Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 2. November 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Improving the accuracy of sentiment analysis using slang words lexicon and spelling correction

I Komang Surya Adinandika[a1], I Gusti Agung Gede Arya Kadyanan[a2]

[a]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Bali, Indonesia
[1]kmsurya.adi44@gmail.com
[2]gungde@unud.ac.id

Abstract

Text pre-processing has long been a research subject to improve accuracy of Natural Language Processing models. In this paper we propose a technique for text sentiment classification with extra steps on text pre-processing using slang word lexicon and spelling correction to annotate non-formal Indonesian text and normalize them. This study aims to improve the accuracy of sentiment analysis models by strengthening text pre-processing methods. We compared the performance of these preprocessing methods using 2 popular classification algorithms: Support Vector Machine (SVM) and Naïve Bayes, and 3 different feature extraction methods: term presence, Bag of Words, and TF-IDF. Model was trained and tested with 1705 datasets of twitter posts from Indonesian users about Covid 19.

Keywords: *pre-processing, slang, lexicon, Indonesian, spelling, correction*

1. Introduction

Manusia sebagai makhluk sosial, selalu berinteraksi dengan manusia lainnya. Interaksi tersebut dapat berbentuk sebagai suara, teks maupun gerakan. Pada zaman yang modern ini, komunikasi dalam bentuk teks menjadi pilihan pertama sebagai media interaksi. Interaksi dengan media teks pada zaman ini telah berkembang, diawali dengan penggunaan SMS hingga media sosial yang berbasis internet. Internet sebagai pintu gerbang komunikasi manusia, memungkinkan manusia dapat berinteraksi dimanapun dan kapan saja secara bebas.

Kebebasan ini mendorong banyak individu berekspresi dengan bahasa yang mereka gunakan dalam berinteraksi. Bahasa Indonesia umumnya terbagi menjadi 2 bentuk, yaitu formal dan non-formal [1], dimana pada umumnya bahasa non-formal tersebut digunakan dalam percakapan sehari-hari dan media sosial (Facebook, Twitter, Instagram dan lainnya) [2]. Dari penelitian [1], masyarakat Indonesia menggunakan 67.18% bahasa formal, 12.23% bahasa non-formal dan 20.57% tak diketahui. Dimana dari data yang tidak diketahui tersebut dapat berupa kata dari bahasa lain, singkatan ataupun *slang* (kata gaul). Dari data tersebut, dapat dikatakan karena tidak sesuai dengan EYD, kata tersebut merupakan kata non-formal. Kemudahan penggunaan dan kesederhanaan bahasa tersebut membuat bahasa "*slang*" sering digunakan dalam komunikasi [3].

Lebih dari sepertiga penggunaan bahasa tersebut menggunakan bahasa non-formal. Hal ini merupakan salah satu permasalahan dalam *Natural Language Processing.* Diluar bahasa yang tidak ada di kamus, bahasa "*slang*" terkadang memiliki bentuk yang beragam walaupun berasal dari bahasa formal yang sama.

Penelitian ini berfokus pada masalah yang sama dengan penelitian Salsabila et al, [4], dengan memberikan langkah tambahan untuk menghilangkan bahasa "*slang*" menggunakan *spelling correction* SymSpell [5]. Dengan penelitian ini diharapkan dapat meningkatkan akurasi algoritma *Natural language processing* yang dibatasi jumlah data leksikal yang dimiliki. Dimana dengan metode yang diteliti, data leksikal yang dibuat secara manual dapat didukung oleh penggunaan *spelling correction.*

2. Research Methods

Dalam penelitian ini menggunakan pendekatan *simulation-based*, beberapa tahapan yang dilalui dapat dilihat pada Figure 1.

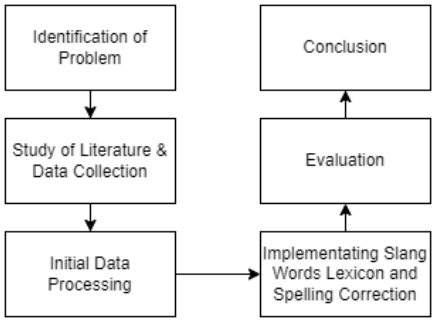

Figure 1. *Research Flowchart*

2.1. *Identification of Problem*

Dalam penelitian ini, masalah yang ingin diselesaikan adalah menemukan metode yang dapat meningkatkan akurasi *Natural Language Processing* dalam menangani bahasa Indonesia non-formal khususnya pada kasus analisis sentimen.

2.2. *Study Literature & Obtaining Data*

Pada fase ini, studi literatur bertujuan untuk memperdalam pengetahuan mengenai permasalahan yang diangkat, kemungkinan penyelesaiannya serta sumber data yang sesuai. Data yang digunakan dalam penelitian ini merupakan sebuah dataset *tweet* atau kiriman post pada sosial media *twitter* yang diperoleh dari dataset "*Indonesian Tweets COVID-19 Handling (2020)"* [6]*.* Dataset tersebut merupakan dataset yang sudah memiliki label sentimen positif dan negatif untuk masing-masing *tweet*-nya sehingga sangat cocok digunakan untuk mensimulasikan metode yang diteliti. Terdapat total 2270 data *tweet* dengan sampel data seperti pada Table 1.

Table 1. Sampel data *tweet*

tweets	sentimen
Pemerintah Indonesia Terlambat Sigap Hadapi Covid-19 https://abroryaguscp.wordpress.com/2020/03/28/pemerintah-indonesia-terlambat-sigap-hadapi-covid-19/Â â€¦	*Negative*
Salah ya. Keliru ya. Dalam mengantisipasi covid-19. Tak mudah merbau keputusan untuk kepentingan seluruh warga bukan hanya di Jakarta, di Jawa tapi warga seluruh Indonesia dari Sabang sampai Merauke. Tapi paling mudah menunjuk pemerintah yang salah	Positive
Menurutnya prediksi ada lebih dari 100.000 orang yang konfirm COVID 19, ayoo lah men.. OKE kita ikutin aturan pemerintah, tp tolong jgn bikin berita konyol yang bs buat orang yg lemah jadi tambah panik!! Kalau semua warga indonesia bisa santuy gpp, beda orang beda mikirnya braaay	Negative

Selain data *tweets,* diperlukan juga data kamus KBBI yang didapat dari halaman web *github.com* mengenai aplikasi *desktop* kbbi-qt [7], data *corpus* Bahasa Indonesia[8] untuk *spelling correction dan* data *Colloquial Indonesian Lexicon* atau kamus "*Alay*" oleh Salsabila et al [4].

2.3. *Initial Data Processing*

Pada fase ini, Data mentah yang didapatkan akan diproses sesuai kebutuhan. Data kata yang berada di KBBI akan dipisahkan dengan deskripsinya dan dihilangkan duplikatnya untuk digunakan sebagai pengecekan kata formal atau non-formal. Data kamus "*Alay*" diubah strukturnya menjadi *json dictionary* untuk mempermudah *lookup* data dan artinya. Serta data *tweet* dihilangkan duplikatnya yang kemudian menyisakan 1705 data *tweet*

2.4. *Implementing Slang Word Lexicon and Spelling Correction*

Pada fase dilakukan implementasi konversi bahasa *slang* menggunakan *Colloquial Indonesian Lexicon,* implementasi *spelling correction* menggunakan algoritma *SymSpell* serta implementasi penggunaannya dalam tahap *pre-processing.*

Sebelum dikirim ke algoritma *classifier,* data akan melewati tahap *pre-processing.* Umumnya tahap *pre-processing* yang digunakan menurut [9] adalah sebagai berikut;

p-ISSN: 2301-5373
e-ISSN: 2654-5101

a. *Data cleaning*, dimana data akan dibersihkan dari karakter unik, *tag, mention* dan sebagainya.
b. *Stopword removal*, dimana kata *stopword* yang tidak memiliki makna yang signifikan terhadap dokumen, dan
c. *Stemming*, dimana kata akan diubah menjadi kata dasar.

Dalam penelitian ini, kombinasi implementasi *slang word correction* dan *spelling correction* dilakukan di antara langkah *a* dan *b*. Hal ini dilakukan karena tahap *b* menggunakan penghapusan berbasis *dictionary,* sehingga bergantung pada kata yang ada pada daftar *stopword*.

Secara garis besar kombinasi *pre-processing* yang dicoba pada penelitian ini ada pada Table 2.

Table 2. Kombinasi *pre-processing*

Kombinasi	Langkah
Unnormalized	Tanpa *spell correction* dan *slang word removal*
M	Hanya *spelling correction*
S	Hanya *slang words removal*
M+S	*Spelling correction* dan *slang word removal* secara berurutan
S+M	*Slang word removal* dan *spelling correction secara berurutan*

Implementasi *pre-processing* yang diteliti adalah sebagai berikut:

a. *Slang Word Correction*
Implementasi *slang word lexicon* dilakukan dengan menggunakan kamus “Alay” yang dikembangkan oleh Salsabila et al [4]. Algoritma yang digunakan dapat dilihat pada Figure 2.

Algorithm 1: Slang Word Correction

```
Result: Sentence with translated slang word
open slang word dictionary;
for each tweet in tweets do
    for each word in tweet do
        if word is in slang word dictionary then
            Replace word with translation;
        end
    end
end
```

Figure 2. *Slang word removal algorithm*

b. Spell Correction

Implementasi *spell correction* dilakukan dengan menggunakan *library SymSpell* [5]. Dibandingkan dengan algoritma lainnya yang menggunakan struktur data *tree* untuk memvalidasi kata dan mencari kata terdekat dengan nilai tertentu, yang biasanya nilai tersebut berupa *Levenshtein distance* atau nilai minimal perubahan yang diperlukan untuk mengubah satu kata menjadi kata lainnya seperti yang ada pada penelitian [10].

Pada penelitian [11] pemilihan kata terdekat berdasarkan nilai minimum *Levenshtein distance* dengan operasi *insert (*penambahan huruf*), delete (*penghapusan huruf*), substitution (*penggantian huruf*) dan transposition (*pengubahan urutan huruf*)*. Namun, pada algoritma *SymSpell (Symmetric Delete spelling correction algorithm)* operasi yang dilakukan hanya berupa *delete* atau penghapusan huruf.

Dengan hanya sebuah operasi yang digunakan, kompleksitas algoritma akan berkurang. Pemilihan algoritma ini dilakukan karena pada umumnya kata non-formal bahasa Indonesia pada komunikasi sehari-hari digunakan karena bentuknya lebih singkat dan lebih cepat [1] yang sebagian besar hanya menggunakan operasi *delete.*

Algoritma yang digunakan dapat dilihat pada Figure 3.

Algorithm 2: Spelling Correction

```
Result: Corrected Sentence from misspelled Word
for each tweet in tweets do
    for each word in tweet do
        if word is not in KBBI then
            text suggestion = lookup word in SymSpell;
            Replace word with text suggestion;
        end
    end
end
```

Figure 3. *Spell correction algorithm*

2.5. *Evaluation*

Pada tahap ini, kombinasi *pre-processing* akan dievaluasi berdasarkan *F1-score*. Uji coba analisis sentimen dilakukan menggunakan kombinasi 2 algoritma *classifier* populer yaitu *Support Vector Machine* dengan *kernel RBF* dan *Naïve Bayes,* dengan 3 metode ekstraksi fitur yang berbeda yaitu *term presence, Bag of Words* dan *TF-IDF.*

3. Result and Discussion

3.1. Correction Evaluation

Pada bagian ini, dilakukan *pre-processing* pada data *tweet* dengan kombinasi yang berbeda. Berikut frekuensi pengubahan kata yang dilakukan berdasarkan kombinasi yang telah dijelaskan sebelumnya.

Table 3. Frekuensi penggantian kata tiap kombinasi *pre-processing*

Kombinasi	Spelling Correction	Slang Correction	Total Correction
Unnormalized	0	0	0
M	**1520**	0	1520
S	0	**4822**	4822
M+S	**1520**	4584	6104
S+M	1509	**4822**	**6331**

Dari Table 3, dapat dilihat kombinasi S+M mengalami total perubahan kata paling banyak, hal ini dapat terjadi karena *slang word correction* merupakan normalisasi teks berbasis kamus. Sehingga ketika pada tahap *spelling correction* dilakukan kemungkinan besar *slang word* sudah diperbaiki.

3.2. Classification Evaluation

Pada bagian ini, dilakukan pelatihan dan pengujian 2 model *machine learning* menggunakan algoritma *Support Vector Machine (SVM)* dan *Naïve Bayes.* Sebelum dimasukan pada algoritma, teks sebelumnya akan dipisahkan menjadi data latih dan uji dengan presentasi 80%:20%. Dan setelahnya memasuki tahap ekstraksi fitur menggunakan 3 metode yang berbeda yaitu *term presence, Bag of Words* dan *TF-IDF.* Hasil *F1-score* dari penelitian ini dapat dilihat pada Table 4 dan 5

Table 4. *F1-score* untuk algoritma *Naïve Bayes*

Kombinasi	*Term Presence*	*BoW*	*TF-IDF*
Unnormalized	**0.82914**	**0.82914**	0.81878
M	0.81609	0.82117	**0.82744**
S	0.81643	0.81913	0.82692
M+S	0.81898	0.81913	0.81456
S+M	0.81609	0.81404	0.81151

Table 5. *F1-score* untuk algoritma *SVM*

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Kombinasi	*Term Presence*	*BoW*	*TF-IDF*
Unnormalized	0.80497	0.83423	**0.85420**
M	**0.82660**	**0.84458**	0.85185
S	0.82117	0.82692	0.85223
M+S	0.82152	0.83169	0.84968
S+M	0.79713	0.83996	0.84695

4. Conclusion

Pada hasil yang sudah dideskripsikan dapat disimpulkan,

1. Penggunaan kombinasi *slang word correction* dan *spelling correction* tidak mengalami peningkatan pada semua kasus. Namun dengan hanya menggunakan tambahan *spelling correction* pada tahap *pre-processing* data, terbukti meningkatkan *F1-score* pada kasus ekstraksi fitur TF-IDF pada algoritma *Naïve Bayes,* dan pada kasus ekstraksi fitur *Term Presence* dan *Bag of Words* pada algoritma SVM walaupun peningkatan keseluruhan tidak terlalu signifikan.
2. Dari data yang bisa dilihat pada Table 4 dan 5 juga bisa dilihat bahwa penggunaan *slang word correction* atau *spelling correction* secara mandiri menghasilkan performa *F1-score* yang sama atau bahkan lebih baik dari penggunaan kedua algoritma tersebut.

References

[1] E. Utami, A. D. Hartanto, S. Adi, R. B. Setya Putra, and S. Raharjo, “Formal and Non-Formal Indonesian Word Usage Frequency in Twitter Profile Using Non-Formal Affix Rule,” *2019 1st Int. Conf. Cybern. Intell. Syst. ICORIS 2019*, pp. 173–176, Aug. 2019, doi: 10.1109/ICORIS.2019.8874908.

[2] R. B. S. Putra and E. Utami, “Non-formal affixed word stemming in Indonesian language,” *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-January, pp. 531–536, Apr. 2018, doi: 10.1109/ICOIACT.2018.8350735.

[3] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, “Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation,” *J. Big Data*, vol. 8, no. 1, pp. 1–16, Dec. 2021, doi: 10.1186/s40537-021-00413-1.

[4] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, “Colloquial Indonesian Lexicon,” *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, pp. 226–229, 2019, doi: 10.1109/IALP.2018.8629151.

[5] W. Garbe, “SymSpell.” 2012. [Online]. Available: hhttps://github.com/wolfgarbe/SymSpell

[6] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, “Tweets Responding to the Indonesian Government’s Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel,” *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 2, pp. 112–122, Oct. 2020, doi: 10.20473/JISEBI.6.2.112-122.

[7] “bgli/kbbi-qt: KBBI Offline Remake with Qt.” https://github.com/bgli/kbbi-qt/ (accessed Oct. 03, 2022).

[8] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” Sep. 2020, doi: 10.48550/arxiv.2009.05387.

[9] V. Gurusamy and S. Kannan, “Preprocessing Techniques for Text Mining,” 2014.

[10] K. Kukich, “Techniques for Automatically Correcting Words in Text,” *ACM Comput. Surv.*, vol. 24, no. 4, pp. 377–439, 1992, doi: 10.1145/146370.146380.

[11] P. Santoso, P. Yuliawati, R. Shalahuddin, and A. P. Wibawa, “Damerau Levenshtein Distance for Indonesian Spelling Correction,” *J. Inform.*, vol. 13, no. 2, p. 11, 2019, doi: 10.26555/jifo.v13i2.a15698.

This page is intentionally left blank