DAN IMPLEMENTASI IDENTIFIKASI PARAFRASA *TWEET* MENGGUNAKAN ALGORITMA BLEU, METEOR DAN EDIT DISTANCE

ANALYSIS AND IMPLEMENTATION OF PARAPHRASE IDENTIFICATION USING BLEU, METEOR AND EDIT DISTANCE ALGORITHM

Dennis Hidayat[1], M. Arif Bijaksana[2]

[1,2]Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

[1]dennishidayat@telkomuniversity.ac.id, [2]arifbijaksana@telkomuniveristy.co.id,

Abstrak

Situs *microblogging* Twitter merupakan contoh nyata dimana sebuah berita yang memiliki informasi dapat ditulis ulang dengan informasi yang sama dan konsep berbeda. Dalam mengenali suatu bentuk parafrasa dapat dilakukan oleh evaluasi manusia, namun identifikasi parafrasa oleh manusia membutuhkan biaya yang besar dan waktu yang lebih lama, hal ini bisa menjadi masalah besar untuk developer.

***Automatic metric* adalah sebuah mesin evaluasi otomatis yang menggunakan fitur-fitur yang dapat digunakan sebagai ekstraksi lingual sehingga menghasilkan nilai (*score*) yang dapat digunakan sebagai ukuran parafrasa dua buah kalimat yang dibandingkan. Pada penelitian ini digunakan tiga algoritma *automatic metric* yaitu BLEU, METEOR, Damerau-Levensthein *Edit Distance* yang menguji nilai parafrasa dari data Twitter yang sama. Selain itu dilakukan analisis terhadap performa algoritma dengan membandingkan nilai korelasi *human judgement* antara BLEU, METEOR, Damerau-Levensthein Edit Distance.**

Dari hasil simulasi yang dilakukan pada penelitian ini, diperoleh akurasi tertinggi dengan menggunakan *metric* METEOR dengan nilai akurasi 0,55 dan F1 sebesar 0,76.

Kata kunci : Identifikasi parafrasa, BLEU, METEOR, Edit Distance

Abstract

Twitter microblogging site is a real example where a news that has information can be rewritten with the same information and different concepts. In recognizing a form of paraphrase can be done by human evaluation, but the paraphrase identification by humans requires a large cost and a longer time, this can be a big problem for developers.

Automatic metric is an automated evaluation engine that uses features that can be used as a lingual extraction to produce a score that can be used as a paraphrase measurement of two sentences compared. In this research, three automatic metric algorithms are BLEU, METEOR, Damerau-Levensthein Edit Distance which test paraphrase values from the same Twitter data. In addition, an analysis of the performance of algorithms by comparing the correlation value of human judgment between BLEU, METEOR, Damerau-Levensthein Edit Distance.

From the simulation results done in this study obtained the highest accuracy by using METEOR metric with an accuracy of 0.55 and F1 of 0.76.

Keywords: Paraphrase identification, BLEU, METEOR, Edit distance

1. Pendahuluan

Natural Language Processing (NLP) adalah area penelitian dan pengaplikasan yang mengekplorasi bagaimana caranya sebuah komputer dapat digunakan dan memanipulasi berupa teks atau suara bahasa alami sehingga berguna untuk beragam hal [1]. NLP mempunyai tujuan untuk melakukan proses pembuatan model komputasi dari bahasa, sehingga dapat terjadi interaksi antara manusia dengan komputer dengan perantara bahasa alami. Secara umum NLP memodelkan pengetahuan terhadap fonetik dan fonologi, morfologi, sintaksis, semantik dan pragmatik pada bahasa [2].

Identifikasi Parafrasa merupakan sebuah percabangan dari studi NLP. Parafrasa menurut Harimurti Kridalaksana merupakan istilah linguistik yang berarti pengungkapan kembali suatu konsep dengan cara lain dalam bahasa yang sama, namun tanpa mengubah maknanya [3]. Laman berita merupakan contoh nyata, sebuah berita yang memiliki informasi dapat ditulis ulang dengan informasi yang sama dan konsep berbeda, walaupun kedua berita tersebut

memiliki elemen leksikal yang berbeda atau mungkin memiliki struktur sintaksis yang berbeda namun memiliki makna yang sama dapat disebut sebagai parafrasa [4]. Dalam mengenali suatu bentuk parafrasa dapat dilakukan oleh evaluasi manusia, namun evaluasi parafrasa oleh manusia membutuhkan biaya yang besar dan waktu yang lebih lama, hal ini bisa menjadi masalah besar untuk *developer* [5].

Automatic metric adalah sebuah mesin evaluasi otomatis yang menggunakan *metric* yang dapat digunakan sebagai ukuran identifikasi parafrasa. Pada beberapa *automatic metric* menggunakan fitur-fitur yang dapat digunakan sebagai ekstraksi lingual sehingga menghasilkan nilai (*score*) yang dapat digunakan sebagai ukuran parafrasa dua buah kalimat yang dibandingkan [6] Pada penelitian ini penulis menggunakan tiga algoritma automatic metric yaitu BLEU, METEOR, Damerau-Levensthein *Edit Distance* yang menguji data testing yang sama. Kemudian hasil dari masing-masing automatic metric akan dihitung nilai akurasinya dengan menggunakan F1-*measure*. *Dataset* yang digunakan berasal dari data Twitter yang berupa *corpus* yang sudah di-format ke dokumen ber-ekstensi *.txt*.

Berdasarkan latar belakang yang telah disampaikan, dilakukan sebuah penelitian Analisis dan Implementasi algoritma BLEU, METEOR, Damerau-Levensthein *Edit Distance* untuk mengidentifikasi parafrasa, selanjutnya hasil dari identifikasi oleh *automatic metric* akan evaluasi dengan hasil korelasi masing-masing algoritma dengan *human judgement* dengan melakukan melakukan perhitungan nilai akurasi dan F1. Dengan melakukan perhitungan nilai akurasi dan F1 dapat menjadi pertimbangan untuk melihat performansi *metric* dimana dapat melihat seberapa akurat *automatic metric* jika dibandingkan dengan penilaian manusia.

2. Dasar Teori

2.1 Identifikasi Parafrasa

Parafrasa dapat dipandang sebagai sebuah teknik analisis data, teknik parafrasa atau bisa disebut teknik ubah wujud, adalah teknik mengubah wujud satuan data lingual yang dianalisis dengan tetap mempertahankan informasi. Penggunaan teknik parafrasa itu dalam analisis selalu mengakibatkan wujud salah satu atau beberapa unsur satuan lingual yang bersangkutan. Teknik parafrasa merupakan salah satu teknik analisis data dalam metode distribusional, di dalam analisis data teknik parafrasa berguna untuk menganalisis satuan-satuan lingual pada tataran sintaksis. [8]

Parafrasis merupakan tindakan atau kegiatan untuk membuat parafrasa, dalam pembuatan parafrasis seseorang harus membaca keseluruhan data lingual dan mengambil kata-kata kunci dari teks tersebut lalu mengembangkan kata kunci tersebut dikembangkan menjadi gagasan pokok, kemudian gagasan pokok dikembangkan menjadi sebuah paragraf dengan cara berbeda. [7]

Dengan menggunakan pemahaman diatas dapat disimpulkan bahwa identifikasi parafrasa merupakan sebuah kegiatan dalam menganalisa kemungkinan parafrasa pada dua data lingual atau lebih yang berbeda.

2.2 Evaluasi Mesin Translasi

Sebagaimana sebuah mesin transalasi yang baik adalah mampu mendekati hasil dari terjemahan manusia, maka mengkorelasikannya dengan evaluasi manual manusia dibutuhkan untuk menganalisis data dari mesin terjemahan. Evaluasi oleh manusia pada mesin penterjemah teks (*Machine Translation*) menimbang banyak aspek pada terjemahannya, yaitu mencakup kecukupan, keakuratan, dan kelancaran (*adequacy*, *fidelity, fluency*) [9]. Walaupun dalam penerapannya evaluasi manusia membutuhkan biaya yang lebih besar dan waktu yang lebih lama. Maka dibutuhkan evaluasi mesin transalasi otomatis yang lebih efektif dari segi biaya dan waktu [10].

Evaluasi mesin transalasi otomatis digunakan dalam bidang *natural language processing* untuk mengukur kualitas terjemahan dari mesin translasi dengan menggunakan metric, metric digunakan sebagai penilaian subjektif dari hasil output mesin translasi. Secara sederhana evaluasi manusia merupakan rujukan terbaik dalam penilaian mesin transalasi maka dari itu tugas utama dari *metric* adalah mengeluarkan *score* (nilai) terhadap hasil mesin translasi, yang mana *score* tersebut dapat dikorelasikan dengan penilaian manusia [11]. Terdapat lima faktor penting lainnya yang menunjukan *metric* evaluasi mesin translasi otomatis tersebut baik yaitu korelasi, konsistensi, sensitivitas, reliabilitas dan generalitas. Setiap metrik yang baik harus berkorelasi tinggi dengan penilaian manusia, harus konsisten memberikan hasil yang serupa dengan sistem evaluasi yang sama dengan teks serupa. harus peka terhadap perbedaan antar sistem yang digunakan. dan dapat diandalkan dimana dalam sistem yang memiliki *score* serupa juga dapat dilakukan dengan cara yang sama. Terakhir, *metric* harus bersifat umum, yaitu dapat bekerja dengan domain teks yang berbeda, dalam berbagai tugas dan skenario [12] [13].

2.3 *Bilingual Evaluation Understudy* (BLEU)

Dasar dari algoritma BLEU adalah melakukan perbandingan *n-gram* pada kalimat hipotesis dengan *n-gram* pada kalimat referensi dan dan menghitung jumlah kecocokan kata pada setiap kalimat, semakin banyak kecocokan menunjukan semakin baik nilai parafrasanya. Modifikasi nilai presisi *n-gram* adalah cara untuk melakukan pengukuran *precision* salah satunya dengan perbandingan *unigram* dengan menghitung jumlah total kata yang cocok dari kalimat hipotesis dengan kata dari kalimat referensi dan dibagi dengan jumlah total kata dari kalimat hipotesis, seperti terlihat pada persamaan (2,1).

$$Pn = \frac{\sum_{C \in \{Kandidat\}} \cdot \sum_{n-gram \in C} \cdot \text{Jumlah}_{potongan}(n-gram)}{\sum_{C' \in \{Kandidat\}} \cdot \sum_{n-gram' \in C} \cdot \text{Jumlah}_{potongan}(n-gram')} \qquad (2,1)\ [11]$$

Dalam menghindari ketidakefektifan dari kemungkinan munculnya kalimat dengan nilai presisi tinggi yang kemungkinan muncul akibat dari kalimat hipotesis yang pendek dibandingkan dengan kalimat referensi yang lebih panjang pada BLEU dengan menggunakan konstanta *brevity penalty* yaitu penguraian eksponen pada r/c dimana r adalah jumlah total kata dari kalimat referensi dan c adalah jumlah total kata dari kalimat hipotesis dengan dikalikan rata-rata geometri dari nilai modifikasi n-gram. Seperti terlihat pada persamaan (2,2).

$$\text{BP} = \begin{cases} 1 & if\ c>r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \text{ maka, BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{N} w_n \log p_n) \qquad (2,2)\ [16]$$

2.4 *Machine Translation Evaluation System* (METEOR)

METEOR menggunakan dan menekankan *recall*. Sebuah properti penting yang telah dikonfirmasi oleh bebeberapa metric karena memiliki korelasi tinggi dengan peniliaian manusia. METEOR juga membahas permasalahan perbedaan terjemahan dengan memanfaatkan pencocokan kata fleksibel yang memungkinkan varian morfologi dan sinonim agar dapat diperhitungkan dalam korespondensi yang sah. Fitur lain dalam METEOR adalah parameter yang memungkinkan penyetelan parameter bebas *metric* dalam mencari nilai-nilai yang menghasilkan korelasi optimal dengan penilaian manusia.

METEOR mengevaluasi dengan menghitung skor menggunakan *word-to-word matching* yaitu pencocokan kata ke kata yang eksplisit antar kalimat hipotesis dan kalimat referensi. Jika terdapat beberapa referensi maka dilakukan penghitungan skor antar kalimat hipotesis dengan kalimat referensi secara independen dan menggunakan *score* hasil dengan nilai tertinggi yang dipilih. Untuk setiap kalimat referensi yang akan dibandingkan *The Meteor Matcher* akan digunakan *alignment* (penyelarasan) untuk membuat penyelarasan kata antara hipotesis dan referensi. Untuk melakukan penyelarasan dilakukan pemetaan antar *unigram* sedemikian rupa, dalam penyelarasan sebuah unigram tidak bisa melakukan pemetaan lebih dari satu pada kalimat referensi dan kalimat hipotesis itu sendiri. Setelah penyelarasan akhir dilakukan, antara terjemahan hipotesis dan terjemahan referensi, *score* METEOR dihasilkan berdasarkan pada jumlah pemetaan *unigram* yang ditemukan antar kedua buah *string* (m), jumlah total *unigram* pada hipotesis (t) dan jumlah total *unigram* pada referensi (r), kemudian menghitung *unigram precision*, dengan persamaan (2.4). [12]

$$P = \frac{m}{t} \qquad (2,4)\ [9]$$

R = m / r kemudian menghitung parameter F1 yaitu nilai harmonic mean dari P dan R dengan persamaan 2.6

$$R = \frac{m}{r} \qquad (2,5)\ [17]$$

Precision, *recall*, dan *Fmean* semua berdasarkan perbandingan satu buah kata tunggal, untuk penjelasan lebih jauh *unigram* mana pada kedua terjemahan yang masih dalam ururtan yang sama dilakukan penghitungan fragmentasi sebagai berikut. Pertama, urutan pada *unigram* yang cocok pada kedua buah terjemahan dipisah hingga menjadi bagian terkecil sehingga unigram yang cocok pada setiap pecahan berdekatan (di kedua buah *string*) dan pada urutan yang identikal. Jumlah pecahan atau *chunk* (ch) dimana chunk didefinisikan sebagai serangkaian matched yang berdekatan dan identik yang berurutan dalam kedua kalimat dan jumlah kecocokan antara kalimat referensi dengan kalimat hipotesis (m) kemudian digunakan untuk menghitung fraksi (γ) fragmentasi : frag = ch / m [14]. ini kemudian dihitung seperti pada persamaan 2.7.

$$Pen = \gamma \cdot \frac{\#Chunk}{\#Unigram\_matched}^{3} \qquad (2,7)\ [12]$$

Nilai dari γ menentukan penalty maksimal (0 ≤γ ≤1) nilai tersebut menentukan hubungan fungsional antara fragmentasi dan *penalty*-nya. Yang akhirnya akan menghasilkan nilai keselarasan dengan formula yang didefinisikan dengan persamaan 2.8. [17]

$$Score = (1 \cdot Pen) \cdot F_{mean} \quad (2,8)\ [12]$$

2.5 Damerau-Levensthein *Edit distance*

Edit distance merupakan cara untuk menilai seberapa perbedaan antar dua buah *string* dengan menghitung transformasi dari jumlah penyisipan, penghapusan atau subtitusi yang dibutuhkan pada string agar menciptakan kucukupan kesamaan pada *string* yang dituju, semakin besar transformasi yang dibutuhkan maka menunjukan semakin besar jarak antar *string*. Damerau-Levensthein *Edit distance* berusaha untuk melakukan kalkulasi minimum yang dibutuhkan untuk memenuhi kesamaan dua buah kalimat yang dibandingkan dengan melakukan operasi pengapusan, subtitusi, penyisipan dan transposisi. Perhitungan Damerau-Levensthein *Edit distance* dapat diformulasikan dengan persamaan 2.9. [19]

$$Lev_{a,b}(i,j) = \begin{cases} \max(i,j) \\ min \begin{cases} Lev_{a,b}(i-1,j)+1 \\ Lev_{a,b}(i,j-1)+1 \\ Lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} jika \min(i,j) = 0, lainnya \end{cases} \quad (2,9)\ [19]$$

Untuk setiap operasi yang dilakukan sesuai dengan kasus yang cocok dilakukan :
d_(a,b) (i-1,j)+1, untuk setiap operasi penghapusan.
d_(a,b) (i,j-1)+1, untuk setiap operasi penyisipan.
d_(a,b) (i-1,j-1)+1_((a_i≠b_j)), untuk setiap operasi subtitusi.
d_(a,b) (i-2,j-2)+1, untuk setiap operasi transposisi.

Sementara itu untuk mencari nilai *similarity* antara kedua *string* yang dibandingkan dapat menggunakan persamaan (2,10)

$$sim = 1 - \left(\frac{Dis}{MaxLength}\right) \quad (2,10)\ [20]$$

3. Analisis dan Implementasi

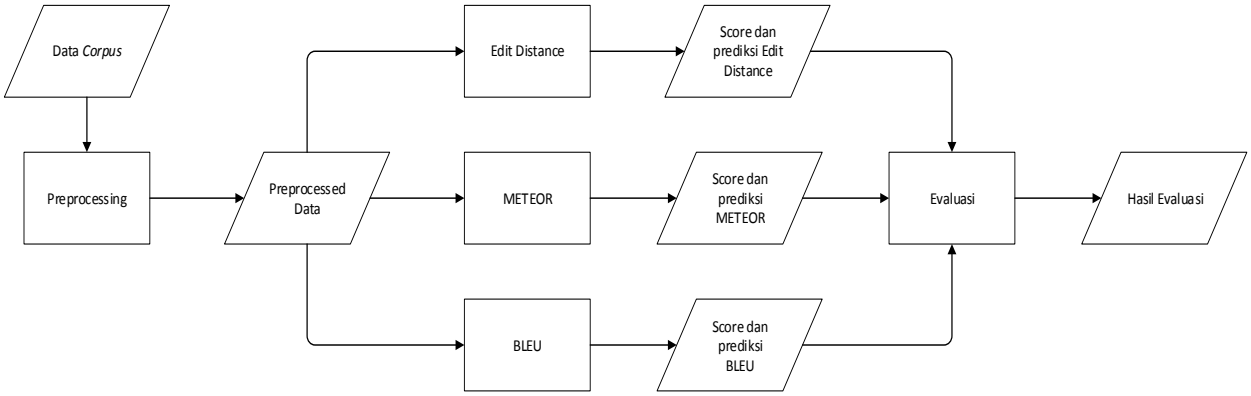
3.1 Perancangan Sistem

Aplikasi yang akan dibangun akan menggunakan data *corpus* yang terdiri dari sepasang kalimat dan penilaian parfrasa dari pakar bahasa pada setiap segmen. Data ini akan disimpan dalam dokumen dengan ekstensi *.txt* dengan format spasi untuk pengenalan pembeda kata dan *tab* untuk setiap label dan *enter* untuk membedakan antar segmen. Langkah pertama yang dilakukan oleh aplikasi adalah dengan *mining* data *corpus* dari dokumen yang disimpan dengan ekstensi *.txt* kemudian dilakukan tokenasasi terhadap data pada tahapan *preprocessing* dengan harapan dapat mengurai *noise* dalam *corpus* dan meningkatkan kualitas data. proses *preprocessing* selanjutnya akan dilakukan *word replacer* untuk mengganti kata menjadi (*lowercase*) dan menhilangkan tanda baca. Tahap selanjutnya adalah melakukan *stemming* sehingga kata dalam *string* menjadi bentuk kata dasar, *stop words* untuk menghilangkan kata penghubung dalam kalimat dan *part-of-speech tagging* untuk memberikan label *noun, pronoun, verb, adjective, adverb* dan *preposition* pada setiap kata. Langkah selanjutnya adalah melakukan identifikasi parafrasa pada data dengan menggunakan *automatic metric*.

Hasil dari tahap *preprocessing* akan di gunakan untuk perhitungan *score* pada *automatic metric* yang dilakukan secara independen. *Score* dari *metric* digunakan sebagai ukuran dalam menilai parafrasa, BLEU akan melakukan *scoring* dengan penghitungan *n-gram*(s) pada kalimat hipotesis dan kalimat referensi setiap segmen menggunakan *modified n-gram(s) precision* dan mencari nilai *brevity penalty* lalu melakukan perhitungan *score* dengan menggunakan logaritma BLEU. Pada METEOR dilakukan perhitungan score dengan langkah pemetaan kata identik dengan metode *word-to-word matching*. Damerau-Levensthein *Edit Distance* melakukan kalkulasi minimum yang

dibutuhkan untuk memenuhi kesamaan dua buah kalimat yang dibandingkan dengan melakukan operasi penghapusan, subtitusi, penyisipan dan transposisi.

Penilaian parafrasa atau non-parafrasa pada tiap metric akan dilakukan perbandingan dengan *human judgement* pada tahap evaluasi. Maka secara keseluruhan keluaran dari aplikasi ini adalah berupa nilai prediksi parafrasa atau non-parafrasa serta akurasi dan F1 dari setiap *metric*. Evaluasi sistem dilakukan untuk mencari nilai performansi dari ketiga algoritma *automatic metric* yang digunakan. Pada pengujian ini performansi sistem akan dievaluasi menggunakan nilai *precision, recall, accuracy* dan F1-*measure. Flowchart* perancangan sistem dapat terlihat pada gambar 1.



Gambar 1. Flowchart Perancangan Sistem

3.2 Perhitungan BLEU *Score*

Untuk mendapatkan nilai identifikasi parafrasa hasil dari hasil *preprocessing* akan diolah oleh BLEU terlebih dahulu. Dimana BLEU melakukan perhitungan segmen per segmen jumlah kecocokan *n-gram(s)* kalimat 1 pada *n-gram(s)* kalimat 2 dengan menggunakan *modified n-gram(s) precision*. Setiap kemunculan, maka *score precision* pada *n-gram(s)* tertentu akan bertambah satu. selanjutnya, dilakukan perhitungan *Breavity Penalty* dan selanjutnya nilai BP tersebut ditambahkan dan dirata-ratakan, untuk menghasilkan nilai BP. selanjutnya adalah perhitungan *score* (nilai) BLEU dengan memasukkan ke dalam persamaan BLEU.

3.3 Perhitungan METEOR *Score*

Data hasil *preprocessing* yang disimpan dalam *variable* dipanggil untuk diolah METEOR. Setiap segmen pada *corpus data* dilakukan pemetaan kalimat 1 dan kalimat 2 dimana keluarannya berupa *alignment* dari pasangan *unigram* yang cocok berdasarkan bentuknya yang identik. Dari proses pemetaan tersebut didapatkan jumlah pasangan *matched unigram* dan *chunks* untuk mendapatkan nilai *score* menggunakan rumus METEOR.

3.4 Perhitungan Damerau-Levensthein *Edit distance Score*

Algoritma *Edit Distance* digunakan untuk mencari kecocokan antara dua buah *string*. Setiap *string* dimasukkan ke dalam matriks dua dimensi dengan panjang baris dan kolom matriks tergantung pada panjang *string* yang akan dibandingkan. Selanjutnya, algoritma melakukan transformasi yang dibutuhkan untuk mencari kesamaan antar string dengan melakukan subtitusi, penghapusan, penyisipan dan transposisi dari transformasi yang dilakukan didapatkan nilai jarak. Kemudian untuk melakukan similarity dapat dilakukan dengan menggunakan rumus *similariy* sehingga didapatkan nilai akhir antara 0 untuk tidak sama sekali dan 1 untuk kemiripan sempurna.

3.5 Perancangan evaluasi Sistem

Hasil dari *judgement* ketiga algoritma dikorelasikan *human judgement* dengan mencari nilai *precision* dan *recall* untuk mendapatkan nilai akurasi dan F1. Setiap segmen mempunyai sepasang kalimat yang telah di-*scoring* oleh algoritma dengan pengetahuan tersebut bisa didapatkan nilai akurasi dengan rumus yang ada dan dibandingkan dengan nilai parafrasa dari *human judgement.* Alur perangan evaluasi terlihat pada gambar 3.6.

3.6 Pengujian Algoritma

Implementasi identifikasi parafrasa dilakukan dengan algoritma BLEU, METEOR dan Damerau-Levensthein *Edit Distance* untuk melihat bagaimana hasil identifikasi parafrasa yang dihasilkan oleh *automatic metric* untuk dibandingkan dengan *human judgement* menghasilkan *score* seperti terlihat pada tabel 1.

Tabel 1 Pengujian Identifikasi Parafrasa Dengan Algoritma METEOR

Kalimat 1	**Kalimat 2**	***Score* dan status parafrasa**	***Score* dan status parafrasa**	***Score* dan status parafrasa**	**Status parafrasa oleh pakar bahasa**
EJ Manuel the 1st QB to go in this draft	But my bro from the 757 EJ Manuel is the 1st QB gone	0,88/YA	0, 57/YA	0,88/YA	YA
EJ Manuel the 1st QB to go in this draft	Can believe EJ Manuel went as the 1st QB in the draft	0,87/YA	0, 71/YA	0,87/YA	YA
EJ Manuel the 1st QB to go in this draft	EJ MANUEL IS THE 1ST QB what	0,88/YA	0, 66/YA	0,88/YA	YA
EJ Manuel the 1st QB to go in this draft	EJ da 1st QB off da board	0,88/YA	0, 50 /YA	0,88/YA	TIDAK
EJ Manuel the 1st QB to go in this draft	Manuel is the 1st QB to get drafted	0,84/YA	0, 66/YA	0,84/YA	YA
EJ Manuel the 1st QB to go in this draft	My boy EJ Manuel being the 1st QB picked	0,91/YA	0, 46/YA	0,91/YA	YA
EJ Manuel the 1st QB to go in this draft	Not surprised EJ Manuel was 1st QB taken	0,82/YA	0, 72. /YA	0,82/YA	YA
EJ Manuel the 1st QB to go in this draft	WOW EJ MANUEL FSU 1ST QB TAKEN	0,85/YA	0, 76 /YA	0,85/YA	YA
EJ Manuel the 1st QB to go in this draft	Wow EJ Manuel 1st QB taken in the draft	0, 22/ TIDAK	0, 53. /YA	0, 53. /YA	YA
EJ Manuel the 1st QB to go in this draft	if EJ is the 1st QB off the board	0, 24/ TIDAK	0, 50 /YA	0, 50 /YA	YA

Berikut ini merupakan hasil evaluasi untuk mencari keakurasian algoritma METEOR, BLEU dan Edit Distance dalam identifikasi parafrasa dua buah kalimat. Pengujian dilakukan dengan menggunakan jumlah dataset yang berbeda. score seperti terlihat pada tabel 2 berikut.

Tabel 2 Range Standar Deviasi dari Gambar

Keterangan	Nilai					
metric	METEOR		BLEU		Edit Distance	
Jumlah Data uji	100	13064	100	13064	100	13064
Precision	0,73	0,67	0,66	0,79	0,35	0,29
Recall	0,59	0,46	0,054	0,02	0,91	0,92
Akurasi	0,77	0,55	0,64	0,05	0,35	0,44
F1	0,65	0,76	0,10	0,70	0,51	0,30

4. Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa:

1. Algortima BLEU, Meteor dan Damereu-levensthein Edit distance dalam mengidentifikasi parafrasa dua buah tweet dapat diimplementasikan pada data Twitter menggunakan bahasa pemprograman phyton versi 2.7.
2. Nilai akurasi dan F1 menggunakan algortima BLEU, METEOR dan Damereu-levensthein Edit distance dalam mengidentifikasi parafrasa dua buah tweet secara berturut-turut adalah 0,05 dan 0,70 lalu 0,55 dan 0,76 lalu 0,44 dan 0,30.
3. Dengan menggunakan nilai akurasi dan F1 perbandingan algoritma BLEU, METEOR dan Damereu-levensthein Edit distance. Diperoleh bahwa METEOR memiliki nilai akurasi terbaik dan BLEU memiliki nilai akurasi terburuk.

4.2 Saran

Adapun saran untuk pembangunan sistem dalam penelitian yang serupa agar dapat meningkatkan efisiensi sitem adalah sebagai berikut:

1. Menggunakan data *corpus* uji yang lebih baik.
2. Melakukan *stemming* dan *stop words* lebih baik mungkin dapat meningkatkan akurasi.
3. Mengklasifikasikan data menggunakan pendekatan yang ada untuk meningkatkan performansi sistem.
4. Menggunakan metode *n-gram* untuk pengklasifikasian data di dalam pencocokan agar frasa dapat tertangani.
5. Hasil Identifikasi Parafrasa dengan automatic metric yang telah dibangun dapat dibandingkan dengan hasil pengukuran SemEval-2015 Task 1 untuk mengetahui seberapa efektif dan efisien sistem yang telah dibangun.

Daftar Pustaka:

[1] C. G. G, "Natural Language Processing," in *University of Strathclyde*, Glasgow.

[2] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," in *Artificial Intelligence Research 57*, Israel, 2016.

[3] H. Kridalaksana, Kamus Linguistik, Jakarta: Gramedia Pustaka Utama, 2008.

[4] J. Cordeiro, G. Dias and P. Brazdil, "A Metric for Paraphrase Detection," in *Computing in the Global Information Technology*, Nice, 2007 .

[5] P. Koehn, "Statistical Significance Test for Machine Translation Evaluation," in *Massachusetts Institute of Technology*, Cambridge, MA.

[6] D. M. McKeown, S. Cochran, D. T. Bulwinkle, W. Harvey, C. McGlone and J. A. Shufelt, "PERFORMANCE EVALUATION FOR AUTOMATIC FEATURE EXTRACTION," in *international Archives of Photogrammetry and Remote Sensing. Vol. XXXIII*, Amsterdam , 2000.

[7] Sudaryanto, Linguistik : Esai Tentang Bahasa dan Pengantar ke Dalam Ilmu Bahasa, Yogyakarta: Gadjah Mada Press, 1983.

[8] Sudaryanto, Metode dan Teknik Analisis Bahasa, Yogyakarta: Masyarakat Linguistik Indonesia Komisariat Universitas Gadjah Mada Press, 1985.

[9] S. J. White, T. O'Connell and S. F. O'Mara, *The ARPA MT Evaluation Methodologies:Evolution, Lessons, and Future Approaches*, 1994.

[10] A. Agarwal and A. Lavie, "Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output," in *Third Workshop on Statistical Machine Translation*, Colombus, Ohio, 2008.

[11] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *40th Annual Meeting of the Association for omputational Linguistics (ACL)*, Philadelphia, 2002.

[12] S. Benerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Carnegie Mellon University*, Pittsburgh.

[13] P. Koehn, "Statistical Significance," Massachusetts Institute of Technology, Cambridge. MA.

[16] S. L. Hadla, M. T. Hailat and N. M. Al-Kabi, "Comparative Study Between METEOR and BLEU Methods of MT: Arabic into English Translation as aCase Study," *International Journal of Advanced Computer Science and Applications*, 2015.

[17] M. Denkowski and A. Lavie, "Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, p. pages 250–253, June 2010.

[19] D. Jurafsky and J. H. Martin, "Spelling Correction and the Noisy Channel," *Speech and Language Processing*, 2016.