Identifikasi Tensi dan Ujaran Kebencian di Twitter terkait Opini Politik di Indonesia

Della Fitrayani Budiono
Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sebelas Maret
dellafitrayani@student.uns.ac.id

Lia Ristiana
Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sebelas Maret
lia.ristiana@student.uns.ac.id

Maulia Harjono
Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sebelas Maret
maulia.harjono@student.uns.ac.id

***Abstract*— Although the rise of social media provides the idea of freedom of speech, it also opens a new problem of increases in tension and hate speech presented in differing political opinion. Several studies relating to the abuse of freedom of speech, i.e. racism and hate speech against certain individual and group using a number of machine learning techniques have been done on various languages in the world. In Indonesia, the President often becomes the target of critics and hate speech done by parties of opposite interest. This study investigates the possibility of identifying which opinion on Twitter, in the form of *tweet*s delivered by its users, may cause tension and lead to hate speech against the President. We approach our idea using text mining, probabilistic method, and machine learning method with a dataset containing 574 *tweet*s of various degrees of tension and hatred collected from Twitter. Results of this research indicate that Support Vector Machine – Recursive Feature Elimination (SVM-RFE), a machine learning method, outperforms the result of Multinomial Naive Bayes probabilistic approach. Visualizations of top words found on each class also show a deeper understanding of the data.**

Keywords*— *hate speech, machine learning, social media analysis, tension, text mining

I. PENDAHULUAN

Meningkatnya penggunaan media sosial juga turut meningkatkan penyalahgunaan media sosial di masyarakat, termasuk di Indonesia. Salah satu bentuk penyalahgunaan ini adalah penyebaran ujaran kebencian atau yang sering disebut dengan *hate speech* terhadap individu atau kelompok tertentu. Di Indonesia beberapa kasus ujaran kebencian yang terjadi di Twitter dan Facebook tidak hanya dianggap meresahkan oleh masyarakat, namun juga telah ditanggapi serius oleh pihak kepolisian karena dianggap melanggar Undang-Undang yang berlaku, yaitu UU ITE [1][2]. Salah satu contoh kasus yang sempat hangat di Indonesia adalah terungkapnya sindikat *Saracennews* pada Agustus 2017 yang diduga telah mengunggah konten provokatif dalam bentuk kata-kata, narasi, dan *meme* di media sosial untuk mengarahkan pandangan negatif terhadap kelompok masyarakat lain [3]. Data tahun 2017 dari Polri sendiri mencatat bahwa 80% kejahatan siber yang terjadi di Indonesia merupakan kasus ujaran kebencian yang meningkat tajam terutama menjelang pilkada 2017 [4].

Media sosial seperti Twitter, Instagram, dan Facebook umumnya menekankan adanya kebebasan berekspresi dalam bentuk teks, gambar atau *meme*, dan video. Salah satu ekspresi yang sering disampaikan melalui media sosial adalah ekspresi opini politik. Contoh terbaru mengenai hal ini adalah Donald Trump dan Hillary Clinton pada Pemilihan Umum Presiden Amerika Serikat 2016. Kasus ini menjadi perbincangan hangat di Twitter dan menarik sejumlah peneliti untuk mengidentifikasi rasisme yang terkandung pada *tweet* politis mengenai kedua kandidat [5].

Twitter merupakan salah satu media sosial paling popular dan berpengaruh di internet saat ini. Di Indonesia, hingga pertengahan tahun 2015, data terakhir menunjukkan terdapat sekitar 50 juta pengguna, di mana 77% di antaranya aktif menggunakan Twitter setiap hari [6]. Penyampaian informasi di Twitter diekspresikan dalam bentuk teks dengan limit sebesar 140 karakter yang disebut dengan *tweet*. Sayangnya, kebebasan menggunakan internet dan berpendapat bagi setiap individu dapat menjadikan media sosial seperti Twitter mudah disalahgunakan. Twitter menyebutkan sejumlah aturan yang wajib diikuti penggunanya pada artikel *terms of service* di situsnya, salah satunya yaitu tidak melakukan *hateful conduct* atau mempromosikan *hateful speech*, yaitu suatu tindakan kekerasan yang bersifat mengancam atau menyerang orang lain dalam basis ras, etnis, nasionalitas, dan sebagainya [7]. Selain masalah *hate speech*, masalah yang juga muncul beriringan dengan kasus ujaran kebencian atau *hate speech* adalah deteksi timbulnya tensi atau yang sering disebut *tension detection*. Tensi sendiri didefinisikan sebagai "kejadian apa saja yang cenderung menunjukan bahwa hubungan antarindividu atau antarkelompok telah memburuk dan cenderung akan meningkat ke kelompok dari individu-individu yang bersangkutan" [8].

Di Indonesia sendiri, menilik kasus yang terjadi di lapangan, Presiden merupakan salah satu target ujaran kebencian yang sering ditemui di sosial media [2]. Melihat besarnya pengaruh Twitter terhadap opini politik publik ini, penulis tertarik untuk mengidentifikasi tensi dan ujaran kebencian terhadap Presiden Republik Indonesia dari opini politis yang ditemui di Twitter. Presiden Indonesia Joko Widodo sebagai objek penelitian selanjutnya akan dirujuk sebagai objek pada makalah ini.

Bab II makalah ini membahas penelitian terkait. Bab III menjelaskan metode pengumpulan data. Bab IV menjelaskan metode penelitian. Bab V memaparkan hasil dan analisis penelitian. Bab VI memuat kesimpulan.

II. PENELITIAN TERKAIT

Sejumlah penelitian mengenai *tension detection* dan *hate speech* di media sosial telah dilakukan pada sejumlah bahasa di dunia. Di antaranya yaitu tension detection terhadap dua publik figur di persepakbolaan Inggris terlibat kasus rasisme oleh Burnap et al (2014) yang menimbulkan sejumlah komentar kontroversial di Twitter yang menimbulkan tensi sosial. Penelitian mengklasifikasikan *tweet* yang diperoleh dari kejadian ini menjadi empat kelas, yaitu 0 yang berarti *irrelevant*, 1 berarti *some tension*, 2 berarti *tension*, dan 3 berarti *high tension*. Penelitian ini dilakukan menggunakan sejumlah metode salah satunya yaitu Naive Bayes. Namun, penelitian tidak menitikberatkan pada visualisasi hasil [9].

Penelitian lainnya telah dilakukan oleh Warner & Hirschberg (2012) yang meneliti hate speech di media sosial melalui *tweet* mengenai anti-semitisme dengan mengklasifikasikan *tweet* secara biner. Hasilnya, klasifikasi menggunakan text mining dan SVM menyatakan bahwa text mining dengan unigram memberikan hasil paling maksimal dibandingkan bigram atau trigram dengan akurasi sebesar 94% [10]. Penelitian lainnya yaitu identifikasi rasisme pada *tweet* opini politik yang bersangkutan dengan Pemilu 2016 Amerika Serikat oleh Lozano et al (2017) [5].

Oleh karena itu, fokus serta tujuan dari paper ini adalah mengidentifikasi tweet berbahasa Indonesia yang mengandung suatu tensi dan ujaran kebencian dengan menggunakan metode machine learning Multinomial Naive Bayes, Support Vector Machine, dan Support Vector Machine – Recursive Feature Elimination sebagai metode klasifikasinya.

III. DATASET TWITTER

Pembuatan dataset dilakukan dengan mengumpulkan data dari Twitter berupa *tweet* yang mengandung kata kunci yang berhubungan dengan objek. Pengumpulan dataset ini dilakukan pada rentang waktu tertentu, yang dimulai dari tanggal 1 Agustus hingga 19 Agustus 2017. Dari proses ini diperoleh sebanyak 574 *tweets*. Himpunan *tweet* yang diperoleh kemudian dilabeli berdasarkan tingkat kekasarannya.

Dengan mengikuti definisi dari *hate speech* dan *tension* yang telah disebutkan sebelumnya, penulis memberikan tiga (3) macam kelas label terhadap *tweet* yang diperoleh. Label 0 (tidak relevan) meliputi *tweet* mengenai objek yang tidak termasuk ke dalam kategori ujaran kebencian dan *tweet* yang tidak memiliki keterkaitan dengan politik atau Pilpres 2019 mendatang. Label 1 (tensi rendah) adalah *tweet* opini yang bersifat negatif, baik yang berasal dari individu maupun opini pasif seperti teks berita di mana konten dari teks yang bersangkutan memiliki kecenderungan menimbulkan tensi. Label 2 (tensi tinggi/ujaran kebencian) adalah *tweet* yang mengandung opini negatif individu terhadap objek yang disertai dengan makian atau kata-kata kasar terhadap objek.

TABLE I. CONTOH PENGKLASFIKASIAN *TWEET*

No	Isi *Tweet*	Kelas
1	Patung Lilin Jokowi di Hongkong Kini Kenakan Kemeja Batik.	0
2	@liputan6dotcom Hidup pak @jokowi maju Indonesia ku	0
3	9266Mungkin Jokowi buat negara jadi seperti ini karena ingin Balas dendam pada negara ini dan umat islam,dia kan diduga keturunan PKI boyolali?	1
4	Zaman rezim Jokowi kok makin hari makin memperlihatkan sikapnya yg kontra sama Islam???.	1
5	@jokowi Sy sdh gak ada respect sama situ bung... smg pemilu depan situ bener2 sudah gak ada lagi muncul di arena politik or berita2 media.amiin 3×	1
6	@anti_densus88 @DivHumasPolri kontol tunduk sama @jokowi pki planga plongo yang akan tumbang	2
7	@Freyra_ @aldhiraaa @jokowi Hehe....pahala apa yg diambil dari anjing terkutuk penghina islam bgtu, mending doakan biar dikutuk jd monyet	2

Tabel 1 menggambarkan proses pengkatogerian *tweet* menjadi kelas 0, 1, dan 2. *Tweet* pertama dikategorikan sebagai kelas 0 karena merupakan artikel berita netral yang dianggap tidak relevan terhadap topik penelitian. *Tweet* kedua merupakan tanggapan bersentimen positif, sehingga masuk ke kelas 0. Sedangkan *tweet* ketiga dan keempat merupakan *tweet* bersentimen negatif yang mengandung tuduhan bahwa objek termasuk pada golongan masyarakat tertentu, sehingga berpotensi menimbulkan tensi. *Tweet* kelima adalah opini negatif terhadap objek. Sedangkan *tweet* keenam dan ketujuh selain berpotensi menimbulkan tensi karena isi opini yang cenderung negatif, namun juga mengandung kata-kata makian, sehingga berpotensi dianggap sebagai ujaran kebencian.

Dari tahap ini diperoleh 183 data berlabel 0, 284 data berlabel 1, dan 107 data berlabel 2, yang berjumlah total sebanyak 574.

IV. METODE PENELITIAN

Penelitian ini dilakukan dengan menggunakan teknik *text mining* dan data mining dengan dataset yang diperoleh dari Twitter sebagaimana telah dijelaskan pada bab sebelumnya. Tahapan penelitian terdiri dari *preprocessing* dengan teknik *text mining*, klasifikasi dengan teknik data mining, dan terakhir visualisasi.

A. Preprocessing

Dataset dimasukkan ke dalam tahap *preprocessing* dengan teknik *text mining* yaitu *natural language processing*. Tahapan ini dijelaskan sebagai berikut.

1) Twitter attributes removal: yaitu membersihan teks dari atribut yang umumnya ditemui di Twitter namun dianggap tidak memberikan kontribusi terhadap klasifikasi, seperti hashtag (#), mention (@).

2) Punctuation removal: yaitu penghilangan tanda baca seperti titik (.), koma (,), tanda tanya (?), tanda seru (!), dan sebagainya.

3) Numbers removal: yaitu menghapus angka dari teks.

4) Hyperlink removal: yaitu menghapus tautan dari teks.

5) Special characters removal: yaitu menghapus karakter yang tidak termasuk dalam ASCII.

6) Extra white spaces removal: yaitu menghapus spasi ekstra yang ada pada teks.

7) Normalisasi: yaitu mengubah kata-kata tidak baku, dalam kasus ini yaitu pada Bahasa Indonesia, menjadi bentuk standarnya.

8) Transforming regular expressions: yaitu melakukan transformasi kata-kata tidak baku, kata-kata asing, dan sebagainya yang mengandung karakter berulang ke bentuk formalnya.

9) Stopwords removal: yaitu penghapusan kata-kata pada teks yang terlalu sering muncul namun tidak memiliki makna atau kontribusi khusus terhadap makna teks secara umum. Dalam Bahasa Indonesia, contohnya meliputi: yang, ke, dari, dan sebagainya.

10) Stemming: yaitu pengembalian kata-kata dalam Bahasa Indonesia yang ada pada teks ke bentuk kata dasarnya.

B. TF-IDF (Term Frequency – Inverse Document Frequency)

Term Frequency - Inverse Document Frequency (TF-IDF) adalah salah satu metode yang umum dipakai di pengolahan *Natural Language Processing*. Metode ini bekerja dengan cara menentukan frekuensi relatif suatu kata pada suatu dokumen tertentu dengan membandingkan inverse dari rasio kata tersebut terhadap keseluruhan dataset atau korpus. Hasilnya akan menunjukkan seberapa relevan sebuah kata terhadap dokumen tertentu. Kata-kata yang sering muncul di satu atau sekelompok kecil dokumen akan memiliki nilai TF-IDF yang tinggi dibandingkan kata-kata yang sering muncul di keseluruhan korpus [11].

TF-IDF dapat direpresentasikan dalam persamaan matematika secara formal sebagai berikut. Jika D adalah keseluruhan dataset, w adalah sebuah kata, dan sebuah dokumen direpresentasikan dengan d di mana d ε D, maka persamaan dari TF-IDF dapat digambarkan sebagai berikut.

$$w_d = f_{w,\,d} * \log\left(|D|/f_{w,\,D}\right) \qquad [11]$$

Dari perhitungan menggunakan persamaan di atas akan diperoleh nilai TF-IDF dari semua kata yang terkandung pada korpus. Nilai TF-IDF yang diperoleh disimpan dalam bentuk vektor matriks. Dataset terdiri dari 574 *tweet*s dari 3 kelas yang secara total mengandung 1913 fitur.

C. Klasifikasi Multinomial Naive Bayes (MNB)

Metode *baseline* yang umum digunakan adalah Naive Bayes. Naïve Bayes mengasumsikan bahwa keberadaan fitur tertentu dalam sebuah kelas tidak terkait dengan keberadaan fitur lainnya. [12] Salah satu pengembangan dari Naïve Bayes adalah Multinomial Naïve Bayes. MNB diterapkan dengan menambahkan nilai satu yang disebut sebagai smoothing untuk menghindari probabilitas nol dari *term* baru pada *testing set* yang tidak ada dalam *training set*.[13]

D. Klasifikasi Support Vector Machine (SVM)

Support Vector Machine adalah salah satu metode *machine learning* yang bekerja dengan mekanisme *over-fitting internal* yang kuat sehingga dapat bekerja lebih baik pada data dengan dimensi yang tinggi [14]. Salah satu bentuk pengembangan SVM adalah SVM-RFE, yaitu metode klasifikasi dengan menggunakan SVM yang dikombinasikan dengan metode feature selection Recursive Feature Elimination. SVM-RFE mengurutkan fitur berdasarkan eliminasi *backward data* hingga akurasi klasifikasi tertinggi didapatkan [15].

E. Visualisasi

Untuk mempermudah dalam melihat kondisi fitur (kata) pada dataset, akan ditampilkan visualisasi yang terdiri dari 2 macam. Visualisasi pertama adalah dengan grafik yang menampilkan 15 fitur (kata) yang berkontribusi bagi tiap kelas. Sedangkan visualisasi kedua adalah dengan menampilkan wordcloud, yang berisi kata-kata yang paling sering muncul (baik itu unigram, bigram, dan lain-lain) pada tiap kelas. Visualisasi ini juga tentunya akan sangat membantu untuk mengidentifikasi error pada proses klasifikasi, karena dengan adanya gambaran persebaran fitur (kata) pada tiap kelas bisa mempermudah dalam mengidentifikasi penyebab adanya data yang salah diklasifikasikan.

V. HASIL DAN ANALISIS HASIL

A. Hasil Klasifikasi

Uji coba dataset menggunakan metode Multinomial Naive Bayes menghasilkan akurasi sebesar 65.8%. *Confusion Matrix* dari klasifikasi tersebut bisa dilihat di Tabel II.

TABLE II. CONFUSION MATRIX DARI MULTINOMIAL NAIVE BAYES

Actual class	**Predicted class**		
	0	1	2
0	91	92	0
1	11	273	0
2	3	90	14

Selanjutnya, ujicoba dataset menggunakan metode Support Vector Machine (dengan tuning: *kernel = rbf*, C = 12.0, *gamma* = 1) menghasilkan akurasi sebesar 78.3%. *Confusion Matrix* dari klasifikasi tersebut bisa dilihat di Tabel III.

TABLE III. CONFUSION MATRIX DARI SUPPORT VECTOR MACHINE

Actual class	Predicted class		
	0	1	2
0	128	53	2
1	33	251	0
2	4	29	74

Berdasarkan hasil ujicoba di atas dapat dilakukan analisis sebagai berikut:

1. Pada klasifikasi data kelas 0, MNB cenderung tidak bisa membedakan kelas 0 dan kelas 1. Sedangkan klasifikasi SVM cenderung memberikan hasil yang lebih baik pada kelas 0.
2. Pada klasifikasi data kelas 1, MNB dan SVM cenderung memberikan hasil yang sama baiknya. Keduanya mampu mengklasifkasikan sebagian besar data kelas 1 secara benar.
3. Pada klasifikasi data kelas 2, MNB melakukan banyak kesalahan dengan mengklasifikasikan data ke kelas 1, namun hal serupa tidak terjadi pada SVM.

Dari ketiga analisis tersebut, bisa disimpulkan bahwa secara umum, SVM dianggap memberikan hasil klasifikasi yang lebih baik daripada MNB pada kasus ini. Hal ini mungkin disebabkan karena kelas 1 lebih banyak direpresentasikan dalam dataset, sehingga Naive Bayes yang sifatnya probabilistik cenderung mengklasifikasikan *instances* dari kelas 0 dan kelas 2 ke kelas yang dominan.

Selain melakukan uji coba dataset menggunakan MNB dan SVM, juga dilakukan uji coba menggunakan *Support Vector Machine - Recursive Feature Elimination* (SVM-RFE) dengan *parameter tuning*: C = 5.0 yang menghasilkan akurasi sebesar 78.74% dengan jumlah fitur optimal (terbaik) yang digunakan adalah sebesar 298 fitur. Grafik hubungan antara akurasi (10 Fold CV) dan jumlah fitur yang digunakan bisa dilihat pada Gambar.

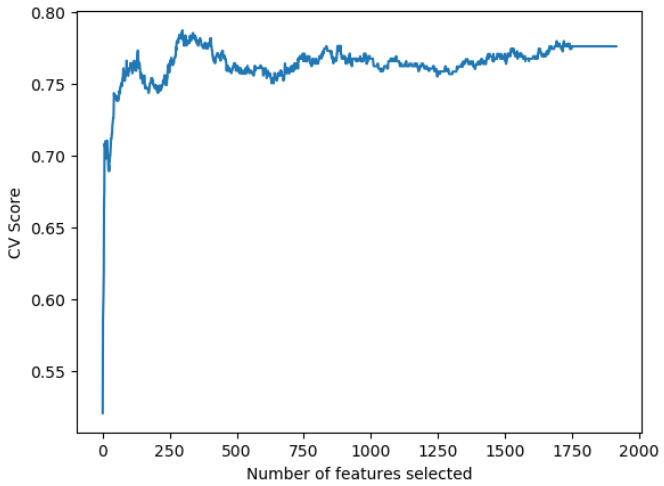

Fig. 1. Grafik hubungan antara akurasi (CV Score) dan jumlah fitur yang digunakan

Confusion Matrix dari klasifikasi menggunakan SVM-RFE bisa dilihat pada Tabel.

TABLE IV. CONFUSION MATRIX DARI SUPPORT VECTOR MACHINE – RECURSIVE FEATURE ELIMINATION

Actual class	Predicted class		
	0	1	2
0	120	62	1
1	32	249	3
2	3	21	83

Dengan menggunakan metode klasifikasi SVM dan *feature selection* RFE, diperoleh hasil yang sedikit lebih baik dibandingkan dengan SVM. Sekilas, akurasi yang diberikan tidak berbeda jauh. Namun dengan SVM-RFE yang hanya menggunakan 298 fitur (dari 1913 original fitur), *classifier* ternyata mampu memberikan akurasi yang sama baiknya Sehingga dapat disimpulkan bahwa meskipun SVM dan SVM-RFE memberikan akurasi yang hampir sama, SVM-RFE dianggap lebih efisien karena berhasil memangkas 85% fitur yang ada dalam proses klasifikasi.

B. Analisis Error

Kesalahan klasifikasi dapat diakibatkan oleh berbagai penyebab, yaitu dari segi dataset (*tweet*) dan fitur.

Dari segi dataset (*tweet*), error klasifikasi bisa disebabkan oleh:

- Adanya *tweet* yang bersifat lelucon, contoh: *tweet* 'Makan indomie pake nasi adalah hak setiap warga negara. Pak Jokowi harus menjamin ini.' harusnya diklasifikasi sebagai kelas 0, namun diprediksi sebagai kelas 1. Hal ini karena *classifier* belum bisa membedakan lelucon dengan teks biasa.
- Adanya *tweet* yang tergolong sarkasme, contoh: *tweet* 'Kerja menko jaman jokowi berat sih, ngurusin kerjaan presiden sementara presidennya jalan bagi2 sepeda' harusnya diklasifikasi sebagai kelas1, namun diprediksi sebagai kelas 0. Selain teks yang berkonotasi sarkastis, masalah lainnya adalah adanya fitur 'presiden', dan 'jokowi' yang cenderung lebih banyak muncul di kelas 0 (penjelasan bisa dilihat di bagian visualisasi).
- Konteks dari *tweet* yang tidak menyinggung (menghina) objek, contoh: *tweet* 'Megawati: Orang sebut Jokowi Diktator Harus Bisa Buktikan – Kompas TV' harusnya diklasifikasi sebagai kelas 0, namun diprediksi sebagai kelas 1. Hal ini mungkin disebabkan oleh mesin yang kesulitan mengambil konteks dari *tweet*, karena konsep dari TF-IDF hanya mempertimbangkan kemunculan fitur tanpa mempertimbangkan makna. Sehingga ketika dihadapkan pada *tweet* dengan konteks menyinggung pihak lain (yang menghina objek), *classifier* kesulitan melakukan prediksi dengan benar. Untuk kasus *tweet* tersebut, adanya kata 'Jokowi' dan 'diktator' memungkinkan *tweet* tersebut diprediksi ke dalam kelas

1, meskipun sebenarnya *tweet* tersebut tidak sedang menyinggung (menghina) objek.

Dari segi fitur, error klasifikasi bisa disebabkan oleh:

- Adanya fitur yang sangat memberikan kontribusi besar terhadap suatu kelas, contoh: *tweet* 'Jokowi: Kalau Palestina tertekan, statement kita keras' harusnya diklasifikasi sebagai kelas 0, namun diprediksi sebagai kelas 1. Hal ini mungkin disebabkan adanya kata yang berkesan negatif seperti kata 'tertekan', dan 'keras'. Padahal pada teks ini, kalimat yang berkesan negatif berasal dari objek yang berkomentar tentang suatu topik, dan bukan kalimat yang memberitakan objek secara negatif.
- Adanya fitur yang sangat jarang muncul, sehingga tidak memberikan kontribusi yang besar dalam proses klasifikasi, contoh; *tweet* '@anti_densus88 @DivHumasPolri kontol tunduk sama @jokowi pki planga plongo yang akan tumbang' harusnya diklasifikasi sebagai kelas 2, namun diprediksi sebagai kelas 1. Hal ini mungkin disebabkan oleh adanya kata makian yang jarang muncul di dataset seperti 'kontol', 'plonga', dan 'plongo' meskipun sebenarnya kata-kata tersebut sangat mengarah ke salah satu kelas (yaitu kelas 2).

C. *Visualisasi*

Visualisasi akan dibagi menjadi 2 macam, yaitu grafik 15 fitur (kata) teratas yang berkontribusi bagi tiap kelas, dan wordcloud dari tiap kelas.

Visualisasi bagian pertama adalah dengan grafik yang menampilkan 15 fitur (kata) teratas yang berkontribusi bagi tiap kelas. Tingkat pentingnya suatu data (berkontribusi bagi suatu kelas), diukur dengan menggunakan *weight* yang diperoleh pada SVM dengan kernel linear. Seperti yang telah dikatakan sebelumnya, konsep dari SVM adalah membuat hyperplane yang menggunakan *support vectors* untuk memaksimalkan jarak antara 2 kelas. *Weight* yang diperoleh merepresentasikan koordinat vektor yang orthogonal terhadap *hyperlane*, dan arahnya menunjukkan kelas yang diprediksi. Nilai absolut dari weights itulah yang digunakan untuk menentukan pentingnya suatu fitur dalam memisahkan data (antara kedua kelas).

15 fitur (kata) teratas yang berkontribusi terhadap kelas 0, adalah sebagai berikut.

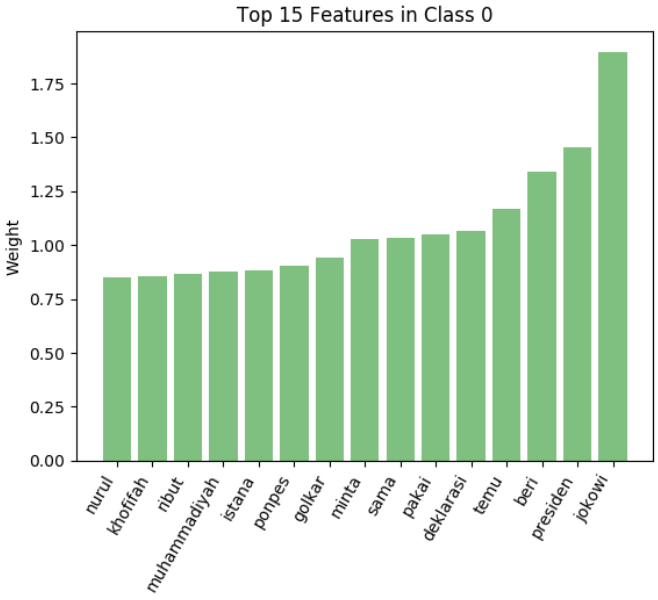

Fig. 2. Grafik dari 15 fitur (kata) teratas yang berkontribusi terhadap kelas 0

Sedangkan 15 fitur (kata) teratas yang berkontribusi terhadap kelas 1, adalah sebagai berikut.

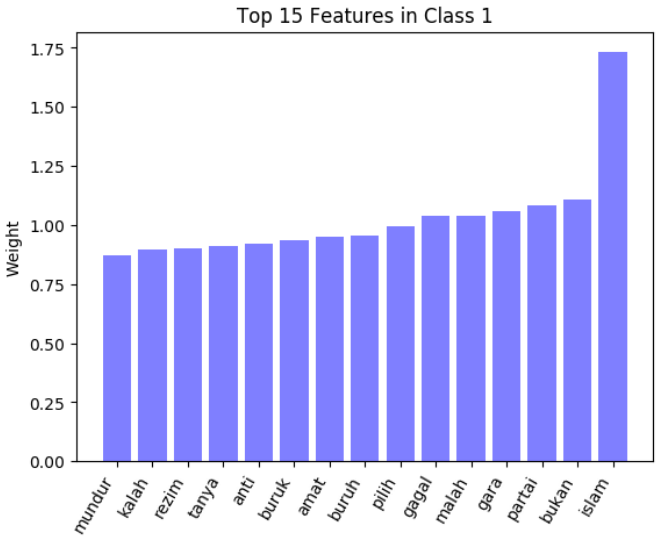

Fig. 3. Grafik dari 15 fitur (kata) teratas yang berkontribusi terhadap kelas 1

Sedangkan 15 fitur (kata) teratas yang berkontribusi terhadap kelas 2, adalah sebagai berikut.

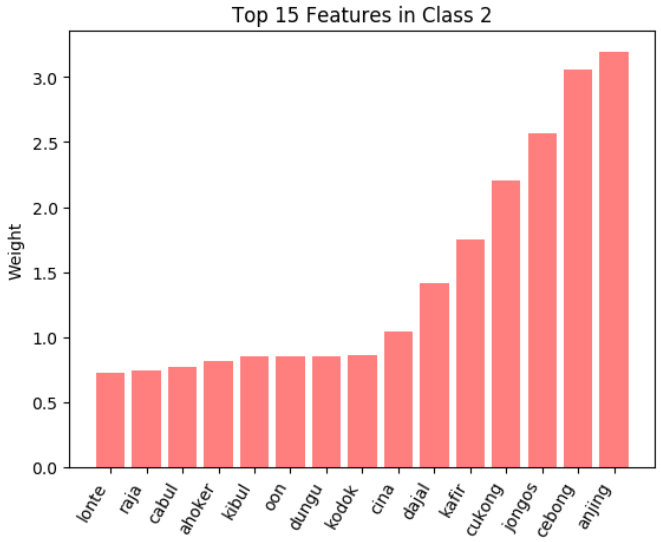

Fig. 4. Grafik dari 15 fitur (kata) teratas yang berkontribusi terhadap kelas 2

Dari visualisasi, bisa dilihat bahwa istilah ‘presiden’, ‘Jokowi’, dan ‘beri’ umumnya muncul di *tweet* kelas 0 seperti digambarkan pada Gambar 2. Istilah yang sering muncul pada kelas ini cenderung netral atau bersifat positif. Sementara ‘islam’, ‘partai’, ‘anti’, dan ‘rezim’ cenderung muncul di kelas 1 sebagaimana dapat dilihat pada Gambar 3. Analisis yang mungkin dari munculnya kata-kata ini pada kasus ujaran kebencian adalah orang-orang cenderung mengaitkan objek sebagai rezim yang anti terhadap golongan tertentu. Isi *tweet* yang bersangkutan cenderung provokatif dan bersifat menuduh, sehingga berpotensi menimbulkan tensi dan ujaran kebencian. Sedangkan pada kelas 2 seperti dilihat pada Gambar 4, istilah yang sering muncul adalah kata-kata hinaan seperti ‘anjing’, ‘dajjal’, dan ‘kafir’, selain itu muncul pula istilah ‘ahoker’. Dari kasus ini dapat dianalisis bahwa orang-orang yang mengutarakan ujaran kebencian terhadap objek dengan menggunakan istilah-istilah hinaan juga cenderung melakukan labelisasi ekstrem dengan istilah keagamaan. Selain itu, munculnya istilah ‘ahoker’ pada kelas ini menunjukkan bahwa orang-orang yang bersangkutan, selain berpersepsi negatif terhadap objek, juga cenderung berpersepsi negatif terhadap tokoh politik teretntu lainnya.

Sedangkan visualisasi kedua adalah dengan *wordcloud*, yang menampilkan kata-kata yang paling sering muncul pada setiap kelas. Visualisasi ini dilakukan dengan tidak melibatkan kata ‘Jokowi’ yang muncul di setiap kelas, karena ‘Jokowi’ sendiri merupakan search query pada proses pengambilan data (di Twitter), sehingga pastinya kata tersebut akan muncul berkali-kali di tiap kelas, yang tentunya tidak informatif. Kata-kata yang dimaksud bisa bersifat, unigram, bigram, dan seterusnya.

Wordcloud dari kelas 0 adalah sebagai berikut.

Fig. 5. Wordcloud dari kelas 0

Wordcloud dari kelas 1 adalah sebagai berikut.

Fig. 6. Wordcloud dari kelas 1

Wordcloud dari kelas 2 adalah sebagai berikut.

Fig. 7. Wordcloud dari kelas 2

Dari ketiga wordcloud tersebut, bisa dilihat bahwa kata-kata yang sering muncul di kelas 0,1, maupun 2 berkorespondensi dengan grafik 15 fitur yang berkontribusi bagi tiap kelas. Hal tersebut menunjukkan bahwa kinerja dari TF-IDF dan weighting pada Support Vector Machine sudah cukup baik dalam mengambil informasi fitur yang penting bagi tiap kelas.

VI. KESIMPULAN

Timbulnya tensi akibat perbedaan opini politik yang juga mengakibatkan munculnya ujaran kebencian di media sosial merupakan salah satu kasus yang tengah marak di Indonesia.

Identifikasi tensi dan ujaran kebencian yang ditujukan pada individu atau kelompok tertentu diharapkan dapat menjadi langkah awal pencegahan penyalahgunaan media sosial. Dengan menggunakan *text mining* dan metode probabilistik dan *machine learning*, eksperimen dilakukan pada 574 *tweet*s untuk mengidentifikasi tensi dan ujaran kebencian terhadap Presiden Republik Indonesia. Eksperimen yang dilakukan menunjukkan bahwa metode Support Vector Machine (SVM) dengan feature selection Recursive Feature Elimination (RFE) memberikan hasil yang terbaik (jika dibandingkan dengan Multinomial Naive Bayes (MNB) dan SVM biasa tanpa feature selection) dengan akurasi 78.7%. Selain itu, hasil visualisasi dengan menggunakan grafik 15 fitur (kata) teratas yang berkontribusi bagi tiap kelas dan *wordcloud*, juga bisa memberikan gambaran dataset lebih mendalam.

REFERENCES

[1] Polisi Tangkap Pria di Koja Terkait Hate Speech. Detikcom. https://news.detik.com/berita/d-3568740/polisi-tangkap-pria-di-koja-terkait-hate-speech (diakses terakhir 9 September 2017).

[2] Buat Meme Menghina Presiden di Facebook, Seorang Pemuda Ditangkap. Kompas. http://regional.kompas.com/read/2017/06/09/16265591/buat.meme.menghina.presiden.di.facebook.seorang.pemuda.ditangkap (diakses terakhir 9 September 2017).

[3] Kasus Saracen: Pesan kebencian dan hoax di media sosial 'memang terorganisir'. BBC. http://www.bbc.com/indonesia/trensosial-41022914 (diakses terakhir 9 September 2017).

[4] Polri: 80 Persen Kejahatan Siber Didominasi Ujaran Kebencian. Detikcom. https://news.detik.com/berita/d-3517151/polri-80-persen-kejahatan-siber-didominasi-ujaran-kebencian (diakses terakhir 9 September 2017).

[5] Lozano, E., Cedeño, J., Castillo, G., Layedra, F., Lasso, H., & Vaca, C. (2017, April). Requiem for online harassers: Identifying racism from political *tweet*s. In eDemocracy & eGovernment (ICEDEG), 2017 Fourth International Conference on (pp. 154-160). IEEE.

[6] Twitter Rahasiakan Jumlah Pengguna di Indonesia. CNN Indonesia. https://www.cnnindonesia.com/teknologi/20160322085045-185-118939/twitter-rahasiakan-jumlah-pengguna-di-indonesia/ (diakses terakhir 9 September 2017).

[7] The Twitter Rules, Twitter Support. https://support.twitter.com/articles/18311 (diakses terakhir 9 September 2017).

[8] Dyfed Powys Police, Community Tension Force Policy Document, 2008.

[9] Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., & Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. Technological Forecasting and Social Change, 95, 96-108.

[10] Warner, W., & Hirschberg, J. (2012, June). Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media (pp. 19-26). Association for Computational Linguistics.

[11] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).

[12] Zia, T., Akram, M. S., Nawaz, M. S., Shahzad, B., Abdullatif, A. M., Mustafa, R., Lali, M. I. (2016, November). Identification Of Hatred Speeches On Twitter. In Proceedings of 52nd The IRES International Conference, Kuala Lumpur, Malaysia.

[13] Goel, A., Gautam, J., & Kumar, S. (2016, October). Real time sentiment analysis of tweets using Naive Bayes. In Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on (pp. 257-261). IEEE.

[14] Ismail, H., Harous, S., & Belkhouche, B. (2016). A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis. *Research in Computing Science*, 110, 71-83.

[15] Bahatti, L., Bouattane, O., Echhibat, M. E., & Zaggaf, M. H. (2016). An Efficient Audio Classification Approach Based on Support Vector Machines. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 7(5), 205-211.

[16] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.