Pembangunan Model untuk Aplikasi Detektor Kampanye

Gisela Supardi[1], Andika Kusuma[2]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]13515009@std.stei.itb.ac.id, [2]13515033@std.stei.itb.ac.id

***Abstract*—Twitter merupakan salah satu media sosial yang populer di kalangan masyarakat Indonesia. Mendekati Pemilu Presiden tahun 2019, Twitter sering digunakan menjadi sarana kampanye masing-masing kubu. Pembangunan aplikasi pendeteksi kampanye dilakukan untuk mengetahui mencari tahu kepopuleran Twitter sebagai media kampanye pendukung masing-masing kubu dan sarana awal pendeteksian kemungkinan pelanggaran peraturan larangan kampanye pada masa tenang. Selain memisahkan cuitan kampanye dan nonkampanye, aplikasi berbasis Natural Language Processing ini juga dapat menentukan kubu yang didukung oleh cuitan yang telah terdeteksi sebagai kampanye. Untuk melakukan deteksi pada cuitan, perlu dibangun sebuah model klasifikasi. Eksperimen ini bertujuan untuk mencari algoritme pembangunan model yang paling tepat. Model klasifikasi final menggunakan gabungan dari tiga buah model klasifikasi yang didapat dengan algoritme *decision tree*, *support vector machine*, dan *multi layer perceptron*. Hasil eksperimen menunjukkan nilai akurasi yang baik terhadap data uji, namun masih belum dapat mengklasifikasi data di luar korpus dengan baik.**

***Keywords*—Algoritme paling tepat, deteksi kampanye, Natural Language Processing, Pemilu Presiden 2019, Twitter**

I. PENDAHULUAN

Twitter merupakan salah satu media sosial yang sering dipergunakan masyarakat Indonesia. Mendekati penghujung tahun 2018, cuitan masyarakat Indonesia pada aplikasi Twitter kian banyak diwarnai dengan isu-isu seputar Pemilu Presiden. Banyaknya cuitan terkait topik ini tidak mengherankan mengingat pelaksanaan Pemilu semakin dekat, yakni pada tahun 2019 mendatang. Masa-masa ini merupakan waktu yang tepat bagi kedua pasangan calon (paslon) presiden dan wakil presiden untuk menarik hati masyarakat. Para pendukung masing-masing kubu juga terlihat gencar membela paslon jagoannya.

Melihat tren ini, kami memutuskan untuk membuat suatu aplikasi yang dapat mendeteksi cuitan dengan konotasi kampanye. Tujuan pembangunan aplikasi ini adalah untuk mencari tahu kepopuleran Twitter sebagai media kampanye pendukung masing-masing kubu. Di sisi lain, aplikasi ini juga dapat dimanfaatkan untuk mendeteksi kemungkinan adanya pelanggaran salah satu peraturan Pemilu, yaitu larangan kampanye pada masa tenang. Sebagai informasi, masa tenang Pemilu di Indonesia ialah sejak tiga hari sebelum hari pemungutan suara.

II. KAKAS YANG DIGUNAKAN

Dalam proses pembuatan aplikasi deteksi kampanye, terdapat beberapa kakas yang digunakan antara lain.

1. Sastrawi
   Sastrawi merupakan kakas pemprosesan bahasa alami khususnya Bahasa Indonesia. Sastrawi menyediakan fitur untuk melakukan *stemming* kata. *Stemming* adalah proses untuk menghilangkan imbuhan pada sebuah kata, seperti pe-, me-, -an, dst. *Stemming* bertujuan untuk menyamakan fitur pada kata berimbuhan maupun tidak berimbuhan karena keduanya mengandung makna yang mirip.
2. NLTK
   NLTK merupakan kakas pemprosesan bahasa Indonesia. NLTK menyediakan corpus kata yang termasuk ke dalam *stopwords* bahasa Indonesia. *Stopwords* adalah kata-kata yang sering muncul dan kurang memberikan makna dalam sebuah kalimat, seperti ‘saya’, ‘dan’, ‘atau’. *Stopwords* bertujuan untuk menghilangkan fitur-fitur yang sering muncul padahal kurang membawa informasi penting.
3. Scikit-Learn
   Scikit-Learn merupakan kakas pembelajaran mesin. Scikit-Learn menyediakan model-model klasifikasi yang dapat digunakan dalam pembelajaran mesin. Model yang digunakan pada aplikasi pendeteksi kampanye ini adalah model *Decision Tree* dan *Support Vector Machine*.
4. Tweepy
   Tweepy merupakan kakas untuk melakukan pemanenan dan *streaming* data Twitter. Tweepy menyediakan fitur untuk melakukan pengambilan cuitan Twitter berdasarkan suatu filter *query*, tempat, waktu, dan lain-lain. Tweepy digunakan untuk pengambilan data cuitan.
5. Numpy
   Numpy merupakan kakas untuk melakukan pengolahan data larik. Numpy digunakan untuk memudahkan pengolahan larik pada pengambilan fitur.

6. Django
   Django merupakan sebuah *framework* untuk membuat aplikasi *web*. Django menyediakan banyak fitur untuk melakukan pembuatan API Server dan juga *frontend* secara mudah. Django digunakan untuk pembuatan *backend* API dan juga *frontend* tampilan aplikasi pendeteksi kampanye.
7. Python
   Python merupakan bahasa pemrograman yang digunakan untuk *framework* Django. Selain itu, Python juga digunakan untuk kode eksperimen, pemanenan data.

III. Data Yang Digunakan

Pada eksperimen penentuan model terbaik, data yang digunakan diambil dari cuitan Twitter menggunakan kakas Tweepy. Cuitan dipanen menggunakan filter berupa geocode agar dapat memperoleh cuitan yang berasal dari Indonesia, terkhusus Pulau Jawa. Adapun koordinat yang digunakan adalah koordinat untuk Gunung Tidar dengan radius 600 kilometer. Gunung Tidar digunakan sebagai titik tengah karena beberapa sumber mengatakan bahwa gunung tersebut merupakan titik tengah Pulau Jawa. Cuitan juga dipanen menggunakan filter untuk menghilangkan cuitan dengan media gambar maupun video.

Total data berukuran 2411 cuitan dengan rincian 10% digunakan untuk data tes dan sisanya digunakan sebagai data latih. Data yang dipilih sebagai korpus adalah data yang cuitan yang mayoritas merupakan Bahasa Indonesia. Cuitan dengan Bahasa Inggris atau bahasa daerah tidak dimasukkan sebagai korpus. Oleh karena itu, pada proses deteksi juga diasumsikan bahwa masukan berupa Bahasa Indonesia. Selain cuitan dengan bahasa yang berbeda, cuitan yang hanya berupa mention atau link juga dihapus dari korpus.

IV. Eksperimen

Pada pembuatan aplikasi pendeteksi kampanye, dilakukan dua buah eksperimen. Eksperimen pertama dilakukan untuk menentukan model terbaik dalam modul normalisasi penggunaan angka di dalam kata. Eksperimen ini bertujuan untuk membuat model terbaik dalam penentuan hasil normalisasi kata dengan campuran angka di dalamnya. Sebagai contoh, normalisasi kata ‘t4’ akan menghasilkan kata ‘tempat dan kata ‘h3m4t’ akan menjadi kata ‘hemat’. Model yang digunakan pada modul normalisasi ini adalah model *Decision Tree*.

Eksperimen kedua dilakukan untuk menentukan model terbaik dalam klasifikasi kubu kampanye. Hasil klasifikasi berupa label 0 (non-kampanye), 1 (kubu 01), dan 2 (kubu 02). Eksperimen pencarian fitur terbaik model klasifikasi dilakukan dengan menggunakan skema validasi 10-*fold*. Model klasifikasi yang dibangun adalah model *Decision Tree*, *Multilayer Perceptron*, dan *Support Vector Machine*. Strategi pencarian parameter terbaik yang digunakan adalah *Grid Search*. Model klasifikasi akhir adalah hasil *voting* dari ketiga model yang dibangun.

V. Hasil

Hasil eksperimen model normalisasi angka menghasilkan model *Decision Tree* untuk melakukan klasifikasi jenis pelafalan angka ketika dicampur dengan kata lain. Hasil eksperimen model klasifikasi kampanye adalah tiga buah model klasifikasi yaitu model *Decision Tree*, *Multilayer Perceptron*, dan *Support Vector Machine*. Model *Decision Tree* memberikan akurasi 88.8428% pada data latih dan 89.6694% pada data tes. Model *Support Vector Machine* memberikan akurasi 91.2863% pada data latih dan 92.1488% pada data tes. Model *Multilayer Perceptron* memberikan akurasi 93,2688% pada data latih dan 91.7355% pada data tes. Model klasifikasi kampanya yang digunakan merupakan campuran ketiga model klasifikasi tersebut dengan bobot masing-masing model 0.33.

Untuk model klasifikasi, hasil eksperimen memang menunjukkan nilai akurasi yang cukup baik pada data latih maupun data tes untuk ketiga model. Namun, model klasifikasi yang dihasilkan masih belum dapat mengenali cuitan di luar data tes secara sempurna. Hal ini kemungkinan besar disebabkan oleh kurang meratanya pengambilan cuitan pada saat pemanenan data. Pemanenan data sendiri dilakukan dengan menggunakan kakas Tweepy yang melakukan pemanenan pada waktu yang berdekatan. Selain itu, jumlah korpus latih juga masih sangat sedikit sehingga belum dapat mencakup semua jenis cuitan yang mungkin. Ditambah lagi, tingginya kreativitas pengguna Twitter dalam menghasilkan cuitan kampanye secara tidak langsung menurunkan kinerja aplikasi ini.

VI. Simpulan

Hasil eksperimen model normalisasi angka sudah dapat berfungsi dengan baik dan cepat. Di sisi lain, hasil eksperimen model klasifikasi kampanye menunjukkan nilai akurasi yang baik terhadap data uji, namun masih belum dapat mengklasifikasi data di luar korpus dengan baik. Faktor-faktor penyebabnya antara lain ialah ukuran korpus yang kurang besar, persebaran data yang kurang merata, serta bahasa satire yang dilontarkan pengguna Twitter dalam membahas topik Pemilu.

VII. Saran Pengembangan

Aplikasi pendeteksi kampanye ini masih memiliki beberapa kekurangan. Untuk pengembangan aplikasi lebih lanjut, penambahan ukuran korpus data dapat dilakukan untuk meningkatkan kinerja aplikasi. Selain itu, pengambilan data yang lebih tersebar juga dapat memberikan hasil yang lebih baik.

VIII. Ucapan Terima Kasih

Pertama-tama kami ingin mengucapkan terima kasih kepada Tuhan Yang Maha Esa yang hanya karena berkat penyertaannyalah kami dapat menyelesaikan eksperimen

ini. Kemudian tidak lupa kami mengucapkan terima kasih kepada dosen-dosen pengampu mata kuliah IF4072 Pemrosesan Teks dan Suara Bahasa Alami, yaitu Ibu Dessi Puji Lestari ST,M.Eng.,Ph.D. dan terkhusus Ibu Dr. Eng. Ayu Purwarianti ST.,MT. Harlili S., M.Sc., selaku dosen pengampu materi teks yang telah membimbing kami dalam belajar sehingga memungkinkan kami untuk melaksanakan eksperimen ini. Selanjutnya ucapan terima kasih juga ingin kami sampaikan kepada orang tua kami masing-masing yang telah mendukung perkuliahan kami selama ini. Terakhir, saya juga ingin mengucapkan terima kasih kepada teman-teman Teknik Informatika 2015 terkhusus kepada Ida Ayu Putu Ari C, Rachel Sidney D., Erick Wijaya, dan Christopher Clement Andreas yang telah memberi dukungan baik secara moral maupun sosial sehingga eksperimen ini dapat kami selesaikan tepat pada waktunya.

IX. REFERENSI

[1] F. Hidayatullah dan M. R. Maarif. "Pre-processing Tasks in Indonesian Twitter Messages". *International Conference on Computing and Applied Informatics*. IOP: 2016.