Perbaikan Ejaan Kata pada Dokumen Bahasa Indonesia dengan Metode *Cosine Similarity*

Muhammad Fachrurrozi[1], Anne Agustina Manik[2]

[1,2] *Jurusan Teknik Informatika Universitas Sriwijaya*
Kampus Unsri Indralaya Ogan Ilir

Jl. Raya Pelembang Prabumulih KM 32 Indralaya OI Sumsel

obetsobets@gmail.com, manik.anne@yahoo.co.id

Abstrak— **Kesalahan ejaan kata dalam penulisan dokumen Bahasa Indonesia sering dijumpai sehingga sulit memahaminya. Penggunaan teknologi dalam memperbaiki kesalahan kata (*spelling checker*) telah banyak dilakukan. Pada penelitian ini dilakukan perbaikan kata pada dokumen bahasa Indonesia berbasis kemiripan kata menggunakan metode *n-gram* dan *cosine similarity*. Proses dimulai dengan melakukan pembentukan data latih dengan metode *n-gram* dalam pemotongan sejumlah kata. Pada proses pengujian dilakukan tahapan pra proses terlebih dahulu dan dilakukan pengecekan kata berdasarkan kamus kata dan data latih yang ada. Kata yang diasumsi salah dilakukan perbaikan kata dengan mencari kemiripan katanya dengan metode *n-gram* dan *cosine similarity*. Hasil kemiripan kata yang tertinggi disesuaikan dengan data latih, bila tidak sesuai maka kata dengan kemiripan tertinggi dianggap kata benar yang dilakukan perbaikan. Pada penelitian ini hasil percobaan dari 3 tingkatan kesalahan kata yaitu 20 %, 50 %, dan 70 % dengan masing-masing 20 dokumen menghasilkan perbaikan kata yang tepat diatas 70 %. Hasil penelitian dapat dilihat bahwa perbaikan kata sangat bergantung pada kamus kata trigram dan latih yang ada. Ini menunjukkan bahwa metode n-gram dan cosine similarity baik dalam penelitian ini.**

Kata Kunci— **spelling checker, n-gram, cosine similarity.**

I. PENDAHULUAN

Pembentuk struktur Bahasa Indonesia secara semantik adalah subjek, predikat, objek dan keterangan (SPOK). Kesalahan makna tulisan salah satunya disebabkan oleh kesalahan dalam penulisan ejaan kata. Kesalahan dalam penulisan ini disebabkan kesalahan dalam pengetikan dokumen yang tidak sesuai dengan ejaan kata yang tepat. Dalam suatu dokumen terdiri dari kumpulan beberapa kalimat. Kalimat adalah satuan bahasa terkecil dalam wujud lisan atau tulisan yang mengungkapkan pikiran yang utuh [1]. Kalimat terdiri dari kumpulan urutan kata yang berkesinambungan antara yang satu dengan yang lain, yang menyampaikan suatu informasi sehingga diperlukan ejaan kata yang tepat dalam urutan kata dalam kalimat.

Ejaan kata yang tepat dibentuk berdasarkan proses morfologi yang tepat. Perbaikan dokumen yang dikerjakan merupakan hal yang penting dilakukan untuk menghindari dalam kesalahan ejaan kata. Namun dalam pelaksanaanya hal ini jarang dilakukan dikarenakan pengecekan secara berulang-ulang yang memakan waktu dan tenaga untuk mendapatkan ejaan kata yang benar.

Salah satu solusi untuk menyelesaikan masalah ini adalah dengan menggunakan sistem yang mampu memperbaiki kata secara otomatis. Pada penelitian ini metode yang digunakan dalam perbaikan kata pada dokumen dengan menggunakan *N-gram* dan *cosine similarity*. Penelitian ini bertujuan untuk menerapkan metode yang digunakan dalam perbaikan kata pada dokumen bahasa Indonesia serta melihat akurasi metode tersebut dalam sistem ini. Dari sistem yang dibuat ini diharapkan dapat menghasilkan ejaan kata yang tepat pada dokumen bahasa Indonesia dengan metode *Cosine Similarity* sesuai dengan data asli yang diasumsi tidak ada yang salah.

II. TINJAUAN PUSTAKA

A. *Spelling Checker*

Spelling checker adalah proses pemeriksaan kata untuk mendeteksi kata yang salah eja dan memberikan kandidat kata yang benar. Kesalahan ejaan terdapat dua kategori yaitu [2]:

a. Kesalahan non kata

Kesalahan ejaan non kata adalah kesalahan yang berfokus pada kata yang terbentuk umumnya oleh kesalahan pengetikan. Kesalahan ejaan non kata ini menghasilkan kata- kata yang tidak masuk akal.

b. Kesalahan kata yang sebenarnya

Kesalahan kata yang sebenarnya adalah kesalahan yang menekankan pada penanganan penempatan kata dalam kalimat. Kesalahan kata yang sebenarnya menghasilkan kata sah lainnya.

Desain dari proses Spelling checker yang dilakukan yaitu [2]:

1. Melakukan pengolahan pra proses pada teks
2. Lalu memeriksa setiap kata apakah kata salah atau tidak.
3. Selanjutnya tahapan memperbaiki kata untuk mendapatkan kata yang benar.

B. *N-Gram*

N-gram adalah potongan sejumlah n karakter dari sebuah string yang diaplikasikan untuk pembangkitan kata atau karakter. Metode n-gram ini digunakan untuk mengambil

potongan-potongan karakter huruf sejumlah n dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen [3].

N-gram dibedakan berdasarkan jumlah potongan karakter sebesar n. Penambahan garis bawah (blank) pada awal dan akhir kata digunakan untuk membantu menentukan kondisi awal dan akhir kata. Sebagai contoh : kata ”ROSE” dapat diuraikan ke dalam beberapa n-gram berikut (”_” merepresentasikan blank) :

uni-grams : R,O,S,E
bi-grams : _R,RO,OS,SE,E_
tri-grams : _RO,ROS,OSE,SE_
quad-grams : _ROS, ROSE, OSE_
quint-grams : _ROSE, ROSE_

Pada pembangkitan kata metode *N-gram* digunakan untuk mengambil potongan kata sejumlah n dari kalimat yang secara kontinuitas dibaca dari teks sumber hingga akhir [4].

B. *Cosine Similarity*

Cosine Similarity adalah ukuran kesamaan antara dua vektor n dimensi dengan mencari cosinus dari sudut antara dua vektor. Pada metode *cosine similarity* tidak melihat dari panjang pendeknya dokumen melainkan dari nilai term masing- masing. Berikut adalah persamaan dari metode *Cosine Similarity* [5]:

$$\text{Similarity} = \cos(A,B) = \frac{|A \cap B|}{\sqrt{|A|}\sqrt{|B|}} \quad (1)$$

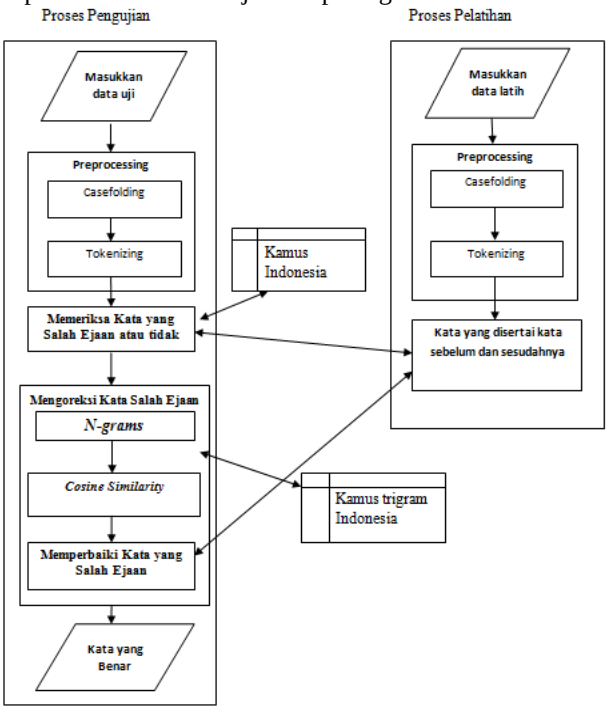
A dan B adalah kata yang dihitung kemiripannya, yang telah di lakukan pemotongan karakter (*ngrams*) dari kata A dan B terlebih dahulu. Nilai kesamaan kosinus antara 0 dan 1. Dua vektor dikatakan sama jika membentuk sudut 0^0 atau nilai kosinusnya 1.

Langkah- langkah perhitungan similarity dengan metode cosine adalah :

1. Hitung jumlah irisan dari *ngams* kata A dan B :
$$Intersection = |A \cap B| \quad (2)$$
2. Kemudian hitung semua product menggunakan persamaan :
$$Product = |A|.|B| \quad (3)$$
3. Kemudian hitung tingkat kemiripan pada tiap kata dengan menggunakan persamaan :
$$\text{Similarity} = \cos(A,B) = \frac{|A \cap B|}{\sqrt{|A|}\sqrt{|B|}} \quad (4)$$

III. METODE PENELITIAN DAN HASIL PEMBAHASAN

Dalam melakukan perbaikan kata pada dokumen terdapat beberapa langkah yang dilakukan. Pertama adalah melakukan proses praproses. Dokumen yang telah melewati tahapan praproses ini dicek yaitu dalam tahapan pengecekan kata. Pengecekan kata ini dilihat berdasarkan data latih dengan metode *n-gram* per kata dan kamus. Setelah melakukan tahapan pengecekan kata, kata yang salah ejaan dilakukan perbaikan kata. Kata yang salah ejaan ini kemudian dicari kandidat katanya dengan *N-gram* dan selanjutnya didapatkan kandidat katanya dan dilakukan perhitungan kemiripan katanya dengan metode *Cosine Similarity*. Setelah didapatkan nilai kemiripan kataya, maka diambil nilai kemiripan kata yang tertinggi, kemudian disesuaikan kembali kata sebelum dan sesudahnya untuk mendapatkan kata yang benar, bila kata dengan kemiripan kata tertinggi tidak sesuai dengan data latih kata tersebut dianggap kata perbaikan yang benar . Rancangan sistem perbaikan kata ditunjukkan pada gambar :



A. *Proses Pengujian*

Pada proses pengujian tahapan dilakukan yaitu praproses teks, memeriksa kata salah ejaan dan memperbaiki kata salah ejaan. Pada saat pemeriksaan kata salah ejaan dan perbaikan kata menggunakan hasil dari proses pelatihan berupa kata, kata sebelum dan kata sesudahnya untuk disesuaikan. Pada saat pemeriksaan kata salah ejaan bila kata pada dokumen uji sesuai dengan hasil dari data latih maka dianggap kata benar dan bila tidak dilakukan pengecekan ke kamus kata. Pada saat perbaikan kata bila kandidat kata bersesuaian dengan hasil kata dari data latih maka kata itu dilakukan perbaikan kata.

B. *Proses Pelatihan*

Tahapan proses pelatihan yaitu praproses teks dan pemotongan per kata untuk mendapatkan kata, kata sebelum dan kata sesudah yang digunakan dalam proses pengujian.

IV. HASIL DAN PEMBAHASAN

Sistem ini diuji berdasarkan data uji dengan tiga tingkatan kesalahan kata dengan data yang sama yaiu tingkat 20 %, 50 %, dan 70 %. Dalam data uji kesalahan kata yang ada pada dokumen data uji yaitu kesalahan kata berupa penghilangan satu huruf dalam suatu kata (contoh : ginja seharusnya ginjal), penambahan satu huruf dalam suatu kata (contoh : kaesehatan seharusnya kesehatan), dan penukaran posisi huruf dalam suatu kata (contoh : kesehaatn seharusnya kesehatan).

A. *Hasil Percobaan 1*

Pada tabel ditunjukkan hasil pengujian dokumen dimana pada saat tahapan perbaikan digunakan *treshold* dalam mendapatkan kandidat kata dan kemudian disesuaikan dengan data latih, kandidat kata yang sesuai dengan data latih dianggap sebagai perbaikan kata benar dan bila tidak sesuai dianggap sebagai kata benar namun tidak dilakukan perbaikan.

TABEL I
HASIL PERCOBAAN PERBAIKAN DOKUMEN PADA TINGKAT 20%

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 20%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
1	1.txt	11	120	9,16 %
2	2.txt	7	133	5 %
3	3.txt	0	160	0 %
4	4.txt	3	117	2,5 %
5	5.txt	9	111	7,5 %
6	6.txt	11	109	9,16%
7	7.txt	8	152	5 %
8	8.txt	6	124	4,6 %
9	9.txt	5	115	4,16 %
10	10.txt	4	116	3,33 %
11	11.txt	3	127	2,3 %
12	12.txt	4	116	3,33 %
13	13.txt	5	145	3,33 %
14	14.txt	1	119	0,83 %
15	15.txt	0	120	0 %
16	16.txt	2	118	%
17	17.txt	3	120	1,6 %
18	18.txt	8	112	6,66 %
19	19.txt	4	116	3,33 %
20	20.txt	4	116	3,33 %

TABEL III
HASIL PERCOBAAN PERBAIKAN DOKUMEN PADA TINGKAT 50%

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 50%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
1	1.txt	30	220	12 %
2	2.txt	10	240	4 %
3	3.txt	12	238	4,8 %
4	4.txt	9	241	3,6 %
5	5.txt	10	240	4 %
6	6.txt	9	241	3,6 %
7	7.txt	14	286	4,6 %
8	8.txt	9	291	3 %
9	9.txt	3	297	1 %
10	10.txt	10	290	3,33 %
11	11.txt	4	296	1,3 %
12	12.txt	9	291	3 %
13	13.txt	14	286	4,6 %
14	14.txt	7	243	2,8 %
15	15.txt	10	290	3,33 %
16	16.txt	12	268	4,28 %
17	17.txt	4	246	1,6 %

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 50%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
18	18.txt	13	237	5,2 %
19	19.txt	7	293	2,33 %
20	20.txt	7	243	2,8 %

TABEL IIIII
HASIL PERCOBAAN PERBAIKAN DOKUMEN PADA TINGKAT 70%

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 70%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
1	1.txt	28	322	8 %
2	2.txt	24	326	6,85 %
3	3.txt	36	314	10,28 %
4	4.txt	14	336	4 %
5	5.txt	11	339	3,14 %
6	6.txt	17	333	4,85 %
7	7.txt	17	383	4,25 %
8	8.txt	14	406	3,33 %
9	9.txt	6	414	1,42 %
10	10.txt	8	412	1,9 %
11	11.txt	5	395	1,25 %
12	12.txt	8	412	1,9 %
13	13.txt	19	401	4,52 %
14	14.txt	3	347	0,85 %
15	15.txt	4	416	0,95 %
16	16.txt	4	396	1 %
17	17.txt	2	348	0,57 %
18	18.txt	2	348	0,57 %
19	19.txt	4	416	0,95 %
20	20.txt	7	343	2 %

Pada tabel hasil percobaan tabel 1, tabel 2 dan tabel 3 menunjukkan bahwa hasil akurasi dari percobaan rata-rata dibawah 10 %. Minimnya perbaikan kata yang tepat ini dikarenakan dalam penyesuaian kandidat kata yang tidak ada dalam data latih dan adanya kesalahan kata yang berurutan. Kata yang ada dalam data latih sangat sedikit yang bersesuaian dengan kandidat kata yang ada dan jika ditemukan kesalahan kata yang berurutan menyebabkan tidak dapat dalam penyesuaian ke dalam data latih. Oleh karena itu dalam perbaikan ini sangat dipengaruhi oleh data latih yang ada.

B. Hasil Percobaan 2

Pada tabel ditunjukkan hasil pengujian dokumen dengan data yang sama, namun dalam percobaan ini saat melakukan tahapan perbaikan nilai yang digunakan nilai kemiripan yang tertinggi dan saat penyesuaian dengan data latih, walaupun kata dengan kemiripan tertinggi tidak sesuai dengan data latih maka kata itu dianggap sebagai kata perbaikan yang benar.

TABEL IVV
HASIL PERCOBAAN PERBAIKAN DOKUMEN PADA TINGKAT 20%

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 20%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
1	1.txt	93	27	77,5 %
2	2.txt	110	30	78,57 %
3	3.txt	118	42	73,75 %
4	4.txt	93	27	77,5 %
5	5.txt	77	43	64,16 %
6	6.txt	98	22	81,66 %
7	7.txt	114	46	71,25 %
8	8.txt	97	33	74,61 %
9	9.txt	90	30	75 %
10	10.txt	86	34	71,66 %
11	11.txt	88	52	62,85 %
12	12.txt	84	36	70 %
13	13.txt	105	45	70 %
14	14.txt	85	35	70,83 %
15	15.txt	85	35	70,83 %
16	16.txt	84	36	70 %
17	17.txt	86	34	71,66 %
18	18.txt	85	35	70,83 %
19	19.txt	96	24	80 %
20	20.txt	88	32	73,33 %

TABEL V
HASIL PERCOBAAN PERBAIKAN DOKUMEN PADA TINGKAT 50%

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 50%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
1	1.txt	179	71	71,6 %
2	2.txt	192	58	76,8 %
3	3.txt	199	73	79,6 %
4	4.txt	203	47	81,2 %
5	5.txt	180	70	72,4 %
6	6.txt	220	30	88 %
7	7.txt	229	71	76,33 %
8	8.txt	230	70	76,66 %
9	9.txt	236	64	78,66 %
10	10.txt	219	81	73 %
11	11.txt	214	86	71,33 %
12	12.txt	214	86	71,33 %
13	13.txt	210	90	70 %
14	14.txt	180	70	72 %
15	15.txt	222	78	74 %
16	16.txt	203	77	72,5 %
17	17.txt	207	43	82,8 %
18	18.txt	208	42	83,2 %
19	19.txt	217	83	72,33 %
20	20.txt	185	65	74 %

TABEL VI
HASIL PERCOBAAN PERBAIKAN DOKUMEN PADA TINGKAT 70%

No	Hasil Percobaan Perbaikan Dokumen pada Tingkat 70%			
	Nama Dokumen	Ketepatan Kata	Kata yang Salah	Presentase
1	1.txt	275	75	78,57 %
2	2.txt	271	79	77,42 %
3	3.txt	254	106	72,57 %
4	4.txt	280	70	80 %
5	5.txt	258	92	73,71 %
6	6.txt	277	73	79,14 %
7	7.txt	292	108	73 %
8	8.txt	326	94	77,61 %
9	9.txt	327	93	77,85 %
10	10.txt	313	107	74,52 %
11	11.txt	287	113	71,75 %
12	12.txt	296	124	70,47 %
13	13.txt	299	121	71,19 %
14	14.txt	250	100	71,42 %
15	15.txt	322	98	76,66 %
16	16.txt	302	98	75,5 %
17	17.txt	300	50	85,71 %
18	18.txt	309	41	88,28 %
19	19.txt	332	88	79,04 %
20	20.txt	260	90	74,28 %

Pada table 4,5,6 hasil percobaan pada tingkat masing-masing kesalahan dengan masing- masing tingkatan sebanyak 20 data uji hasil akurasi yang didapatkan rata- rata di atas 70 %. Pada percobaan ini digunakan nilai kemiripan tertinggi dan bila tidak bersesuaian dengan data latih dianggap kata benar untuk mengatasi pada percobaan pertama. Dari data tersebut terdapat juga ketidak tepatan dalam memperbaiki kata disebabkan karena adanya kata bukan bahasa Indonesia. Banyaknya kesalahan kata juga mempengaruhi dalam hasil perbaikan kata. Kesalahan suatu kata yang diperbaiki ditentukan dari besar kemiripan kata dan data latih. Pada saat kata yang salah diperbaiki dipengaruhi oleh kamus trigram dalam mencari kemiripan kata dan data latih dalam menyesuaikan kemiripan kata. Namun dalam percobaan ini hasil dari kemiripan kata yang tertinggi yang tidak bersesuaian tetap dianggap sebagai kata benar, karena semakin mirip kata tersebut mendekati 1, maka kata tersebut semakin benar. Oleh karena itu dalam perbaikan kata dalam suatu dokumen ini, tidak hanya memerlukan data latih yang bagus namun juga kamus trigram yang lengkap untuk mendapatkan hasil kemiripan kata.

V. PENUTUP

Pada penelitian ini bahwa penelitian ini dengan menggunakan metode *n-gram* dan *cosine similarity* dapat dikatakan baik. Pada penelitian ini sangat bergantung pada kamus trigram dalam mencari kemiripan kata dan data latih yang ada. Pada ketidaktepatan dalam memperbaiki kata dikarenakan adanya kata yang bukan bahasa Indonesia dan nilai kemiripan kata di bawah batas yang ditentukan.

REFERENSI

[1] Widyaningsih, N. (2010). *Kalimat dalam Bahasa Indonesia*. Tersedia : http://lecturer.ukdw.ac.id/othie/PengertianKalimat.pdf

[2] Soleh,M. Y., & Purwarianti, A. (2011). A Non Word Error Spell Checker for Indonesian using Morphological Analyzer and HMM. *International Conference onElectrical Engineering and Informatics.*

[3] Permadi, Y. (2008). Kategorisasi Teks Menggunakan N-Gram Untuk Dokumen Berbahasa Indonesia. Bogor : Institut Pertanian Bogor.

[4] Sugianto, S.A & Liliana & Rostianingsih, S. (2014). *Pembuatan Aplikasi Predictive Text Menggunakan Metode N-gram Based*. Surabaya : Universitas Kristen Petra.

[5] Butgereit, L., & Botha, R.,. (2013). A Comparison of Different Calculations for N-Gram similarities in a Spelling Corrector for Mobile Instant Messagiing Language. *ACM 978-1-4503-2112-9*.