



UNIVERSITAS INDONESIA

**Deteksi Twitter Buzzer pada Proses Kampanye
Pemilihan Presiden Republik Indonesia Tahun 2019**

Laporan Akhir

Kelompok Kepala DSA:

Alif Ahsanil Satria - 1606882540

Fari Qodri Andana - 1606875964

Febrian Fikar Farras - 1606875850

M. At Thoriq - 1606918484

Valerysa Regita - 1606875900

**FAKULTAS ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
DEPOK
2019**

DAFTAR ISI

DAFTAR ISI	2
BAB 1	3
PENDAHULUAN	3
1.1 Deskripsi Masalah	3
1.2 Nilai Kebaruan	3
1.3 Manfaat	4
1.4 Batasan	4
BAB 2	5
METODOLOGI	5
2.1 Koleksi Data	5
2.2 Fitur Awal	6
2.3 Seleksi Fitur	7
2.4 Algoritma	8
BAB 3	9
ANALISIS	9
3.1 Eksperimen	9
3.2 Analisis Hasil Eksperimen	10
3.3 Keluaran Program	10
BAB 4	11
PENUTUP	11
DAFTAR PUSTAKA	12

BAB 1

PENDAHULUAN

1.1 Deskripsi Masalah

Pemilihan presiden (pilpres) Republik Indonesia (RI) tahun 2019 merupakan topik yang hangat diperbincangkan pada saat ini, dan bahkan sudah mulai ramai sejak 2018. Peran media sosial, seperti Twitter dan Facebook, saat ini telah menjadi alat penting untuk kegiatan politik mengingat kemampuan penyebaran informasinya yang cepat. Salah satu contoh penggunaan media sosial dalam kontestasi politik yaitu pemilihan presiden AS tahun 2008 yang mana semua kandidatnya menggunakan sosial media seperti Facebook untuk mengambil sejumlah pendukung. Menurut Cornfield, kesuksesan besar Barack Obama dikaitkan dengan strategi "politik *online*" yang dicapai oleh tim kampanyenya [1]. Pada pilpres RI tahun 2019, kedua pasangan calon juga menggunakan media sosial sebagai media untuk menjual keunggulan masing-masing calon. Kampanye yang dilakukan di media sosial ini tidak lepas dari peran pihak pendukung yang sengaja meramaikan untuk menyebarkan lebih luas, yang kemudian kami katakan selanjutnya sebagai *buzzer*, substansi kampanye tersebut¹. Pihak pendukung yang memang sengaja membantu menyebarkan lebih luas konten kampanye salah satu calon pendukung dapat menjadi data yang mengganggu atau *noise* apabila ingin melakukan analisis sentimen yang berkaitan dengan calon presiden.

1.2 Nilai Kebaruan

Proyek ini bertujuan untuk mendapatkan pandangan baru terkait fitur apa saja yang dapat digunakan untuk mendeteksi *buzzer* pada sosial media Twitter. Penelitian sebelumnya telah menggunakan pendeteksian *buzzer* sebagai salah satu proses pembersihan data untuk menganalisis sentimen dua pasangan calon presiden RI tahun 2014 [2]. Namun, proses yang dilakukan belum dilaporkan secara rinci terkait pemilihan fitur. Oleh karena itu, proyek ini memfokuskan kepada eksplorasi fitur yang dapat membantu penelitian selanjutnya untuk mendeteksi akun *buzzer* Twitter.

¹ <http://www.rmol.co/read/2014/04/23/152340/Pasukan-Nasi-Bungkus-ItuPunya-Jokowi-atau-Prabowo>

1.3 Manfaat

Memberikan referensi *framework* deteksi akun *buzzer* untuk melakukan analisis sentimen menggunakan data Twitter. Keluaran dari proyek ini adalah sebuah model *machine learning* yang bisa menunjukkan akun twitter *buzzer*. Jika dikembangkan keluaran dari proyek ini bisa diterapkan ke berbagai bidang contoh nya pendeteksian akun spam pada kesempatan/ situasi politik.

1.4 Batasan

Ruang lingkup dari proyek ini adalah menunjukkan model yang memberikan nilai akurasi paling baik dari tiga algoritma yang kami gunakan yaitu SVM RBF Kernel, SVM Linear Kernel, dan Linear Regression untuk mengklasifikasi akun *buzzer*. Definisi *buzzer* yang kami gunakan adalah akun yang secara berturut-turut mengunggah tweet yang berkaitan dengan pasangan calon nomor 01 dan pasangan calon nomor 02. Batas antara *buzzer* atau *non-buzzer* adalah aktivitas keseharian. Apabila terdapat akun yang memiliki pembahasan di luar topik yang terkait dengan pasangan calon, akun tersebut kami definisikan sebagai *non-buzzer*.

BAB 2

METODOLOGI

2.1 Koleksi Data

Kami mengumpulkan sekitar 2500 *tweet* politik selama masa kampanye berlangsung yaitu sejak 29 September 2018 sampai 13 April 2019 menggunakan Twitter Search API . Untuk mengumpulkan *tweet* itu, kami menggunakan empat kata kunci yang merepresentasikan dua pasangan calon dan satu kata kunci yang merepresentasikan netral. Pengambilan kata kunci didasari dari hashtag trending pada Twitter yang merepresentasikan tiga pihak tersebut. Dari banyaknya kemungkinan kata kunci, kami secara acak mengambil 5 kata kunci. Berikut kata kunci yang kami gunakan.

Representasi Pihak	Kata Kunci
Joko Widodo	"saya01", "jokowi2periode"
Prabowo Subianto	"2019gantipresiden", "01maincurang"
Netral	"pemiludamai"

Setelah mendapatkan 2500 *tweet* yang merepresentasikan tiga pihak, kami mengambil fitur dari akun yang mengunggah *tweet* tersebut. Akun yang berhasil kami dapatkan sebanyak 1017 akun. Lalu, masing-masing akun yang kami dapatkan diperiksa dan diberikan label *buzzer* atau non-*buzzer* secara manual. Hasil proses label yang kami lakukan terdapat 504 akun non-*buzzer*, 489 akun *buzzer*, dan 24 akun yang tidak ditemukan. Akun yang tidak ditemukan ini merupakan akun yang sudah dimatikan oleh pemilik. Dua puluh empat akun tersebut kami hilangkan dari *dataset* sehingga terdapat 993 data yang kemudian dijadikan sebagai data untuk mendeteksi *buzzer* Twitter.

2.2 Fitur Awal

Penentuan fitur yang kami ambil didasarkan pada penelitian sebelumnya terkait pendeteksi *bot* pada Twitter [3]. Kami mengambil beberapa fitur secara acak yang dapat kami aplikasikan dalam proyek ini. Berikut 11 fitur yang kami gunakan.

Tabel 2.1 Fitur Awal

No	Nama Fitur	Jenis Data	Deskripsi
1	<i>Most_active_days</i>	nominal	Fitur ini menentukan hari yang paling sering digunakan pengguna untuk mengunggah <i>tweet</i> .
2	<i>Most_active_hours</i>	nominal	Fitur ini menentukan waktu dalam format jam yang paling sering digunakan pengguna untuk mengunggah <i>Tweet</i> .
3	<i>Freq_of_most_hashtag</i>	numerikal	Fitur ini menentukan frekuensi terbanyak penggunaan <i>hashtag</i> dalam 500 <i>tweet</i> terbaru.
4	<i>Location</i>	nominal	Fitur ini menentukan apakah akun mencantumkan lokasi atau tidak.
5	<i>Verified</i>	nominal	Fitur ini menentukan apakah akun sudah terverifikasi oleh Twitter atau tidak.
6	<i>Followers_count</i>	numerikal	Fitur ini menentukan jumlah pengikut akun.
7	<i>Friends_count</i>	numerikal	Fitur ini menentukan jumlah yang diikuti akun.
8	<i>Favorites_count</i>	numerikal	Fitur ini menentukan jumlah <i>tweet</i> yang disukai akun.
9	<i>Statuses_count</i>	numerikal	Fitur ini menentukan jumlah <i>tweet</i> yang diunggah oleh akun.
10	<i>Month_created</i>	nominal	Fitur ini menentukan bulan saat akun dibuat.
11	<i>Year_created</i>	nominal	Fitur ini menentukan tahun saat akun dibuat.

Dari 11 fitur tersebut, fitur yang bertipe nominal akan di-*encode* menjadi beberapa fitur dengan masing-masingnya 1 *domain*. Selain itu, fitur yang bertipe numerik akan dilakukan *scaling* menggunakan *z-score*. Setelah dilakukan *encode*, jumlah fitur menjadi sebanyak 64 fitur.

2.3 Seleksi Fitur

Kami menggunakan Mutual Information untuk menghitung dependensi antara setiap fitur dengan label. Dari 64 fitur yang dihitung nilai MI-nya dengan label, kami hanya menggunakan 5, 10, 15, 20, dan 25 fitur teratas sebagai eksperimen kami. Berikut 25 fitur tersebut ;

1. freq_of_most_hashtag : 0.05331559049244583;
2. year_2008 : 0.043791358234855915;
3. favourites_count : 0.03886014405079119;
4. hour_21 : 0.0347004129789914;
5. hour_5 : 0.030872979346599383;
6. year_2010 : 0.028334397158559232;
7. location_1 : 0.026989276024333764;
8. month_6 : 0.021668068844014732;
9. month_9 : 0.020090822038762335;
10. year_2012 : 0.018259750981194145;
11. hour_1 : 0.017981767526687564;
12. location_0 : 0.0166995069449416;
13. hour_23 : 0.016586308333431443;
14. month_4 : 0.015320322182071022;
15. year_2011 : 0.015120296712667791;
16. hour_20 : 0.014636691018166115;
17. hour_16 : 0.013678067906250035;
18. month_3 : 0.01344071052149265;
19. statuses_count : 0.012928982814989665;
20. year_2019 : 0.012403013540483121;
21. most_active_day_7 : 0.011501199853389377;
22. month_12 : 0.011480939566253179;

- 23. year_2016 : 0.011117816888257748;
- 24. month_8 : 0.010490830168591536;
- 25. month_2 : 0.010112104166710711.

Dari 25 fitur dengan nilai MI tertinggi, kami melakukan variasi deteksi buzzer dengan menggunakan 5, 10, 15, 20, dan 25 fitur.

2.4 Algoritma

Kami menggunakan 3 algoritma diantaranya SVM RBF Kernel, SVM Linear Kernel, dan Logistic Regression. Pemilihan SVM dan Regression disebabkan data kami terdiri dari 2 jenis data yaitu numerik dan nominal. Evaluasi yang kami gunakan diantaranya nilai akurasi, presisi, dan *recall*.

BAB 3

ANALISIS

3.1 Eksperimen

Dalam melakukan eksperimen, kami mencoba berbagai model serta berbagai fitur dengan *training* dan *testing splitting* menggunakan cross-validation sebesar 10 untuk *testing*, sisanya untuk training (digilir). Model *machine learning* yang kami coba adalah SVM dengan kernel RBF, SVM dengan kernel linear, dan Logistic Regression. Lalu, kami mencoba 5, 10, 15, 20, dan 25 fitur terpenting berdasarkan skor *mutual information* nya untuk dicoba diklasifikasikan pada masing-masing algoritma *machine learning* kami. Berikut hasil eksperimen kami (akurasi, *precision*, dan *recall* (*precision* dan *recall* untuk *positive buzzer*) pada masing-masing sel didapat dengan cara merata-ratakan hasil *cross-validation* nya).

No	Classifier	Jumlah Fitur	Akurasi	Precision	Recall
1	SVM RBF Kernel	5	0.5531	0.5688	0.3255
		10	0.6119	0.6366	0.5480
		15	0.6099	0.6348	0.544
		20	0.6049	0.6312	0.534
		25	0.6029	0.6303	0.53
2	SVM Linear Kernel	5	0.5346	0.5358	0.2425
		10	0.5897	0.6100	0.4685
		15	0.5907	0.6100	0.4705
		20	0.5869	0.6093	0.4725
		25	0.5869	0.6093	0.4725
3	Logistic Regression	5	0.5377	0.5458	0.2525
		10	0.5888	0.6150	0.4745
		15	0.5918	0.6190	0.4785
		20	0.5859	0.6101	0.4785
		25	0.5859	0.6101	0.4785

3.2 Analisis Hasil Eksperimen

Sebelum mengulas hasil eksperimen terlebih dahulu, kita akan mengupas apa itu *precision* dan *recall*.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

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

Precision merupakan proporsi tebakan “positif” yang berhasil ditebak dengan tepat. Sebagai contoh, misalnya ada 15 hasil prediksi user (dalam kasus *buzzer detection*). Prediksi user 1-10 bernilai *positive* (1) dan 11-15 bernilai *negative* (0). Namun, pada label *gold standard* (hasil anotasi manusia), ternyata user 1-6 itu *positive* dan 7-15 *negative*. Dari 10 yang ditebak *positive*, ada 6 yang tepat. Artinya *precision untuk label positive buzzer* sebesar 6/10.

Recall merupakan proporsi *actual* “positif” yang berhasil ditebak dengan tepat. Sebagai contoh, misalnya ada 15 hasil prediksi user (masih dalam kasus *buzzer detection*). Prediksi user 1-4 bernilai *positive* (1) dan 5-15 bernilai *negative* (0). Namun, pada label *gold standard* (hasil anotasi manusia), ternyata user 1-11 itu *positive* dan 12-15 *negative*. Dari 11 *actual positive*, yang berhasil ditebak hanya 4. Artinya *recall untuk label positive buzzer* sebesar 4/11.

Akurasi paling tinggi dari tiga algoritma SVM RBF Kernel, SVM Linear Kernel, dan Linear Regression berada di sekitar 60%. Ada dua hal yang menjadi penyebab rendahnya nilai akurasi ini. Pertama, proses memberikan label *buzzer* dan *non-buzzer*. Pada proses ini, kemungkinan *human-error* sangat tinggi mengingat terdapat 5 individu yang berpartisipasi.

Batasan antara *buzzer* dan *non-buzzer* juga sukar diidentifikasi karena dinamika kampanye pada media sosial tidak lagi hanya pada 1 topik, misalnya hanya keunggulan pasangan calon tertentu. Topik yang dibahas oleh dua kubu kandidat pasangan calon pada pilpres tahun 2019 sangat bervariasi, kubu nomor 01 dengan citra yang nasionalis dan kubu nomor 02 dengan citra yang bernuansa islami. Topik pembahasan pada akun-akun yang kami dapatkan memang terbatas pada dua citra tersebut yang kemudian menjadi sulit untuk mengidentifikasi akun yang sering mengunggah tweet terkait dengan sisi nasionalis atau islami merupakan pendukung 01 atau 02 secara berturut-turut.

Penyebab selanjutnya adalah variasi fitur yang kurang. Saran untuk penelitian selanjutnya adalah lebih banyak untuk eksplorasi fitur yang mungkin diambil dan dijadikan data untuk klasifikasi.

3.3 Keluaran Program

Keluaran program ini adalah pesan pada konsol terkait nilai akurasi, presisi, recall dari 3 algoritma dengan variasi fitur yang berbeda.

BAB 4

PENUTUP

Pendeteksi buzzer dengan nilai akurasi tertinggi di sekitar 60% sudah cukup baik. Hal yang perlu diperhatikan ketika ingin mengaplikasikan kerangka kerja deteksi buzzer ini ialah dengan memahami secara komprehensif konteks topiknya. Pada konteks pemilihan presiden RI tahun 2019, dinamika serta topik yang menjadi konten kampanye sangat bervariasi sehingga sangat sulit untuk mengidentifikasi akun buzzer dan non-buzzer. Selain itu, eksplorasi lebih jauh terkait fitur yang ingin digunakan juga diperlukan untuk menambah pandangan terhadap proses pendeteksian buzzer.

DAFTAR PUSTAKA

- [1] M. Cornfield, "Yes, it did make a difference," Media & Politics, June 2008. [Online]. Available: Retrieved July 16, 2015, from <http://takingnote.tcf.org/2008/06/yes-it-did-make.html>
- [2] Ibrahim, Mochamad & Abdillah, Omar & F. Wicaksono, Alfian & Adriani, Mirna. (2015). Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. 1348-1353. 10.1109/ICDMW.2015.113.
- [3] Kantepe, Mucahit & Can Ganiz, Murat. (2017). Preprocessing framework for Twitter bot detection. 630-634. 10.1109/UBMK.2017.8093483.