Gambar 4.

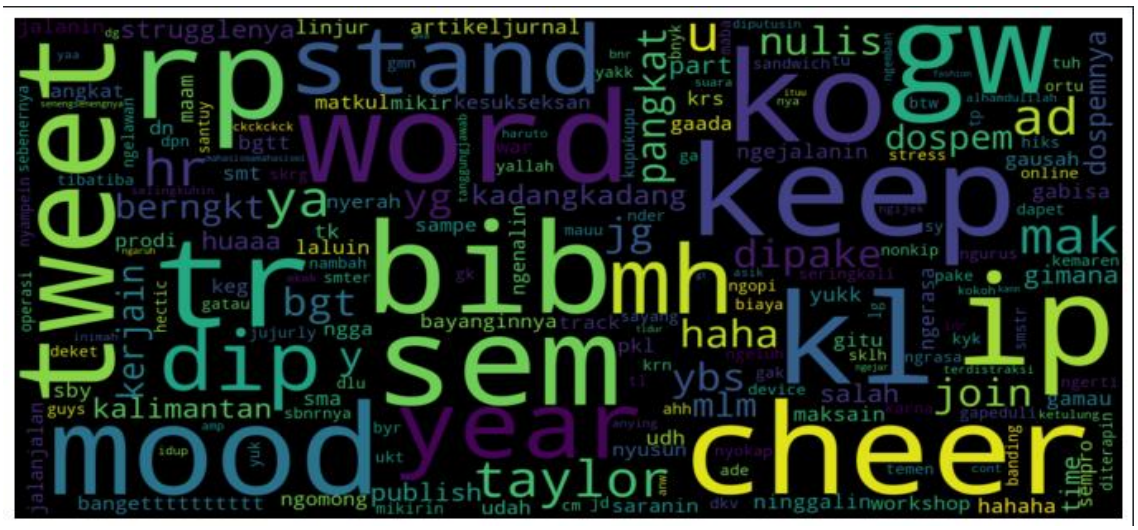
Gambar 4. Hasil Identifikasi *Slang*

Melalui analisis *wordcloud* pada Gambar 3, dapat dilihat bahwa data *slang* yang diperoleh dari proses identifikasi *slang* dengan penerapan teknik *stemming* dan kamus kata dasar menunjukkan variasi yang lebih kaya dan beragam. Penggunaan metode *stemming* dalam penelitian ini bertujuan untuk menemukan akar kata atau bentuk dasar dari kata-kata dalam kalimat yang dianalisis. Selanjutnya hasil stemming tersebut diarahkan untuk dibandingkan dengan sekumpulan kata-kata dalam kamus kata dasar yang telah ditetapkan sebelumnya. Penerapan pendekatan ini memungkinkan penemuan berbagai variasi *slang* yang terdapat dalam kalimat yang sedang diuji karena proses *stemming* membantu mengidentifikasi beragam bentuk turunan atau variasi morfologi dari kata *slang* tersebut. Kamus kata dasar yang digunakan sebagai acuan dalam proses perbandingan memberikan kerangka kerja yang solid dan terstandarisasi dalam mengklasifikasikan kata-kata sebagai *slang* atau bukan. Jumlah kemunculan *slang* hasil identifikasi dengan proses *stemming* dan pencarian pada kamus kata dasar dapat diperhatikan pada Gambar 4.



Gambar 4. Grafik *Slang* Hasil Stemming

Data slang yang berhasil terkumpul melalui kedua metode identifikasi *slang* kemudian dilakukan penggabungan sehingga membentuk satu kesatuan data *slang* yang lebih lengkap. Proses penggabungan ini juga melibatkan langkah-langkah untuk menangani duplikasi data *slang* yang terdeteksi. Ketika terdapat kata slang yang muncul lebih dari satu kali dalam hasil identifikasi dari kedua metode hanya satu kemunculan dari kata *slang* tersebut yang dipilih untuk dimasukkan ke dalam data slang yang disatukan. Proses pengambilan salah satu dari kata slang yang memiliki kemunculan ganda ini bertujuan untuk memastikan bahwa data slang yang disajikan dalam satu kesatuan tetap berkualitas, terstruktur, dan tidak redundan. Selain itu langkah ini juga berperan dalam menjaga konsistensi dan validitas data slang yang digunakan dalam analisis selanjutnya. Dengan demikian data *slang* yang telah diolah dan dipersiapkan secara cermat menjadi lebih representatif dan memungkinkan analisis yang lebih akurat terhadap keberagaman serta distribusi penggunaan slang dalam teks yang dianalisis. Jumlah *slang* total yang diperoleh yaitu sebanyak 3239. Hasil dari slang yang dikumpulkan dapat diperhatikan pada Gambar 5.



Gambar 5. *Slang* Hasil Penggabungan

E. Normalisasi *Slang*

Normalisasi *slang* dengan *Pretrain* model *fasttext* menghasilkan kata – kata dengan ketekatan yang paling tinggi dengan *slang. Fasttext* juga mampu mencari alternatif kata yang paling cocok untuk menggantikan kata-kata *slang* yang teridentifikasi. Penggunaan representasi vektor kata dalam pemrosesan bahasa memungkinkan model untuk mencocokkan *slang* dengan kata-kata baku yang memiliki kemiripan semantik yang tinggi. Hasil dari normalisasi yang berhasil dilakukan dengan pretrain model *fasttext* dapat diperhatikan pada Tabel 4.

Tabel 4. Normalisasi Slang

No	*Slang*	*Normalized*
1	jg	juga
2	join	ikut
2	nulis	nulis
4	y	ya
5	haha	tertawa
6	yg	yang
7	bgt	banget
8	berngkt	berangkat
9	mlm	malam
10	dospem	dosen pembimbing

Dari hasil normalisasi yang dilakukan tidak semua slang dapat normalisasi. Masih banyak *slang* yang tidak bisa dinormalisasi dengan baik oleh model *fasttext.* Dari hasil normalisasi yang dilakukan masih terdapat 1329 kata *slang* yang tidak berhasil dinormalisasi dengan benar. Data yang tidak mampu dinormalisasi dengan benar dapat diperhatikan pada Tabel.

No	*Slang*	*Normalized*
1	dipake	dipake.
2	pangkat	pangkatnya
3	ybs	ybs.
4	nulis	nulis-nulis
5	kerjain	ngerjain
6	kadangkadang	Kadangkadang
7	artikeljurnal	jurnaljurnal
8	hahaha	haha
9	ninggalin	tinggalin
10	to	To

Hasil dari penelitian menunjukkan bahwa dengan menggunakan algoritma *fasttext*, sebagian besar *slang* dapat dinormalisasi dengan baik. Namun ditemukan bahwa ada beberapa kasus *slang* yang sulit atau bahkan tidak dapat sepenuhnya dinormalisasi. Jumlah data *slang* yang tidak berhasil dinormalisasi adalah sebesar 41%. Masih banyaknya *slang* yang tidak bisa dinormalisasi diakibatkan oleh performa model *fasttext*. Keberhasilan dan kinerja algoritma *fasttext* sangat bergantung pada dua aspek penting yaitu jumlah data pelatihan yang memadai dan kualitas variasi slang yang tercakup dalam data tersebut. Jika data pelatihan yang digunakan terbatas dalam mencakup variasi kata slang yang beragam, algoritma kemungkinan tidak akan mampu mengenali dan mengatasi secara efektif berbagai variasi slang yang ada. Hal ini dapat menyebabkan kegagalan dalam proses normalisasi *slang*, di mana beberapa kata slang mungkin tidak dapat sepenuhnya dinormalisasi atau bahkan mungkin disalahartikan.

Faktor lain yang mungkin menjadi penyebab kesalahan dalam normalisasi yaitu beberapa kata slang dalam bahasa mungkin memiliki karakteristik yang menarik, di mana beberapa di antaranya bisa menjadi homonim yaitu kata-kata dengan bunyi yang sama tetapi memiliki makna yang berbeda atau

polisemi yaitu kata-kata yang memiliki banyak makna yang berbeda. Kehadiran homonim dan polisemi dalam slang dapat menjadi tantangan bagi algoritma yang digunakan untuk memahami dan menormalisasi teks yang mengandung istilah-istilah ini. Penggunaan slang yang ambigu, di mana satu kata slang bisa memiliki banyak arti berbeda tergantung pada konteksnya bisa membuat algoritma fasttext atau algoritma pemrosesan bahasa alami lainnya menghadapi kesulitan dalam menentukan makna yang sesuai. Ketika algoritma dihadapkan pada situasi semacam ini kesalahan interpretasi dapat terjadi mengakibatkan normalisasi yang tidak tepat atau pemahaman yang salah terhadap makna sebenarnya dari kata slang tersebut.

4. SIMPULAN DAN SARAN

Dari penelitian yang dilakukan untuk upaya melakukan identifikasi dan normalisasi bentuk dari *slang* pada data *twiter* ditemukan bahwa untuk melakukan normalisasi pada data *slang* diperlukan pengenalan dan pelatihan tersendiri untuk model *fasttext* agar bisa melakukan normalisasi *slang* ke bentuk yang lebih standar dengan lebih baik. Proses identifikasi *slang* pada kalimat juga berperan penting dalam proses normalisasi *slang.* Hasil penelitian menunjukkan bahwa penggunaan metode *stemming* dan kamus kata dasar yang dikombinasikan dengan metode pencarian dengan kamus *slang* merupakan pendekatan yang efektif dalam mengidentifikasi berbagai variasi slang yang ada. Penerapan metode ini membantu mendeteksi akar kata dari slang yang kemudian dibandingkan dengan kamus kata dasar untuk mengklasifikasikan kata-kata sebagai slang atau tidak. Meskipun demikian metode pencarian *slang* dengan memanfaatkan kamus harus sering diperharui karena *slang* terus berubah dan berkembang seiring waktu dan tren penggunaan bahasa. Hasil penelitian yang dilakukan, disarankan untuk melakukan optimasi pada algoritma *fasttext* agar dapat melakukan normalisasi dengan lebih baik. Model *fasttext* memerlukan pengenalan dan pelatihan yang lebih khusus. Oleh karena itu disarankan untuk melakukan eksplorasi lebih lanjut dalam pengembangan model *fasttext* yang dioptimalkan untuk normalisasi *slang*. Proses pelatihan ini dapat melibatkan data pelatihan yang berfokus pada berbagai variasi *slang* yang relevan dengan platform *twitter*. Selain itu mungkin dapat dilakukan metode normalisasi yang berbeda seperti metode *Bidirectional Encoder Representations from Transformers* (BERT) yang sering digunakan untuk *lexical normalization* pada teks bahasa inggris. Mungkin dapat disarankan juga untuk melakukan perbaruan pada kamus slang yang digunakan dalam proses pencarian slang. Hal ini akan memastikan bahwa kamus slang selalu terkini dan mencakup kata-kata slang terbaru, sehingga meningkatkan akurasi dan relevansi dalam identifikasi slang.

UCAPAN TERIMAKASIH

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa atas rahmat dan anugerah-Nya. Pada kesempatan ini penulis hendak menyampaikan ucapan terimakasih kepada:

1. Dosen ilmu komputer dan teknologi rekayasa perangkat lunak, serta pembimbing yang banyak mengarahkan dan memberikan masukan ke penulis sampai revisi artikel ini.
2. Orang tua dan saudara penulis yang telah memberikan dukungan dan doa dalam pembuatan artikel ini
3. Teman-teman penulis yang telah saling mendukung dan memotivasi satu sama lain dalam pembuatan artikel ini

Besar harapan kami artikel ini akan memberikan manfaat bagi pembaca ke depannya.

DAFTAR PUSTAKA

[1] W. Trimastuti, “an Analysis of Slang Words Used in Social Media,” *J. Dimens. Pendidik. dan Pembelajaran*, vol. 5, no. 2, pp. 64–68, 2017, doi: 10.24269/dpp.v5i2.497.

[2] M. Oktaviana, Z. A. Achmad, H. Arviani, and K. Kusnarto, “Budaya komunikasi virtual di Twitter dan Tiktok: Perluasan makna kata estetik,” *Satwika Kaji. Ilmu Budaya dan Perubahan Sos.*, vol. 5, no. 2, pp. 173–186, 2021, doi: 10.22219/satwika.v5i2.17560.

[3] R. C. Cenderamata, “Abreviasi dalam Percakapan Sehari-Hari di Media Sosial: Suatu Kajian Morfologi,” *Metahumaniora*, vol. 8, no. 2, p. 238, 2018, doi: 10.24198/mh.v8i2.20699.

[4] S. Irawan, I. N. Sudika, and R. Hidayat, “Karakteristik Bahasa Gaul Remaja sebagai Kreativitas Berbahasa Indonesia pada Komentar Status Inside Lombok di Instagram,” *J. Bastrindo*, vol. 1, no. 2, pp. 201–213, 2020, doi: 10.29303/jb.v1i2.44.

[5] Z. R. N. S. Prasetija, A. Romadhony, and E. B. Setiawan, “Analisis Pengaruh Normalisasi Teks pada Klasifikasi Sentimen Ulasan Produk Kecantikan,” *e-Proceeding Eng.*, vol. 9, no. 3, pp. 1769–1775, 2022, [Online]. Available:

https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18184/17795.

[6] R. Riyaddulloh and A. Romadhony, “Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: Tweet Produk Gadget Pada Twitter,” *eProceedings Eng.*, vol. 8, no. 4, pp. 4216–4228, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15246/14969.

[7] M. A. Nur and N. Wardhani, “Optimasi Normalisasi Kata Pada Data Twitter Untuk Meningkatkan Akurasi Analisis Sentimen (Studi Kasus Respon Masyarakat Terhadap Layanan Teman Bus),” *J. Fokus Elektroda*, vol. 07, no. 04, pp. 237–243, 2022, [Online]. Available: https://elektroda.uho.ac.id/index.php/journal/article/view/21%0Ahttps://elektroda.uho.ac.id/index.php/journal/article/download/21/15.

[8] F. Zuhad and N. Wilantika, “Perbandingan Penggunaan Kamus Normalisasi dalam Analisis Sentimen Berbahasa Indonesia,” *J. Linguist. Komputasional*, vol. 5, no. 1, pp. 13–23, 2022.

[9] T. Malik Iryana and P. Pandu Adikara, “Analisis Sentimen Masyarakat Terhadap Mass Rapid Transit Jakarta Menggunakan Metode Naïve Bayes Dengan Normalisasi Kata,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 6, pp. 2548–964, 2021, [Online]. Available: http://j-ptiik.ub.ac.id.

[10] K. K. Agustiningsih, E. Utami, and M. A. Alsyaibani, “Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings,” *J. Ilmu Komput. dan Inf.*, vol. 15, no. 1, pp. 39–46, 2022, doi: 10.21609/jiki.v15i1.1044.

[11] P. Mojumder, M. Hasan, M. F. Hossain, and K. M. A. Hasan, “A study of fasttext word embedding effects in document classification in bangla language,” *Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST*, vol. 325 LNICST, no. March 2021, pp. 441–453, 2020, doi: 10.1007/978-3-030-52856-0_35.

[12] A. G. D’Sa, I. Illina, and D. Fohr, “BERT and fastText Embeddings for Automatic Detection of Toxic Speech,” *Proc. 2020 Int. Multi-Conference Organ. Knowl. Adv. Technol. OCTA 2020*, 2020, doi: 10.1109/OCTA49274.2020.9151853.

[13] L. Mutawalli, M. T. A. Zaen, and W. Bagye, “KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto),” *J. Inform. dan Rekayasa Elektron.*, vol. 2, no. 2, p. 43, 2019, doi: 10.36595/jire.v2i2.117.