Sistem Perbaikan Kata Tidak Baku Bahasa Indonesia Menggunakan Metode *Crowdsourcing*

http://dx.doi.org/10.28932/jutisi.v5i3.1983

Danny Sebastian ✉ #1, Kristian Adi Nugraha#2

#*Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana*
Jl. Dr. Wahidin Sudirohusodo 5-25 Yogyakarta 55224
[1]danny.sebastian@staff.ukdw.ac.id
[2]adinugraha@staff.ukdw.ac.id

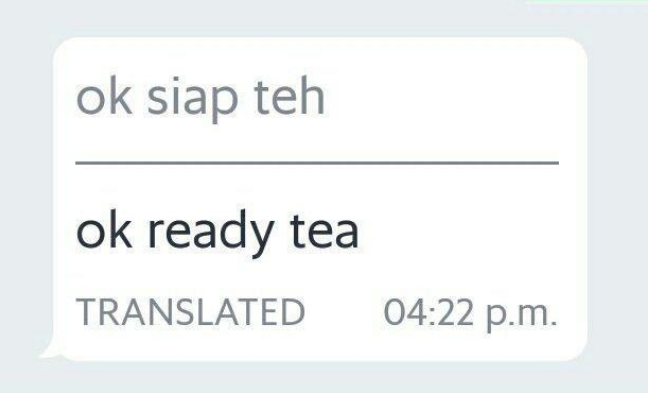
Abstract — **Most languages have two forms of words: standard and not-standard. Standard words usually used in formal conversation, while non-standard words used in informal conversation. The massive use of non-standard words in the text-based communication through social media application causing a new problem for its text-processing features such as language translator or artificial intelligence (AI) conversation program (chatbots). Our research purpose is to create a system that can convert any non-standard Indonesian word into standard Indonesian word using crowdsourcing method, a community-based method to create a database that contains non-standard Indonesian words and their associated standard words. From the result, only 59.67% of testing data can be converted perfectly into a standard form of Indonesian language. The rest sentences had several problems so it can't properly be processed by the system.**

Keywords— **Indonesian language; normalization; crowdsourcing;**

I. PENDAHULUAN

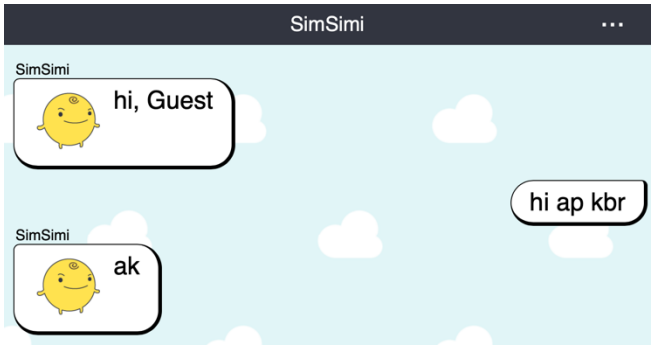
Pada umumnya, kata yang digunakan dalam sebuah bahasa dapat dikategorikan menjadi dua jenis, yaitu kata baku dan kata tidak baku. Demikian juga pada Bahasa Indonesia, kata baku merupakan kata yang digunakan untuk percakapan yang bersifat resmi atau formal. Sedangkan kata tidak baku merupakan kata yang digunakan dalam situasi yang tidak resmi atau informal. Kata tidak baku terbagi menjadi beberapa kategori di dalamnya, salah satunya adalah *slang*. *Slang* merupakan jenis bahasa informal yang digunakan di dalam lingkup internal kelompok sosial tertentu. Namun karena keberadaan media seperti televisi atau aplikasi media sosial, tidak jarang penggunaan bahasa *slang* tersebut akhirnya menyebar ke orang-orang di luar kelompok tersebut, bahkan hingga ke tingkat nasional. Sehingga bahasa *slang* yang awalnya hanya digunakan di dalam lingkup internal dapat dijumpai dalam komunikasi sehari-hari di tengah masyarakat, baik untuk komunikasi secara lisan maupun secara tertulis melalui media sosial. Selain bahasa *slang*, bahasa daerah yang digunakan bersamaan dengan Bahasa Indonesia juga dikategorikan sebagai bahasa tidak baku, karena penggunaan bahasa campuran tersebut hanya dapat dilakukan untuk percakapan yang bersifat informal.

Penggunaan bahasa *slang* pada media sosial menimbulkan permasalahan baru bagi aplikasi atau mesin pengolah teks, seperti aplikasi translasi antar bahasa (mis. *Google Translate*) dan aplikasi pencarian (*search engine*). Hal tersebut dapat terjadi karena kata-kata dalam bahasa *slang* tidak dapat dikenali oleh basis *data* pada aplikasi atau mesin pengolah teks tersebut. Solusi yang dapat digunakan untuk mengatasi permasalahan tersebut adalah dengan menambahkan *data* kata-kata *slang* pada basis *data*, sehingga kata-kata tersebut akan dapat dikenali oleh aplikasi atau mesin pengolah teks. Namun permasalahan berikutnya adalah kosa kata pada bahasa *slang* selalu bertambah setiap waktu, sehingga perlu dilakukan pembaruan *data* setiap kali menemukan kosa kata baru. Selain bahasa *slang*, penggunaan Bahasa Indonesia yang dicampur dengan bahasa lain seperti Bahasa Inggris atau Bahasa Daerah juga dapat menimbulkan kesalahan interpretasi, contohnya seperti kesalahan transalsi pada sebuah aplikasi ditunjukkan pada Gambar 1. Karena pada dasarnya, mesin pengolah teks yang ada saat ini, contohnya Google Translate, hanya dapat memproses satu jenis bahasa saja di dalam sebuah teks.



Gambar 1. Kesalahan translasi pada sebuah aplikasi

Permasalahan lain yang disebabkan oleh penggunaan kata-kata tidak baku juga dialami oleh aplikasi-aplikasi pembalas pesan otomatis berbasis kecerdasan buatan atau biasa disebut sebagai *chatbots*. Karena *chatbots* hanya mampu membalas pesan dengan masukan kata-kata yang telah ditentukan sebelumnya, di mana sebagian besar kata-kata tersebut merupakan kata baku. Sehingga ketika *chatbots* menerima masukan berupa kata tidak baku, maka kata tersebut tidak dapat dikenali oleh *chatbots*, akibatnya *chatbots* tidak akan dapat membalas pesan dengan baik seperti ditunjukkan pada gambar 2.



Gambar 2. Kesalahan jawaban pada *chatbots*

Berdasarkan permasalahan yang telah dikemukakan sebelumnya, penulis memiliki gagasan untuk membangun sebuah sistem yang dapat menerjemahkan kata tidak baku ke dalam bentuk baku, sehingga kalimat yang mengandung kata tidak baku tersebut dapat digunakan untuk keperluan formal. Kata-kata yang dianggap tidak baku oleh penulis adalah kata-kata yang tidak terdapat pada Kamus Besar Bahasa Indonesia (KBBI). Kata-kata tersebut meliputi bahasa *slang*, singkatan, kata-kata dalam bahasa daerah, dan kata-kata dalam bahasa asing. Cara yang ditempuh oleh penulis untuk mendapatkan basis *data* mengenai kata-kata tidak baku adalah dengan melibatkan peran responden untuk memperbarui basis *data* tersebut, cara ini disebut dengan *crowdsourcing*. Mula-mula, sistem akan menerima input berupa kalimat mentah yang berisi kata baku maupun tidak baku. Kemudian sistem akan melakukan pengecekan untuk setiap kata yang ada pada kalimat tersebut, apakah kata yang hendak dicek terdapat pada basis *data* Kamus Besar Bahasa Indonesia (KBBI) atau tidak. Jika tidak, maka kata tersebut akan dikategorikan sebagai kata tidak baku dan akan dimasukkan ke dalam daftar antrian terjemahan. Seluruh kata yang terdapat pada daftar antrian ini akan ditanyakan kepada orang-orang yang telah terdaftar pada sistem sebagai responden, kemudian responden tersebut memasukkan bentuk baku dari kata-kata tidak baku yang ditanyakan. Karena menggunakan metode *crowdsource*, maka sistem ini terbuka untuk umum, artinya siapapun dapat berpartisipasi sebagai responden untuk menerjemahkan kata-kata tidak baku menjadi kata baku. Hasil terjemahan kata baku dari kata tidak baku akan diolah lebih lanjut menggunakan pendekatan statistika sebelum dimasukkan ke dalam basis *data*. Hal ini bertujuan untuk mengantisipasi apabila terdapat dua atau lebih responden yang memasukkan terjemahan kata baku berbeda untuk kata tidak baku yang sama, sehingga sistem harus dapat menentukan kata baku mana yang akan digunakan untuk menerjemahkan kata tidak baku tersebut. Sistem ini tidak hanya menangani kata tidak baku saja, namun kata informal lain seperti singkatan juga turut ditangani. Karena penggunaan singkatan dalam komunikasi tertulis juga dapat dikategorikan sebagai bahasa yang tidak baku, sehingga hal ini juga akan ditangani oleh sistem. Responden dapat mengkategorikan kata-kata yang tidak baku sebagai *slang*, kata asing, kata singkatan, atau tidak tahu. Selain menyimpan kata asli dan kata dalam bentuk baku, sistem juga akan menyimpan kata-kata asli yang sudah berada dalam bentuk baku. Kata-kata ini nantinya akan digunakan sebagai kata kunci dalam menentukan bentuk baku mana yang akan digunakan apabila kata tersebut memiliki lebih dari satu bentuk baku. Harapan penulis, sistem ini dapat digunakan untuk membantu masyarakat dalam mengolah informasi dalam bentuk teks Bahasa Indonesia dari bentuk tidak baku menjadi bentuk baku, sehingga teks tersebut lebih mudah diolah lebih lanjut karena teks sudah dalam bentuk formal atau resmi.

II. TINJAUAN PUSTAKA

Slang merupakan permasalahan yang muncul pada pengolahan teks, dan sampai saat ini belum ada sumber yang lengkap untuk kosa kata *slang* [1]. Penelitian ini bertujuan untuk mengumpulkan kosa-kata yang termasuk *slang* menggunakan *data* dari twitter. Frameworks yang digunakan terdiri dari 3 langkah utama, yaitu pembersihan token menggunakan *word shrinking algorithm*, *auto correction*, dan pengecekan kata dengan korpus kata baku dari *Indonesian Online Dictionary*. Penelitian ini menemukan bahwa metode *word shrinking algorithm* dan *auto correction* menghasilkan *success rate* sebesar 98.97%.

Media sosial memiliki informasi yang sangat banyak dan menarik untuk menjadi sumber *data information extraction* dan *text mining*. Konten yang berasal dari media sosial memiliki banyak sekali *noise* dan menghambat proses pengolahan teks secara otomatis. Secara umum, penduduk Indonesia menggunakan 5 gaya penulisan konten pada media sosial twitter oleh orang Indonesia, yaitu menggunakan Bahasa yang baku, Bahasa daerah, Bahasa asing, singkatan, dan *slang* [2]. Penelitian yang dilakukan oleh Ahmad Fathan Hidayatullah juga menemukan 16 karakteristik tweet orang Indonesia yang merupakan kombinasi dari 5 gaya penulisan konten media sosial oleh orang Indonesia.

Penelitian yang dilakukan oleh Bo Han, Paul Cook, dan Timothy Baldwin melakukan normalisasi token menggunakan metode *contextual and word similarity* [3]. Kelemahan dari penelitian ini adalah masih belum dapat

melakukan normalisasi kata yang bersifat ambigu atau tidak dapat melihat konteks.

Penelitian yang megolah kata *slang* juga pernah dilakukan oleh Dian Sa'adillah Maylawati, Wildan Budiawan Zulfikar, Cepy Slamet, Muhammad Ali Ramdhani, Yana Aditia Gerhana pada penelitiannya yang berjudul "An Improved of Stemming Algorithm for Mining Indonesian Text with *Slang* on Social Media". Penelitian ini mengusulkan sebuah metode untuk melakukan stemming ke kumpulan kata dari media sosial yang mengandung kata *slang* [4]. Metode yang ditawarkan adalah algoritma *stemming* dikembangkan dari Porter *Stemmer*. Metode yang ditawarkan menghasilkan akurasi sebesar 88.65%.

Dataset adalah sekumpulan *data* yang sudah diverifikasi kebenarannya dan dapat digunakan dalam penelitian sebagai sumber *data* yang valid [5]. Penelitian yang dilakukan oleh Antonius R dan Yuan L menghasilkan *dataset* status dan komentar yang sudah diberi label sentiment (positif, negatif, atau netral). Metode yang digunakan pada penelitian ini adalah *crowdsource labelling* menggunakan *weightened majority voting*. *Dataset* yang dihasilkan sejumlah 3400 komentar dari 68 status, dan divalidasi dengan tingkat konfiden 95% dan validitas sebesar 95.3%.

Penelitian terkait pembentukan *dataset* juga pernah dilakukan menggunakan metode *TF-IDF* dan *Cosine Similarity* [6]. Metode TF-IDF digunakan untuk megolah *data* yang berasal dari media social twitter. Setelah diujikan ke kelompok *data* dengan jumlah tertentu, penelitian ini menghasilkan akurasi tertinggi sebesar 64% untuk jumlah tweet sebesar 200 tweet.

Crowdsourcing method adalah metode untuk memecahkan permasalahan yang tidak dapat diselesaikan oleh komputer dengan cara meminta banyak manusia untuk mengerjakan suatu *task*/pekerjaan [7]. Pekerjaan yang didistribusikan, biasanya merupakan pekerjaan yang membutuhkan *supervise* manusia. Dalam satu dekade terakhir, metode *crowdsource* menjadi populer karena adanya teknologi *internet* yang memungkinkan manusia bekerja tanpa terbatas jarak.

Berdasarkan tinjauan pustaka yang sudah dilakukan, maka diperoleh kesimpulan bahwa metode *crowdsourcing* dapat digunakan untuk memperbaiki kata yang tidak baku. Perbedaan penelitian aktual dengan penelitian terdahulu dapat dilihat pada TABEL I dan TABEL II.

TABEL I
PERBEDAAN PENELITIAN AKTUAL DENGAN PENELITIAN TERDAHULU

Judul	*Generating Indonesian slang lexicons from twitter*	*Language Tweet Characteristics of Indonesian Citizens*	*Lexical Normalization for Social Media Text*	Sentipol: *Dataset* Komentar pada Kampanye Pemilu Presiden Indonesia 2014 dari Facebook Page
Oleh	Wahyu Muliady & Harya Widiputra	Ahmad Fathan Hidayatullah	Bo Han, Paul Cook, & Timothy Baldwin	Antonius Rachmat & Yuan Lukito
Sumber	*International Conference on Uncertainty Reasoning and Knowledge Engineering*	*International Conference on Science and Technology*	*ACM Transactions on Intelligent Systems and Technology*	Prosiding Konfrensi Nasional Teknologi Informasi dan Komunikasi (KNASTIK 2016)
Tahun	2012	2015	Januari 2013	2016
Metode	*Word skhrinking Algorithm & Auto Correction*	-	*Contextual and word similarity*	*Crowdsourcing menggunakan Weighted Majority Voting*
Obyek	*Data* tweet	*Data* tweet	New York Times, *data* tweet, SMS *corpus*	*Data* komentar dari *facebook page*

TABEL II
PERBEDAAN PENELITIAN AKTUAL DENGAN PENELITIAN TERDAHULU (LANJUTAN)

Judul	*An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media*	Pembentukan *Dataset* Topik Kata Bahasa Indonesia pada Twitter Menggunakan *TF-IDF* & *Cosine Similarity*	*Challenges in Data Crowdsourcing*	Pembentukan *Dataset Slang* Kata Bahasa Indonesia Menggunakan Metode Crowdsourcing
Oleh	Dian Sa'adillah Maylawati, Wildan Budiawan Zulfikar, Cepy Slamet, Muhammad Ali	Kristian Adi Nugraha & Danny Sebastian	H Gracia-Molina, Monas Joglekar, Adam Marcus, Aditya Parameswaran, & Vasilis Verroios	Danny Sebastian & Kristian Adi Nugraha

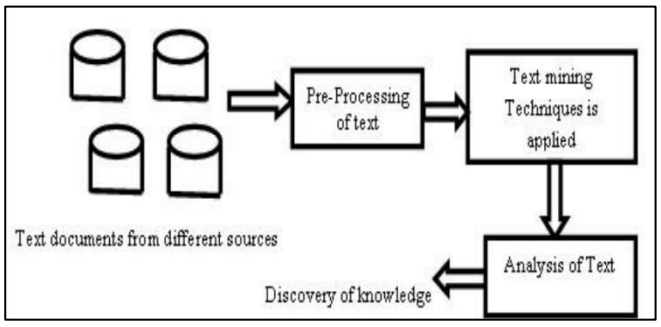
	Ramdhani, Yana Aditia Gerhana			
Sumber	*The 6th International Conference on Cyber and IT Service Management*	Jurnal Teknik Informatika dan Sistem Informasi (JuTISI 2018)	*IEEE Transaction on Knowledge and Data Engineering*	-
Tahun	2018	Desember 2018	2016	2019
Metode	*Improved Porter Algorithm*	*TF-IDF & Cosine Similarity*	*Crowdsourcing*	*Crowdsourcing*
Obyek	*Data* teks media sosial	*Data* tweet akun berita	-	*Data* komentar Instagram

III. LANDASAN TEORI

A. Text Mining

Text data mining sering disingkat dengan *text mining*, merupakan bidang ilmu yang mempelajari *information retrieval*, *text analysis*, *information extraction*, *clustering*, *categorization*, *visualization*, *database technology*, *machine learning*, dan *data mining* [6]. Dua proses dalam *text mining* adalah *text refining* dan *knowledge destillation*. *Text refining* merupakan proses awal dalam *text mining*, dimana proses ini merubah data mentah atau tidak terstruktur menjadi data terstruktur. Sedangkan *knowledge destillation* merupakan proses mengolah data menjadi sebuah pengetahuan.

Dalam beberapa tahun terakhir, banyak sekali penelitian *text mining* yang menggunakan sumber data dari media sosial [8] [9] dan *e-commerce* [10]. *Text mining* terdiri dari 3 langkah utama, yaitu langkah *pre-processing*, kemudian dilanjutkan dengan mengaplikasikan algoritma *text mining*, dan diakhiri dengan proses analisis. Langkah *text mining* dapat dilihat pada Gambar 3.



Gambar 3. Proses *text mining*
Dikutip dari: Vijayarani, S., Ms J. Ilamathi, and Ms Nithya, 2015, *Preprocessing Techniques for Text Mining – An Overview* [11]

B. Text Preprocessing

Langkah pertama dalam *text mining* adalah *text pre-processing*. Langkah *text pre-processing* bertujuan untuk mempersiapkan data teks menjadi token yang siap diolah pada langkah selanjutnya, langkah implementasi algoritma *text mining*. Langkah dari *text-preprocessing* sangat berpengaruh terhadap keberhasilan algoritma *text mining* pada tahap selanjutnya [12].

Beberapa hal yang dilakukan pada tahap text *pre-processing* adalah tokenisasi, pembersihan karakter spesial, menyetarakan kapitalisasi huruf, proses *filtering*, *stemming*, dan lematisasi [10] [11]. Tokenisasi adalah proses mengubah dokumen teks menjadi token atau kata [13]. Proses *filtering* pada *text pre-processing* dilakukan untuk menghilangkan token yang tidak penting dan dapat mengurangi keberhasilan algoritma *text mining*. Salah satu proses *filtering* adalah *stopwords removal* atau proses menghilangkan *stopword* [12]. Selain meningkatkan tingkat keberhasilan dari algoritma *text mining*, proses *filtering* juga digunakan untuk mengurangi dimensi token yang diolah sehingga dapat mempercepat waktu pemrosesan algoritma *text mining* [11].

Stemming adalah metode *text pre-processing* yang digunakan untuk menghasilan *stem/root*/kata dasar dari sebuah *token* [11] [14]. Sebagai contoh: token 'mencoba' akan dirubah menjadi 'coba'. Tujuan dari *stemming* adalah mengurangi jenis token yang diproses dalam *text mining* dengan cara menggabungkan beberapa token yang memiliki makna yang sama. Beberapa metode stemming adalah Porter *Stemmer Algorithm* [15], Nazief & Adriani *Stemmer* [16], Lovins [17], Lucene *Stemmer* [18] dan lain sebagainya. Sedangkan lematisasi merupakan proses mengubah sebuah kata menjadi bentuk yang sesuai (*lemma*), sehingga dapat dikelompokkan dengan kata lain yang sama [12]. Lematisasi bertujuan untuk mengubah *infinite tense* dan *noun* menjadi sebuah kata dalam Bahasa inggris yang sama. Dalam penelitian ini, lematisasi tidak digunakan karena dalam Bahasa Indonesi tidak memiliki bentuk khusus yang perlu diproses seperti dalam Bahasa Inggris [19].

C. Specific Text Preprocessing

Text pre-processing juga terus berkembang, ada *text pre-processing* yang bersifat spesifik untuk karakteristik data teks yang akan diproses [20]. Dalam kasus media sosial, ada beberapa karakter spesial yang tidak dapat dihapus secara langsung, seperti tanda '@' untuk menandakan *mention*, tanda '#' untuk menandakan *hashtag*. Atau untuk lebih

p-ISSN : 2443-2210
e-ISSN : 2443-2229

spesifik lagi, pada twitter terdapat istilah re-tweet atau RT [21].

Secara umum, media sosial akan mengandung karakter emoticon atau emoji dan alamat URL. Emoticon juga dapat diolah untuk menghasilkan informasi emosi dari teks [22].Pengolahan emoticon biasanya dilakukan menggunakan metode *corpus based* [11] [20].

D. Karakteristik konten media sosial orang Indonesia

Konten yang ada di media sosial mengandung banyak *noise* karena terkadang ditulis dalam bahasa yang tidak formal [20] [23] [24]. Singkatan, kesalahan ketik, dan bahasa *slang* adalah contoh *noise* yang sering terjadi pada konten media sosial [25]. Secara umum ada 5 gaya penulisan konten media sosial yang dilakukan oleh orang Indonesia, yaitu menggunakan kata-kata yang baku, menggunakan kata-kata dalam bahasa daerah, menggunakan kata bahasa asing, menggunakan kata singkatan, menggunakan kata-kata gaul [2]. Contoh penggunaan gaya penulisan orang Indonesia di media sosial dapat dilihat pada TABEL III. Pada praktiknya, penggunaan kelima gaya penulisan tersebut dapat dikombinasikan satu sama lain. Seperti penggunaan kata-kata baku yang ditambahkan dengan beberapa kata *slang*, atau penggunaan kata baku, kata *slang*, dan kata singkatan dalam sebuah konten teks. Dalam beberapa kasus, proses melakukan normalisasi teks tersebut dilakukan dengan cara mengganti kata yang tidak baku dengan kata yang baku, seperti kata 'bisaaaaa' menjadi 'bisa' [26].

TABEL III
CONTOH GAYA PENULISAN PADA MEDIA SOSIAL ORANG INDONESIA
(SUMBER: W. Muliady and H. Widiputra, "*Generating Indonesian Slang Lexicons from Twitter*," in *International Conference on Uncertainty Reasoning and Knowledge Engineering*, 2012) [1]

Karakteristik	Contoh
Kata baku	"belum daftar ospek satupun!". Kalimat ini dibentuk dari kata bahasa Indonesia yang baku
Kata bahasa daerah	"**mari** ospek **loro**". Kalimat ini dibentuk dari bahasa daerah, yaitu bahasa jawa
Kata bahasa asing	"**fix** ospek jalan kaki…". 'fix' merupakan bahasa asing.
Kata singkatan	"**brpa** hari pra ospek?". 'brpa' merupakan singkatan dari 'berapa' "**tgl** 20 persiapan ospek". 'tgl' merupakan singkatan dari 'tanggal'
Kata *slang*	"**hayuuk** minggu depan …". 'hayuk' merupakan kata *slang* yang berasal dari kata baku 'ayo'.

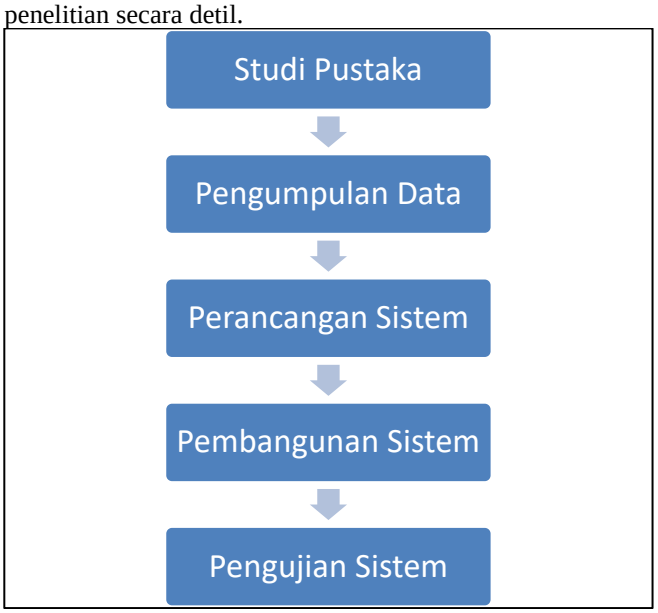
E. Crowdsourcing Method

Crowdsourcing method adalah metode memecahkan permasalahan dengan cara membagi pekerjaan ke banyak orang [7]. Metode *crowdsourcing* memungkinkan kita membagi pekerjaan melalui perangkat komputer dan *internet*. Metode *crowdsource* memunkinkan pemrosesan yang hanya dapat dilakukan oleh manusia, seperti membersihkan data tidak terstruktur [27]. Data tersebut dapat digunakan menjadi latih data untuk *machine learning*.

Tetapi metode *crowdsource* memiliki kelemahan, seperti lambat dan mahalnya pekerja manusia. Selain itu mungkin terjadi kesalahan/*error* pada saat melakukan pemrosesan. Oleh karena itu, *crowdsourcing method* yang baik memerlukan metode untuk mengevaluasi dan melakukan *filter* jawaban yang tidak sesuai.

IV. METODE PENELITIAN

Langkah penelitian terbagi kedalam 6 tahap utama seperti pada Gambar 4. Pada bagian ini dijelaskan langkah penelitian secara detil.



Gambar 4. Langkah Penelitian

A. Studi Pustaka

Langkah awal penelitian ini adalah melakukan studi pustaka mengenai metode yang sudah pernah pernah dilakukan oleh peneliti lain untuk mengolah kata-kata yang tidak baku. Selain itu peneliti juga mempelajari karakteristik data komentar Instagram sebagai sumber data yang akan diproses.

B. Pengumpulan Data

Penelitian ini menggunakan data yang diambil dari komentar Instagram. Akun yang dipilih adalah 10 selebgram yang berasal dari Indonesia dengan jumlah *follower* di atas satu juta, dengan pertimbangan semakin banyak jumlah follower maka semakin variatif data komentar yang ada di dalamnya. Selain memiliki jumlah *follower* di atas satu juta, akun tersebut harus membuka fitur komentar pada akun Instagram, karena sebagian besar akun menutup fitur tersebut agar orang lain tidak bisa memberikan komentar. Sepuluh akun selebgram yang menjadi sumber data adalah:

1. Atta Halilintar (@attahalilintar)
2. Raisa Andriana (@raisa6690)
3. Raditya Dika (@raditya_dika)
4. Ayu Tingting (@ayutingting92)
5. Laudya Cynthia Bella (@laudyacynthiabella)
6. Deddy Corbuzier (@mastercorbuzier)
7. Prilly Latuconsina (@prillylatuconsina96)
8. Luna Maya (@lunamaya)
9. Agnes Monica (@agnezmo)
10. Chelsea Olivia (@chelseaoliviaa)

Dari masing-masing akun selebgram tersebut, diambil 10 posting terakhir. Semua komentar diambil dari masing-masing posting tersebut. Semua komentar yang berisi nomor *handphone* dan alamat *website* akan dianggap sebagai *spam*/iklan, sehingga akan dihilangkan dari sumber data. TABEL IV merupakan contoh data yang dihapus karena berisikan nomor *handphone* dan alamat *website*.

TABEL IV
CONTOH DATA KOMENTAR YANG DIANGGAP SEBAGAI SPAM DAN DIHAPUS

No	Komentar spam
1	Awesome kak... Ingin pasang berita kabar? Kunjungi aja @halokuindo web kami **www.haloku.com**. Kami juga lagi ngadain give away lo...
2	Yuk gabung dengan hubungi **0812-9850-5883** solusi kantor virtual & pendirian PT/CV paling terjangkauâ€¦ Dijamin puas dan bersahabat!
3	MELAYANI : PINDAHAN RUMAH,KANTOR,APARTMEN, BARANG LOGISTIK,PUING TANAH DLL. soal harga flexibel nego aja JAWA BALI BORNEO SUMATRA (WA **089622944558**)
4	Cek postingan terbaru! Open order jasa sewa & hias seserahan. Lengkap dengan mahar juga loh. Pricelist: **0857 2628 9579** Thankyou!
5	Paketan bali murah cuma 1.5an aja cus, di ceki ceki @babehwisata **0812.9769.1160**

C. Perancangan Sistem

Setelah data dikumpulkan, penulis melakukan perancangan sistem yang akan dibangun. Terdiri dari perancangan *database* dan perancangan antar muka aplikasi. Perancangan *database* dilakukan untuk menghasilkan struktur data komentar Instagram dan jawaban dari responden. Hasil perancangan antarmuka halaman *crowdsource* dapat dilihat pada Gambar 5.

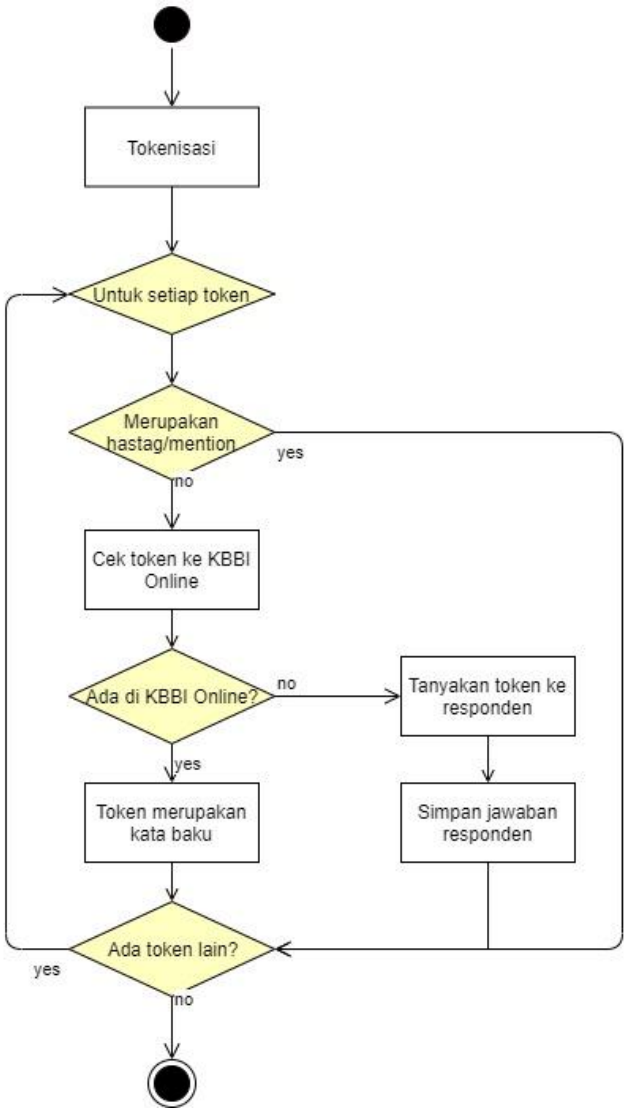
Gambar 5. Antarmuka halaman *crowdsourcing*

D. Pembangunan Sistem

Tahap selanjutnya adalah pembangunan sistem. Sistem dibangun dengan *platform website* dan menggunakan Bahasa pemrograman PHP. *Platform website* dipilih karena aplikasi website akan terhubung ke jaringan *internet*, sehingga proses komunikasi untuk pengumpulan data dapat dilakukan dengan lebih mudah.

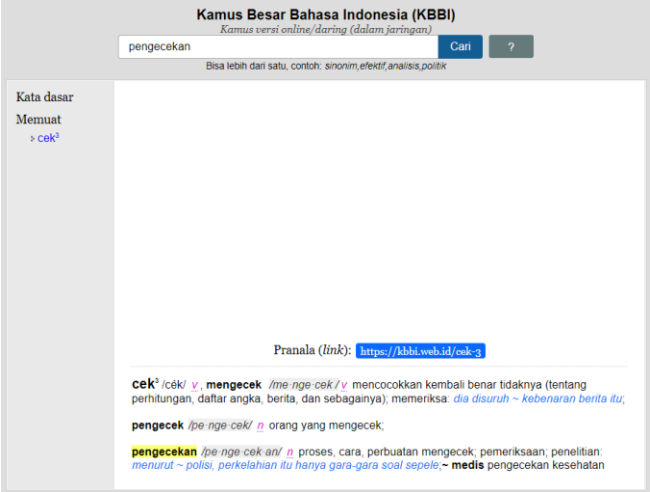
Fungsi utama dari sistem yang dibangun adalah *crowdsourcing*. Proses *crowdsource* dapat dilihat pada Gambar 6.

p-ISSN : 2443-2210
e-ISSN : 2443-2229



Gambar 6. Proses pengolahan data menggunakan *crowdsourcing*

Langkah pertama dimulai dengan melakukan tokenisasi kalimat kedalam bentuk token. Pada tahap ini dilakukan pembersihan karakter spesial dan tanda baca. Kemudian dari setiap token yang dihasilkan, sistem akan melakukan pengecekan awal, apakah token tersebut merupakan *hastag* atau *mention*. Apabila token tersebut merupakan *hastag* atau *mention*, maka tidak akan diproses lebih lanjut. Sedangkan bila token tersebut bukan *hastag* atau *mention*, maka sistem akan melakukan pengecekan ke website "Kamus Besar Bahasa Indonesia Online" atau "KBBI Online" (https://kbbi.web.id/) dengan menggunan PHP CURL. Contoh halaman KBBI Online dapat dilihat pada Gambar 7. Apabila token tersebut terdaftar di website "KBBI Online", maka token tersebut akan ditandai sebagai kata baku. Sedangkan apabila kata tersebut tidak terdaftar di KBBI Online, maka sistem akan menampilkan form untuk menanyakan ke responden, apakah bentuk baku dari token tersebut? Setiap jawaban responden akan disimpan oleh sistem. Sistem secara otomatis melakukan perhitungan voting jawaban dari setiap responden dan menentukan bentuk baku dari kata/token yang tidak terdaftar di "KBBI Online".



Gambar 7. Halaman KBBI Online

Sebagai contoh, diambil teks "Ciee yg pose muka nya udah berubah semenjak di komen mas panji" dari Gambar 5. Sistem melakukan tokenisasi dan menghasilkan 12 kata. Kemudian sistem melakukan perulangan untuk masing-masing token, misal token 'Ciee', token akan diperiksa ke website "KBBI Online". Pada token ini, "KBBI Online" tidak memberikan kembalian, yang artinya token tidak terdaftar sebagai kata baku di "KBBI Online". Token ini ditanyakan ke responden. Idealnya responden akan menjawab token 'Ciee' merupakan kata *slang*, dan sistem akan menanyakan apakah responden mengetahui kata baku dari token tersebut. Apabila responden mengetahui bentuk baku dari token tersebut, maka responden akan diminta untuk mengisikan kata baku dari token yang ditanyakan. Pada Gambar 5 dicontohkan responden tidak mengetahui bentuk baku dari kata 'Ciee'. Untuk pengolahan token-token lain, dapat dilihat pada TABEL V

TABEL V
CONTOH PENGOLAHAN TOKEN PADA PROSES *CROWDSOURCING*

Token	Pengecekan KBBI Online	Simpulan
Ciee	Tidak ada	Sesuai voting responden *crowdsourcing*
yg	Tidak ada	Sesuai voting responden *crowdsourcing*
pose	Ada	Kata baku
muka	Ada	Kata baku
nya	Ada	Kata baku
udah	Tidak ada	Sesuai voting responden *crowdsourcing*
berubah	Ada	Kata baku
semenjak	Ada	Kata baku
di	Ada	Kata baku
komen	Tidak ada	Sesuai voting responden *crowdsourcing*
mas	Ada	Kata baku
panji	Ada	Kata baku

Selanjutnya sistem akan menghitung setiap jawaban dari responden. Sistem melihat jumlah voting dari setiap

responden dan mengambil simpulan berdasarkan voting terbanyak. Simpulan yang diambil adalah 'jenis kata' dan 'perbaikan bentuk baku'. TABEL VI menunjukkan contoh voting jawaban dan simpulan yang diambil oleh sistem dengan 3 responden.

TABEL VI
CONTOH PENGOLAHAN VOTING RESPONDEN PADA PROSES *CROWDSOURCING*

Token	Resp 1	Resp 2	Resp 3	Simpulan
Ciee	*slang*	*slang*	tidak tahu	*Slang* (3 vote)
	tidak tahu	tidak tahu	tidak tahu	tidak tahu bentuk baku (3 vote)
yg	singkatan	singkatan	singkatan	singkatan (3 vote)
	yang	yang	yng	yang (2 vote)
udah	singkatan	*slang*	*slang*	*slang* (2 vote)
	sudah	sudah	tidak tahu	sudah (2 vote)
komen	singkatan	singkatan	singkatan	singkatan (3 vote)
	komentar	komentar	komentar	komentar (3 vote)

Setelah data dikumpulkan, penulis melakukan validasi data secara manual. Penulis mengelompokkan jawaban responden yang memiliki arti yang sama, tetapi dituliskan secara berbeda oleh responden. Misal kata 'aq', responden menjawab 'aku' dan 'saya'. Kedua token tersebut akan memiliki arti yang sama dan oleh penulis akan disamakan menjadi 'saya'. Penulis juga memperbaiki kesalahan pengetikan seperti pada contoh diatas, token 'yg' memiliki 2 jenis respon yaitu 'yang' dan 'yng'. Respon 'yng' dapat dianggap sebagai kesalahan pengetikan pada responden.

E. Pengujian Sistem

Pengujian sistem dilakukan dengan cara memverifikasi hasil pelabelan data yang berhasil dikumpulkan dengan metode *crowdsourcing*. Terdapat tiga kategori label yaitu kata *slang*, singkatan, dan kata dalam bahasa asing (termasuk bahasa daerah). Dari kategori tersebut, penulis akan melakukan analisa kata-kata yang tidak berhasil dilabelkan dengan baik, yaitu apabila sebuah kata tidak baku masuk ke dalam dua atau lebih kategori dengan jumlah vote masing-masing yang cukup besar. Sistem juga memiliki satu kategori tambahan yaitu kategori 'tidak tahu', di mana kategori ini berisi kata-kata tidak baku yang dinilai salah oleh responden, namun responden tidak tahu bagaimana pembetulan kata-kata tersebut.

V. ANALISIS & PEMBAHASAN

A. Pengumpulan data menggunakan Crowdsourcing

Pengujian dilakukan dengan menggunakan data uji yang terdiri dari 15.687 *data* yang telah dikumpulkan sebelumnya. Data uji berasal dari *data* komentar instagram yang berasal dari akun-akun selebgram yang berasal dari Indonesia dengan jumlah follower di atas satu juta orang. Dari setiap akun, diambil sepuluh *post* terakhir untuk tiap akun, kemudian seluruh komentar dari setiap *post* akan diambil untuk dijadikan data uji. Total seluruh komentar yang berhasil dikumpulkan berjumlah 15.687 data. Dari seluruh data komentar tersebut, komentar-komentar yang mengandung alamat URL, alamat *e-mail*, dan nomor telepon akan dihapus, karena jika sebuah komentar mengandung satu atau lebih komponen tersebut, maka dapat dipastikan bahwa komentar tersebut merupakan iklan yang tidak perlu diolah lebih lanjut. Dari seluruh *data* yang berhasil dikumpulkan, setelah melalui tahap pembersihan *data*, maka terkumpul *data* komentar sebanyak 5.739 *data*. Dari seluruh *data* tersebut, penulis hanya mengambil komentar yang memiliki jumlah kata di atas 10 kata dengan jumlah sebanyak 2567 komentar. Tujuannya adalah agar komentar yang diolah merupakan kalimat utuh dengan struktur yang lengkap, sehingga peluang untuk menemukan kata-kata yang tidak baku juga akan semakin besar.

Setelah data dibersihkan, penulis melakukan proses *crowdsourcing* sesuai dengan proses pada Gambar 6. Proses *crowdsourcing* dilakukan pada bulan Juli 2019 sampai dengan Agustus 2019. Responden terdiri dari 17 orang yang berusia antara 18-30 tahun, dengan latar belakang pendidikan terakhir SMA, SMK, D3 dan S1. Penulis memastikan setiap responden dapat berkomunikasi dengan Bahasa Indonesia.

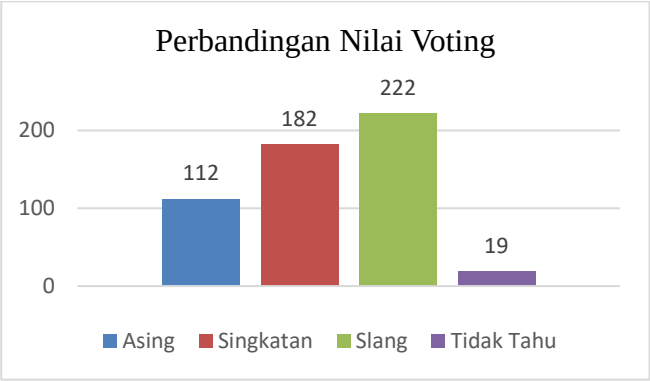
Terdapat beberapa permasalahan selama proses pengumpulan data menggunakan metode *crowdsourcing*, antara lain:

- Terdapat beberapa kali kegagalan pada saat proses pengecekan kata baku ke sistem "KBBI Online", sehingga kata yang seharusnya merupakan kata baku dianggap tidak baku oleh sistem dan ditanyakan ke responden. Berdasarkan *log file* yang dihasilkan, hal ini terjadi karena adanya kegagalan komunikasi ke sistem "KBBI Online" menggunakan PHP CURL. Jawaban responden kata baku tersebut dihapus dari *dataset* yang dihasilkan dan dianlisis.
- Terdapat beberapa kata yang mengekspresikan sesuatu dan tidak ada di "KBBI Online", seperti kata orang tertawa: 'hahaha', 'hahahah', 'wkwkwk', dan lain sebagainya. Ada banyak kata yang menunjukkan sebenarnya kata tersebut adalah sama arti, tetapi sistem memproses data tersebut sebagai kata yang berbeda. Hal ini juga menyebabkan responden juga kebingungan harus memberikan respon seperti apa.
- Responden melakukan *misstype*/salah ketik untuk kata perbaikan. Seperti kata singkatan 'utk', responden memahami bahwa perbaikan katanya adalah 'untuk', tetapi responden mengisi dengan 'untul'. Pada kasus ini, penulis tidak melakukan perbaikan pengetikan, dimana data tetap digunakan, karena kesalahan pengetikan ini dapat diselesaikan dengan metode *voting* pada *crowdsourcing*.

p-ISSN : 2443-2210
e-ISSN : 2443-2229

B. Analisis data hasil Crowdsourcing

Berdasarkan hasil pengujian, penulis mendapatkan 309 jenis kata tidak baku. Setiap kata tidak baku dapat memiliki satu atau lebih bentuk baku. Misalnya kata tidak baku 'bio' yang memiliki bentuk baku 'biodata' dan 'biografi'. Setiap kata bentuk baku memiliki nilai voting yang menggambarkan berapa banyak responden yang setuju dengan bentuk baku tersebut. Misalnya kata baku 'biodata' memiliki nilai voting singkatan sebesar 2. Artinya 2 orang menyetujui bahwa 'biodata' merupakan kepanjangan dari kata singkatan 'bio'. Total perbandingan keseluruhan nilai voting ditunjukkan pada Gambar 8.



Gambar 8. Grafik total perbandingan keseluruhan nilai voting

Berdasarkan hasil pengujian yang telah dilakukan, kata paling banyak yang dikategorikan sebagai kata tidak baku adalah kata hubung seperti 'yang' dan kata tunjuk seperti 'di'. Hampir seluruh kata hubung 'yang' ditulis sebagai singkatan menjadi kata 'yg', demikian juga kata tunjuk 'di' ditulis sebagai singkatan menjadi 'd'. Beberapa contoh *dataset* yang berhasil didapatkan dapat dilihat pada TABEL VII.

TABEL VII
CONTOH *DATASET* YANG DIHASILKAN OLEH SISTEM *CROWDSOURCE*

Kata tidak baku	Bentuk baku	
	Responden	Simpulan penulis
ama	sama, dengan	dengan
aja	saja	saja
aq	aku, saya	saya
banget, bangt	sekali, sangat	sekali
bb	berat badan	berat badan
cumak, cuman	cuma, hanya	hanya
bs, bsa	bisa	bisa
carik, cr	cari	cari
chanel	channel, saluran	saluran
dah	udah, sudah	sudah
diamah	kalau dia	kalau dia
doang, doank	saja	saja
emang, emmg	memang	memang
ga, g, gak, engga	tidak	tidak
gibah	gosip	gosip
gaess, gaes	teman, teman-teman	teman
gimana, gmn	bagaimana	bagaimana
gua, gw, gue	saya, aku	saya
hape	handphone, telepon genggam	telepon genggam
i	saya	saya
iyaa, iye	iya	iya
klo, klw, kl	kalau	kalau
langsunggg	langsung	langsung
makasi, makasih	terima kasih	terima kasih
mantul	mantap betul, mantap sekali	mantap sekali
mingguu, mggu	minggu	minggu
mintak, mnta	minta, meminta	minta
nanya	bertanya, tanya	tanya
nyampek	sampai	sampai
pengen, pgn	ingin, menginginkan	ingin
seken	bekas, second	bekas
sbenernya	sebenarnya	sebenarnya
sdhkah	sudahkah, sudah	sudahkah
syetan	setan	setan
thank, thanks, thks	terima kasih, thank you	terima kasih
thn, thun	tahun	tahun
tp, tpi	tapi, tetapi	tapi
u	kamu	kamu
udah, udh	udah, sudah	sudah
yaa	ya	ya

Penulis melakukan pengujian menggunakan 300 kalimat uji yang telah dipersiapkan sebelumnya. Berdasarkan hasil pengujian tersebut, hanya 179 kalimat (59.67%) yang dapat diproses dengan sempurna. Sisa kalimat sebanyak 121 kalimat (40.33%) memiliki beberapa kondisi yang tidak dapat ditangani oleh sistem, sehingga perbaikan kalimat yang dihasilkan oleh sistem dinilai kurang tepat.

Penulis menemukan beberapa jenis kata tidak baku yang tidak dapat ditangani oleh sistem, salah satunya adalah kata dengan huruf yang berulang hingga beberapa kali seperti ditunjukkan pada tabel TABEL VIII. Kata-kata jenis ini menjadi sulit ditangani karena jumlah huruf yang berulang bisa berbeda antara satu kata dengan yang lain, seperti kata 'apaaa' dan 'apaaaaaaaaa', sementara keduanya memiliki bentuk baku yang sama yaitu 'apa'. Untuk memperbaiki kata-kata tersebut tidak dapat dilakukan hanya dengan mereduksi jumlah huruf yang berulang menjadi satu huruf saja, karena terdapat beberapa kata baku yang memiliki dua huruf berulang seperti kata 'maaf', jika tetap menggunakan cara tersebut maka kata 'maaaaaaaaaaf' akan berubah menjadi kata 'maf'. Dibutuhkan pendekatan lain untuk menentukan berapa jumlah akhir huruf berulang yang hendak direduksi agar mendapatkan kata baku yang tepat.

TABEL VIII
CONTOH KATA TIDAK BAKU DENGAN HURUF BERULANG

Kata Tidak Baku	Bentuk Baku
maaaaaaaaaaf	maaf
tidaaaaaaak	tidak
apaa	apa
ashiaaaaaaaap	siap
aaammmiiinnn	amin
langsunggg	langsung

Permasalahan lain yang ditemukan oleh penulis adalah adanya beberapa kata yang merupakan kata *slang*, namun kata tersebut memiliki bentuk kata baku dengan arti yang berbeda. Beberapa contoh kata-kata jenis tersebut ditunjukkan pada TABEL IX. Salah satu contohnya adalah kata 'cabut', di mana kata ini merupakan kata baku di dalam Bahasa Indonesia yang memiliki arti 'menarik hingga terlepas'. Namun jika digunakan sebagai bahasa *slang*, kata 'cabut' memiliki arti lain yaitu 'pulang'. Demikian juga dengan kata 'ember' yang memiliki arti 'wadah', jika digunakan dalam bahasa *slang* memiliki arti 'memang'. Selain itu terdapat kata-kata dalam bahasa lain seperti Bahasa Jawa yang menyerupai kata dalam Bahasa Indonesia, namun memiliki arti berbeda. Contohnya adalah kata 'tak' yang biasa digunakan sebagai pengganti kata 'saya', misalnya 'tak pikir' yang berarti 'saya pikir' dan 'tak kira' yang berarti 'saya kira'. Kata 'tak' sendiri dalam Bahasa Indonesia merupakan kata baku yang memiliki arti 'tidak'. Agar dapat mengatasi masalah-masalah tersebut, sistem harus memiliki kemampuan untuk membaca konteks isi dari kalimat yang sedang diproses.

TABEL IX
CONTOH KATA TIDAK BAKU YANG MENYERUPAI KATA BAKU

Kata Tidak Baku	Bentuk Baku
cabut	pulang
ember	memang benar
tak	saya

Permasalahan lain yang serupa dengan kata-kata pada TABEL X adalah adanya beberapa kata yang merupakan nama, namun bentuknya menyerupai singkatan. Salah satu contohnya adalah kata 'bca' yang merupakan nama sebuah bank. Apabila kata ini diolah menggunakan basis *data* yang ada, maka kata ini akan dianggap sebagai sebuah kata singkatan, kemudian kata tersebut akan diterjemahkan menjadi kata 'baca'. Contoh kata lain dengan permasalahan serupa ditunjukkan pada TABEL X.

TABEL X
CONTOH KATA BAKU YANG DIANGGAP TIDAK BAKU

Kata Asli	Obyek yang memungkinkan	Bentuk Baku oleh Sistem
bca	Bank BCA	baca
bri	Bank BRI	beri
tpi	Nama stasiun TV Swasta	tapi

Permasalahan lain juga ditemukan pada kata tidak baku yang terdapat salah pengetikan, seperti "*give away*". "*give away*" merupakan kata kesatuan yang memiliki arti 'hadiah' atau 'pemberian hadiah'. Di sini sistem menanyakannya sebagai 2 kata yang terpisah, '*give*' dan '*away*'. Sehingga responden kesulitan untuk memberikan kata baku yang sesuai. Hasil yang didapatkan oleh sistem, '*give away*' diartikan secara terpisah oleh responden, kata '*give*' memiliki arti 'memberi', sedangkan kata '*away*' memiliki arti 'pergi'. Contoh lain adalah "nge gym", sistem memisahkan kedua kata tersebut menjadi 'nge' dan 'gym', sehingga responden memberikan jawaban secara terpisah. TABEL XI menunjukkan beberapa contoh kasus yang terjadi untuk permasalah tersebut.

TABEL XI
CONTOH KATA YANG SALAH DITOKENISASI OLEH SISTEM

Kata salah tokenisasi oleh sistem	Tokenisasi sistem	Bentuk baku seharusnya
Buat **give away** dong om yang menang jadi member gratis gym disitu om wkwkwkwkwk @mastercorbuzier	2 kata, yaitu '*give*' dan '*away*'	Hadiah/pemberian hadiah
thanks u so much, master !!!! saya harap selalu ada buku free seperti ini aaammmiiinnn	4 kata, yaitu 'thanks', 'u', '*so*' dan '*much*'	Terima kasih banyak
Saya udah latihan lama tapi kenapa gk berbentuk **six pack** ya bantu jawab bro	2 kata, yaitu 'six' dan 'pack'	-
Nge gym dalam islam itu di larang @mastercorbuzier soalnya menyakiti diri sendiri	2 kata, yaitu 'nge' dan 'gym'	Berolah-raga
Intinya biar **Stand out** dimana pun berada harus pakai sesuatu yg berbeda, mulai dari style yg **antimainstream**	2 kata, yaitu '*stand*' dan '*out*'	menonjol
	1 kata, yaitu 'antimainstream'	tidak sesuai trend

VI. KESIMPULAN & SARAN

Secara keseluruhan sistem telah mampu mengumpulkan daftar kata tidak baku beserta bentuk baku dari kata tersebut dengan baik. Namun terdapat beberapa potensi permasalahan yang belum dapat ditangani oleh sistem dengan baik, sehingga hanya 59.67% dari data uji yang dapat ditangani oleh sistem dengan baik. Beberapa masalah tersebut antara lain seperti kata tidak baku dengan huruf yang terduplikasi, kata tidak baku yang menyerupai kata baku, serta kata baku yang menyerupai kata tidak baku. Untuk mengatasi permasalahan tersebut, dibutuhkan metode tambahan agar sistem mampu memahami konteks kalimat, sehingga sistem mampu menentukan apakah kata-kata tidak baku di dalam sebuah kalimat benar-benar merupakan kata tidak baku, maupun untuk menentukan kata baku yang paling tepat untuk kata-kata di dalam kalimat tersebut. Salah satu metode yang dapat digunakan untuk menyelesaikan masalah tersebut adalah dengan menggunakan metode *n-gram*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Fakultas Teknologi Infomasi Universitas Kristen Duta Wacana dan LPPM Universitas Kristen Duta Wacana yang telah mendukung kegiatan penelitian ini sehingga dapat terlaksana dengan baik. Penelitian ini didasari oleh kontrak 079/D01/LPPM/2019 antara penulis dengan LPPM Universitas Kristen Duta Wacana.

DAFTAR PUSTAKA

[1] W. Muliady and H. Widiputra, "Generating Indonesian Slang Lexicons from Twitter," *International Conference on Uncertainty Reasoning and Knowledge Engineering,* pp. 123-126, 2012.

[2] A. F. Hidayatullah, "Language Tweet Characteristics of Indonesian Citizens," *2015 International Conference on Science and Technology (TICST),* pp. 397-401, 2015.

[3] B. Han, P. Cook and T. Baldwin, "Lexical Normalization for Social Media Text," *ACM Transactions on Intelligent Systems and Technology,* vol. 4, no. 1, pp. 5-32, 2013.

[4] D. S. Maylawati, W. B. Zulfikar, C. Slamet, M. A. Ramdhani and Y. A. Gerhana, "An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media," *International Conference on Cyber and IT Service Management,* pp. 1-6, 2018.

[5] A. Rachmat and Y. Lukito, "Sentipol: Dataset Sentimen Komentar Pada Kampanye Pemily Presiden Indonesia 2014 dari Facebook Page," *Konfrensi Nasional Teknologi Informasi dan Komunikasi (KNASTIK 2016),* pp. 218-228, 2016.

[6] K. A. Nugraha and D. Sebastian, "Pembentukan Dataset Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 4, no. 3, pp. 376-386, 2018.

[7] H. Gracia-Molina, M. Joglekar, A. Marcus, A. Parameswaran and V. Verroios, "Challenges in Data Crowdsourcing," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 4, pp. 901-911, 2016.

[8] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business horizons,* vol. 53, no. 1, pp. 59-68, 2010.

[9] A. R. Chrismanto and Y. Lukito, "Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes," *Jurnal Informatika dan Sistem Informasi,* vol. 2, no. 2, pp. 26-34, 2016.

[10] D. Sebastian, "Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari beberapa E-marketplace," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 5, no. 1, pp. 51-61, 2019.

[11] S. Vijayarani, J. Ilamathi and Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks,* vol. 5, no. 1, pp. 7-16, 2015.

[12] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919,* 2017.

[13] S. A. Salloum, M. Al-Emran, A. A. Monem and K. Shaalan, "Using text mining techniques for extracting information from research articles," *Intelligent Natural Language Processing: Trends and Applications,* pp. 373-397, 2018.

[14] J. Asian, H. E. Williams and S. M. M. Tahaghoghi, "Stemming Indonesian," *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science,* vol. 38, pp. 307-314, 2005.

[15] M. F. Porter, "An Algorithm for Suffix Stripping," *Electron. Libr. Inf. Syst.,* vol. 40, no. 3, pp. 211-218, 2006.

[16] A. G. Jivani, "A Comparative of Stemming Algorithms," *IJCTA,* vol. 2, no. 6, pp. 1930-1938, 2011.

[17] J. B. Lovins, "Development of a stemming algorithm," *Mech. Transl. Comput. Linguist.,* vol. 11, pp. 22-31, 1968.

[18] F. Z. Tala, *A study of stemming effects on information retrieval in Bahasa Indonesia,* Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands, 2003.

[19] K. A. Nugraha and D. Sebastian, "Analisis Trend Akun Media Sosial Twitter Menggunakan TF-IDF dan Cosine Similarity," *Prosiding Seminar Nasional ReTII ke-13 2018,* pp. 103-110, 2018.

[20] A. Hidayatullah and M. Ma'arif, "Pre-processing task in Indonesian Twitter Messages," *Journal of Physics: Conference Series,* vol. 801, no. 1, p. 012072, 2017.

[21] I. Hemalatha, G. S. Varma and A. Govardhan, "Pre-processing the informal text for efficient Sentiment Analysis," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. 1, no. 2, pp. 58-61, 2012.

[22] W. Wolny, "Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms," *International Conference On Information Systems Development,* pp. 476-483, 2016.

[23] E. Tuncdemir, A. Akbarov, K. Gonen and H. Aydogan, "Discourse of chatting habit on writing," *International Journal of Linguistics, Literature and Translation,* vol. 3, no. 1, pp. 575-583, 2014.

[24] F. Liu, F. Weng and X. Jiang, "A broad-coverage normalization system for social media language," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers,* vol. 1, pp. 1035-1044, 2012.

[25] T. Baldwin, M. C. d. Marneffe, B. Han, Y.-B. Kim, A. Ritter and W. Xu, "Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition," *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text,* pp. 126-135, 2015.

[26] M. A. Fauzi, . R. F. N. Firmansyah and T. Afirianto, "Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion," *TELKOMNIKA,* vol. 16, no. 3, pp. 1345-1350, 2018.

[27] N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi and S. Merugu, "A web of concepts," *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,* pp. 1-12, 2009.