Penentuan Fitur Supervised Learning dalam Identifikasi Kalimat Sitasi pada Makalah Ilmiah

Determining Supervised Learning Feature in Identifying Citing Sentence in Scientific Paper

Rian Putra Mantovani[1], Yuliant Sibaroni S.Si., M.T[2] Annisa Aditsania, S.Si., M.Si.[3]
[1]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[2]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[3]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[1]mantovanirian@gmail.com, [2]ysibaroni@gmail.com, [3]annisaaditsania@gmail.com

Abstrak — Kalimat sitasi berperan penting dalam penulisan jurnal ilmiah. Kalimat sitasi dapat diidentifikasi dengan mengekstraksi fiturnya. Pada penelitian ini digunakan 5 fitur utama dan juga akan dikombinasikan. Fitur-fitur yang kita gunakan adalah unigram, bigram, proper noun, cue phrase, dan pronoun. Untuk mengklasifikasi kita menggunakan *Naive Bayes* (NB) dan *support vector machine* (SVM). Penelitian ini menggunakan 500 makalah ilmiah yang diambil dari acl-arc. Hasil dari penelitian ini adalah fitur yang terbaik untuk mengidentifikasi kalimat sitasi adalah "*Proper Noun*, dan *Cue Phrase*" dengan 59,069% f-measure, dan 92,157% akurasi, jika menggunakan *naive bayes*, dan 51,234% *f-measure*, dan 92,503% akurasi jika menggunakan SVM.

Kata Kunci —*supervised learning*, ekstraksi fitur, identifikasi, kalimat sitasi

Abstract— Citing sentence plays a role at scientific paper. Citing sentence could be identified with extracting it features. In this paper we extract and combine 5 features. The features that we used are n-gram, proper noun, cue phrase, and pronoun. To classify citing sentence we use naive bayes and support vector machine(SVM). The results of this research is the best feature to identify citing sentence is "proper noun, and cue phrase" with 59,069% of f-measure, and 92,157% of accuracy, if we use naive bayes classifier, and 51,234% of f-measure and 92,503% of accuracy, if we use SVM classifier.

Keywords— supervised learning, feature extraction, identifying, citing sentence

1. Pendahuluan

Sitasi adalah daftar pustaka dari sejumlah dokumen yang dirujuk atau yang dikutip oleh sebuah dokumen dan setiap daftar pustaka dokumen tersebut dimuat dalam bibliografi dokumen yang mengutip, yang secara khusus mengkaji pengarang dan karya-karya lain. Bisa juga di definisikan untuk menunjukkan asal-usul atau sumber suatu kutipan, mengutip pernyataan atau menyalin/mengulang pernyataan seseorang dan mencantumkannya di dalam suatu karya tulis yang dibuat, namun tetap mengindikasikan bahwa kutipan tersebut itu adalah pernyataan orang lain [1].

Kalimat sitasi sering digunakan pada artikel atau paper ilmiah dengan maksud mereferensikan pekerjaan ilmiah seseorang yang telah dipublikasikan seperti, jurnal, buku, paper, dan makalah. Salah satu contoh sitasi adlah jika sebuah kalimat diikuti dengan kurung siku maka kalimat tersebut adalah sebuah kalimat sitasi [2]. Kalimat sitasi juga mempunyai peran penting dalam perkembangan ilmu pengetahuan. Ketika sebuah referensi muncul di sebuah artikel ilmiah, biasanya diikuti dengan teks yang berisi tentang peranan penting paper yang telah dikerjakan [3].

Pengidentifikasian kalimat sitasi berguna untuk membantu para akademisi untuk mengecek kembali jurnal ilmiah yang ditulis. Pengidentifikasian sitasi berguna untuk ilmuwan untuk melihat seberapa jauh sebuah penelitian telah berkembang.

Jurnal ilmiah menjadi aset suatu institusi atau lembaga pendidikan sebagai wadah untuk mempublikasikan hasil penelitian dari dosen, mahasiswa, dan praktisi. Jika penelitian pengidentifikasian sitasi didalami lebih lanjut, ilmuwan dapat mengetahui apakah makalah ilmiah yang dia tulis sering direferensikan. Semakin sering sebuah penelitian direferensikan oleh penelitian lain maka penelitian tersebut berguna dalam pengembangan topik yang sedang diteliti.

2. *Classifier yang digunakan*

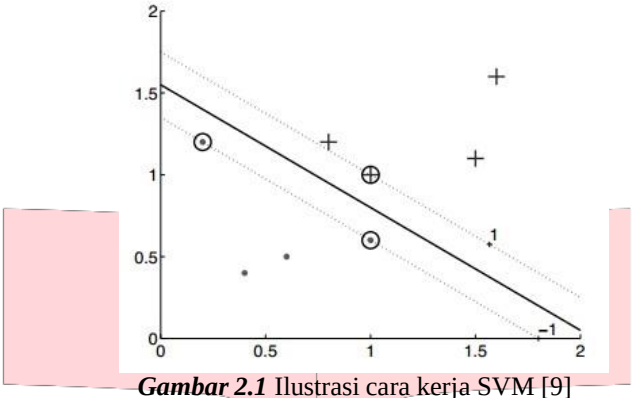
Dalam penelitian ini ada dua *classifier* yang digunakan dalam penulisan kalimat sitasi dan dijelaskan juga oleh Kumar et al [2] *classifier* tersebut antara lain SVM dan *naive bayes*. Pada penelitian ini *classifier* menggunakan aplikasi weka yang diimplementasikan di dalam aplikasi yang peneliti buat.

2.1 Support Vector Machine

Pada [12] dijelaskan bahwa SVM pertama kali diperkenalkan oleh Guyon Boser dan Vapnik, SVM adalah salah satu cara untuk melakukan klasifikasi dengan meggunakan teknik regresi. Dengan kata lain SVM adalah

alat prediksi klasifikasi dan regresi yang menggunakan teori dari machine learning untuk memaksimalkan akurasi dengan cara menghidari overfit pada data secara automatis. S*upport vector machine* dapat juga didefinisikan sebagai sistem yang menggunakan ruang hipotesis dari fungsi linier yang didepat dari algoritma yang diturunkan dari statistical learning.

Ide dari SVM ini adalah membagi dua buah data sehingga terbentuk dua buah kelas. Tetapi untuk membagi data menjadi dua kelas hyperplane yang memungkinkan cukup banyak dan harus ditentukan hyperplane yang terbaik untuk membagi kelas tersebut. SVM adalah salah satu algoritma klasifikasi terfavorit karena selain mempunyai dasar teori yang kuat SVM juga mengklasifikasi lebih akurat dari algoritma lainnya di banyak bidang seperti klasifikasi web page dan bioinformatika. Dalam gambar 2.1, support vector digambarkan dengan data yang dilingkari [9].



Gambar 2.1 Ilustrasi cara kerja SVM [9]

Hyperplane yang paling optimal adalah yang memiliki margin terbesar. Dengan margin yang besar, maka kita dapat menghindari kesalahan dalam pengklasifikasian data. Contohnya, jika terdapat sebuah data baru pada kelas "+" yang posisinya berada sedikit lebih dekat ke kelas "•", maka data tersebut masih dapat diklasifikasikan ke kelas yang tepat. Jika kita memiliki suatu data $\mathcal{X} = \{$ dengan 2 kelas berbeda, yaitu dan , di mana $= +1 \in$ dan $= -1$ jika $\in$, persamaan garis hyperplane dapat dinyatakan sebagai berikut:

$$^{T} + _{0} = 0 \quad (2.1)$$

$$= \sum_{i} _{i} \quad (2.2)$$

$$= 1,2,\ldots,N$$

keterangan:

$_{i}$: Lagrange multiplier

: label kelas, bernilai 1 untuk $\in$ atau -1 jika $\in$

: fitur dari data

N : jumlah data

Nilai $_{i}$dapat dicari dengan memaksimalkan persamaan berikut:

$$L_{d} = -\frac{1}{2}\| \|^{2} - \sum_{i} _{i} ( )^{T} + \sum_{i} _{i} \quad (2.3)$$

$$\text{s.t. } \sum_{i} _{i} = 0 \text{ dan } _{i} \geq 0, \forall$$

Setelah semua nilai $_{i}$ didapatkan, beberapa akan memiliki nilai $_{i} = 0$ dan hanya sebagian kecil yang memiliki nilai $_{i} > 0$. Kumpulan dengan nilai $_{i} > 0$ adalah yang disebut dengan support vector. Kemudian nilai bisa didapatkan dengan rumus berikut:

$$_{0} = - ^{T} \quad (2.4)$$

Untuk menguji kelas data *testing* dapat menggunakan rumus berikut:

$$( ) = (^{T} + _{0}) \quad (2.5)$$

jika $( ) = 1$ pilih , atau jika $( ) = -1$ pilih .

2.2 Naive Bayes

Naive bayes adalah salah satu jenis bayesian yang sering digunakan pada klasifikasi dan *clustering*, tetapi masih banyak kegunaan dari *naive bayes* yang belum dieksploitasi. *Naive Bayes* merepresentasikan distribusi

yang beragam dari banyak komponen dimana setiap komponen pada semua variabel dianggap independen antara satu dan lainnya [7]. *Naive bayes* juga merupakan metoda klasifikasi yang berakar pada teorema Bayes.

Ciri utamanya adalah asumsi yang naif akan independesi atas kejadian. Pada teorema Bayes, jika terdapat dua kejadian yang terpisah (misal A dan B) maka teorema Bayes dapat dirumuskan sebagai berikut:

$$P(A \mid B) = \frac{P(A)}{P(B)} P(B \mid A) \tag{2.7}$$

Keterangan:

$P(A \mid B)$: Peluang kejadian A bila B terjadi

$P(A)$: Peluang kejadian A

$P(B)$: Peluang kejadian B

$P(B \mid A)$: Peluang kejadian B bila A terjadi.

Teorema Bayes sering juga dikembangkan karena berlakunya hukum probabilitas total menjadi seperti berikut:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{\sum_{i=1}^{n} P(A_i \mid B)} \tag{2.8}$$

Keterangan:

$P(A \mid B)$: Peluang kejadian *A* bila *B* terjadi

$P(B)$: Peluang kejadian *B*

$P(B \mid A)$: Peluang kejadian *B* bila *A* terjadi.

$P(A_i \mid B)$: Peluang kejadian *A* ke-i bila *B* terjadi, untuk $i = 1 \ldots n$

Untuk menjelaskan bahwa teorema *naive bayes*, perlu diketahui bahwa proses klasifikasi memerlukan petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis maka teorema diatas dapat disesuaikan menjadi:

$$P(C \mid F_1 \ldots . F_n) = \frac{P(C)P(F_1 \ldots . F_n \mid C)}{P(F_1 \ldots . F_n)} \tag{2.9}$$

Dimana variabel C merepresentasikan kelas, sementara $F_1 \ldots . F_n$ merepresentasikan petunjuk yang akan dignakan untuk melakukan klasifikasi. Rumus tersebut menjelaskan peluang masuknya sampel dengan karakteristik tertentu.

Untuk mencari kelas data yang kita test maka kita akan membandingkan posterior (nilai peluang suatu sampel berada di kelas C) untuk masing masing kelas. Nilai posterior dapat diperoleh dengan cara berikut:

$$C_{NB} = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(f_i \mid c) \tag{2.10}$$

Dengan c adalah variabel kelas yang tergabung pada suatu himpunan kelas C [12].

3. Data

Data yang digunakan adalah data dari http://web.eecs.umich.edu/~lahiri/acl_arc.html sebanyak 500 paper. Pada setiap paper yang digunakan bagian judul sampai abstract, dan reference dihapus agar pengekstraksian kalimat lebih baik.

4. Ekstraksi Fitur

Ekstraksi fitur merupakan pengambilan ciri/*feature* dari suatu kalimat untuk dianalisis dan diproses pada pelatihan. Ekstraksi fitur didapat dari melihat bentuk-bentuk umum yang ada di kalimat sitasi dan pengambilan dari penelitian sebelumnya. Penelitian ini menggunakan 5 fitur yaitu unigram, bigram, *proper noun*, *pronoun*, *cue phrase*

4.1 N-Gram

Dengan mempelajari corpus yang tersedia akan didapatkan kata yang berpengaruh didalam penulisan kalimat sitasi di *paper* ACL-ARC *Anthology Corpu*. Penggunaan n-Gram tidak terbatas dalam penghitungan kata, tetapi bisa juga digunakan dalam penghirungan karakter. Pada penelitian ini penulis menggunakan pembagian n-gram untuk kata, dan menghitung 5 dan 10 kata yang paling sering digunakan di dalam penulisan kalimat sitasi.

4.1.1 Unigram

Fitur ini digunakan karena penggunaan kata yang berulang dalam membahas sebuah topik. Dalam pengklasifikasian kalimat sitasi, unigram juga digunakan dalam penelitian oleh Tarun Kumal et al., [2]. Fitur unigram memuat kata-kata yang ada di sebuah kalimat. Setelah menghilangkan *stop words*, fitur ini membuat setiap kata yang terekstrak menjadi atribut untuk p. Contoh penggunaan unigram "*The classification model is trained …*" dapat direperesentasikan dalam unigram *classification*, *model*, *trained* (*the* dan *is* adalah *stop words*) [2].

4.1.2 Bigram

Bigram adalah penggabungan dua kata dalam sebuah kalimat. Penggunaan bigram dapat membantu proses pengklasifikasian kalimat sitasi karena pengulangan kata yang sama dalam membahas sebuah topik. Contoh penggunaan bigram, "*The classification model is trained …*" dapat direperesentasikan dalam bigram "*classification model*", "*model trained*" (*the* dan *is* adalah *stop words*) [2].

Fitur ini dipilih karena frase kata cukup baik dibandingkan trigram, karena masih banyak kata yang mempunyai dua frase, sehingga bigram dapat membantu pengklasifikasian kalimat sitasi.

4.2 *Proper Noun*

Proper noun adalah nomina atau kata benda yang menunjukkan orang, tempat, sesuatu atau kejadian tertentu yang sudah melekat dan punya arti khusus pada nomina tersebut dan selalu diawali dengan huruf besar [2].

Fitur ini dipilih oleh penulis dikarenakan didalam kalimat sitasi banyak memuat proper noun yang menginformasikan nama orang, nama organisasi, lokasi, sistem, waktu, uang, dan persentase. Sehingga berpengaruh dalam pengidentifikasian kalimat sitasi. Untuk mengkategorikan kata ke dalam proper noun menggunakan stanford NER tagger.

4.3 Pronoun

Pronoun adalah kata yang dapat menggantikan suatu kata benda atau frasa kata benda. Fungsi dari pronoun sendiri adalah menghindari pengulangan kata atau frasa kata yang telah disebut sebelumnya[6].

Pemilihan fitur ini dikarenakan menurut penglihatan penulis, banyak kalimat sitasi yang merujuk seseorang atau sebuah organisasi, bahkan sebuah sistem.

4.4 *Cue Phrase*

Cue phrase sering digunakan untuk pendekatan ekstraksi informasi dari sebuah artikel. Penggunaan cue phrase (kata acuan) sering dilakukan di kategori-kategori yang bersifat retoris. Contoh frasa cue phrase dapat diliihat pada tabel berikut [14].

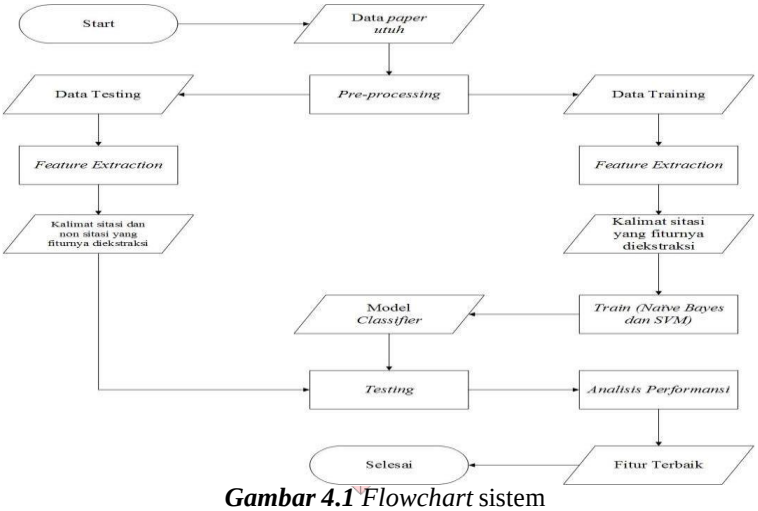
Dan berikut adalah cara penulisan cue phrase menurut [14]:

Tabel 3.3 Tipe *cue phrase*

Kategori *Cue Phrase*	***Regular Expression Cue Phrase***
PROBLEM	(focused in|used in) |(in these methods| in their algorithm|based on)

METHOD	(<PRP | NNP>(tested | employed | used | use |trained | examined | extracted | constructed |analyzed | described) (described in) <REF|NNP>)|(found|shown|demonstrated)|(has|have)been(shown|demonstrated) <PRP|NNP>(found|show|demonstrate)
CONCLUSSION	((agree | found | suggest | find | argue | showed| observed | shown |shows)\s (that))(precision | recall | reporting | report <NUMBER>)

5. Eksperimen



Gambar 4.1 *Flowchart* sistem

Secara garis besar ada 4 poin penting dalam penelitian ini yaitu *preprocessing*, pengekstraksian fitur, pembangunan *classifier*, dan pengujian model.

5.1 Preprocessing

Penelitian ini menggunakan data dari http://web.eecs.umich.edu/~lahiri/acl_arc.html sebanyak 10 paper sebagai *training* dan 1 paper sebagai *testing*. Preprocessing yang akan dilakukan adalah mempartisi makalah menjadi kalimat kalimat dan membuat record yang berisi kalimat dan kelas kalimat (kalimat sitasi atau bukan). Lalu menghapus tanda kalimat sitasi, menghilangkan stop words, dan mendefinisikan kalimat sitasi sebagai positive instance dan yang bukan kalimat sitasi sebagai negative instance. Pelabelan kalimat sitasi dilakukan menggunakan regular expression dan pengamatan penulis sehingga tidak terlalu lama dalam pelabelan sitasi. Hasil dari preprocessing adalah tabel kalimat yang telah dilabeli kalimat sitasi atau bukan untuk data training.

Tabel 5.1 Regex Kalimat Sitasi

Regular Expression	Tipe Kalimat Sitasi
\\(\\d{4}\\)	*Textual*
\\[\\d{1,3}\\].*	*Numbered*
\\[\\d{1,3}\\]	*Numbered*
\\([A-Z]^\\(]+\\d{4}\\)	*Textual*
et al	*Prosaic*
\\,\\s\\d{4}\\)	*Numbered*

5.2 Ekstraksi Fitur

Pengekstraksian fitur akan dilakukan terhadap setiap record yang telah dibuat agar dapat digunakan supervised learning untuk proses pengklasifikasiannya. Fitur yang akan digunakan pada tugas akhir ini diambil

dari hasil penelitian makalah yang telah dipublikasikan dengan hasil yang paling baik yaitu *N-Gram*, *Proper Noun* [2], *Cue Phrase*[14], dan fitur yang diteliti yaitu pronoun. Pada penelitian ini juga menggabungkan fitur-fitur yang digunakan untuk analisis

5.3 Membangun *Classifier*

Classifier akan dibangun dari data *training*, untuk menguji classifier yang dibangun penulis menggunakan *10-cross-fold validation*.

6. Hasil

Penelitian ini menggunakan dua skenario untuk menganalisis fitur. Skenario pertama adalah menggunakan 5 kata yang paling sering muncul di kalimat sitasi, dan yang kedua adalah menggunakan 10 kata yang paling sering muncul di kalimat sitasi

6.1 N-gram Menggunakan 5 Kata Terbanyak pada Kalimat Sitasi

Tabel 6.1 Hasil Skenario Pertama (Menggunakan 5 kata yang paling sering muncul)

Fitur	*Naive Bayes*		*SVM*	
	F-Measure	***Word Accuracy***	***F-Measure***	***Word Accuracy***
1-Gram	0%	91,412%	0%	91,412%
2-Gram	0%	91,412%	0%	91,412%
Cue Phrase	0%	91,412%	0%	91,412%
Pronoun	0%	91,412%	0%	91,412%
Proper Noun	57,913%	92,126%	51,234%	92,503%
1-Gram dan 2-Gram	0%	91,410%	0%	91,412%
1-Gram dan Proper Noun	58,282%	91,120%	51,234%	92,503%
2-Gram dan Cue Phrase	0,037%	91,405%	0%	91,412%
2-Gram dan Pronoun	0%	91,412%	0%	91,412%
2-Gram dan Proper Noun	57,945%	92,126%	51,234%	92,503%
Pronoun dan Cue Phrase	0,46%	91,412%	0%	91,412%
Proper Noun dan Cue Phrase	**59,069%**	**92,157%**	**51,234%**	**92,503%**
Proper Noun dan Pronoun	58,455%	92,11%	51,234%	92,503%
1-Gram, 2-Gram, dan Pronoun	0%	91,401%	0%	91,412%
1-Gram, 2-Gram, dan Proper Noun	58,313%	92,225%	51,234%	92,503%
2-Gram, Proper Noun, Pronoun	58,438%	92,099%	51,234%	92,503%
2-Gram, Proper Noun, dan Cue Phrase	59,040%	92,155%	51,234%	92,503%
1-Gram, 2-Gram, Proper Noun, dan Cue Phrase	58,013%	92,229%	51,234%	92,503%
1-Gram, 2-Gram, Cue Phrase	1,403%	91,37%	0%	91,412%
1-Gram, 2-Gram, Proper Noun, Cue Phrase, Pronoun	58,013%	92,229%	51,234%	92,503%
2-Gram, Proper Noun, Cue Phrase, Pronoun	57,290%	92,216%	51,234%	92,503%

6.2 N-gram Menggunakan 10 Kata Terbanyak pada Kalimat Sitasi

Tabel 6.2 Hasil dari Skenario dua (Menggunakan 10 kata terbanyak untuk N-gram)

Fitur	*Naive Bayes*		*SVM*	
	F-Measure	***Word Accuracy***	***F-Measure***	***Word Accuracy***
1-Gram	0%	91,412%	0%	91,412%
2-Gram	0%	91,412%	0%	91,412%
Cue Phrase	0%	91,412%	0%	91,412%
Pronoun	0%	91,412%	0%	91,412%
Proper Noun	57,913%	92,126%	51,234%	92,503%
1-Gram dan 2-Gram	0%	91,41%	0%	91,412%
1-Gram dan Proper Noun	58,543%	92,128%	51,234%	92,503%
2-Gram dan Cue Phrase	0,042%	91,405%	0%	91,412%
2-Gram dan Pronoun	0%	91,412%	0%	91,412%

2-Gram dan Proper Noun	59,02%	92,146%	51,234%	92,503%
Pronoun dan Cue Phrase	0,46%	91,392%	0%	91,412%
Proper Noun dan Cue Phrase	**59,069%**	**92,157%**	**51,234%**	**92,503%**
Proper Noun dan Pronoun	58,455%	92,110%	51,234%	92,503%
1-Gram, 2-Gram, dan Pronoun	0,292%	91,376%	0%	91,412%
1-Gram, 2-Gram, dan Proper Noun	58,500%	92,106%	51,234%	92,503%
2-Gram, Proper Noun, Pronoun	58,481%	92,128%	51,234%	92,503%
2-Gram, Proper Noun, dan Cue Phrase	59,02%	92,146%	51,234%	92,503%
1-Gram, 2-Gram, Proper Noun, dan Cue Phrase	56,979%	92,138%	51,234%	92,503%
1-Gram, 2-Gram, Cue Phrase	2,346%	91,279%	0%	91,412%
1-Gram, 2-Gram, Proper Noun, Cue Phrase, Pronoun	57,706%	92,173%	51,234%	92,503%
2-Gram, Proper Noun, Cue Phrase, Pronoun	57,975%	92,128%	51,234%	92,503%

7. Analisis

Pada proses pengidentifikasian jika memilih fitur yang kurang tepat tidak akan mempengaruhi akurasi, karena perbandingan data yang cukup drastis pada training. Proses learning yang dilakukan menjadi lebih sulit dikarenakan pola kalimat sitasi dan non-sitasi yang hampir mirip.

Berdasarkan hasil dari tabel diatas, maka didapatkan fitur yang terbaik dalam identifikasi kalimat sitasi yaitu fitur gabungan dari *proper noun* dan *cue phrase* dengan *f-measure* sebesar 59,069%, dan akurasi sebesar 92,157% menggunakan naive bayes. Pada SVM fitur ini memiliki *f-measure* sebesar 51,234% , dan akurasi sebesar 92,503%. Fitur ini akan digunakan pada pengujian pada paper lain.

Beberapa fitur yang memiliki hasil presisi, recall, dan f-measure 0% terjadi karena gagalnya pembelajaran oleh sistem. Penyebab gagalnya adalah adanya fitur yang sama di kalimat sitasi dan non sitasi. Pada penelitian ini penggunaan N-Gram dengan 5 kata terbanyak pada kalimat sitasi, dan 10 kata terbanyak pada kalimat sitasi tidak berperngaruh. Karena pada fitur N-Gram(1-Gram, dan 2-Gram) gagal mempelajari pola dari kalimat sitasi yang disebabkan kata-kata tersebut yang tidak unik.

8. Kesimpulan

1. Berdasarkan hasil analisis performansi dari fitur yang telah dibuat didapatkan fitur yang paling baik untuk pengklasifikasian kalimat sitasi mempunyai *f-measure* sebesar 59,069% dan akurasi sebesar 92,157% menggunakan *naive bayes*, dan akurasi sebesar 92,503% dan *f-measure* 51,234% menggunakan SVM.
2. Fitur yang paling baik dalam pengklasifikasian kalimat sitasi adalah fitur gabungan dari *proper noun*, dan *cue phrase.*

DAFTAR PUSTAKA:

[1] repository.usu.ac.id/bitstream/123456789/34673/4/Chapter%20II.pdf (diakses pada tanggal 11 Maret 2015 pukul 22.24)

[2] Kumar, T, Sugiyama, K.,Kan, M., and Tripathi, R., 2010 , "*Identifying Citing Sentences in Research Papers Using Supervised Learning*",

[3] Abu-Jbara, Amjad, and Radev, Dragomir, "*Reference Scope Identification in Citing Sentences*"

[4] Powley. Bred., and Robert Dale.,"Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification", *In Large Scale Semantic Access to Content (Text, Image, Video and Sound)* (pp. 618-632), 2007.

[5] https://docs.oracle.com/javase/tutorial/essential/regex/intro.html diakses 21 Des. 15 23:27

[6] http://www.iusb.edu/~libg/instruction/helpguide/handouts/2005Boolean.s html *diakses 22 Des 15 0:37*

[7] Dougherty, G., 2013, "*Pattern Recognition and Classification*". Springer.

[8] Jakkula, Viramaditya, *"Tutorial on Support Vector Machine", 2003*

[9] Alpaydin, E., 2010, "*Introduction to Machine Learning*", Massachusetts Institute of Technology.

[10] Lowd, Daniel, and Domingos, Pedro: " *Naive Bayes Models for Probability*"

[11] Natalius, Samuel, 2010"Metoda Naive Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen"

[12] Ammar, Shadiq: *"Keoptimalan Naïve Bayes Dalam Klasifikasi"*

[13] Widyantoro, D.,and Amin, I.,2014, "Citation Sentence Identification and *Classification for Related Work Summarization*"

[14] McCallum, A., Peng, F., 2004, "*Accurate Information Extraction from Research Paper using Conditional Random Fields*", Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL).