Deteksi Gempa Berdasarkan Data Twitter Menggunakan Decision Tree, Random Forest, dan SVM

Rendra Dwi Lingga P., Chastine Fatichah, dan Diana Purwitasari
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: chastine@cs.its.ac.id

***Abstrak*—Twitter merupakan salah satu media sosial yang cukup populer saat ini. Pengguna aktif Twitter mencapai kurang lebih 400 juta orang. Fitur utama yang paling penting dari Twitter yaitu layanan yang bersifat real-time dimana pengguna dapat menuliskan catatan singkat tentang apa yang terjadi secara langsung. Sebagai contoh, ketika terjadi bencana alam(gempa bumi) di suatu tempat, banyak pengguna aktif twitter menulis informasi berupa (tweet) tentang gempa bumi yang sedang berlangsung melalui Twitter. Hal ini memungkinkan dibuatnya sebuah metode yang mendeteksi terjadinya gempa atau tidak dengan melakukan observasi melalui tweet yang ada.**

Dalam tugas akhir ini dibuat sebuah metode klasifikasi untuk membedakan antara tweet yang mengandung informasi gempa yang sesungguhnya (gempa positif) dan tweet yang mengandung informasi gempa namun memiliki arti lain (gempa negatif).

Setelah dilakukan klasifikasi menggunakan Decision Tree, Random Forest dan Support Vector Machine (SVM). Hasil yang didapat memberikan nilai akurasi Support Vector Machine (SVM) secara keseluruhan lebih baik daripada Decision Tree dan Random Forest dengan persentase gempa yang dideteksi oleh sistem (Recall) didapatkan nilai 86.3%.dengan precision sebesar 88.7%. Namun jika dilihat dari terdeteksinya gempa oleh sistem tanpa dirata-rata, Random Forest memiliki persentase recall sebesar 96.7% lebih baik daripada Decision Tree dan Random Forest.

***Kata Kunci*—Twitter, Deteksi Kejadian, Gempa.**

I. PENDAHULUAN

TWITTER merupakan salah satu media sosial yang cukup terkenal. Pada tahun 2016, jumlah pengguna aktif Twitter mencapai 400 juta orang. Fitur utama dari Twitter yaitu tweet dimana pengguna dapat menuliskan catatan singkat tentang apa yang terjadi di lingkungan sekitarnya secara langsung. Twitter bersifat real-time sehingga membuat Twitter menjadi media sosial yang menarik digunakan untuk berbagai metode event detection, salah satunya bencana alam gempa bumi [1].

Deteksi tentang adanya gempa adalah hal yang cukup krusial, bahkan bisa menyelamatkan nyawa. Bila seseorang mengetahui akan ada gempa beberapa detik saja sebelum gempa terasa, maka waktu itu bisa digunakan untuk hal penting seperti mematikan saluran gas.

Data tweet yang didapatkan dari Twitter, bisa menjadi dasar untuk mendeteksi terjadinya gempa bumi. Akibat data tweet yang banyak mengandung kata tidak baku, maka perlu dibuat sebuah model klasifikasi untuk menentukan apakah tweet benar-benar memberi informasi tentang adanya gempa saat ini.

Dari permasalahan yang telah diipaparkan di atas, dalam tugas akhir ini akan diimplementasikan sebuah metode klasifikasi deteksi gempa di wilayah Indonesia berdasarkan informasi yang terdapat pada Twitter. Dengan adanya metode ini diharapkan dapat memberikan informasi dini yang akurat terkait ada atau tidaknya gempa.

II. TINJAUAN PUSTAKA

A. Twitter

Twitter adalah sosial media yang memberikan layanan pada penggunanya untuk berbagi hal-hal yang sedang terjadi saat itu juga. Twitter menanyakan satu pertanyaan, “What’s happening? “(Apa yang sedang terjadi?), kepada penggunanya. Jawaban yang diberikan oleh pengguna tidak boleh lebih dari 140 karakter [1]. Salah satu aspek penting dari Twitter adalah karakteristiknya yang bersifat real-time. Sebagai contoh, saat terjadi gempa, banyak pengguna mengirim jawaban ke Twitter (tweets) yang berhubungan dengan gempa. Hal ini memungkinkan dilakukan deteksi terhadap gempa dengan menginspeksi berbagai tweet yang masuk ke Twitter.

B. Gempa

Gempa bumi adalah getaran di permukaan bumi yang diakibatkan oleh pelepasan energi dari dalam secara tiba-tiba sehingga menciptakan gelombang seismik. Gempa bumi di Indonesia biasa diakibatkan oleh pertemuan antara plat tektonik [2].

Kekuatan dari gempa ditentukan dari amplitude getaran gempa yang diukur dari seismograph, serta jarak seismograph tersebut dari pusat gempa. Kekuatan dari gempa biasa diukur dalam skala Magnitude, yang memperlihatkan seberapa besar energi yang dikeluarkan oleh gempa.

Indonesia merupakan salah satu negara yang cukup sering terkena gempa bumi, hal ini diakibatkan oleh letak geografis

dari Indonesia. Gempa yang memakan banyak korban di Indonesia terjadi pada tahun 2006 di pulau Jawa serta pada tahun 2004 di Aceh yang juga disertai oleh gelombang tsunami [3].

C. *Support Vector Machine*

SVM atau *Support Vector Machine* adalah sebuah metode pembelajaran *supervised* yang menganalisis data dan mengidentifikasi pola. SVM biasa digunakan dalam klasifikasi data menjadi kelas tertentu [4]. SVM bekerja dengan cara memetakan data *training* yang telah memiliki label menjadi titik-titik di ruang. Titik-titik tersebut dipetakan sedemikian rupa sehingga jarak dari titik terdekat antara masing-masing kelas *traning data* menjadi sejauh mungkin.

SVM pada dasarnya adalah sebuah model klasifikasi linear, namun bisa digunakan untuk melakukan klasifikasi non-linear dengan *kernel trick*. *Kernel trick* dalam SVM adalah cara untuk melakukan klasifikasi data non-linear dengan cara mentransformasikan ruang data asli menjadi ruang data dengan dimensi yang lebih tinggi. *Kernel* yang biasa digunakan dalam SVM antara lain *Gaussian Radial Basis Function* (RBF), Polynomial, dan Hyperbolic.

D. *Random Forest*

Random Forest adalah sebuah metode bisa yang digunakan untuk klasifikasi, regresi, ataupun tujuan lainnya. Random Forest bekerja dengan cara membangun lebih dari 1 Decision Tree secara random saat training. Hasil yang diberikan oleh Random Forest untuk klasifikasi adalah modus dari decision tree-decision tree nya. Sementara nilai yang diberikan untuk regresi adalah mean [4].

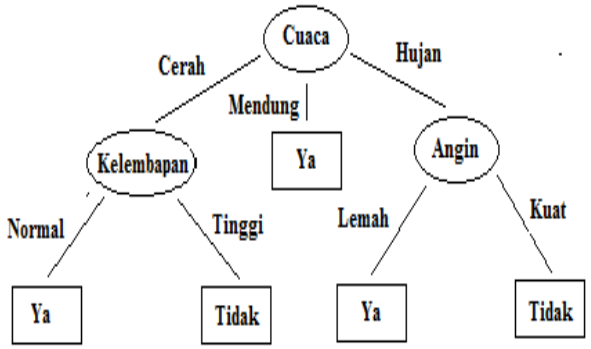
Dengan membuat banyak Decision Tree secara random, maka sebenarnya banyak dari pohon-pohon yang dibuat oleh metode Random Forest menjadi kurang berguna. Namun Random Forest mampu menjadi sebuah metode klasifikasi yang cukup baik, karena beberapa Decision Tree yang ikut dibuat saat konstruksi, ternyata memiliki kemampuan prediksi yang baik. Saat dilakukan pemilihan untuk menentukan klasifikasi secara keseluruhan, pohon-pohon yang buruk akan membuat prediksi yang acak dan saling bertentangan, sehingga jawaban dari beberapa decision tree yang merupakan prediktor yang baik akan muncul sebagai jawaban.

Random Forest pertama kali dipublikasikan secara resmi oleh Leo Breiman pada tahun 2001 [4]. Random Forest dikembangkan untuk memperbaiki metode-metode Decision Tree yang rawan overfitting. Dalam perkembangannya Random Forest menjadi salah satu metode yang populer di bidang machine learning. Hal ini diakibatkan oleh mudahnya penggunaan Random Forest, yang mampu mencapai akurasi tinggi tanpa perlu melakukan banyak parameter tuning [5].

E. *Decision Tree*

Decision Tree adalah sebuh metode klasifikasi yang dibangun untuk mendapatkan sebuah kesimpulan dari sejumlah data. Penarikan kesimpulan dibuat dalam bentuk pohon, dimana nantinya hasil kesimpulan berbentuk hierarki pohon yaitu dari akar, batang dan daun yang merepresentasikan hasil keputusan yang dibuat.

Sebuah *node* keputusan (misalnya, Cuaca) memiliki dua cabang atau lebih (misalnya, cerah, mendung dan hujan). *Nod*e daun (misalnya, *Play*) merupakan klasifikasi atau keputusan. *Node* keputusan paling atas di pohon adalah yang sesuai dengan prediktor terbaik disebut *node* akar [6].



Gambar 1. Ilustrasi Decision Tree

Penerapan *Decision Tree* dapat dilihat pada Gambar 1. Pada gambar tersebut dalam dilihat bawah *Decision Tree* memiliki 2 macam *node* yaitu *node* keputusan (Cuaca) dan *node* daun (*Play*=Ya atau *Play*=Tidak).

F. *Cross Validation*

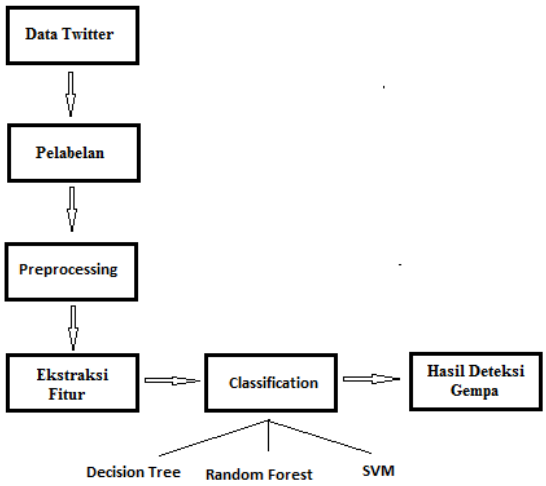
Cross-validation merupakan salah satu teknik untuk menilai keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Pembuatan model biasanya bertujuan untuk melakukan prediksi maaupun klasifikasi terhadap suatu data baru yang boleh jadi belum pernah muncul di dalam dataset. Data yang digunakan dalam proses pemgbangunan model disebut data training., sedangkan data yang digunakan untuk menilai model disebut sebagai data test.

Salah satu metode cross-validation yang populer dalah K-Fold Cross-validation. Dalam teknik ini dataset dibagi menjadi sejumlah K-buah partisi secara acak. Kemudian dilakukakn sejumalh K-kali eksperimen, dimana masing-masing eksperimen menggunakan data partisi ke-K sebagai data testing dang memanfaatkan sisa partisi lainnya sebagai data training.

III. ANALISIS DAN PERANCANGAN

A. *Deskripsi Umum Sistem*

Rancangan tugas akhir deteksi gempa berdasarkan data Twitter dapat dilihat pada Gambar 2.



Gambar 2. Arsistektur Umum Sistem

1) Pelabelan

Proses dimulai saat sejumlah data yang sudah didapat dari Twitter diberi label class gempa atau tidak gempa yang dilakukan secara manual.

2) Preprocessing

Data yang sudah diberi label akan dilakukan *preprocessing*. Pertama dilakukan tokeninasi yaitu proses pemisahan kata dari data tweet menjadi token-token/bagian-bagian tertentu. Tokenisasi didapatkan berdasarkan spasi, enter, tab, titik, koma, dsb seperti pada Tabel 1 :

Tabel 1.
Tokenisasi

No	Dokumen	Tokenisasi
1	Barusan saja kota tarakan dilanda gempa	Barusan, saja, kota, tarakan, dilanda, gempa
2	Barusan sekitar jam 5 Wita ada gempa kah di bali	Barusan, sekitar, jam 5, wita, ada, gempa, kah, di, bali
3	barusan terasa gempa di cibubur, pusat gempa terjadi dimana ya	Barusan, terasa, gempa, di, cibubur, pusat, gempa, terjadi, dimana, ya
4	Barusan terasa gempa di wilayah banda aceh dan sekitar.	Barusan, terasa, gempa, di, wilayah, banda, aceh, dan, sekitar

Data yang telah memalui tahapan tokenisasi selanjutnya akan diproses untuk dilakukan penyaringan kata yang muncul dalam jumlah besar/umum atau kata yang tidak baku dan tidak memiliki makna(*stopword*) seperti pada Tabel 2 :

Tabel 2.
Stopword Removal

No	Dokumen	Stopword Removal
1	**Barusan**, **saja**, kota, tarakan, dilanda, gempa	kota, tarakan, dilanda, gempa
2	**Barusan**, **sekitar**, jam 5, **wita**, **ada**, gempa, **kah**, **di**, bali	jam 5, gempa, bali
3	**Barusan**, **terasa**, gempa, **di**, cibubur, pusat, gempa, terjadi, dimana, **ya**	gempa, cibubur, pusat, gempa, terjadi
4	**Barusan**, **terasa**, gempa, **di**, wilayah, banda, aceh, **dan**, sekitar	Gempa, wilayah, banda, aceh, sekitar

Selanjutnya data akan diproses untuk menghilangkan angka seperti pada Tabel 3 :

Tabel 3.
Hilangkan Angka

No	Dokumen	Hilangkan Angka
1	kota, tarakan, dilanda, gempa	kota, tarakan, dilanda, gempa
2	**jam 5**, gempa, bali	Jam, gempa, bali
3	gempa, cibubur, pusat, gempa, terjadi	gempa, cibubur, pusat, gempa, terjadi
4	Gempa, wilayah, banda, aceh, sekitar	Gempa, wilayah, banda, aceh, sekitar

Selanjutnya data akan diproses dengan cara normalisasi data agar menghilangkan kata yang berulang seperti pada Tabel 4 :

Tabel 1.
Normalisasi

No	Dokumen	Normalisasi
1	kota, tarakan, dilanda, gempa	kota, tarakan, dilanda, gempa
2	gempa, bali	gempa, bali
3	**gempa**, cibubur, pusat, **gempa**, terjadi	gempa, cibubur, pusat, terjadi
4	Gempa, wilayah, banda, aceh, sekitar	Gempa, wilayah, banda, aceh, sekitar

Dalam proses preprocessing juga akan dilakukan menghilangkan karakter-karakter seperti ('"+=!&?*^~#-_) seperti pada Tabel 5 :

Tabel 2.
Hilangkan Karakter

No	Dokumen	Hilangkan Karakter
1	Barusan gempa**?**	Barusan gempa
2	Barusan gempa. okebye-_-	Barusan gempa. Okebye
3	Barusan gempa**...**	Barusan gempa
4	barusan jogja gempa kan**!**	barusan jogja gempa kan

B. Ekstraksi Fitur

Data akan dilakukan pembobotan menggunakan *Term Frequency (TF). Term Frequency (TF)* merupakan *frequency* kemunculan *term* atau kata dalam dokumen. Contoh seperti pada Tabel 6 :

Dokumen 1 : Saya belajar menghitung nilai tf.

Dokumen 2 : Tf merupakan frekuensi kemunculan term pada dokumen.

Tabel 3.
Perhitungan Term Frequency

Term	Dokumen 1	Dokumen 2
Saya	1	0
frekuensi	0	1
Belajar	1	0
Hitung	1	0
Nilai	1	0
Tf	1	1
Pada	0	1
Term	0	1
Dokumen	0	1

Setelah proses selesai kemudian dilakukan eksport matriks dalam format .xls. Matriks yang dieksport memiliki ukuran n x m, dimana n merupakan jumlah dari data dan m merupakan jumlah dari term.

C. Perbandingan Klasifikasi Decision Tree, Random Forest, dan SVM

Pada tahap ini, dilakukan klasifikasi untuk mendapatkan *tweet* yang dianggap sebagai sinyal positif. Dalam kasus ini *tweet* yang dianggap sinyal positif adalah *tweet* yang dimaksudkan penggunanya untuk memberikan informasi

tentang gempa yang baru dialaminya. *Tweet* yang mengandung kata gempa dalam arti lain, dianggap sebagai sinyal negatif.

Data yang digunakan adalah data tweet yang mengandung kata gempa dari 20 September 2014 sampai 25 September 2014, 20 Desember 2014 sampai 30 Desember 2014, serta 1 sampai 15 Januari 2015.

Proses klasifikasi menggunakan aplikasi WEKA dengan metode *Decision Tree, Random Forest,* dan SVM. *cross-validation* yang digunakan adalah *cross-validation* 3, 5 dan 10. Setelah didapatkan hasil dari *cross-validation* selanjutnya akan dibandingkan hasil dari *TP rate, FP rate, precision, recall dan F-Measure* dari ketiga metode.

IV. IMPLEMENTASI

Dalam Klasifikasi ini dibangun dengan menggunakan Bahasa pemrograman php, aplikasi yang digunakan Xampp, Notepad++, Excel, dan Weka. Xampp digunakan sebagai program yang menjalankan modul-modul dalam bahasa php. Notepad++ dan Excel digunakan untuk pengeditan file data yang digunakan dan file data yang sudah diolah. Sedangkan Weka digunakan untuk menjalankan metode klasifikasi Decision Tree, Random Forest, dan SVM dan nantinya luaran dari hasil klasifikasi akan dibandingkan metode klasifikasi mana yang lebih baik.

Dalam membangun aplikasi perangkat keras yang digunakan adalah komputer laptop Samsung RV418 dengan spesifikasi prosesor intel Core i3 dan memori 4GB.

V. PENGUJIAN DAN EVALUASI

Perangkat yang digunakan adalah laptop dengan prosesor Intel® Core™ i3-2310M dengan kecepatan 2.10GHz dan memori 4 gigabyte RAM. Uji coba dilakukan di sistem operasi Microsoft Windows 8.1 64bit dengan kakas bantu XAMPP dan WEKA.

Uji coba dan evaluasi diimplementasikan menggunakan metode Decision Tree, Random Forest, dan SVM untuk klasifikasi tweet gempa. Pengujian akan dilakukan dengan melihat nilai dari TP Rate, FP Rate, Precision, Recall, dan F-Measure kemudian hasil yang keluar akan dibandingkan metode mana yang lebih baik dalam mendeteksi gempa dilihat dari nilai Recall.

Tabel 7.
Nilai *Recall* dari Sistem dalam Mendeteksi Gempa

3-fold cross validation	Gempa		
	DT	RF	SVM
Recall	46.7%	**96.7%**	70.0%

Dari Tabel 7 persentase Recall Random Forest dalam mendeteksi class gempa lebih baik daripada Decision Tree dan SVM dengan nilai persentase tertinggi menggunakan 3-fold cross-validation sebesar 96.7%.

Tabel 8.
Perbandingan *weighted avg*

3-fold Cross validation	Decision Tree	Random Forest	SVM
TP Rate	77.5%	51.3%	87.5%
FP Rate	34.8%	30.6%	19.5%
Precision	79.7%	73.9%	88.6%
Recall	77.5%	51.3%	**87.5%**
F-Measure	75.5%	46.2%	87.0%

Dari Tabel 8 dapat terlihat bahwa secara keseluruhan nilai *Recall* metode SVM dalam mendeteksi gempa menggunakan 3-fold *cross-validation* lebih baik daripada *Decision Tree* dan *Random Forest* dengan persentase 87.5%.

VI. KESIMPULAN

Dari hasil selama proses perancangan, implementasi, serta pengujian dapat diambil kesimpulan sebagai berikut:

1. Dalam melakukan klasifikasi tweet yang mengandung informasi gempa, metode Random Forest memiliki akurasi Recall sebesar 96.7%, lebih baik bila dibandingkan dengan Decision Tree dan SVM
2. Secara rata-rata dalam melakukan klasifikasi tweet yang mengandung informasi gempa atau tidak, metode SVM memiliki akurasi Recall sebesar 87.5%, lebih baik bila dibandingkan dengan Decision Tree dan Random forest yang masing-masing nilai rata-ratanya 77.5% dan 51.3%.

DAFTAR PUSTAKA

[1] Twitter, "What is Twitter | About," 2014. [Online]. Available: https://about.twitter.com/what-is-twitter.

[2] Geoscience Australia, "What is an Earthquake?," 2013. [Online]. Available: http://www.ga.gov.au/scientific-topics/hazards/earthquake/basics/what.

[3] Albaqir. Haidar M, "Aplikasi Deteksi Lokasi Gempa dan Peringatan Dini Berdasarkan Data Twitter menggunakan Random Forest dan Partikel Filter." 2015.

[4] A. S. Nugroho, "Support Vector Machine, Teori dan Aplikasinya dalam Bioinformatika," IlmuKomputer.com, 2003.

[5] L. Breiman, "Random Forests," *Machine Learning*, 2001.

[6] S.Sayad, "Decision Tree Classification,". [Online]. Available: http://www.saedsayad.com/decision_tree.htm

[7] T. Cheese, "Random Forest | Kaggle," 5 11 2014. [Online]. Available: https://www.kaggle.com/wiki/RandomForests.