ANALISA MODIFIKASI ALGORITMA *STEMMING* UNTUK KASUS *OVERSTEMMING*

Stephanie Betha Rossi Hersianie[1]
[1] Teknik Komputer, Universitas Wiralodra, Indonesia
[1]ntephbetha@gmail.com

ABSTRAK

Overstemming merupakan pemenggalan kata ke bentuk asal (*root word)* yang berlebihan. Hal ini menyebabkan kata tersebut bermakna sangat berbeda dengan kata asal. Namun, stem yang dihasilkan sama bentuknya. Untuk mengatasi permasalahan tersebut, penelitian sebelumnya telah menerapkan algoritma *stemming* dengan tabel aturan kata. Namun kekurangan dari tabel aturan kata ini adalah kesulitan dalam menambahkan jenis kata yang mengalami overstemming. Oleh karena itu, penelitian ini bertujuan untuk memodifikasi algoritma overstemming tersebut. Penelitian ini akan menggabungkan algoritma stemming (*hybrid stemming*) yaitu algoritma look-up table, tabel aturan kata dan algoritma *stemming Porter* yang biasa digunakan. Dataset yang digunakan dalam pengujian adalah atribut judul pada dokumen publikasi ilmiah. Hasil pengujian menunjukkan bahwa modifikasi algoritma stemming menghasilkan recall sebesar 89, 9%.Saran untuk penelitian selanjutnya adalah pengujian dapat dilakukan menggunakan atribut lainnnya pada dokumen publikasi.

Kata Kunci : *Stemming*, Modifikasi *stemming, Hybrid Stemming, Stemming Porter*

ABSTRACT

Overstemming is the splitting of words into excessive root words.This causes the word to have a very different meaning from the original word. However, the resulting stem is the same shape. To solve this problem, previous research has implemented a stemming algorithm with a word rule table. But the drawback of this word rule table is the difficulty in adding overstemming types of words. Therefore, this study aims to modify the overstemming algorithm. This research will combine stemming algorithm (hybrid stemming), namely look-up table algorithm, word rule table and Porter's commonly used stemming algorithm. The dataset used in testing is the title attribute in scientific publication documents. The test results show that the stemming algorithm modification results in a recall of 89.9%. Suggestions for further research are that testing can be done using other attributes in the publication document.

Keywords: *Stemming, Stemming Modification, Hybrid Stemming, Stemming Porter*

I. PENDAHULUAN

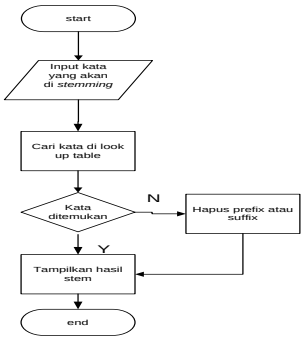
Stemming memiliki beberapa manfaat, di antaranya yaitu untuk mengembalikan kata ke bentuk asalnya, menambah nilai recall dan berfungsi dalam pada proses pencarian kata. Selain itu, tujuan dari proses stemming adalah memperkecil perubahan dan bentuk turunan dari suatu kata ke bentuk kata dasarnya [1]. Misalnya, kata " cars, car's, car " memiliki bentuk dasar yang sama yaitu "car". Permasalahan utama dalam proses stemming adalah bagaimana cara memperoleh kata dasar yang benar dari suatu kata yang telah mengalami perubahan bentuk [2], [3].

Overstemming merupakan salah satu kasus yang terjadi pada proses stemming. Overstemming merupakan pemenggalan kata ke bentuk asal (*root word)* yang berlebihan. Hal ini menyebabkan suatu kata yang bermakna sangat berbeda, dapat menghasilkan stem yang sama [4]. Misalnya, kata "generalized, general, generous, generating" menghasilkan stem "gener". Penelitian sebelumnya [5] telah mengatasi kasus *overstemming* menggunakan algoritma *stemming Porter* dan membuat tabel aturan kata. Sedangkan, penelitian [6] telah melakukan stemming menggunakan *look up table. Look up table* ini sebagai tempat menyimpan *root words* beserta bentuk kata imbuhannya tersebut pada database [7].

Tabel aturan kata pada penelitian [5] memiliki kekurangan yaitu kesulitan dalam menambahkan jenis kata yang mengalami overstemming. Oleh karena itu, penelitian ini bertujuan untuk menangani overstemming dengan melakukan perubahan (modifikasi) pada proses stemming. Proses perubahan (modifikasi) stemming dilakukan dengan menggabungkan algoritma look-up table yang berisi tabel aturan kata dan algoritma stemming Porter yang sudah dikenal [8], [9]. Tabel aturan kata tersebut dimasukkan ke dalam database sehingga dapat memudahkan proses update jenis kata yang mengalami overstemming. Dataset yang digunakan pada penelitian ini adalah atribut judul dari dokumen publikasi. Pengujian dilakukan dengan mengolah atribut judul tersebut melalui proses stemming biasa dan proses modifikasi stemming. Kedua proses tersebut akan menghasilkan perbandingan nilai recall pada judul [10], [11].

II. METODE PENELITIAN

Penanganan kasus overstemming ini dilakukan dengan menggabungkan algoritma look-up table dan tabel aturan kata dari. Look up table ini berisi kamus kata [1], [11]. Kamus kata tersebut memiliki daftar kata yang berasal dari tabel aturan kata. Selain itu, penelitian ini juga menggunakan *algoritma affix removal stemming Porter* untuk menangani kata yang tidak mengalami *overstemming*. Algoritma stemming Porter digunakan karena tabel aturan kata yang dihasilkan dari penelitian sebelumnya merupakan hasil dari adaptasi algoritma stemming Porter. Algoritma *look-up table* dilakukan dengan memasukan kata-kata ke database secara manual. Ketika pengguna memasukan kata infleksi (perubahan bentuk kata yang tidak mengubah arti kata tersebut), maka stemmer akan mencari keberadaan [12].



Gamabar. Proses Algoritma *Look up Table* [13]

Ada beberapa usulan dalam aturan kata untuk menangani kasus *overstemmin*g [14], [15]. Berikut ini adalah beberapa tabel aturan kata tersebut :

1. Aturan pertama mengatasi kata bentuk jamak pada kata tidak beraturan yang tidak ditangani pada algoritma *stemming* Porter.

Tabel 1. Aturan Pertama Daftar Kategori Kasus yang ditangani

Kategori	*Original Words*	Aturan	Hasil *Stem* Modifikasi	Total Kata
Kata yang berakhiran children adalah bentuk jamak dari child	*child/children*	Mengubah children menjadi child	*child/child*	6 kata
Kata yang berakhiran -men adalah bentuk jamak dari -man	*dryman/ drymen*	mengubah –men menjadi -man	*dryman/ dryman*	429 kata
Kata yang berakhiran -ci adalah bentuk jamak dari -cus	*abacus/ abaci*	mengubah –cus menjadi -ci	*abaci/ abaci*	35 kata
Kata yang berakhiran –eaux adalah bentuk jamak dari -eau	*plateau/ plateaux*	mengubah –eaux menjadi -eau	*plateau/ plateau*	29 kata
Kata yang berakhiran –mata adalah bentuk jamak dari -ma	*automa/ automata*	mengubah –mata menjadi -ma	*automa/ automa*	108 kata
Kata yang berakhiran –trices adalah bentuk jamak dari -trix	*matrix/ matrices*	mengubah -trices menjadi -trix	*matrix/ matrix*	21 kata
Kata yang berakhiran -ses adalah bentuk jamak dari -sis	*analysis /analyses*	mengubah -sis menjadi -s	*analys/ analys*	492 kata

2. Aturan kedua yaitu mengatasi kata-kata yang memiliki stem "gener" dan kata yang memiliki awalan "gene". Beberapa kata

yang ada pada Tabel 2 sebelum mengalami modifikasi algoritma stemming, akan mengalami kasus *overstemming.* Semua kata tersebut menghasilkan stem yang sama yaitu "gener". Berikut ini adalah daftar kata yang memiliki stem "gener" dengan perubahan hasil stem-nya.

Tabel 2. Aturan Kedua Daftar Kategori Kasus yang ditangani

Kata	Hasil *stem* sebelum modifikasi	Hasil *stem* modifikasi	Total Kata
generate	*gener*	*generat*	76 Kata
generates	*gener*		
generated	*gener*		
generating	*gener*		
general	*gener*	*general*	
generally	*gener*	*generic*	
generic	*gener*		
generically	*gener*	*generous*	
generous	*gener*		

3. Aturan ketiga menangani kata dengan yang akhiran –y dan tidak terdapat huruf vokal pada kata tersebut serta kata tersebut tidak terselesaikan oleh algoritma *stemming* Porter.

Tabel 3. Aturan Ketiga Daftar Kategori Kasus yang ditangani

Original Words	Hasil *stem* sebelum modifikasi	Hasil Stem Modifikasi	Total Kata
cry/cries/cried/crying	*cry/cri/cri/cry*	*cry/cry/cry/cry*	20 kata

4. Aturan keempat menanggulangi kata dengan akhiran –s (bukan –ss) dan bentuk *participle* dari kata tersebut. Algoritma stemming Porter akan menghapus akhiran –s pada kata dengan akhiran –s. Misalnya, kata "*focus*" menjadi "*focu*". Namun, apabila kata berakhiran –s tersebut berada dalam bentuk *past* atau *present participle* maka *stemming* Porter hanya akan menghapus akhiran –ed atau –ing. Misalnya, kata "*focused*", "*focusing*" menjadi "*focus*" dan "*focus*".

Tabel 4. Aturan Keempat Daftar Kategori Kasus yang ditangani

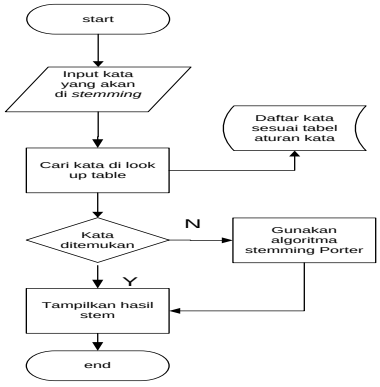
Original Words	Hasil *stem* sebelum modifikasi	Aturan	Hasil *Stem* Modifikasi	Total Kata
focus/ focuses/ focused/ focusing	*focu/focus/ focus/ focus*	mengubah -sed menjadi –s mengubah –sing menjadi -s	*focu/focu/focu/ focu*	28 kata

5. Aturan kelima algoritma stemming Porter tidak menanggulangi present atau past participle derivations. Misalnya, kata "studiedly" menghasilkan stem "studiedli" seharusnya "study". Aturan kelima akan mengatasi kata-kata pada kategori tersebut.

Tabel 5. Aturan Kelima Daftar Kategori Kasus yang ditangani

Kategori	*Original Words*	Hasil *stem* sebelum modifikasi	Aturan	Hasil Stem	Total Kata
Kata yang berakhiran -iedly atau -iedness berelasi dengan kata yang berakhiran -ied	*Study /studied/ studiedness/studiedly*	*studi/studi/ studied /studiedli*	mengubah –ly menjadi -ied mengubah -ss menjadi -ied	*study/ study /study /study*	13 kata
Kata yang berakhiran -edly atau -edness berelasi dengan kata yang berakhiran -ed	*Amaze /amazed /amazedly /amazedness*	*amaz/amaz/ amazedli/ amazed*	mengubah –ly menjadi -ed mengubah -ss menjadi -ed	*amaz/ amaz/ amaz /amaz*	439 kata

Tabel aturan kata memuat aturan pemotongan kata pada proses stemming. Dataset kata pada tabel aturan kata ini diperoleh dari website morewords [16]. Website ini digunakan sebagai pilihan kamus kata karena memiliki kata yang bersumber dari Enahanced North American Benchmark Lexicon (ENABLE2K). Jumlah kata yang dimiliki oleh situs ini lebih dari 173.528. Kemudian, daftar kata tersebut beserta kata hasil bentukannya dimasukan ke dalam tabel di database. Alur proses usulan metode modifikasi algoritma *stemming* untuk menangani kasus *overstemming* adalah sebagai berikut :



Gambar 2. Proses Modifikasi Algoritma *Stemming*

P-ISSN: 2621-8070
E-ISSN:2686-3219

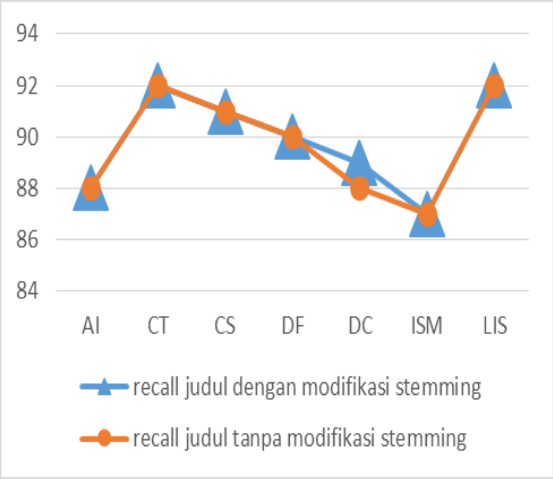
III. HASIL DAN PEMBAHASAN

Pengujian modifikasi algoritma stemming ini dilakukan menggunakan atribut judul pada dokumen publikasi. Dataset dokumen publikasi diperoleh dari DBLP Computer Science Bibliography, serta dokumen publikasi dari STEI-ITB. Pengujian menggunakan dokumen publikasi sebanyak 3.012. Dokumen publikasi ini sudah dikelompokkan ke dalam 7 bidang penelitian yaitu *AI and Image Processing (AI), Computational Theory and Mathematics (CT), Computer Software (CS), Data Format (DF), Distributed Computing (DC), Information System (ISM), Library and Information Science (LIS).* Tabel 6 menunjukkan hasil kata yang mengalami *overstemming* dan tertangani oleh modifikasi algoritma stemming pada atribut judul.

Tabel 6. Kata Jadul Overstemming yang ditangani

Attribut	Jumalah Kata Overstemming	Total Kata	Presentasi (%)
Judul	1430	17743	8, 06

8,06%. Tabel di atas menunjukkan bahwa kemungkinan jenis aturan kata masih kurang mampu untuk menangani kasus overstemming pada atribut judul. Jenis aturan kata yang lebih bervariasi memungkinkan kasus tersebut lebih banyak yang tertangani. Gambar 3 merupakan hasil recall penggunaan atribut judul pada modifikasi algoritma stemming dan stemming tanpa modifikasi.



Gambar 3. Perbandingan nilai recall pada judul

Gambar 3 menunjukkan bahwa recall pada kasus overstemming yang ditangani dengan modifikasi algoritma stemming memiliki nilai yang hampir sama besarnya dengan algoritma stemming tanpa modifikasi. Kedua proses tersebut menghasilkan nilai rata-rata recall sebesar 89,9% dan 89,7%. Perbedaan nilai recall kata judul terjadi pada kategori DC. Nilai recall kategori DC pada stemming dengan modifikasi memiliki nilai yang lebih tinggi dibandingkan recall kategori DC pada stemming tanpa modifikasi.

Kasus overstemming yang ditanggulangi oleh modifikasi algoritma stemming tidak terlalu mengakibatkan adanya peningkatan hasil *recall* dari proses algoritma stemming tanpa modifikasi. Penyebabnya adalah atribut judul hanya memiliki sedikit variasi kata yang terkandung pada aturan stemming dengan modifikasi. Selain itu, kata yang mengalami stemming dengan modifikasi kemungkinan termasuk kata yang bersifat umum (memiliki nilai *IDF* kecil) pada judul sehingga tidak terlalu mempengaruhi peningkatan *recall*. Kata yang bersifat umum artinya kata yang memiliki frekuensi kemunculan di beberapa kategori, sehingga menghasilkan nilai *IDF* yang kecil. Selain itu, jenis aturan kata kurang banyak variasinya, sehingga kurang dapat menangani kasus overstemming pada data yang digunakan. Hal ini menyebabkan tidak terlalu mempengaruhi peningkatan recall.

IV. KESIMPULAN

Kasus overstemming yang ditanggulangi menggunakan modifikasi algoritma stemming, dapat meningkatkan hasil recall pada penggunaan atribut judul namun peningkatan recall tersebut tidak terlalu signifikan. Penyebabnya adalah masih banyaknya kata yang belum dapat ditanggulangi oleh tabel aturan kata. Atribut judul hanya memiliki sedikit variasi kata yang terkandung pada tabel aturan kata tersebut. Selain itu, kata yang mengalami stemming dengan modifikasi kemungkinan termasuk kata yang bersifat umum (memiliki nilai *IDF* kecil) pada judul sehingga tidak terlalu mempengaruhi peningkatan recall. Kekurangan dari modifikasi algoritma ini adalah pengguna harus mengupdate jenis kata yang dapat mengalami overstemming. Saran untuk penelitian selanjutnya adalah dataset pengujian dapat

ditambahkan dengan menggunakan atribut lain pada dokumen publikasi, misalnya, abstrak. Hal ini bertujuan untuk menambah jumlah dan jenis kata yang kemungkinan dapat mengalami overstemming.

DAFTAR PUSTAKA

[1] M. A. Nq, L. P. Manik, and D. Widiyatmoko, "Stemming Javanese: Another Adaptation of the Nazief-Adriani Algorithm," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 627–631, Dec. 2020.

[2] S. Yuan and H. Sun, "A general adaptive finite element eigen-algorithm stemming from Wittrick-Williams algorithm," *Thin-Walled Structures*, vol. 161, p. 107448, Apr. 2021.

[3] U. Tukeyev, A. Turganbayeva, B. Abduali, D. Rakhimova, D. Amirova, and A. Karibayeva, "Lexicon-free stemming for Kazakh language information retrieval," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–4, Oct. 2018.

[4] M. N. Kassim, M. A. Maarof, A. Zainal, and A. A. Wahab, "Word stemming challenges in Malay texts: A literature review," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, vol. 4, no. c, pp. 1–6, May. 2016.

[5] R. B. Setya Putra, E. Utami, and S. Raharjo, "Accuracy Measurement on Indonesian Non-formal Affixed Word Stemming With Levenhstein," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, pp. 486–490, Jul. 2019.

[6] R. B. S. Putra and E. Utami, "Non-formal affixed word stemming in Indonesian language," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, vol. 2018–Janua, pp. 531–536, Mar. 2018.

[7] K. Vijayl, "Genera Tion of Caption Selection for News Images Using Stemming Algorithm," pp. 536–540, 2015.

[8] S. Ernawati, E. R. Yulia, Frieyadie, and Samudi, "Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, no. Citsm, pp. 1–5, Aug. 2018.

[9] 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," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, no. Citsm, pp. 1–6, Aug. 2018.

[10] H. A. Almuzaini and A. M. Azmi, "Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020.

[11] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, "Comparative Study of Arabic Stemming Algorithms for Topic Identification," *Procedia Computer Science*, vol. 159, pp. 794–802, 2019.

[12] W. Hare, C. Planiden, and C. Sagastizábal, "A derivative-free vu-algorithm for convex finite-max problems," *Optimization Methods and Software*, vol. 35, no. 3, pp. 521–559, 2020.

[13] A. Muklason, R. G. Irianti, and A. Marom, "Automated Course Timetabling Optimization Using Tabu-Variable Neighborhood Search Based Hyper-Heuristic Algorithm," *Procedia Computer Science*, vol. 161, pp. 656–664, 2019.

[14] A. Jabbar, S. Iqbal, A. Akhunzada, and Q. Abbas, "An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 3079, no. May, pp. 1–21, May 2018.

[15] G. R. Hall and K. Taghva, "Using the Web 1T 5-Gram Database for Attribute Selection in Formal Concept Analysis to Correct Overstemmed Clusters," in *2015 12th International Conference on Information Technology - New Generations*, Apr. 2015, pp. 651–654, Apr. 2015.

[16] Moreword, "Enahanced North American Benchmark Lexicon (ENABLE2K)," 2020, [Online]. Available: https://www.morewords.com.