RISET UNGGULAN UNIVERSITAS

LAPORAN AKHIR PENELITIAN

Pengujian Pengaruh Normalisasi Kata Tidak Baku pada Algoritma Klasifikasi

TIM PENGUSUL

Danny Sebastian, S.Kom.,MM,MT
Kristian Adi Nugraha, S.Kom.,MT

Informatika

UNIVERSITAS KRISTEN DUTA WACANA

Oktober 2021

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

RINGKASAN

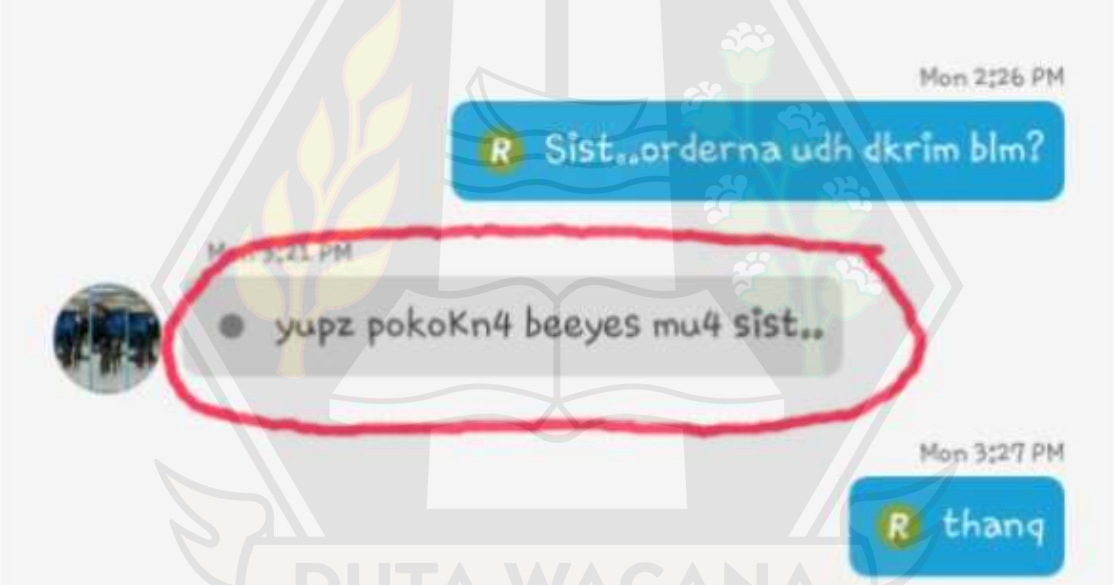
Media sosial atau social media merupakan sumber media komunikasi yang sangat kaya namun mengandung banyak akta tidak baku dan emoticon. Saat ini banyak aplikasi yang mencoba mengolah dokumen teks yang ada di media sosial. Kesulitan muncul ketika dokumen teks mengandung kata-kata yang tidak baku seperti, mengandung singkatan, slang, dan lain sebagainya. Dari metode normalisasi kata tidak baku yang ada saat ini, masih belum ada metode yang memberikan hasil normalisasi yang sangat tepat. Penelitian ini bertujuan untuk menguji seberapa besar pengaruh normalisasi kata tidak baku dan membangun korpus data kata tidak baku dan bentuk bakunya.

Kata Kunci: *text-mining, pre-processing,* normalisasi

BAB 1
PENDAHULUAN

1.1 Latar Belakang

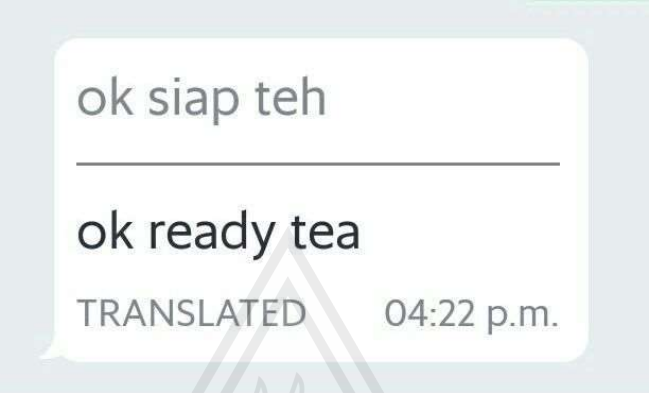
Media sosial adalah sumber informasi yang sangat kaya, tetapi banyak mengandung kata-kata yang tidak baku (Rezeki & Sagala, 2019) (Hidayatullah, 2015) (Wisnu, et al., 2020), emoticon (Zhao, et al., 2012), dan lain sebagainya. Seperti pada Gambar 1.1, terlihat sebuah kalimat yang menggunakan kata tidak baku pada komunikasi media sosial. Terlihat orang pertama, dengan *bubble chat* berwarna biru, menggunakan beberapa kata singkatan, seperti udh, dkrim, blm. Dan menggunakan beberapa kata tidak baku, seperti orderna dan thanq. Sedangkan orang kedua, dengan *bubble chat* berwarna abu-abu, berkomunikasi menggunakan kata yang dituliskan dengan angka.



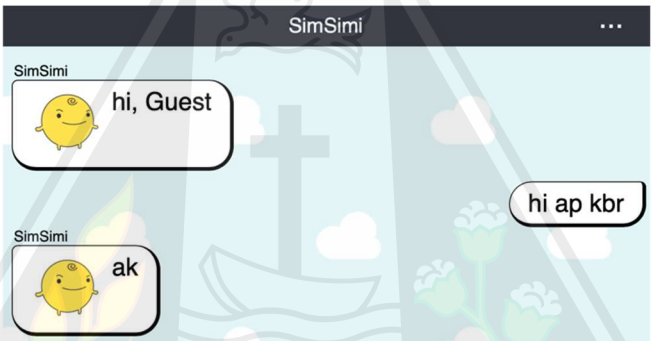
Gambar 1.1 Contoh kalimat tidak baku yang digunakan untuk berkomunikasi pada media sosial (Trimastuti, 2017)

Saat ini banyak sekali aplikasi yang mencoba menggunakan teks dari media sosial sebagai bahan untuk aplikasi yang dibangunnya, seperti pada chatbot, system translasi otomatis, dan lain sebagainya. Gambar 1.2 menunjukkan sebuah percakapan pada aplikasi ojek online. Aplikasi ojek online tersebut melakukan translasi secara otomatis ke Bahasa inggris. Terlihat kesalahan translasi pada kata “teh”, yang seharusnya berarti “bibi/ibu” pada Bahasa Sunda, tetapi

diterjemahkan ke “tea” atau kata objek minuman. Gambar 1.3 menunjukkan kesalahan *response* oleh aplikasi chatbot otomatis. Komunikasi dilakukan menggunakan kata singkatan.



Gambar 1.2 Komunikasi pada salah satu aplikasi ojek online (Sebastian & Nugraha, 2019)



Gambar 1.3 Komunikasi pada salah satu chatbot (Sebastian & Nugraha, 2019)

Sudah ada beberapa metode yang melakukan normalisasi pada kata tidak baku, seperti *crowdsourcing* (Sebastian & Nugraha, 2019) dan *stemming* (Mutiara, et al., 2021). Metode stemming dituliskan dapat meningkatkan akurasi dari algoritma *Support Vector Machine*. Sedangkan hasil normalisasi metode *crowdsourcing* masih belum diujikan ke algoritma *text mining*. Saat ini pengaruh normalisasi kata tidak baku masih belum jelas pengaruh terhadap algoritma *text mining*. Karena hal itu, masih belum belum banyak *developer* yang menggunakan metode normalisasi untuk aplikasi yang mereka buat atau mengembangkan metode normalisasi kata tidak baku.

1.2 Rumusan Masalah

Berdasarkan latar belakang permasalahan yang telah dikemukakan sebelumnya, permasalahan yang akan diteliti dalam penelitian ini adalah:

1. Bagaimana pengaruh normalisasi kata tidak baku pada beberapa algoritma klasifikasi teks.

BAB 6
KESIMPULAN DAN SARAN

6.1 Kesimpulan

Kesimpulan dari penelitian ini:

- Penelitian normalisasi kata tidak baku perlu terus dikembangkan, karena banyaknya aplikasi pengolahan teks automatis yang dikembangkan.

6.2 Saran

Saran untuk penelitian selanjutnya:

- Menambahkan korpus data
- Melakukan pengujian menggunakan algoritma clustering

DAFTAR PUSTAKA

Agnihotri, D., Verma, K. & Tripathi, P., 2014. *Pattern and cluster mining on text data.* s.l., s.n.

Ahmad, S. & Varma, R., 2018. Information extraction from text messages using data mining techniques. *Malaya Journal of Matematik,* pp. 26-29.

Hanafiah, N. et al., 2017. Text normalization algorithm on twitter in complaint category. *Procedia computer science,* Volume 116, pp. 20-26.

Hidayatullah, A. F., 2015. *Language tweet characteristics of Indonesian citizens.* s.l., IEEE.

Hidayatullah, A. F. & Ma'arif, M. R., 2017. Pre-processing Tasks in Indonesian Twitter Messages. *Journal of Physics: Conference Series,* 801(1).

Inzalkar, S. & Sharma, J., 2015. A survey on text mining-techniques and application. *International Journal of Research In Science & Engineering,* Volume 24, pp. 1-14.

Kurbatow, A., 2015. *The research of text preprocessing effect on text documents classification efficiency.* s.l., IEEE, pp. 653-655.

Maylawati, D. S. & Saptawati, G. P., 2016. *Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang.* s.l., s.n., pp. 1-6.

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

Mutiara, A., Wibowo, E. P. & Santosa, P. I., 2021. Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data,* 8(1), pp. 1-16.

Nugraha, K. A. & Sebastian, D., 2018. *Analisis Trend Akun Media Sosial Twitter Menggunakan TF-IDF dan Cosine Similarity.* Yogyakarta, s.n.

Nugraha, K. A. & Sebastian, D., 2018. Pembentukan Dataset Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity. *Jurnal Teknik Informatika dan Sistem Informasi,* 4(3), pp. 376-386.

Rahman, T., Agustin, F. E. M. & Rozy, N. F., 2019. *Normalization of Unstructured Indonesian Tweet Text For Presidential Candidates Sentiment Analysis.* s.l., IEEE, pp. 1-6.

Rezeki, T. I. & Sagala, R. W., 2019. Slang Words Used by Millenial Generation in Instagram. *Jurnal Serunai Bahasa Inggris,* 11(2), pp. 74-81.

Santoso, V. I., Virginia, G. & Lukito, Y., 2017. Penerapan Sentiment Analysis Pada Hasil Evaluasi Dosen Dengan Metode Support Vector Machine. *Jurnal Transformatika,* 14(2), pp. 72-76.

Sebastian, D., 2019. Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari beberapa E-marketplace. *Jurnal Teknik Informatika dan Sistem Informasi,* 5(1), pp. 51-61.

Sebastian, D., 2019. Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari beberapa E-marketplace. *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI),* 5(1), pp. 52-61.

Sebastian, D. & Nugraha, K. A., 2019. Sistem Perbaikan Kata Tidak Baku Bahasa Indonesia Menggunakan Metode Crowdsourcing. *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI),* 5(3).

Sebastian, D. & Nugraha, K. A., 2019. Sistem Perbaikan Kata Tidak Baku Bahasa Indonesia Menggunakan Metode Crowdsourcing. *Jurnal Teknik Informatika dan Sistem Informasi,* Desember, 5(3), pp. 386-396.

Sebastian, D. & Nugraha, K. A., 2019. *Text Normalization for Indonesian Abbreviated Word Using Crowdsourcing Method.* Yogyakarta, IEEE, pp. 529-532.

Sebastian, D. & Nugraha, K. A., 2019. *Text Normalization for Indonesian Abbreviated Word Using Crowdsourcing Method.* Yogyakarta, IEEE, pp. 529-532.

Singh, T. & Kumari, M., 2016. Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science,* Volume 89, pp. 549-554.

Trimastuti, W., 2017. An analysis of slang words used in social media. *Jurnal Dimensi Pendidikan dan Pembelajaran,* 5(2), pp. 64-68.

Vijayarani, S., Ilamathi, J. & Nithya, 2015. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks,* 5(1), pp. 7-16.

Wisnu, H., Afif, M. & Ruldevyani, Y., 2020. *Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes.* s.l., IOP Publishing.

Yao, Z. & Ze-wen, C., 2011. *Research on the construction and filter method of stop-word list in text preprocessing.* s.l., IEEE, pp. 217-221.

Zhao, J., Dong, L., Wu, J. & Xu, K., 2012. *Moodlens: an emoticon-based sentiment analysis system for chinese tweets.* s.l., ACM, pp. 1528-1531.