ISSN: 2477-5126
e-ISSN: 2548-9356

SUPPORT VECTOR MACHINE UNTUK IDENTIFIKASI BERITA HOAX TERKAIT VIRUS CORONA (COVID-19)

Rani Kurnia Putri[1], Muhammad Athoillah[2]
[1]Program Studi Pendidikan Matematika, Fakultas Sains dan Teknologi, Universitas PGRI Adi Buana Surabaya
[2]Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas PGRI Adi Buana Surabaya
Jl. Dukuh Menanggal XII, Surabaya, 60234, Indonesia
email: [1]Rani@unipasby.ac.id, [2]athoillah.muhammad@gmail.com

***Abstract* – Covid-19 or Coronavirus is a virus resulting from the evolution of similar viruses, namely MERS-Cov and SARS-CoV which was first discovered in Wuhan, one of the largest metropolitan cities in China on December 31, 2019 and has claimed millions of victims during 2020. Certainly, Covid-19 is the main topic of discussion in various news media, both in Indonesia and the world. However, ironically, with so many news circulating, not a few that news are hoax news or news that cannot be justified for its truth. Identification of hoax news in cyberspace has actually been carried out by the internet community and published on the turnbackhoax.id webpage, yet the identification method used on that page is still done manually. As a result, if the information grows a lot, it will be more difficult and troublesome. The identification of hoax news can be categorized into classification problems that can be solved with various algorithms, including Support Vector Machine (SVM). SVM is an algorithm that works by defining the boundaries between classes with the maximum distance obtained from the closest data by measuring the hyperplane margins between classes. Consequently, the result of class separation is better. In this framework, an automatic system has been built that can identify news that is included in the hoax category or not using SVM algorithm, which then the validation process is carried out by k-fold cross validation method. The results indicated that the system was able to identify news well, as evidenced by the average values of F-Measure, Recall and Precision, respectively, were 78.02%, 78.18% and 78.96%.**

***Abstrak* – Covid-19 atau biasa disebut Virus Corona, merupakan virus hasil dari evolusi virus sejenis yaitu MERS-Cov dan SARS-CoV yang pertama kali diketahui muncul di kota Wuhan, salah satu kota metropolitan terbesar di Cina pada 31 Desember 2019 dan telah memakan jutaan korban selama tahun 2020. Disepanjang tahun tersebut tentunya Covid-19 menjadi bahasan utama di berbagai media berita, baik di Indonesia maupun dunia. Ironisnya, dengan banyaknya berita yang beredar, tidak sedikit berita yang muncul adalah berita hoax atau berita tidak dapat dipertanggungjawabkan kebenarannya. Identifikasi berita hoax di dunia maya sebenarnya telah dilakukan oleh komunitas internet dan dipublikasikan pada laman turnbackhoax.id. Hanya saja, metode identifikasi yang dilakukan pada laman tersebut masih dilakukan secara manual, sehingga jika informasi semakin berkembang dan banyak, tentunya akan semakin sulit dan merepotkan. Identifikasi berita hoax secara otomatis dapat dikategorikan ke dalam masalah klasifikasi yang tentunya dapat di selesaikan dengan berbagai macam algoritma, diantanya Support Vector Machine (SVM). Algoritma SVM mendefinisikan terlebih dahulu batas antar kelas dengan jarak optimal yang didapat dari data terdekat dengan cara mengukur margin hyperplane antar kelas sehingga pemisahan kelas yang dihasilkan menjadi lebih baik. Pada penilitian ini telah dibangun sebuah sistem otomatis yang dapat mengidentifikasi berita yang termasuk dalam kategori hoax atau tidak dengan memanfaatkan algoritma SVM yang selanjutnya proses validasinya dilakukan dengan metode k-fold cross validation. Hasil penelitian memperlihatkan bahwa sistem yang dibangun mampu mengidentifikasi berita dengan baik, dibuktikan dengan rata-rata nilai Presisi, Recall dan F-Measure secara berturut adalah 78,96%, 78,18% dan 78,02%.**

***Kata Kunci* – Berita Hoax, Virus Corona, Klasifikasi, Suport Vector Machine, Text Mining.**

*) **penulis korespondensi**: Rani Kurnia Putri
Email:Rani@unipasby.ac.id

I. PENDAHULUAN

Virus Corona telah berhasil menghebohkan seluruh dunia tidak hanya di dunia kedokteran tapi juga berdampak besar disemua sektor kehidupan seperti Pendidikan, Ekonomi, bahkan Politik Pemerintahan. Coronavirus 2019 (Covid-19) merupakan virus hasil dari evolusi virus sejenis yaitu MERS-Cov dan SARS-CoV[1]. Coronavirus pertama kali diketahui muncul di kota Wuhan, salah satu kota metropolitan terbesar di Cina pada 31 Desember 2019 yang kemudian menyebar dengan sangat capat ke berbagai daerah bahkan negara di dunia [2]. Menurut organisasi kesehatan dunia WHO pada laman web resminya, sebagian besar orang yang telah terinfeksi Coronavirus akan mengalami gejala klinis seperti sesak napas, sakit, nyeri, serta sakit tenggorokan. Berdasarkan sumber yang sama pula, data terbaru menyebutkan bahwa sampai dengan 19 Januari 2021 atau setahun lebih semenjak dari kasus pertama ditemukan, jumlah kasus konfirmasi positif Covid-19 di seluruh dunia telah mencapai 93,956,883 jiwa dengan total kematian yang diakibatkannya mencapai 2,029,084 jiwa [3].

Disepanjang tahun 2020, Coronavirus tentunya menjadi bahasan utama di berbagai media berita, baik di Indonesia maupun dunia, tidak hanya media mainstream seperti koran dan televisi, berita tentang coronavirus juga menjadi topik ‘panas’ di dunia maya. Ironisnya, dengan banyaknya berita yang beredar, tidak sedikit berita yang muncul adalah berita hoax. Dalam Kamus Besar Bahasa Indonesia (KBBI), definisi dari kata hoaks (hoax) adalah informasi bohong[4], yang artinya berita tersebut tidak dapat dipertanggungjawabkan kebenarannya. Di Indonesia sendiri, berita hoax sangat mudah ditemukan dalam berbagai media facebook, aplikasi-aplikasi chatting dan berbagai situs web. Dari berbagai jenis berita

hoax yang beredar, topik sosial politik, sara dan kesehatan menjadi topik terbanyak yang ditemukan adapun beberapa alasan kenapa berita hoax sangat mudah berkembang di Indonesia diantaranya adalah berita hoax sudah menjadi sarana dari para kelompok tertentu untuk mencapai tujuan, sehingga penyebaran hoax kini menjadi lebih terstruktur dan masiv terlebih dengan kondisi masyarakat kita yang kurang sadar literasi sehingga seringkali abai dengan perlunya memvalidasi kebenaran dari suatu berita[5]. Berita bohong tentunya dapat membahayakan masyarakat, penelitian menunjukkan bahwa berita hoax atau bohong dapat dijadikan alat untuk melakukan kejahatan, ujaran kebencian dan kejahatan lain yang tidak hanya merugikan perseorangan namun juga masyarakat luas pada umumnya[6]. Oleh sebab itu, dengan berkembangnya berita hoax di media sosial yang begitu masif di berbagai media online saat ini dan yang akan datang, keberadaan sebuah sistem yang mampu mendeteksi kebenaran sebuah berita secara otomatis sangatlah diperlukan. Identifikasi berita hoax sebenarnya telah dilakukan oleh komunitas internet dan dipublikasikan pada laman *turnbackhoax.id*, hanya saja, metode identifikasi yang dilakukan pada laman tersebut masih dilakukan secara manual, sehingga jika informasi semakin berkembang dan banyak, tentunya akan semakin sulit dan merepotkan. Laman turnbackhoax.id selama ini dikelola oleh masyarakat anti hoax Indonesia atau MAFINDO. Sedangkan sumber berita pada laman tersebut berasal dari berbagai laporan di jejaring media sosial Facebook dalam forum bernama FAFHH (forum anti fitnah hasut dan hoax)[7].

Text mining adalah bagian dari kelompok ilmu *Natural Language Processing* (NLP) yang bertujuan untuk meng-ekstraksi dan menemukan makna yang terkandung dalam teks tak terstruktur secara otomatis[8]. Identifikasi berita dapat didefinisikan sebagai kemampuan sistem dalam meng-ekstraksi dan mengklasifikasikan teks berita tersebut dalam sebuah kategori secara otomatis. Umumnya, istilah kategorisasi atau klasifikasi teks banyak digunakan dalam banyak judul penelitian, namun pada dasarnya yang para peneliti lakukan adalah mengklasifikasikan dokumen. Karena pada dasarnya setelah data teks di proses dan di transformasikan ke dalam format numerik, metode *data mining* kembali berlaku [9]. Sejauh ini banyak algoritma/metode *data mining* yang telah dikembangkan untuk memecahkan masalah klasifikasi, diantaranya adalah *Support Vector Machine* (SVM). Richhariya dkk dengan algoritma yang disebut *universum support vector machine based recursive feature elimination* (USVM-RFE) yang merupakan pengembangan dari algoritma SVM untuk mediagnosis penyakit Alzheimer's[10]. Dalam bidang Natural Language Processing (NLP), SVM juga banyak digunakan diantaranya adalah untuk analisa sentiment masyarakat terhadap berita-berita dengan kasus tertentu di media sosial seperti *cyberbullying*[11], *Hates speech* [12], politik [13] dan lain sebagainya. Berbeda dari metode klasifikasi pada umumnya yang berusaha menemukan *hyperplane* (garis yang memisahkan) antar kelas, dalam algoritma SVM, garis *hyperplane* terbaik dicari pada ruang input. Selain itu, SVM merupakan algoritma yang dimulai dengan mendefinisikan terlebih dahulu batas antar kelas dengan jarak paling maksimal yang didapat dari data terdekat, adapun jarak tersebut diperoleh dengan cara mengukur *margin hyperplane* antar kelas. *Margin* ialah jarak diantara titik paling dekat dari masing-masing kelas dengan *hyperplane*, sehingga pemisahan kelas yang dihasilkan menjadi lebih baik[14].

Didasari oleh latar belakang tersebut, maka pada penelitian ini telah dibangun sebuah sistem yang mampu untuk mengidentifikasi teks berita hoax atau non-hoax secara otomatis dengan memanfaatkan algoritma *Support Vector Machine*.

II. PENELITIAN YANG TERKAIT

Teks mining merupakan bagian dari Natural Language Processing (NLP) yang bertujuan untuk mengekstrak dan menemukan makna teks tidak terstruktur secara otomatis [15]. Berbagai riset terkait dengan text mining belakangan ini semakin banyak. Sebagai contoh, riset yang dilakukan oleh Yuliang Li dkk [16] dimana mereka menganalisis 152 catatan Pengobatan Tradisional Cina (TCM) diabetes Tipe-2 melalui metode teks mining dengan tujuan mengidentifikasi jenis obat, resep ataupun formula yang nantinya direkomendasikan untuk para pasien tersebut menggunakan Algoritme FP-Growth. Atau contoh lainnya adalah penelitian yang dilakukan oleh Peyman Beyranvand dkk [17] dimana mereka menggunakan teknik text mining untuk mengidentifikasi dan mengklaim pendaftaran pelanggan adalam suatu lembaga sehingga proses pendaftaran menjadi lebih efisien. Kemudian contoh lainnya juga adalah penelitian yang dilakukan oleh De Lima dkk [18] yang menggunakan teks mining sebagai metode untuk mengidentifikasi perkembangan teknologi panel surya yang berkembang di masyarakat saat ini. Selain itu, penelitian terkait dengan text mining yang menggunakan Support Vector Machine sebagai algoritmanya masih menarik dan banyak dilakukan hingga saat ini. Misalkan, penelitian yang telah dilakukan oleh Rajvanshi dkk [19] yang membandingkan hasil klasifikasi teks antara algoritma Naïve Bayes dan algoritma SVM. Atau penelitian serupa yang dilakukan oleh Zhiquan Wang dkk [20] dan Yuling Chen dkk [21] yang menggunakan algoritma SVM dan CNN untuk menganalisis sentimen publik terhadap berita yang sedang populer di internet.

III. METODE PENELITIAN

A. Data Penelitian

Data yang dipakai dalam penelitian ini ialah data berita hoax dan non-hoax baik dari portal berita online ataupun yang tersebar di berbagai media sosial. Berita tersebut sebelumnya telah diferifikasi kebenarannya oleh tim MAFINDO berdasarkan hasil diskusi dan pencarian fakta yang dilakukan oleh para anggota yang hasilnya kemudian dipublikasikan pada laman *turnbackhoax.id*[22]. Data dalam penelitian ini diambil dari publikasi di situs *turnbackhoax.id* pada rentang waktu 1 Januari 2020 sampai dengan 29 Agustus 2020 yang kemudian dibagi menjadi dua kelompok dataset yaitu dataset untuk proses training sebesar 80% dan sisanya 20% untuk proses testing. Untuk memvalidasi hasil klasifikasi dari sistem identifikasi berita hoax yang telah dibangun, digunakan metode *k-fold cross validation*[23], dimana data kemudian dibagi menjadi 5 bagian yang selanjutnya diuji secara berulang dengan saling silang kombinasi data training dan testing sehingga setiap

proses uji data yang diuji merupakan hasil kombinasi data training dan data testing yang berbeda-beda.

B. Pra-Proses Klasifikasi

Pra-proses ini merupakan tahap pemrosesan data teks sebelum diklasifikasikan dengan algoritma SVM, proses ini perlu dilakukan agar hasil yang didapatkan optimal. Pra-pemrosesan teks dimulai dengan analisis leksikal, dalam tahap ini data teks dianalisis untuk kemudian diubah dari urutan karakter menjadi urutan token. Selanjutnya, teks tersebut dibersihkan dengan menghapus "Stopwords" yang ada didalamnya, "Stopwords" merupakan kata-kata yang biasanya frekuensi kemunculannya tinggi namun dianggap tidak memiliki makna serta tidak berpengaruh saat digunakan dalam proses pencarian, seperti kata hubung "tapi", "yang", "dan", "akan", "atau", dan kata lainnya. Penghapusan stopwords bertujuan untuk mengurangi skala indeks data yang nantinya berpengaruh pada kecepatan dan peforma dari sistem yang dibuat. Setelah penghapusan "Stopwords", Langkah selanjutnya adalah proses Stemming, pada proses stemming ini suatu kata akan diproses kembali untuk menghilangkan prefiks dan sufiksnya sehingga kata tersebut menjadi kata dasar sehingga nantinya dapat meningkatkan kinerja pencarian kata[24].

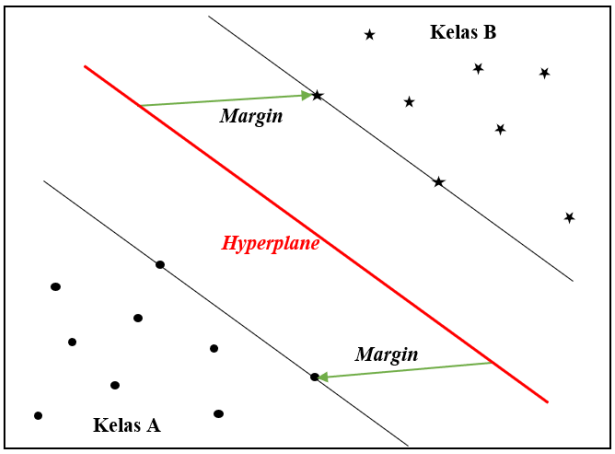
Tahap berikutnya dari pra-proses adalah mentransformasikan teks menjadi bentuk numerik, salah satunya adalah dengan metode *Term Frequency - Invers Document Frequency* (TF-IDF) yang dirumuskan dengan persamaan berikut:

$$TF - IDF = TF \log\left(\frac{n}{DF}\right)$$

Dimana *TF* merupakan frekuensi kemunculan sebuah *term* (kata) dalam dokumen dan *DF* merupakan jumlah dokumen yang memuat *term (*kata*)*. Sedangkan *n* adalah jumlah dari keseluruhan dokumen[25]. TF-IDF adalah metode pemberian nilai bobot pada hubungan sebuah kata (istilah) dengan sebuah dokumen. Nilai dari TFIDF merupakan hasil pengukuran statistik yang digunakan untuk mengevaluasi seberapa signifikan nilai dari sebuah kata dalam suatu dokumen atau sekelompok kata. Untuk satu dokumen miaslnya, setiap kalimat diasumsikan sebagai dokumen. Kemudian frekuensi kemunculan *term* (kata) dalam suatu dokumen tertentu menunjukkan betapa siginifikan pentingnya kata tersebut dalam dokumen. Berapa kali sebuah dokumen berisi kata itu menunjukkan seberapa umum kata itu. Nilai bobot dari suatu kata akan lebih besar jika kata tersebut muncul di banyak dokumen dan sebaliknya [26].

C. Proses Klasifikasi

Setelah data diolah sedemikian rupa pada tahap pra-proses, langkah selanjutnya dari penelitian ini adalah proses klasifikasi. Proses ini terbagi menjadi 2 tahapan utama seperti yang telah banyak disebutkan sebelumnya, yaitu training dan testing. Pada saat training, SVM akan mencari *hyperplane* terbaik pada ruang input yang didapat dengan cara menentukan definisi dari batas kelas dengan jarak maksimal yang diperoleh dari data-data terdekat. Sedangkan jarak maksimal yang dimaksudkan adalah jarak yang didapat dengan cara mengukur *margin hyperplane* antar kelas, adapun *margin* merupakan jarak antar titik paling dekat dengan *hyperplane* dari masing-masing kelas [27] (lihat Gambar 1 untuk lebih jelasnya).



Gambar 1. Ilustrasi Pemisahan Kelas pada SVM

Pada Gambar 1, terlihat pula bahwa pada pemisahan kelas SVM terdapat dua garis pemisah pada bidang input, sepasang bidang sejajar tersebut dituliskan dengan persamaan sebagai berikut :

$$x_i.w + b \geq -1 \text{ saat } y_i = -1$$
$$x_i.w + b \geq +1 \text{ saat } y_i = +1$$

Dengan w merupakan normal bidang sedangkan b merupakan posisi bidang relatif terhadap pusat dari koordinat. Sedangkan data pada bidang pemisah inilah yang kemudian dikenal dengan *support vector*. Permasalahan bidang pemisah ini dapat dioptimalkan dengan formula *lagrangian* dan *quadratic programing*[28], Setelah didapatkan solusi dari optimalisasinya, kemudian kelas data uji x dapat ditetapkan berdasarkan nilai dari persamaan fungsi keputusan berikut:

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i y_i x_i x_d + b$$

x_i merupakan *support vector* dengan *ns* adalah jumlah dari *support vector* tersebut, dan x_d merupakan data yang nantinya diklasifikasikan.

Pada dasarnya, SVM adalah algoritma yang hanya dapat aplikasikan untuk mengklasifikasikan data yang bersifat linier, oleh karena itu, untuk dapat mengklasifikasikan data yang sifatnya non-linier perlu dilakukan penyesuaian dengan menambahkan fungsi kernel didalamnya. Kernel adalah suatu fungsi yang mendefinisikan transformasi pemetaan tiap data pada *input space* ke dalam ruang vektor yang baru dengan dimensi lebih tinggi, sehingga pada akhirnya kedua kelas dapat dipisahkan oleh garis *hyperplane*[29]. Dengan fungsi kernel ini, fungsi hasil dari algoritma SVM dapat dituliskan seperti berikut:

$$f(x_d) = \sum_{i=1,x_i \epsilon SV}^{ns} \alpha_i y_i K(x_i, x_d) + b$$

Dalam penelitian ini, Kernel Linier adalah fungsi kernel yang dipakai [29], Adapun rumusan dari kernel tersebut adalah:

$$K(x_i, x_j) = x_i x_j^T$$

Setelah Model klasifikasi SVM dibangun, proses selanjutnya digunakan proses pengujian data untuk mengidentifikasi data teks yang diujikan nantinya berupa teks berita berita hoax atau non-hoax. Selain itu, untuk menyamakan bobot dan struktur datanya, sebelum data uji dimuat ke dalam sistem dan diperiksa oleh Algoritma SVM, data tersebut haruslah diproses terlebih dahulu sesuai dengan tahapan yang telah disebutkan sebelumnya pada pra-proses data klasifikasi. Berikut adalah alur algoritma dari sistem yang dibuat:

Algoritma Identifikasi hoax dengan SVM

1. Input Teks

Pra-Proses

2. Lakukan Analisis Leksikal
3. Hapus Stopwords
4. Lakukan Stemming
5. Transformasi teks ke dalam numerik dengan TF-IDF

Proses Training Model SVM

6. Tentukan *input X* , target *Y*
7. Masukkan fungsi Kernel Linear
8. Hitung matriks Kernel *K* yaitu perkalian *dot product* dari vektor *input* sesuai dengan fungsi Multi Kernel yang telah ditentukan.
9. Temukan solusi optimal untuk nilai α.
10. Dapatkan data *support vector* yang memenuhi
11. Dapatkan bias yang diperoleh dari $b = y_i - wx_i$.

Proses Identifikasi dengan SVM

12. Input Teks Testing
13. Lakukan **Pra-Proses**
14. Uji dengan **Model SVM**
15. Hasil Hoax/non-Hoax

IV. HASIL DAN PEMBAHASAN

Pengukuran hasil daripada penelitian ini disajikan dengan mempresentasikan nilai Presisi, *Recall* dan *F-Measure* dari hasil uji coba sistem yang telah dibuat. Presisi adalah perhitungan rasio dari prediksi benar positif dibandingkan dengan prediksi positif dari keseluruhan luaran yang dihasilkan, Sebagai ilustrasi, Presisi menjawab pertanyaan "Berapa persentase berita yang benar-benar hoax dari total berita yang diprediksi hoax". Sedangkan *Recall* diartikan sebagai rasio prediksi benar positif dibandingkan dengan jumlah keseluruhan data benar positif, jika diilustrasikan *Recall* menjawab pertanyaan "Berapa persentase berita yang diprediksi hoax dibandingkan dengan semua berita yang sebenarnya hoax". Adapun *F-Measure* atau biasa juga disebut dengan *F1-Score* merupakan pengukuran *harmonic mean* antara nilai Presisi dan Recall[30]. Ketiga nilai tersebut dihitung dengan menggunakan *Confusion Matrix* yang merepresentasikan hasil dari prediksi dan kondisi aktual (jumlah sebenarnya) dari data yang diujikan oleh sistem seperti di bawah ini.

TABEL 1.
CONFUSION MATRIX

		Kelas Sebenarnya	
		Positif	**Negatif**
Kelas Prediksi	**Positif**	True Positive (TP)	False Negative (FN)
	Negatif	False Positive (FP)	True Negative (TN)

$$\text{Presisi} = \frac{(\text{True Positive})}{(\text{True Positive} + \text{False Positive})}$$

$$Recall = \frac{(\text{True Positive})}{(\text{True Positive} + \text{False Negative})}$$

$$F - Measure = 2\,\frac{\text{Presisi} \times Recall}{\text{Presisi} + Recall}$$

Hasil penelitian menunjukkan bahwa sistem memiliki tingkat kinerja yang baik untuk mengidentifikasi berita hoax atau bukan. Hal ini dibuktikan melalui hasil perhitungan skor rata-rata Presisi, Recall dan F-Measure secara berturut adalah 78,96%, 78,18% dan 78,02%. Dengan proses *k-fold cross validation* yang diterapkan dalam penelitian ini, sesuai dengan yang telah disebutkan sebelumnya, uji coba dilakukan secara berulang dengan saling silang kombinasi data training dan testing. Tabel 2 berikut ini mempresentasikan hasil perhitungan performa dari keseluruhan dari uji coba yang telah dilakukan:

TABEL 2.
HASIL UJI COBA

No	**Presisi**	**Recall**	**F-Measure**
1	59,40%	59,09%	58,75%
2	77,50%	77,27%	77,23%
3	89,37%	88,64%	88,58%
4	84,16%	84,09%	84,08%
5	84,38%	81,82%	81,47%
Avg	**78,96%**	**78,18%**	**78,02%**

Dari hasil uji coba yang telah ditunjukkan pada Tabel 2 dapat dilihat bahwa Algoritma SVM yang telah dibangun pada penelitian ini dapat diaplikasikan untuk mengidentifikasi

berita hoax maupun berita non hoax dengan baik, hal ini tentunya terbukti dari nilai Presisi, *Recall* dan *F-Measure* yang baik dari semua hasil eksperimen. Hasil percobaan tersebut juga menunjukkan bahwa sistem identifikasi ini memiliki kemampuan yang stabil dalam mengklasifikasikan berita yang diujikan, hal ini ditunjukkan pada Tabel 2 bahwa selama percobaan semua hasilnya berada di atas 78 %, hasil terbaik diperoleh pada percobaan ke-3. Dengan nilai Presisi 89,37%, *Recall* 88,64% dan *F-Measure* 88,58%. Sebagai perbandingan, hasil terburuk diperoleh saat uji coba pertama dengan nilai Presisi, *Recall* dan *F-Measure* berturut adalah 59,40%, 59,09%, dan 58,75%.

V. KESIMPULAN

Dalam penelitian ini dibangun sebuah sistem otomatis yang mampu untuk mengidentifikasi berita hoax atau tidak menggunakan algortima Support Vector Machine (SVM). Data berita teks yang digunakan dalam penelitian ini merupakan kumpulan berita hoax dan non-hoax yang sebelumnya telah diferifikasi kebenarannya oleh tim MAFINDO berdasarkan hasil diskusi dan pencarian fakta yang dilakukan oleh para anggota yang hasilnya kemudian dipublikasikan pada laman turnbackhoax.id. Data yang diambil dari situs turnbackhoax.id merupakan berita yang dipublikasikan dalam rentang waktu 1 Januari 2020 sampai dengan 29 Agustus 2020 dimana dari data tersebut kemudian dibagi menjadi dua kelompok dataset yaitu dataset untuk proses training sebesar 80% dan sisanya 20% untuk proses testing. Sedangkan untuk validasi hasil dari klasifikasi dari sistem digunakan metode *k-fold cross validation*.

Proses penelitian dimulai dengan pra-proses yang terdiri dari analisis leksikal, penghapusan stopword, stemming dan diakhiri dengan mentransformasikan teks menjadi bentuk numerik dengan metode *Term Frequency - Invers Document Frequency* (TF-IDF). Setelah data terolah dengan baik, maka proses selanjutnya adalah proses training yang dilakukan dengan membuat model klasifikasi menggunakan algoritma SVM. Hasil dari model tersebut kemudian digunakan untuk menguji atau mengidentifikasi teks berita tersebut apakah termasuk dalam kategori berita hoax atau non-hoax, yang disebut dengan proses testing. Hasil dari penelitian menunjukkan bahwa algoritma SVM dapat digunakan untuk mengidentifikasi berita hoax atau tidak dengan baik. Ditunjukan dari rata-rata nilai performanya yang mencapai 78,96% untuk nilai Presisi, 78,18% untuk *Recall* dan nilai *F-Measure* sebesar 78,02%.

UCAPAN TERIMA KASIH

Segenap tim peneliti menyampaikan ucapan terima kasih untuk Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) di Universitas PGRI Adi Buana Surabaya atas dukungan yang telah diberikan melalui program Hibah Adi Buana Tahun 2020 sehingga penelitian ini dapat direalisasikan dengan baik.

DAFTAR PUSTAKA

[1] P. K. S. Chan and M. C. W. Chan, "Tracing the SARS-coronavirus," *J. Thorac. Dis.*, vol. 5 Suppl 2, no. Suppl 2, pp. S118–S121, Aug. 2013, doi: 10.3978/j.issn.2072-1439.2013.06.19.

[2] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. Di Napoli, "Features, evaluation and treatment coronavirus (COVID-19)," in *StatPearls [Internet]*, StatPearls Publishing, 2020.

[3] WHO, "Coronavirus." https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed Jan. 26, 2021).

[4] Kemendikbud, "hoaks @ kbbi.kemdikbud.go.id." https://kbbi.kemdikbud.go.id/entri/hoaks (accessed Jan. 19, 2021).

[5] R. Pakpahan, "Analisis Fenomena Hoax Diberbagai Media Sosial Dan Cara Menanggulangi Hoax," *Konf. Nas. Ilmu Sos. dan Teknol.*, vol. 1, no. 1, 2017.

[6] H. Septanto, "Pengaruh hoax dan ujaran kebencian sebuah cyber crime dengan teknologi sederhana di kehidupan sosial masyarakat," *J. Kalbiscientia J. Sains dan Teknol.*, vol. 5, no. 2, pp. 157–162, 2018.

[7] N. P. S. Meinarni and I. B. A. I. Iswara, "Hoax and its Mechanism in Indonesia," 2018.

[8] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[9] S. M.Weiss, N. Indurkhya, and T. Zhang, *Texts in Computer Science - Fundamentals of Predictive Text Mining*. 2010.

[10] B. Richhariya, M. Tanveer, A. H. Rashid, and A. D. N. Initiative, "Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE)," *Biomed. Signal Process. Control*, vol. 59, p. 101903, 2020.

[11] A. B. Alhamda, "PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES CLASSIFIER PADA KLASIFIKASI TEKS CYBERBULLYING DI KOMENTAR PENGGUNA INSTAGRAM." Institut Telkom Purwokerto, 2019.

[12] G. A. Buntoro, "Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naive Bayes Classifier Dan Support Vector Machine," *J. Din. Inform.*, vol. 5, no. 2, 2016.

[13] A. N. Hidayat, "Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes," *J. Elektron. Sist. Inf. dan Komput.*, vol. 1, no. 1, pp. 12–18, 2015.

[14] M. Athoillah, "Pengenalan Wajah Menggunakan SVM Multi Kernel dengan Pembelajaran yang Bertambah," *J. Online Inform.*, vol. 2, no. 2, p. 84, 2018, doi: 10.15575/join.v2i2.109.

[15] C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432. 2013.

[16] Y. Li and H. Ye, "An Analysis and Research of Type-2 Diabetes TCM Records Based On Text Mining," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 1872–1875, doi: 10.1109/BIBM.2018.8621283.

[17] P. Beyranvand and T. Aytekin, "Automating Customer Claim Registration by Text Mining," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 1–5, doi: 10.1109/ASYU50717.2020.9259889.

[18] A. de Lima, A. Argenta, I. Zattar, and M. Kleina, "Applying Text Mining to Identify Photovoltaic Technologies," *IEEE Lat. Am. Trans.*, vol. 17, no. 05, pp. 727–733, 2019, doi: 10.1109/TLA.2019.8891940.

[19] N. Rajvanshi and K. R. Chowdhary, "Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA," *Int. J. Eng. Res.*, vol. 6, p. 9, 2017.

[20] Z. Wang and Z. Qu, "Research on Web text classification algorithm based on improved CNN and SVM," in *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, 2017, pp. 1958–1961, doi: 10.1109/ICCT.2017.8359971.

[21] Y. Chen and Z. Zhang, "Research on text sentiment analysis based on CNNs and SVM," in *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2018, pp. 2731–2734, doi: 10.1109/ICIEA.2018.8398173.

[22] S. E. Nugroho, "Upaya masyarakat anti fitnah indonesia mengembalikan jatidiri bangsa dengan gerakan anti hoax," *Pros. Konf. Nas. peneliti muda Psikol. Indones.*, vol. 2, no. 1, pp. 1–4, 2017.

[23] D. Suyanto, "Data Mining untuk klasifikasi dan klasterisasi data," *Bandung Inform. Bandung*, 2017.

[24] M. Javed and S. Kamal, "Normalization of unstructured and informal text in sentiment analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 10, pp. 78–85, 2018, doi: 10.14569/IJACSA.2018.091011.

[25] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[26] R. Melita, V. Amrizal, H. B. Suseno, and T. Dirjam, "Penerapan

Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018.

[27] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[28] B. P. Tomasouw and M. I. Irawan, "MULTICLASS TWIN BOUNDED SUPPORT VECTOR MACHINE UNTUK PENGENALAN SUARA," 2012.

[29] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[30] P. Mishra and P. Lotia, "Comparative performance analysis of SVM speaker verification system using confusion matrix," *Int. J. Sci. Res.(IJSR)*, vol. 3, p. 12, 2014.