IDENTIFIKASI KALIMAT SITASI MENGGUNAKAN KOMBINASI METODE SUPPORT VECTOR MACHINE DAN FEATURE SELECTION

Raynaldi Fatih Amanullah[1], Ema Utami[2], Andi Sunyoto[3]

Magister Teknik Informatika
Fakultas Teknik Informatika
Universitas Amikom Yogyakarta

raynaldi.a@students.amikom.ac.id[1], ema@amikom.ac.id[2], andi@amikom.ac.id[3]

Abstrak

Jurnal ilmiah merupakan suatu karya ilmiah yang diterbitkan secara berkala oleh suatu organisasi atau institusi, kecerobohan penulisan dalam karya ilmiah dapat dianggap sebagai bentuk plagiarisme. Sehingga penulisan sitasi dalam karya ilmiah penting untuk diperhatikan, karena sitasi mampu memberikan pengakuan sumber acuan. Metode yang digunakan ialah Support Vector Machine (SVM) dan juga TF-IDF. Adapun dataset yang digunakan ialah CL-SciSumm 2018, yang selanjutnya diseleksi menggunakan TF-IDF guna mengurangi jumlah dimensi suatu dokumen, sehinga data lebih mudah diolah menggunakan SVM. Dari hasil klasifikasi kemudian dianalisa tingkat akurasi atau ketepatan dalam melakukan identifikasi dengan menggunakan skema k-Fold Cross Validation. Hasil penelitian ini menunjukan bahwa penggunaan TF-IDF mampu mendukung metode SVM dalam melakukan identifikasi kalimat sitasi pada dokumen jurnal ilmiah dengan nilai akurasi dan f-measure sebesar 0,52 dan 0,66 dengan nilai k = 5, dari hasil tersebut terjadi kenaikan akurasi sebesar 0,04 dan f-measure sebesar 0,16.

Kata Kunci: jurnal ilmiah, identifikasi sitasi, SVM, TF-IDF.

1. Pendahuluan

Jurnal ilmiah merupakan suatu publikasi atas karya ilmiah yang diterbitkan secara berkala oleh suatu organisasi profesi atau pun institusi akademik yang memuat artikel-artikel yang berisikan laporan sistematis mengenai hasil kajian atau hasil penelitian yang mengandung informasi dan data yang disajikan bagi masyarakat dalam bidang ilmu tertentu (Zifirdaus dan Zifirdaus, 2005; Suryoputro, dkk., 2012).

Ketidakpatuhan penulisan karya ilmiah atau pun publikasi jurnal ilmiah merupakan permasalahan yang senantiasa berkaitan dengan para akademisi dan peneliti (Yahya, 2012). Kecerobohan dalam penggunaan sitasi dapat dianggap sebagai bentuk plagiarisme (Mantovani, 2016). Dalam beberapa kasus, tindakan tersebut justru muncul akibat kesengajaan dan bersifat terencana, yang mengakibatkan rusaknya integritas akademik (Yahya, 2012).

Oleh sebab itu, penulisan sitasi atau sitiran dalam karya ilmiah sangatlah penting (Sophia, 2002). Menulis sitasi berarti memberikan pengakuan terhadap

pengarang, atas ide, gagasan, pendapat atau bahkan teori yang digunakan dalam penyusunan karya ilmiah (Istiana, 2013). Pengidentifikasian kalimat sitasi dapat membantu para akademisi atau pun ilmuwan dalam melakukan peninjauan karya ilmiah yang disusun dan juga untuk melihat perkembangan penelitian tersebut (Mantovani, 2016).

Dalam melakukan identifikasi diperlukan adanya penggalian text atau text mining. Text mining merupakan analisis teks yang sumber datanya berasal dari dokumen dengan tujuan untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan, keterkaitan dan kelas antar dokumen (Feldman dan Sanger, 2007). Selain itu, text mining dapat dilakukan dengan cara mengklasifikasi atau pun hanya dengan melihat jumlah frekuensinya (Pathak, 2014).

Adapun penelitian sejenis antara lain penelitian yang dilakukan Nomoto (2016) dengan memanfaatkan penggabungan algoritma NN (*Neural Network*) dengan TF-IDF (*Term Frequency – Invers Document Frequency*) guna mengidentifikasi hubungan antara *reference paper* (RP) dengan *citation paper* (CP). Dataset yang digunakan ialah CL-SciSumm 2016 dengan rincian 10 RP dan 10 CP. Dari 10 RP yang digunakan, dua diantaranya tidak sesaui dengan *human judgment* sedangkan delapan sisanya sesuai dengan *human judgment*.

Penelitian lain dilakukan oleh Mantovani, dkk., (2016) menggunakan dua buah metode yakni *Naive Bayes* (NB) dan *Support Vector Machine* (SVM), dengan menggunakan *feature selection unigram*, *bigram*, *proper noun*, *cue phrase* dan *pronoun*. Tujuan dalam penelitian ini adalah menentukan fitur yang dapat memksimalkan kinerja *classifer*-nya. Dalam mengukur keakuratan fitur-fiturnya, peneliti menggunakan pengujian *confusion matrix*, dari hasil pengujian menunjukkan dua buah fitur dengan hasil paling maksimal yaitu *proper noun* dan *cue pharse*, dengan nilai *f-measure* 59,069% dan akurasi 92,157% untuk klasifikasi menggunakan NB, dan *f-measure* 51,234% dan akurasi 92,503% dengan menggunakan SVM.

Berdasarkan penelitian diatas NN memiliki keunggulan dalam kemampuan untuk bergeneralisasi, yang bergantung pada ANN sehingga mampu meminimalisir resiko, disisi lain ANN memiliki kelemahan dalam jumlah data training yang relatif banyak (Vapnik, 1999). NB merupakan metode yang mudah diimplementasikan dan memiliki performa yang baik, namun pengklasifikasian NB

didasarkan pada probabilitas bersyarat dari fitur yang terdapat dalam salah satu kelas (Zhang dan Gao, 2011). SVM dapat diterapkan dalam data dengan dimensi yang tinggi dan memiliki tingkat akurasi yang baik, namun SVM sulit digunakan dalam data dengan jumlah yang besar (Nugroho, dkk., 2003).

Feature selection dapat digunakan untuk mengurangi atribut yang kurang relevan (Wang, dkk., 2011) dan juga mampu untuk mengoptimalkan kinerja dari classifier (Koncz dan Paralic, 2011). Algoritma feature selection yang digunakan adalah TF-IDF. TF-IDF merupakan metode yang mampu memberikan bobot terhadap semua kata yang terdapat pada dokumen, dengan memperhitungkan jumlah term atau atribut yang muncul dalam (Suri dan Purnamasari, 2017).

2. Tinjauan Pustaka

2.1. *Text Mining*

Text Mining memberikan solusi dari permasalahan seperti pemrosesan, pengelompokan dan analisa terhadap teks yang tidak terstruktur dalam jumlah yang besar (Maarif, 2015). Sehingga *Text Mining* dapat dikatakan sebagai bagian dari *Data Mining*. Penelitian ini merupakan penelitian yang menekankan analisa teks yang tidak terstuktur dalam suatu dokumen maka dibutuhkanlah *text mining*, sehingga dengan kata lain penelitian ini juga merupakan bagian dari Data Mining.

Tahapan dalam penggalian teks dibagi menjadi tiga, yaitu praproses teks (text preprocessing), transformasi teks (text transformation), dan penemuan pola (pattern discovery) yang dijelaskan secara rinci sebagai berikut (Feldman dan Sanger, 2007):

1. *Text Preprocessing*

   Bertujuan untuk mempersiapkan teks menjadi data yang terstruktur sehingga dapat diproses pada tahap selanjutnya. Secara umum tahap-tahap text preprocessing antara lain:

   a. *Case Folding*, yaitu melakukan konversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar, biasanya dalam bentuk lowercase atau huruf kecil.
   b. *Tokenizing*, yaitu merupakan tahap pemotongan string menjadi bagian-bagian kata, selain itu proses ini juga dapat menghilangkan karakter-karakter spesial seperti titik dua (:), titik koma (;), dan lain-lain.
   c. *Filtering*, yaitu tahap pengambilan kata-kata penting dari tahap tokenizing, biasanya menggunakan stopword removal (membuang kata kurang

penting) atau wordlist (menyimpan kata penting). Stopwords merupakan kata-kata yang tidak terdeskripsi yang dapat dibuang, seperti di, yang, atau, dan lain-lain.

2. *Text Transformation*

   Tahap pengubahan kata-kata penting menjadi kata dasar dengan arti yang serupa namun memiliki bentuk yang berbeda. Sebagai contoh, kata "membela" dari proses filtering apabila dilakukan stemming akan menjadi kata "bela".

3. *Pattern Discovery*

   Proses ini merupakan tahap menemukan pola atau pengetahuan dari keseluruhan teks. Proses ini akan melibatkan pemahaman mengenai machine learning (unsupervised learning dan supervised learning). Perbedaan pada kedua machine learning tersebut hanya terletak pada pelabelannya. Dalam menunjang hasil dari penelitian ini, diperlukanlah *machine learning* atau pembelajaran sistem, dikarenakan dataset dalam penelitian ini terbagi menjadi dua bagian, yakni data *training* dan data *testing*, yang mana dalam melakukan analisa nantinya sistem akan belajar terhadap data *training*, yang selanjutnya sistem tersebut akan mendapatkan pengetahuan guna mengidentifikasi data *testing*. Dengan kata lain, sistem dapat melakakukan identifikasi terhadap data *testing* berdasarkan pengetahuan yang telah diperoleh ketika mengidentifikasi data *training*.

2.2. *Support Vector Machine*

Tujuan *dari* SVM guna menemukan fungsi pemisah (*hyperplane*) dengan margin terbesar, sehingga dapat memisahkan dua kelompok data (Han, dkk., 2012).

Titik data yang terdapat dalam SVM, sebagai contoh adalah $x_i = \{x_1, x_2, \dots, x_n\} \in R^n$, sedangkan kelas datanya $y_i = \in \{-1, +1\}$. Selanjutnya dipasangkan dengan data dan kelas $\{(x_i, y_i)\}_{i=1}^{N}$, sehingga dapat memaksimlakan persamaan 1.

$$Ld = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_i K(x_i, x_j) \qquad (1)$$

Keterangan:

Ld = Dualitas Lagrange Multiplier

n = Banyak Data

α_i = Nilai bobot setiap titik data

C = Nilai Konstanta

$K(x_i, x_j)$ = Fungsi kernel

Dengan syarat, $0 \leq \alpha_i \leq C$ dan $\sum_{i=1}^{N} \alpha_i y_i = 0$. Selanjutnya mencari nilai w dan b, dengan rumus:

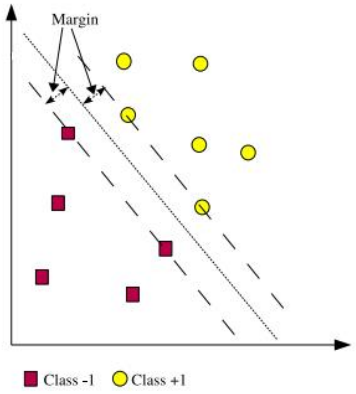
$$w = \sum_{i=1}^{N} \alpha_i \, y_i x_i \qquad (2)$$

$$b = -\frac{1}{2}\,(w.x^{+} + w.x^{-}) \qquad (3)$$

Setelah nilai w dan b diketahui, langkah selanjutnya yaitu menghitung fungsi keputusan klasifikasi *sign* $f(x)$, seperti berikut:

$$f(x) = w.x + b \text{ atau } f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x_j) + b \qquad (4)$$

Klasifikasi linier memisahkan data dengan memaksimalkan margin menggunakan hyperplane sebagai pemisanya, seperti ditunjukan pada Gambar 1.



Gambar 1 Hyperplane Optimum

Sebagaimana yang terlihat pada Gambar 1, metode SVM menunjukkan hyperplane terbaik, dengan letak yang berada tepat di tengah-tengah atau antara kedua kelas. Sedangkan garis putus-putus yang mengelilingi hyperplane disebut dengan margin, yang mana margin tersebut merupakan jarak antara hyperplane dengan data terdekat dari masing-masing class (Faihah, 2010).

2.3. *Feature Selection*

Algoritma feature selection dibedakan menjadi dua tipe, yaitu filter dan wrapper (Chen, dkk., 2005). Menurut (Vercellis, 2009) contoh dari tipe *filiter* adalah Information Gain (IG), Term Frequency – Invers Document Frequency (TF-IDF), dan lain-lain, sedangkan untuk tipe *wrapper* contohnya adalah *foward selection* dan *backward elimination*. Berikut ini merupakan pendekatan-pendekatan feature selection yaitu (Langgeni, dkk., 2010):

1. *Filter Feature Selection*

   Pendekatan yang melakukan pemillihan feature tidak bersamaan dengan pelaksanaan model, pendekatan filter merupakan pendekatan yang paling mudah dalam komputasinya, dikarenakan tidak melibatkan induksi algoritma didalam prosesnya. Penerapan pendekatan filter, cocok untuk data dengan dimensi yang tinggi seperti text mining.

2. *Wrapper Feature Selection*

   Pendekatan wrapper melakukan pemilihan feature dengan menggunkan fungsi evaluasi berdasarkan algoritma learning yang secara bersamaan dengan pelaksanaan pemodelan. Dalam pemilihan feature, terdapat dua komponen utama. Komponen pertama, dilakukan setelah feature space terbentuk, dengan dilakukannya pencarian prosedur yang dapat menghasilkan subset untuk dilakukan evaluasi. Komponen kedua, melakukan evaluasi sebagai ukuran dalam pemilihan subset.

2.4. *Pembobotan Kata (Term Weighting)*

Dalam pencarian informasi dari koleksi dokumen yang memiliki sifat heterogen diperlukan pembobotan term, term tersebut dapat berupa kata, frase atau unit hasil indexing lainnya yang terdapat dalam suatu dokumen yang dapat berfungsi untuk mengetahui konteks dari dokumen tersebut, di karenakan setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, sehingga dilakukanlah term weight (Zafikri, 2008).

Term weighting atau pembobotan term dipengaruhi oleh beberapa hal seperti berikut:

1. Term Frequency (TF)

   TF merupakan salah satu metode yang dapat digunakan untuk menhitung bobot pada masing-masing term dalam text. Dalam metode ini, setiap term

diasumsikan memiliki nilai kepintangan yang sebanding dengan jumlah kemunculan term tersebut. Bobot sebuah term, dirumuskan sebagai berikut:

$$W_{ij} = TF_{ij} \qquad (5)$$

Keterangan:

W_{ij} = Bobot dokumen ke-j terhadap term ke-i

TF_{ij} = Jumlah term i yang muncul pada dokumen d

2. *Invers Document Frequency* (IDF)

IDF merupakan sebuah perhitungan dari terms yang telah didistribusikan pada kumpulan dokumen yang bersangkutan. IDF menunjukan hubungan ketersedian sebuah terms dalam seluruh dokumen. Dimana apabila semakin sedikit jumlah dokumen yang mengandung terms yang dimaksud, maka nilai IDF akan semakin besar, begitu juga sebaliknya (Robertson, 2004).

Untuk menghitung term yang telah didistribusikan pada dokumen-dokumen yang bersangkutan, maka digunakanlah rumus:

$$idf_j = \log \frac{D}{df_j} \qquad (6)$$

Keterangan:

D = Jumlah semua dokumen dalam koleksi

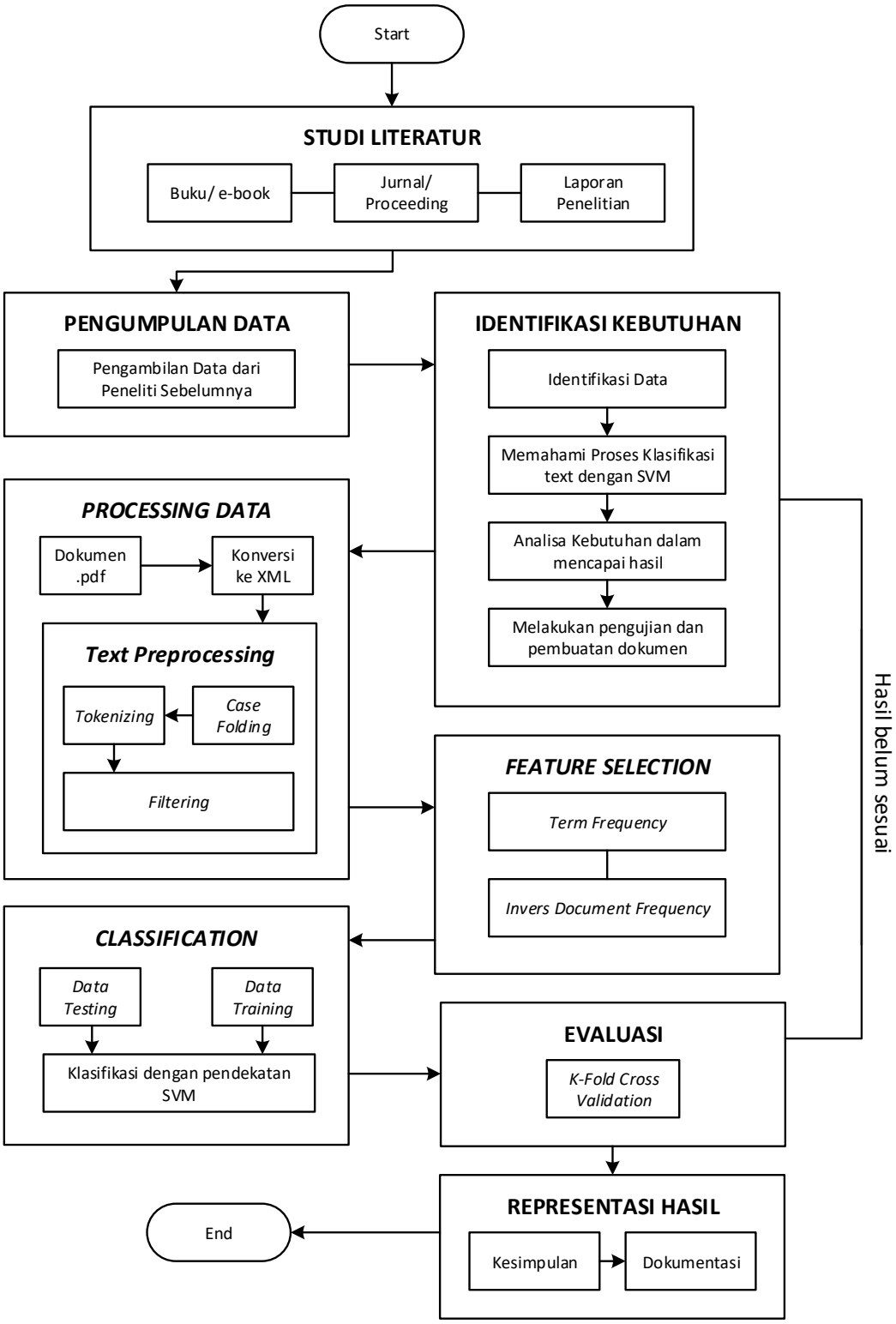
df_j = Jumlah dokumen yang mengandung term

2.5. *Cross Validation*

Pengujian dengan metode *cross-validation* merupakan pengujian yang dilakukan dengan cara membagi data menjadi dua bagian, yakni data *training* dan data *testing*. *K-fold cross validation*, sejak awal secara acak dibagi menjadi subset *k* yang terpisah menjadi k_1, k_2, ..., k_n dengan masing-masing subset memiliki jumlah data yang sama. Data *training* dan data *testing* dilakukan sebanyak *n* kali. Sebagai contoh iterasi ke 1, dengan k_1 sebagai data *testing*, maka k_2,...., k_n akan digunakan sebagai data *training*. Selanjutnya untuk iterasi ke 2, k_2 menjadi data *testing*, sedangkan $k_1, k_3, \dots, k_n$ akan menjadi data *training*, begitu seterusnya hingga k_n (Kamber dan Han, 2006).

3. Metode Yang Diusulkan

Metode yang diusulkan dalam penelitian ini memiliki alur sebagaimana yang ditunjukkan oleh Gambar 2.



Gambar 2. Alur Penelitian

Sebagaimana yang terlihat pada Gambar 2, metode yang diusulkan memiliki bagian-bagian atau proses-proses berikut ini:

1. Studi Literatur

   Pada tahap ini, peneliti melakukan studi literatur dengan membaca buku, jurnal/ karya ilmiah dan laporan penelitian yang terkait dengan topik penelitian.

2. Pengumpulan Data

   Data yang digunakan merupakan data CL-SciSumm 2018, dengan rincian 40 sebagai data training dan 20 sebagai data testing, sehingga secara keseluruhan berjumlah 60 dataset.

3. Identifikasi Kebutuhan

   Dari kedua tahap diatas, selanjut peneliti melakukan identifikasi terhadap kebutuhan-kebutuhan yang diperlukan dalam melakukan penelitian, diantaranya dengan mengidentifikasi data, kemudian mempelajari metode pengklasifikasian teks menggunakan SVM, melakukan analisis terhadap hasil pengujian dan selanjutnya representasi hasil penelitian.

4. *Data Preprocessing*

   Tahap ini berguna untuk melakukan cleaning data, sehingga menjadi data yang dapat diproses atau dapat dijadikan sebagai inputan pada tahap selanjutnya, adapun tahapan yang dilakukan yakni dengan mengubah seluruh kata menjadi huruf kecil (*case folding*), pemotongan string (*tokenizing*), pengambilan kata-kata penting (*filtering*)

5. *Feature Selection*

   Melakukan seleksi fitur berdasarkan term yang terdapat pada masing-masing dokumen, dan dilakukan perhitungan terhadap term tersebut.

6. *Classification*

   Merupakan tahap klasifikasi menggunakan metode SVM berdasarkan data training dan testing yang tersedia.

7. Evaluasi

   Proses pengujian terhadap akurasi yang dilakukan dengan skema k-fold cross validation yang nantinya akan dijadikan sebagai input dalam confusion matrix.

8. Representasi Hasil

   Tahap ini, dilakukan dokumentasi terhadap hasil penelitian atau pun hasil evaluasi yang telah dilakukan sebelumnya, pada tahap ini juga termasuk penulisan laporan tesis dan penyajian data yang dapat berupa diagram atau tabel.

4. Hasil dan Pembahasan

Berdasarkan alur diatas, dilakukanlah dua skenario pengujian, skenario pertama hanya menggunakan metode SVM, sedangkan skenario kedua menggunakan metode SVM+TF-IDF. Hal tersebut dilakukan untuk mengetahui penggunaan metode TF-IDF dalam mendukung metode SVM selain itu untuk mengetahui nilai *k* yang optimum. Oleh karena itu, skema pengujian yang digunakan adalah *k-Fold Cross Validation*. Dari *k-Fold Cross Validation* dihasilkan Confusion Matrix untuk kemudian dihitung nilai akurasi dan F-Measure (dari nilai presisi dan recall) dimana ditunjukkan pada Tabel 1.

Tabel 1. Hasil Akurasi dan *f*-measure Menggunakan Metode SVM tanpa TF-IDF

Nilai *k*	Akurasi	F-Measure
2	0,43	0,43
3	0,44	0,45
4	0,47	0,4
5	**0,48**	**0,50**
6	0,41	0,44
10	0,43	0,44

Berdasarkan Tabel 1, nilai *k* paling tinggi diperoleh pada *k* = 5, dengan nilai akurasi dan *f-measure* adalah 0,48 dan 0,50. Sedangkan untuk skenario pengujian penggunaan SVM dan TF-IDF dapat dilihat pada Tabel 2.

Tabel 2. Hasil Akurasi dan *f-measure* Menggunakan Metode SVM dan TF-IDF

Nilai *k*	Akurasi	F-Measure
2	0,49	0,62
3	0,51	0,54
4	0,45	0,57
5	**0,52**	**0,66**
6	0,46	0,48
10	0,47	0,50

Sama halnya dengan hasil Tabel 1, pada Tabel 2 juga terlihat nilai *k* paling optimum berada pada *k* = 5, dengan nilai akurasi sebesar 0,52 dan *f-measure*

mencapai 0,66. Perbandingan Hasil Akurasi dan *f-measure* ditunjukkan pada Tabel 3.

Tabel 3. Perbandingan Hasil Akurasi dan *f-measure*

Keterangan	Nilai *k*	Akurasi	F-Measure
SVM	5	0,48	0,50
SVM + TF-IDF	5	0,52	0,66

Tabel 3 menunjukkan hasil perbandingan dari dua skenario yang diterapkan, walaupun perbedaanya yang tidak terlalu signifikan, tetapi hasilnya mengalami peningkatan walau pun itu kecil. Dalam hal ini mengapa skenario kedua memiliki nilai akurasi dan *f-measure* yang lebih baik dari pada skenario pertama, hal tersebut dikarenakan sebelum dilakukan klasifikasi, dokumen atau pun teks diseleksi menggunakan metode TF-IDF, masing-masing dokumen baik itu data *training* atau pun data *testing* diseleksi seperti penghapusan teks yang mengandung url, tanda baca, kata sambung, spasi ganda sehingga teks yang akan diidentifikasi jadi lebih sedikit dibanding tanpa menggunakan TF-IDF.

5. Kesimpulan dan Saran

Dari hasil pengujian yang dilakukan, dapat diketahui bahwa penggunaan Metode TF-IDF mendukung metode SVM dalam melakukan identifikasi kalimat sitasi pada dokumen jurnal ilmiah, dengan nilai akurasi dan *f-measure* pada masing-masingnya adalah 0,52 dan 0,66. Dengan kenaikan akurasi sebesar 0,04 sedangkan untuk *f-measure* mengalami kenaikan 0,16. Hasil tersebut tentunya didukung dengan nilai *k* paling optimum yakni *k* = 5. Sedangkan untuk peneliti selanjutnya, dapat menggunakan kombinasi metode yang berbeda guna mengetahui peningkatan akurasi khususnya pada identifikasi kalimat sitasi.

Daftar Pustaka

Chen, J., Ji, D., Tan, C. L., dan Niu, Z. (2005). Unsupervised Feature Selection for Relation Extraction. *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*, 262 - 267.

Faihah, R. T. (2010). *Support Vector Machine (SVM).* Madura: Universitas Trunojoyo.

Feldman, R., dan Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Press.

Han, J., Kamber, M., dan Pei, J. (2012). *Data Mining: Concepts and Techniques, Third Edition.* Waltham: Morgan Kaufmann.

Istiana, P. (2013). Membuat Sitasi dan Daftar Pustaka. *Workshop Literasi Informasi bagi Pustakawan.* Yogyakarta: Universitas Sanata Dharma.

Kamber, M., dan Han, J. (2006). *Data Mining: Concepts and Techniques Second Edition.* San Francisco: Morgan Kaufmann Publisher.

Koncz, P., dan Paralic, J. (2011). An Approach to feature selection for sentiment analysis. *15th International Conference on Intelligent Engineering Systems* (pp. 357 - 362). Solvakia: IEEE Xplore Digital Library.

Langgeni, D. P., Baizal, Z. A., dan W, Y. F. (2010). Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. *Seminar Nasional Informatika 2010* (pp. 1 - 10). Yogyakarta: UPN Veteran Yogyakarta.

Maarif, A. A. (2015). Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah. *Seminar Nasional Teknologi Informasi dan Komunikasi*, 1-8.

Mantovani, R. P. (2016). *Penentuan Fitur Supervised Learning dalam Identifikasi Kalimat Sitasi pada Makalah Ilmiah.* Bandung: Universitas Telkom.

Nomoto, T. (2016). NEAL: A Neurally Enhanced Approach to Linking Citation and Reference. *Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)* (pp. 168 - 174). Newark: Central Europe Workshop Proceedings.

Nugroho, A. S., Witarto, A. B., dan Handoko, D. (2003). Support Vector Machine: Teori dan Aplikasinya dalam Bioinformatika.

Pathak, M. A. (2014). *Begining Data Science with R.* Switzerland: Springer International Publishing Switzerland.

Sophia, S. (2002). *Petunjuk Sitasi Serta Cantuman Daftar Pustaka Bahan Pustaka Online: Seri Pengembangan Perpustakaan Pertanian, No. 25.* Bogor: Departemen Pertanian Bogor.

Suri, D. J., dan Purnamasari, K. K. (2017). Perbandingan Seleksi Fitur untuk Klasifikasi Sentimen SVM Pada Twitter. *Jurnal Ilmiah Komputer dan Informatika*.

Suryoputro, G., Riadi, S., dan Sya'ban, A. (2012). *Menulis Artikel untuk Jurnal Ilmiah.* Jakarta Selatan: Uhamka Press.

Vapnik, V. N. (1999). An Overview of Statistical Learning Theory. *IEEE Transactions On Neural Networks*, 988 - 999.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.* United Kingdom: John Wiley and Sons.

Wang, S., Li, D., Song, X., Wie, Y., dan Li, H. (2011). A Feature Selecetion Method Based on Improved Fisher's Discriminant Ratio for Text Sentiment Classification. *Experty Systems with Applications* (pp. 8696 - 8702). China: Elsevier.

Yahya, I. (2012). Persoalan Sitasi dalam Publikasi Ilmiah dan Usulan Strategi Produktif Penanggulanan Plagiarisme Secara Bersistem di UNS. *Lokakarya Penanggulangan Tindak Plagiasi* (pp. 1 - 7). Surakarta: Universitas Sebelas Maret Surakarta.

Zafikri, A. (2008). *Implementasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi.* Sumatra Utara: Universitas Sumatra Utara.

Zhang, W., dan Gao, F. (2011). An Improvement to Naive Bayes for Text Classification. *Advanced in Control Engineering and Information Science* (pp. 2160 - 2164). China: Elsevier.

Zifirdaus, A., dan Zifirdaus, I. (2005). *Merebut Hati Audiens Internasional: Strategi Ampuh Meraih Publikasi di Jurnal Ilmiah.* Jakarta: Gramedia.