ANALISIS SENTIMEN OPINI PUBLIK BERITA KEBAKARAN HUTAN MELALUI KOMPARASI ALGORITMA *SUPPORT VECTOR MACHINE* DAN *K-NEAREST NEIGHBOR* BERBASIS *PARTICLE SWARM OPTIMIZATION*

Lilyani Asri Utami
Program Studi Sistem Informasi
STMIK Nusa Mandiri Jakarta
Jl. Damai No. 8 Warung Jati Barat Margasatwa Jakarta Selatan
lilyani.lau@nusamandiri.ac.id

Abstract *— Sentiment analysis is a process to determine the content of text-based datasets which are positive or negative. At present, public opinion be an important resource in the decision of a person in finding a solution. Classification algorithms such as Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) is proposed by many researchers to be used in sentiment analysis for review opinion. The problem in this research is the selection of feature selection to improve accuracy values Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) and compare the highest accuracy for sentiment analysis review public opinion about the news of forest fires. The comparison algorithms, SVM produces an accuracy of 80.83% and AUC 0.947, then compared with SVM based on PSO with an accuracy of 87.11% and AUC 0.922. The test result data for K-NN algorithm accuracy was 85.00% and the AUC 0.918, then compared for accuracy by k-NN-based PSO amounted to 73.06% and the AUC 0.500. The results of the testing of the PSO algorithm can improve the accuracy of SVM, but are not able to improve the accuracy of the algorithm K-NN. SVM algorithm based on PSO proven to provide solutions to the problems of classification review news opinion forest fires in order to more accurately and optimally.*

Intisari — Analisis sentimen adalah proses untuk menentukan isi dataset berbasis teks yang positif atau negatif. Saat ini, opini publik menjadi sumber daya penting dalam keputusan seseorang dalam mencari solusi. Algoritma klasifikasi seperti *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (K-NN) diusulkan oleh banyak peneliti untuk digunakan dalam analisis sentimen untuk mengulas pendapat. Masalah dalam penelitian ini adalah pemilihan seleksi fitur untuk meningkatkan akurasi nilai *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (K-NN) dan membandingkan akurasi tertinggi untuk analisis sentimen ulasan opini publik tentang berita kebakaran hutan. Algoritma perbandingan, SVM menghasilkan akurasi 80,83% dan AUC 0,947, kemudian dibandingkan dengan SVM berbasis PSO dengan akurasi 87,11% dan AUC 0,922. Data hasil pengujian untuk algoritma K-NN akurasinya adalah 85,00% dan AUC 0,918, kemudian dibandingkan untuk akurasi K-NN berbasis PSO sebesar 73,06% dan AUC 0.500. Hasil pengujian algoritma PSO dapat meningkatkan akurasi SVM, tetapi tidak dapat meningkatkan akurasi algoritma K-NN. Algoritma SVM berbasis PSO terbukti memberikan solusi terhadap masalah klasifikasi opini berita kebakaran hutan agar lebih akurat dan optimal.

Kata Kunci: Analisis Sentimen, Klasifikasi, *K-Nearest Neighbor, Support Vector Machine, Particle Swarm Optimization.*

PENDAHULUAN

Kemajuan teknologi informasi dan komunikasi jelas memberi dampak pada perubahan gaya hidup masyarakat dunia. Situs internet telah menjadi lautan informasi bagi siapapun untuk mendapatkan informasi mengenai hal apapun. Pemerintah Indonesia pun tanggap akan adanya tuntutan bagi transaksi informasi di dunia maya dengan dibuatnya Undang- undang Republik Indonesia Nomor 11 tahun 2008 tentang Informasi dan Transaksi Elektronik (UU ITE). UU ITE terdiri atas beberapa bab yang di dalamnya membahas segala hal terkait dengan informasi melalui elektronik.

Belum lama ini berita di media dan lini masa sosial media masih ramai dengan berita kabut asap. Di Indonesia, pengguna media sosial mengungkapkan berbagai komentar positif dan negatif mengenai berita yang setiap waktu mengabarkan informasi terkini tentang asap dan kebakaran hutan.

Seseorang yang berkomentar negatif akan berdampak tindakan pidana sebagaimana diatur dalam UU ITE tahun 2008 Pasal 27 ayat (3). Seseorang yang terbukti dengan sengaja menyebarluaskan informasi elektronik yang bermuatan pencemaran nama baik seperti yang dimaksudkan dalam Pasal 27 ayat (3) UU ITE

akan dijerat dengan Pasal 45 Ayat (1) UU ITE, sanksi pidana penjara maksimum 6 tahun dan/atau denda maksimum 1 Milyar Rupiah. Maka dari itu diperlukan suatu sistem yang dapat memfilter atau menyaring kata-kata yang tidak seharusnya dipostingkan.

Meluasnya penggunaan internet telah meningkatkan jumlah informasi yang disimpan dan diakses melalui web dalam kecepatan yang sangat cepat, karena banyaknya data yang terdapat di internet tersebut, tanpa diolah untuk dimanfaatkan lebih dalam maka munculah *Opinion Mining* atau *Sentiment Analysis* yang merupakan cabang penelitian dari *Text Mining.* Fokus dari penelitian *Opinion Mining* adalah melakukan analisis opini dari suatu dokumen teks (Rozi et al., 2012).

Sentiment analysis digunakan untuk mengotomatisasi proses identifikasi pendapat apakah itu adalah pandangan positif atau negatif (Samsudin et al., 2012). Sebuah sistem *sentiment analysis* otomatis telah dilihat sebagai salah satu alat bisnis intelijen yang diinginkan. Sistem ini dapat mengekstrak opini publik tentang topik tertentu, produk atau jasa yang tertanam dalam teks-teks yang tidak terstruktur (Jusoh dan Alfawareh, 2013).

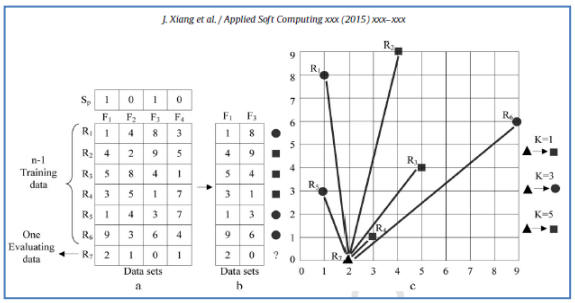
Teknik klasifikasi yang biasa digunakan untuk analisis sentimen *review* diantaranya *Naïve Bayes* (NB), *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN) (Dehkharghani et al., 2014). Eksperimental serta evaluasi menunjukan bahwa SVM, KNN dan NB merupakan tradisional teks klasifikasi. Eksperimen dan evaluasi menunjukan teks klafikasi yang valid (Yao, Min, 2012).

Pada penelitian ini algoritma *Particle Swarm Optimization* digunakan sebagai seleksi fitur untuk *review* opini publik tentang kebakaran hutan dengan metode *Support Vector Machine* dan *K-Nearest Neighbor*.

BAHAN DAN METODE

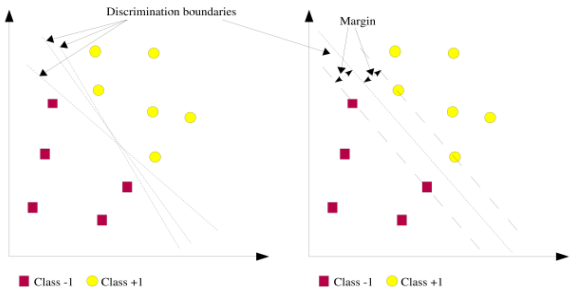
Beberapa peneliti telah menggunakan metode *Support Vector Machine* (SVM) dan *k-Nearest Neighbor* (KNN), namun belum ada dalam penelitiannya yang menggunakan *dataset* berbahasa Indonesia sehingga *preprocessing*nya tentu berbeda dengan teks berbahasa Inggris serta belum menggunakan optimasi dalam hal pemilihan fitur yang sesuai. Dalam penelitian ini, metode klasifikasi akan dikomparasi hasil evaluasinya dan akan menggunakan algoritma *Particle Swarm Optimization* (PSO) untuk menentukan fitur terbaik pada bobot atribut yang sesuai dan optimal sehingga hasil prediksi lebih akurat.

K-Nearest Neighbor adalah metode sederhana namun efektif untuk teks kategorisasi tetapi memiliki beberapa kelemahan yaitu kompleksitas pada *sample* yang komputasi kesamaan besar, *performance* KNN mudah dipengaruhi oleh *sample* tunggal, seperti *noisy sample* dan KNN tidak membangun model klasifikasi karena termasuk ke dalam *lazy learning method* (Jiang et al., 2012). Nilai k yang digunakan menyatakan jumlah tetangga terdekat yang dilibatkan dalam penentuan prediksi label kelas pada data uji. Dari k tetangga terdekat yang terpilih kemudian dilakukan *voting* kelas dari k tetangga dekat tersebut.



Gambar 1. Penerapan nilai k pada KNN
Sumber: Xiang (2015:2)

Support Vector Machine (SVM) merupakan metode *supervised learning* yang menganalisa data dan mengenali pola-pola yang digunakan untuk klasifikasi (Basari et al., 2013). SVM memiliki kelebihan yaitu mampu mengidentifikasi *hyperplane* terpisah yang memaksimalkan margin antara dua kelas yang berbeda (Chou et al., 2014). Namun SVM memiliki kekurangan terhadap masalah pemilihan parameter atau fitur yang sesuai (Basari et al., 2013). Pemilihan fitur sekaligus penyetingan parameter di SVM secara signifikan mempengaruh hasil akurasi klasifikasi (Zhao et al., 2011).

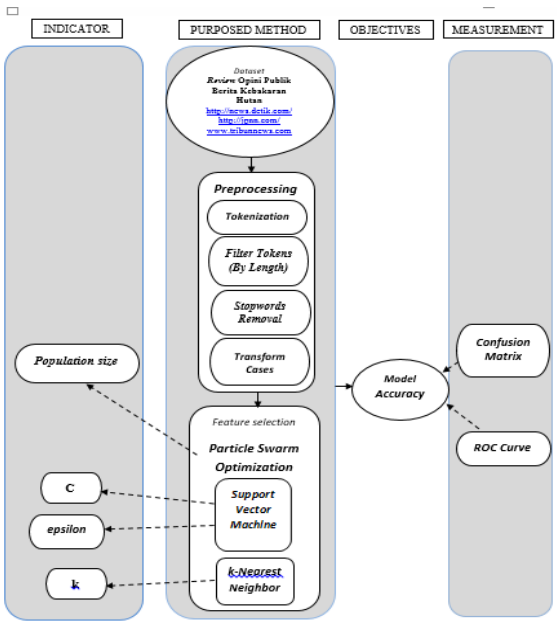


Gambar 2. SVM berusaha menemukan *hyperplane* terbaik yang memisahkan kedua class -1 dan +1
Sumber: Nugroho (2007:2)

Particle Swarm Optimization (PSO) banyak digunakan untuk memecahkan masalah optimasi serta sebagai masalah seleksi fitur (Liu et al.,

2011). Dalam teknik PSO terdapat beberapa cara untuk melakukan pengoptimasian diantaranya meningkatkan bobot atribut *(attribute weight)* terhadap semua atribut atau variabel yang dipakai, menseleksi atribut *(attribute selection)* dan *feature selection*. PSO adalah suatu teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter (Basari et al., 2013).

Confusion matrix adalah alat yang sangat berguna untuk menganalisa seberapa baik pengklasifikasi bias mengenali *tuple* dari *class* yang berbeda (Han dan Kamber, 2007). Kurva ROC akan digunakan untuk mengukur *Area Under Curve* (AUC). Kurva ROC membagi hasil positif dalam sumbu y dan hasil *negative* dalam sumbu x sehingga semakin besar area yang berada di bawah kurva, semakin baik pula hasil prediksi. Kurva *Receiver Operasi Karakteristik* (ROC) digunakan untuk mengevaluasi akurasi *classifier* dan untuk membandingkan klasifikasi yang berbeda model (Vercellis 2009), sehingga semakin besar area yang berada di bawah kurva, semakin baik pula hasil prediksi.



Gambar 3. Kerangka Pemikiran
Sumber: Data Olahan, 2016

Metode Penelitian

Metode penelitian yang penulis gunakan melalui beberapa tahapan sebagai berikut:

1. Pengumpulan Data

Data yang akan penulis gunakan yaitu data *review* opini publik berita kebakaran hutan. Data tersebut penulis peroleh dari news.detik.com, www.jpnn.com, dan www.tribunnews.com dengan *keyword* pencarian "kebakaran hutan di Riau". *Website* tersebut memiliki banyak ulasan mengenai opini publik tentang berita kebakaran hutan, sehingga penulis gunakan untuk mengklasifikasikan data *review* positif dan data *review* negatif sebanyak 360 data yang terdiri dari 180 opini positif dan 180 opini negatif dalam waktu satu bulan, mulai dari tanggal 28 Oktober sampai dengan 28 November 2015.

2. Pengolahan Data Awal

Teks yang belum diolah biasanya memiliki karakteristik dimensi yang tinggi, terdapat *noise* pada data dan terdapat struktur teks yang tidak baik. Untuk itu, dalam pengolahan data awal, *text mining* harus melalui beberapa tahapan yang disebut dengan *preprocessing*. Tahapan *preprocessing* yang dapat dilakukan dalam teks Bahasa Indonesia antara lain:

a. *Tokenize*
   Tokenize merupakan proses untuk memisah-misahkan kata. Proses memotong setiap kata dalam teks dan mengubah huruf dalam dokumen menjadi huruf kecil. Hanya huruf yang diterima, sedangkan karakter khusus atau tanda baca akan dihilangkan.
b. *Filter Tokens (By Length)*
   Filter Token (By Length) merupakan proses mengambil kata-kata penting dari hasil token (Langgeni et al. 2010). Dalam proses ini, kata-kata yang memiliki panjang tertentu akan dihapus.
c. *Stopwords Removal*
   Filter stopwords removal adalah proses menghilangkan kata-kata yang sering muncul namun tidak memiliki pengaruh apapun dalam ekstraksi sentimen suatu *review*. Kata yang termasuk seperti kata penunjuk waktu, kata tanya (Langgeni et al. 2010).
d. *Transform Cases*
   Transform Cases akan mengubah seluruh huruf menjadi huruf kecil atau kapital semua.

3. Metode yang Diusulkan

Metode yang diusulkan penulis menggunakan dua algoritma yaitu *Support Vector Machine* dan *K-Nearest Neighbor* dengan masing-masing menggunakan seleksi fitur *Particle Swarm Optimization* (PSO). Penggunaan *Particle Swarm Optimization* akan menghasilkan akurasi yang lebih tinggi.

4. Eksperimen dan Hasil Pengujian

Eksperimen yang dilakukan peneliti, menggunakan framework RapidMiner 5.3 untuk mengolah data sehingga menghasilkan nilai akurasi yang akurat dan untuk pengujian metode penulis membuat aplikasi menggunakan bahasa pemrograman PHP dan HTML.

5. Evaluasi dan Validasi Hasil

Evaluasi berfungsi untuk mengetahui akurasi dari model algoritma yang diusulkan. Validasi digunakan untuk melihat perbandingan hasil akurasi dari model yang digunakan dengan hasil yang telah ada sebelumnya. Teknik validasi

yang digunakan adalah *Cross Validation*. Akurasi algoritma akan diukur menggunakan *Confusion Matrix* dan hasil perhitungan akan ditampilkan dalam bentuk *Curve ROC (Receiver Operating Characteristic).*

HASIL DAN PEMBAHASAN

Data *training* digunakan pada saat pengujian data yang diambil dari *news.detik.com, jpnn.com,* dan *tribunnews.com*. Pengujian data dilakukan dengan menggunakan *review* opini publik tentang berita kebakaran hutan (360 *data training*, yang terdiri dari 180 *review* negatif dan 180 *review* positif) kemudian dilakukan *testing* dan *training dataset* sehingga didapatkan *accuracy* dan AUC (*Area Under Curve*).

Berikut merupakan tahapan dalam melakukan pengolahan data yaitu:

1. Pengumpulan Data
   Review berita kebakaran hutan masing-masing dikelompokkan dengan cara disimpan ke dalam satu folder yaitu folder positif dan folder negatif, kemudian tiap dokumennya diberikan ekstensi .txt sehingga dapat dibuka dengan aplikasi Notepad maupun Wordpad.
2. Pengolahan Data Awal *(Preprocessing)*
   Berikut merupakan tahapan yang dilakukan dalam *preprocessing:*
   a. *Tokenize*
      Dalam proses *tokenize* ini, semua kata yang ada di dalam tiap dokumen dikumpulkan dan dihilangkan tanda bacanya, serta dihilangkan jika terdapat simbol, karakter khusus atau apapun yang bukan huruf.

Tabel 1. Perbandingan teks sebelum dan sesudah dilakukan proses *Tokenize*

Teks sebelum dilakukan proses *tokenize*	Sy kebetulan berdomisili di palangkaraya, selama 3 hari terakhir keadaan disini tdk bs diungkap kan lagi dengan kata2.. uniknya jokowi pernah kesini sekitar 2 minggu yg lalu, tp bukannya keadaan membaik malah memburuk..
Teks setelah dilakukan proses *tokenize*	Sy kebetulan berdomisili di palangkaraya selama hari terakhir keadaan disini tdk bs diungkap kan lagi dengan kata uniknya jokowi pernah kesini sekitar minggu yg lalu tp bukannya keadaan membaik malah memburuk

Sumber: Data Olahan, 2016

   b. *Filter Tokens (By Length)*
      Dalam proses ini, kata-kata yang memiliki panjang kurang dari 4 dan lebih dari 25 akan dihapus, seperti kata yg, tdk, jd, ga, ane, gan yang merupakan kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan kata sifat yang berhubungan dengan *sentiment*.

Tabel 2. Perbandingan teks sebelum dan sesudah dilakukan proses *Filter Tokens (By Length)*

Teks sebelum dilakukan proses *filter token (by length)*	ga jelas....cuma bs ngomong... jd ketua mpr ga jelas! jd menteri kehutanan ga jelas!
Teks setelah dilakukan proses *filter token (by length)*	jelas cuma ngomong ketua jelas menteri kehutanan jelas

Sumber: Data Olahan, 2016

   c. *Stopwords Removal*
      Dalam proses ini, *Stopwords Removal* yang digunakan adalah operator *Filter Stopwords (Dictionary)* karena *dataset* yang digunakan berbahasa Indonesia, yang sebelumnya penulis telah membuat terlebih dulu daftar kata-kata yang termasuk *stopwords* kemudian *file* nya dimasukkan ke dalam operator tersebut. Dalam proses ini, kata-kata yang tidak relevan akan dihapus, yang merupakan kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan kata sifat yang berhubungan dengan *sentiment*.

Tabel 3. Perbandingan teks sebelum dan sesudah dilakukan proses *Stopwords Removal*

Teks sebelum dilakukan proses *stopwords removal*	Ya betul dinasti yg sekarang memang lebih buruk dari pada dinasti yg th 1997.... bahkan lebih bloon..... yg pilih juga rada bloon....
Teks setelah dilakukan proses *stopwords removal*	betul dinasti buruk dinasti bloon pilih rada bloon

Sumber: Data Olahan, 2016

   d. *Transform Cases*
      Dalam proses ini, kata-kata yang tidak relevan akan diubah, seperti kata yang mengandung huruf besar yang diubah menjadi huruf kecil sehingga dapat saling berhubungan dengan *sentiment*.

Tabel 4. Perbandingan teks sebelum dan sesudah dilakukan proses *Transform Cases*

Teks sebelum dilakukan proses *transform cases*	KABUT ASAP DI SUMSEL AKAN LAMA BERAKHIRNYA KARENA MEREKA BELUM SATU KATA. PEMERINTAH DAERAHNYA MASIH RESISTEN (BELUM BERTAUBAT) TIDAK SEPERTI DI RIAU DIMANA RAKYAT DAN PEMERINTAHNYA SAMA2 MEMPROTES BENCANA INI.
Teks setelah dilakukan proses *transform cases*	kabut asap sumsel berakhirnya pemerintah daerahnya resisten bertaubat riau dimana rakyat pemerintahnya memprotes bencana

Sumber: Data Olahan, 2016

Analisis Evaluasi Hasil dan Validasi Model

Validasi digunakan untuk memperoleh prediksi menggunakan model yang ada dan kemudian membandingkan hasil tersebut dengan hasil yang sudah diketahui, ini mewakili langkah paling penting dalam proses membangun sebuah model.

1. *Support Vector Machine* (SVM)

Nilai *training cycles* dalam penelitian ini ditentukan dengan cara melakukan uji coba memasukkan C dan epsilon. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *training cycles*.

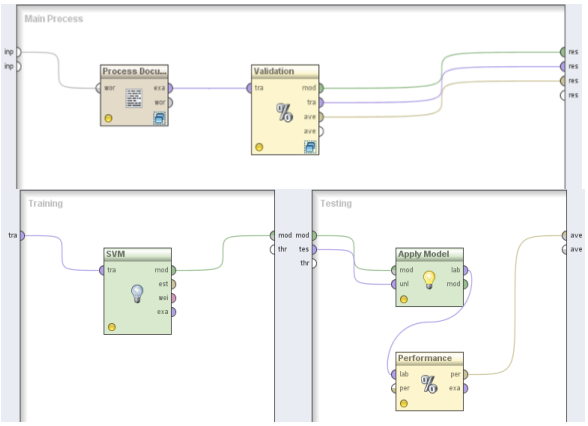
Tabel 5. Eksperimen Penentuan Nilai *Training Cycles* SVM

C	Epsilon	SVM	
		Accuracy	AUC
0.0	0.0	78.33%	0.921
0.1	0.1	80.28%	0.943
0.2	0.2	80.83%	0.944
0.3	0.3	80.28%	0.947
0.4	0.4	79.72%	0.945
0.5	0.5	80.00%	0.947
0.6	0.6	80.28%	0.946
0.7	**0.7**	**80.83%**	**0.947**
0.8	0.8	79.72%	0.943
0.9	0.9	79.72%	0.945
0.0	1.0	50.00%	0.500
1.0	1.0	50.00%	0.500
1.0	0.0	80.83%	0.943

Sumber: Data Olahan, 2016

Hasil pengujian menunjukan bahwa penerapan metode *Support Vector Machine* pada Tabel 5 dengan C = 0.7 dan Epsilon E = 0.7 dihasilkan Accuracy= 80.83% dan AUC= 0.947.

Algoritma *Support Vector Machine* (SVM) pada *framework* RapidMiner dengan desain model berikut ini:



Gambar 4. Desain Model Validasi *Support Vector Machine*

Sumber: Data Olahan, 2016

a. *Confusion Matrix*

Memberikan keputusan yang diperoleh dalam *training* dan *testing, confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek benar atau salah. *Confusion matrix* berisi informasi aktual *(actual)* dan prediksi *(predicted)* pada sistem klasifikasi.

Tabel 6. *Confusion Matrix Support Vector Machine*

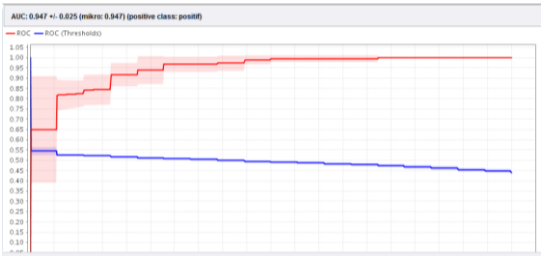
***accuracy*: 80.83% +/- 5.19% (mikro: 80.83%)**			
	***true* negatif**	***true* positif**	***class precision***
pred. negatif	116	5	95.87%
pred. positif	64	175	73.22%
class recall	64.44%	97.22%	

Sumber: Data Olahan, 2016

$$Acc\ (Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} = \frac{175+116}{175+64+5+116} = \frac{291}{360} = 0.8083$$

b. Kurva ROC

Kurva ROC *(Receiver Operating Characteristic)* adalah cara lain untuk mengevaluasi akurasi dari klasifikasi secara visual. Sebuah grafik ROC adalah plot dua dimensi dengan proporsi positif salah pada sumbu X dan positif benar pada sumbu Y. Hasil perhitungan pada kurva ROC, menggambarkan kurva ROC untuk algoritma *Support Vector Machine*. Kurva ROC *Support Vector Machine* dengan nilai AUC (*Area Under Curve*) sebesar 0.947 dimana diagnosa hasilnya *Excellent Classification*.



Gambar 5. Kurva ROC SVM
Sumber: Data Olahan, 2016

2. *Support Vector Machine* berbasis *Particle Swarm Optimization*

Nilai *training cycles* dalam penelitian ini ditentukan dengan cara melakukan uji coba memasukkan C, epsilon dan *population size*. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *training cycles*.

Tabel 7. Eksperimen Penentuan Nilai *Training Cycles* SVM Berbasis PSO

C	Epsilon	SVM		Population Size	C	Epsilon	SVM + PSO	
		Accuracy	AUC				Accuracy	AUC
0.0	0.0	78.33%	0.921	5	0.0	0.0	80.83%	0.923
0.1	0.1	80.28%	0.943	5	0.1	0.1	84.44%	0.917
0.2	0.2	80.83%	0.944	5	**0.2**	**0.2**	**86.11%**	**0.922**
0.3	0.3	80.28%	0.947	5	0.3	0.3	84.17%	0.919
0.4	0.4	79.72%	0.945	5	0.4	0.4	85.00%	0.926
0.5	0.5	80.00%	0.947	5	0.5	0.5	85.28%	0.924
0.6	0.6	80.28%	0.946	5	0.6	0.6	85.28%	0.935
0.7	**0.7**	**80.83%**	**0.947**	5	0.7	0.7	85.83%	0.935
0.8	0.8	79.72%	0.943	5	0.8	0.8	85.28%	0.935
0.9	0.9	79.72%	0.945	5	0.9	0.9	85.56%	0.924
0.0	1.0	50.00%	0.500	5	0.0	1.0	50.00%	0.500
1.0	1.0	50.00%	0.500	5	1.0	1.0	50.00%	0.500
1.0	0.0	80.83%	0.943	5	1.0	0.0	84.44%	0.915

Sumber: Data Olahan, 2016

Hasil terbaik pada eksperimen SVM berbasis PSO di atas adalah C=0.2 dan Epsilon E=0.2 serta population size=5 yang dihasilkan *accuracy*=86.11% dan AUC=0.922. Hal ini menunjukan bahwa dengan menggunakan optimasi *Particle Swarm Optimization* dapat meningkatkan akurasi yang lebih baik.

Hasil pengujian data training metode *Support Vector Machine* berbasis *Particle Swarm Optimization* menggunakan *Set Role* yang berfungsi untuk menentukan field pada kelas kemudian diberikan optimasi menggunakan *Particle Swarm Optimization* agar akurasi yang dihasilkan lebih tinggi. Pengukuran akurasi tersebut, akan dijabarkan melalui Kurva ROC dan *Confusion Matrix* di bawah ini:

a. *Confusion Matrix*

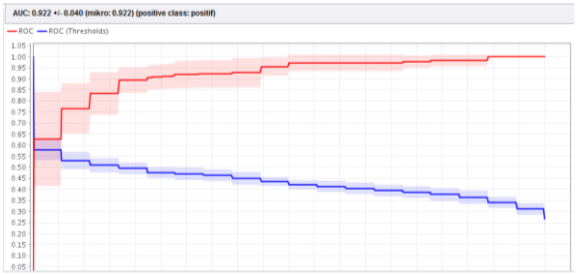
Tabel 8. *Confusion Matrix* SVM Berbasis PSO

***accuracy*: 86.11% +/- 3.04% (mikro: 86.11%)**			
	***true* negatif**	***true* positif**	***class precision***
pred. negatif	157	27	85.33%
pred. positif	23	153	86.93%
class recall	87.22%	85.00%	

Sumber: Data Olahan, 2016

$$\text{Acc (Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{153+157}{153+23+27+157} = \frac{310}{360} = 0.8611$$

b. Kurva ROC



Gambar 6. Kurva ROC SVM Berbasis PSO
Sumber: Data Olahan, 2016

Kurva ROC yang dihasilkan berdasarkan pengujian data pada gambar di atas, menunjukan bahwa ada peningkatan pada akurasi menggunakan *Support Vector Machine* berbasis *Particle Swarm Optimization* sebesar 86.11% dan AUC sebesar 0.922.

3. *K-Nearest Neighbor* (K-NN)

Nilai k yang digunakan menyatakan jumlah tetangga terdekat yang dilibatkan dalam penentuan prediksi label kelas pada data uji. Untuk memperkirakan nilai k yang terbaik, bisa dilakukan dengan menggunakan teknik validasi silang *(Cross Validation)*.

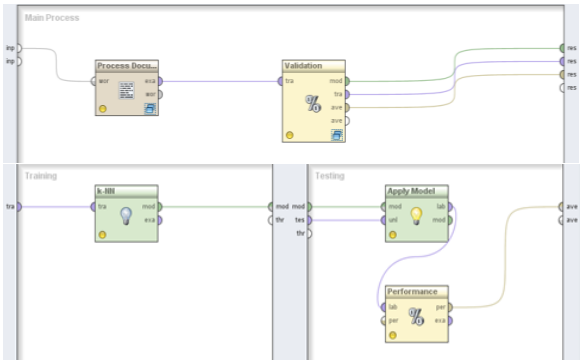
Tabel 9. Eksperimen Penentuan Nilai *Training* K-NN

k	k-NN	
	Accuracy	AUC
1	78.06%	0.500
2	81.39%	0.818
3	81.39%	0.851
4	81.39%	0.882
5	82.50%	0.906
6	**85.00%**	**0.918**
7	83.06%	0.916
8	84.17%	0.918
9	82.22%	0.914
10	83.89%	0.915

Sumber: Data Olahan, 2016

Hasil pengujian menunjukan bahwa penerapan metode *k-Nearest Neighbor* pada Tabel 9 dengan penentuan nilai k=6 menghasilkan *Accuracy*= 85.00% dan AUC= 0.918 adalah nilai yang paling tertinggi.

Algoritma *K-Nearest Neighbor* (K-NN) pada *framework* RapidMiner dengan desain model berikut ini:



Gambar 7. Desain Model Validasi K-NN
Sumber: Data Olahan, 2016

a. *Confusion Matrix*

Tabel 10. *Confusion Matrix* K-NN

***accuracy:* 85.00% +/- 3.56% (mikro: 85.00%)**			
	***true* negatif**	***true* positif**	***class precision***
pred. negatif	151	25	85.80%
pred. positif	29	155	84.24%
class recall	83.89%	86.11%	

Sumber: Data Olahan, 2016

$$\text{Acc (Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{155+151}{155+29+25+151} = \frac{306}{360} = 0.85$$

b. Kurva ROC

Gambar 8. Kurva ROC K-NN
Sumber: Data Olahan, 2016

Kurva ROC tersebut diartikan dengan *False* Positif untuk garis horizontal dan *True* Positif untuk garis vertikal dengan nilai AUC= 0.918.

4. *K-Nearest Neighbor* Berbasis *Particle Swarm Optimization*

Penelitian metode *k-Nearest Neighbor* berbasis PSO, dengan melakukan uji coba nilai k sebagai tetangga terdekat, dan *population size*=5. Adapun hasil dari perhitungannya ditunjukan pada Tabel 11.

Tabel 11. Eksperimen Penentuan Nilai *Training* K-NN Berbasis PSO

k	k-NN		Population Size	k	k-NN & PSO	
	Accuracy	AUC			Accuracy	AUC
1	78.06%	0.500	5	**1**	**73.06%**	**0.500**
2	81.39%	0.818	5	2	62.22%	0.668
3	81.39%	0.851	5	3	71.11%	0.723
4	81.39%	0.882	5	4	70.56%	0.729
5	82.50%	0.906	5	5	66.39%	0.705
6	**85.00%**	**0.918**	5	6	71.67%	0.739
7	83.06%	0.916	5	7	70.28%	0.738
8	84.17%	0.918	5	8	69.17%	0.701
9	82.22%	0.914	5	9	69.44%	0.732
10	83.89%	0.915	5	10	70.56%	0.728

Sumber: Data Olahan, 2016

Hasil perhitungan dari Tabel 11 di atas menunjukkan dengan memasukkan nilai k=1 mendapatkan *Accuracy*=73.06% dan AUC=0.500 adalah nilai yang tertinggi diantara nilai k yang lainnya, namun ternyata terjadi penurunan hasil akurasi pada k-NN sekitar 11% sampai dengan 12% apabila ditambahkan optimasi PSO.

a. *Confusion Matrix*

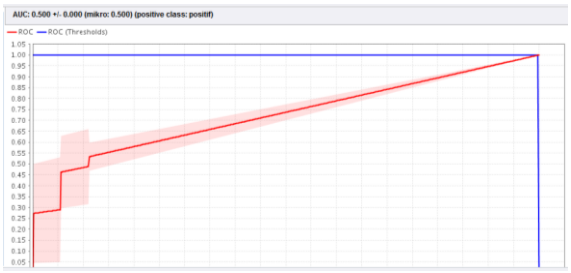
Tabel 12. *Confusion Matrix* K-NN Berbasis PSO

***accuracy:* 73.06% +/-3.52% (mikro: 73.06%)**			
	***true* negatif**	***true* positif**	***class precision***
pred. negatif	175	92	65.54%
pred. positif	5	88	94.62%
class recall	97.22%	48.89%	

Sumber: Data Olahan, 2016

$$\text{Acc (Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{88+175}{88+5+92+175} = \frac{263}{360} = 0.7306$$

b. Kurva ROC



Gambar 9. Kurva ROC K-NN Berbasis PSO
Sumber: Data Olahan, 2016

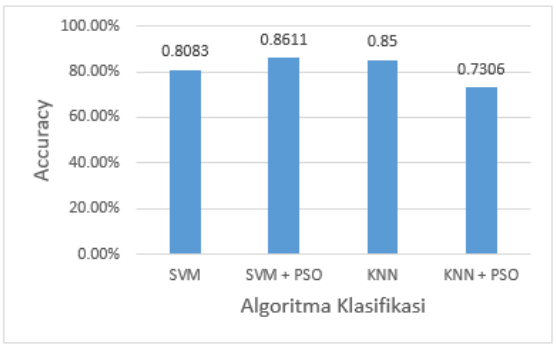
Nilai AUC yang dihasilkan dari Gambar 9 di atas sebesar 0.500, yang termasuk ke dalam *Failure.* Namun ternyata *k-Nearest Neighbor* yang dioptimasi dengan *Particle Swarm Optimization* tidak dapat meningkatkan nilai akurasi yang lebih tinggi dibandingkan dengan metode K-NN saja.

Adapun perbandingan hasil komparasi *Accuracy* dan AUC Algoritma yang telah digunakan sebagai berikut:

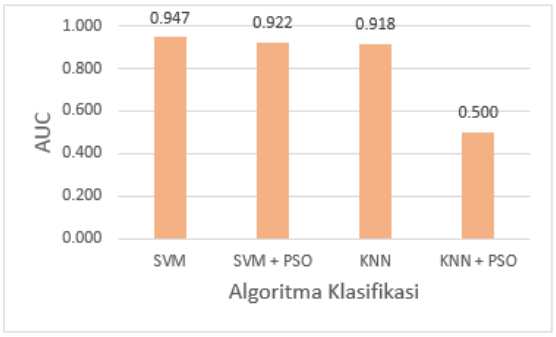
Tabel 13. Komparasi *Accuracy* dan AUC Algoritma Klasifikasi

Algoritma	*Accuracy*	AUC
SVM	80.83%	0.947
SVM + PSO	86.11%	0.922
K-NN	85.00%	0.918
K-NN + PSO	73.06%	0.500

Sumber: Data Olahan, 2016



Gambar 10. Komparasi *Accuracy* Algoritma Klasifikasi
Sumber: Data Olahan, 2016

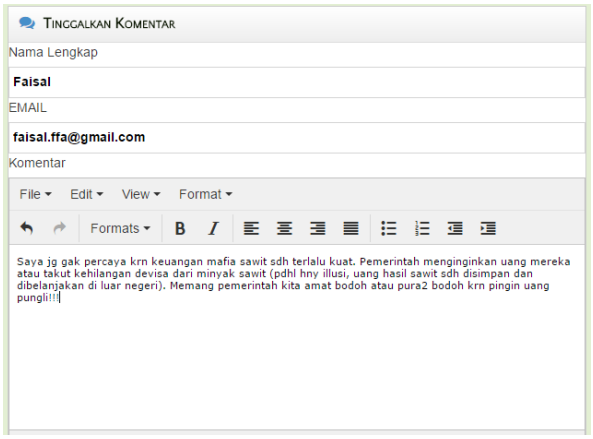


Gambar 11. Komparasi AUC Algoritma Klasifikasi
Sumber: Data Olahan, 2016

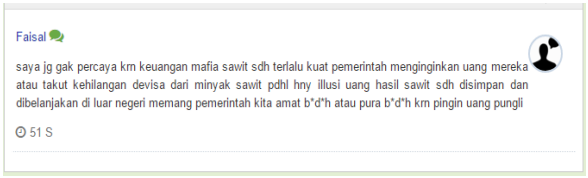
Desain Dan Implementasi

Penulis merancang aplikasi berbasis *website* untuk menguji model dengan menggunakan *dataset* berita kebakaran hutan. Aplikasi dibuat dengan menggunakan bahasa pemrograman PHP dan HTML.

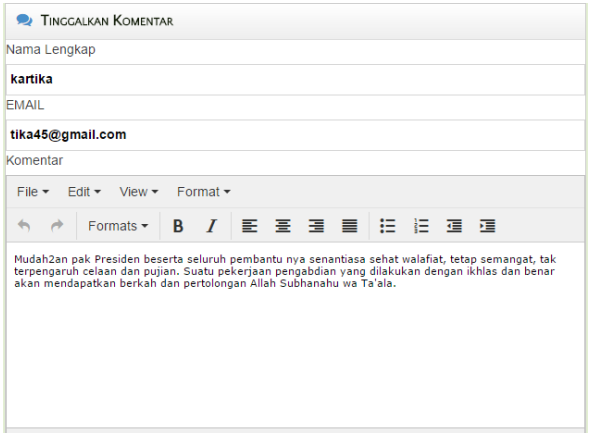
Gambar 12. *Home Page* Opini Publik Berita Kebakaran Hutan
Sumber: Data Olahan, 2016



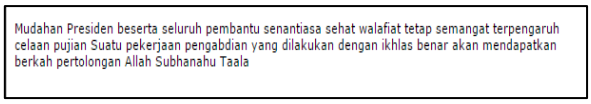
Gambar 13. Tampilan *Preprocessing Tokenize*
Sumber: Data Olahan, 2016



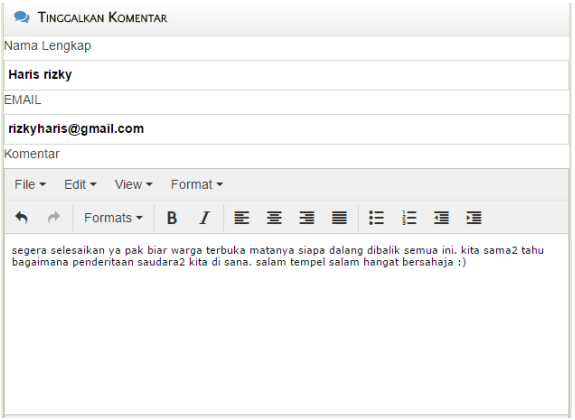
Gambar 14. Tampilan Hasil *Tokenize*
Sumber: Data Olahan, 2016



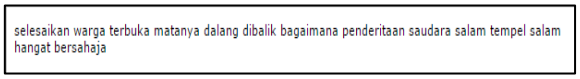
Gambar 15. Tampilan *Preprocessing Filter Tokens (By Length)*
Sumber: Data Olahan, 2016



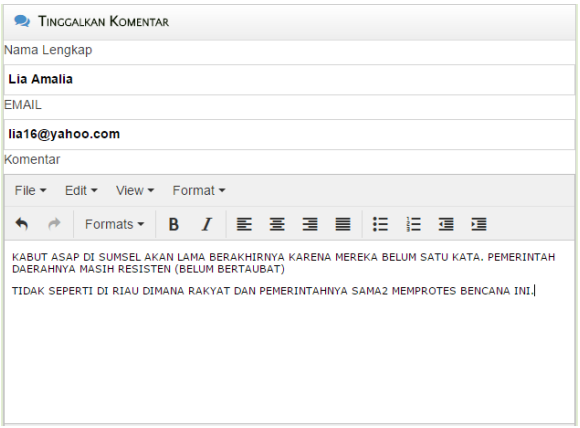
Gambar 16. Tampilan Hasil *Filter Tokens (By Length)*
Sumber: Data Olahan, 2016



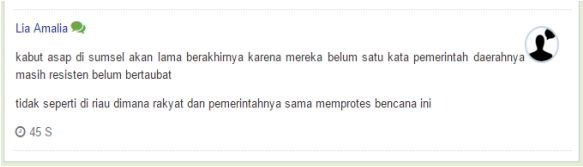
Gambar 17. Tampilan *Preprocessing Stopwords Removal*
Sumber: Data Olahan, 2016



Gambar 18. Tampilan Hasil *Stopwords Removal*
Sumber: Data Olahan, 2016



Gambar 19. Tampilan *Preprocessing Transform Cases*
Sumber: Data Olahan, 2016



Gambar 20. Tampilan Hasil *Transform Cases*
Sumber: Data Olahan, 2016

KESIMPULAN

Berdasarkan pengujian yang telah dilakukan terhadap *review* opini publik berita kebakaran hutan yang dikumpulkan melalui *online news* sebanyak 360 data (180 positif dan 180 negatif) dengan menggunakan metode *Support Vector Machine* (SVM), *Support Vector Machine* berbasis *Particle Swarm Optimization* (SVM+PSO), *k-Nearest Neighbor* (k-NN), dan *k-Nearest Neighbor* berbasis *Particle Swarm Optimization* (k-NN+PSO) maka hasilnya adalah hipotesa awal tidak sesuai dengan hasil akhir. Dalam penelitian ini, hasil perhitungan metode SVM memiliki *Accuracy* sebesar 80.83% dan AUC sebesar 0.947 sedangkan Metode SVM+PSO menghasilkan *Accuracy* sebesar 86.11% dan AUC sebesar 0.922. Pengujian juga telah dibandingkan dengan metode k-NN. Hasil perhitungan yang diperoleh dari pengujian data dengan metode k-NN yaitu *Accuracy* sebesar 85.00% dan AUC sebesar 0.918, kemudian dibandingkan dengan k-NN+PSO menghasilkan nilai *Accuracy* sebesar 73.06% dan AUC sebesar 0.500. Penerapan *Particle Swarm Optimization* (PSO) terbukti dapat meningkatkan akurasi pada klasifikasi *review* opini publik berita kebakaran hutan untuk mengidentifikasi antara *review* positif dan *review* negatif untuk algoritma klasifikasi SVM, sedangkan untuk algoritma k-NN justru menghasilkan akurasi yang lebih rendah dibandingkan algoritma k-NN saja dengan penurunan 11-12%. Hal ini merupakan suatu penemuan dalam penelitian *text mining* ini yang menyimpulkan bahwa optimasi menggunakan PSO belum tentu dapat mengoptimalkan nilai akurasi.

Mengingat banyaknya penelitian terdahulu yang telah menerapkan *text mining* berbahasa Inggris dengan sukses meningkatkan nilai akurasi k-NN menggunakan PSO, maka dapat dianalisa bahwa optimasi PSO pada algoritma k-NN dengan menggunakan *dataset* Bahasa Indonesia belum tentu dapat meningkatkan akurasi. Metode SVM terbukti lebih unggul dalam klasifikasi teks *review* opini berita ini karena SVM bekerja dengan mencari parameter *hyperplane* yang terbaik yaitu nilai C dan Epsilon sehingga ada banyak kemungkinan akurasi dapat optimal, namun waktu pengujian data lebih lama dilakukan oleh SVM+PSO dibandingkan metode KNN+PSO.

Dengan ini dapat disimpulkan bahwa *Support Vector Machine* berbasis *Particle Swarm Optimization* (SVM+PSO) dengan *k-Nearest Neighbor* berbasis *Particle Swarm Optimization* (k-NN+PSO) lebih tinggi nilai akurasi *Support Vector Machine* berbasis *Particle Swarm Optimization* (SVM+PSO) dan PSO tidak dapat

meningkatkan nilai akurasi untuk metode k-NN dalam *dataset* berbahasa Indonesia seperti berita kebakaran hutan dalam penelitian ini.

REFERENSI

Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). *Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization*. *Procedia Engineering*, *53*, 453–462. http://doi.org/10.1016/j.proeng.2013.02.059

Chou, J.-S. S., Cheng, M.-Y. Y., Wu, Y.-W. W., & Pham, A.-D. D. (2014). *Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. Expert Systems with Applications,* 41(8), 3955–3964. http://doi.org/10.1016/j.eswa.2013.12.035

Dehkharghani, R., Mercan, H., Javeed, A., & Saygin, Y. (2014). *Sentimental causal rule discovery from Twitter. Expert Systems with Applications*, *41*(10), 4950–5958. http://doi.org/10.1016/j.eswa.2014.02.024

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). *An improved K-nearest-neighbor algorithm for text categorization*. *Expert Systems with Applications*, *39*(1), 1503–1509. http://doi.org/10.1016/j.eswa.2011.08.040

Jusoh, S., & Alfawareh, H. M. (2013). *Applying fuzzy sets for opinion mining*. *2013 International Conference on Computer Applications Technology (ICCAT)*, 1–5. http://doi.org/10.1109/ICCAT.2013.6521965

Langgeni, D. P., Baizal, Z. K. A., & W, Y. F. A. (2010). *Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection*. In *Seminar Nasional Informatika 2010* (Vol. 2010, pp. 1–10).

Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). *An improved particle swarm optimization for feature selection. Journal of Bionic Engineering*, *8*(2), 191–200. http://doi.org/10.1016/S1672-6529(11)60020-6

Rozi, I. F., Hadi, S., & Achmad, E. (2012). Implementasi *Opinion Mining* (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. *Universitas Stuttgart*, *6*(1), 37–43.

Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. (2012). *Is artificial immune system suitable for opinion mining? Conference on Data Mining and Optimization*, (September), 131–136. http://doi.org/10.1109/DMO.2012.6329811

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. *Business Intelligence: Data Mining and Optimization for Decision Making*. http://doi.org/10.1002/9780470753866

Xiang, J., Han, X., Duan, F., Qiang, Y., Xiong, X., Lan, Y., & Chai, H. (2015). *A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method. Applied Soft Computing*, *31*, 293–307. http://doi.org/10.1016/j.asoc.2015.01.043

Yao, Zhi-Min. (2012), *An Optimized NBC Approach in Text Classification*. *Physics Procedia*, 24, 1910-1914

Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). *Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. Expert Systems with Applications*, *38*(5), 5197–5204. http://doi.org/10.1016/j.eswa.2010.10.041

BIODATA PENULIS

Lilyani Asri Utami, M.Kom. Lahir di Bogor pada tanggal 15 November 1991, lulusan pendidikan Program S2 jurusan Ilmu Komputer – Pasca Sarjana STMIK Nusa Mandiri Jakarta tahun 2016. Bekerja sebagai instruktur di STMIK Nusa Mandiri Jakarta sejak tahun 2014. Sampai saat ini telah mengikuti beberapa kegiatan seminar nasional untuk menambah pengetahuan tentang menulis untuk menuangkan pemikiran dalam rangka melaksanakan Tri Dharma Perguruan Tinggi. Sebuah *prociding* berjudul "Sistem Informasi Administrasi Pasien Pada Klinik Keluarga Depok" pernah dimuat pada Konferensi Nasional Ilmu Pengetahuan dan Teknologi (KNIT) Nusa Mandiri pada tahun 2015. Semoga penelitian ini dapat memberikan manfaat bagi para pembacanya. Demikian dari saya dan terucap terima kasih.