PREDIKSI PROFESI BERDASARKAN MODEL BAHASA PADA *TWEETS*

Hapnes Toba[1)], William Stefanus[2)]

[1), 2)] Jurusan Teknik Informatika, Fakultas Teknologi Informasi, UK. Maranatha Bandung
Jl Suria Sumantri Bandung, 40173
Email : hapnestoba@it.maranatha.edu[1)], william.khiwil@gmail.com[2)]

Abstract

With the advance of social media, people tends to be very reactive on issues which are happening around the globe. Everybody can show their opinions freely, and sometimes uncontrollable, no matters what their job is. This research investigates the tendency of words choice in someone's job based on the style of language he/she used in his/her twitter account. It is assume that most of the people in a specific job has the same language used on social media. The analyses of the study is performed by using Naïve Bayes classifiers for around 30,000 tweets. The text processing are divided into three main parts, i.e.: retrieval and grouping of the data, data processing, and evaluation. The type of jobs which are analyzed, consists of: politicians, actresses/actors, musicians, and students, through their official twitter accounts. The experimental results show that multinomial Bayes classifiers are more reliable than the binomial classifiers. Further investigation shows that the best accuracy is achieved by the unigram model, which has a mean of 0.73±0.127 in a 5 cross validation setting. This fact reveals that there is no direct relatioship between someone's word choice and his/her profession.

Keywords: text analysis, language (n-grams) models, job prediction, Naïve Nayes classification, social media content analysis.

1. Pendahuluan

Seiring dengan berkembangnya penggunaan Internet, semakin banyak pula jejaring sosial yang dapat digunakan untuk saling berkomunikasi, seperti Facebook, Twitter, Ask.fm, Path, dan lain-lain. Dengan munculnya banyak media sosial tersebut, semakin banyak pula orang-orang yang menuliskan opini atau pendapat mereka terhadap suatu masalah yang sedang terjadi di dunia nyata pada media sosial yang ada.

Menarik untuk diteliti bagaimana keterkaitan antara pemilihan kata, berupa model bahasa (*n*-grams) yang digunakan dalam penyampaian opini tersebut dengan jenis pekerjaan seseorang. Dengan menganalisis kemunculan kata pada sebuah jenis pekerjaan, mungkin saja dapat diketahui bagaimana interaksi yang muncul antara orang-orang di dalam pekerjaan tersebut ataupun dengan jejaring di sekitarnya. Pada akhirnya, dapat dibentuk pula semacam model pemilihan kata dalam profesi tertentu.

Di dalam penelitian ini diasumsikan bahwa orang-orang yang terdapat pada suatu golongan pekerjaan yang sama, akan menulis pendapat yang hampir sama terkait dengan suatu masalah yang sedang muncul, dan setiap golongan pekerjaan memiliki kata-kata khusus yang acap kali muncul. Tujuan khusus dari penelitian ini untuk membuat model klasifikasi yang dapat mendeteksi pekerjaan seseorang melalui urutan kata dalam *tweets* yang ditulisnya. Adapun rumusan permasalahan yang secara khusus akan dijadikan sebagai pertanyaan riset adalah: bagaimana membuat model yang dapat mendeteksi pekerjaan seseorang lewat *tweets* yang ditulis dengan dasar hipotesis bahwa dengan kemunculan kata tertentu secara *n*-gram dapat menentukan jenis pekerjaan seseorang

Berdasarkan uraian latar belakang di atas, adapun kebaharuan yang diusulkan melalui penelitian ini adalah:

1. Analisis kemunculan kata-kata yang mendominasi pekerjaan-pekerjaan yang dijadikan obyek penelitian, yaitu yang dianggap seringkali membuat *trending topics* dalam media sosial Twitter. Dalam konteks penelitian ini dipilihlah profesi sebagai berikut: politikus, aktor dan aktris, musisi dan pelajar.
2. Pembentukan dan evaluasi model pemilihan kata untuk pekerjaan-pekerjaan pada nomor satu di atas.

Makalah ini meliputi pembahasan-pembahasan sebagai berikut: kajian literatur terkait analisis tekstual dan model Naïve Bayes, ekstraksi data pada media sosial Twitter, rancangan eksperimen dan evaluasi pembentukan model, serta diakhiri dengan kesimpulan dan usulan riset lebih lanjut.

2. Kajian Literatur

Dalam bagian ini diberikan kajian singkat mengenai model Naïve Bayes dan bagaimana kehadiran kata-kata dalam sebuah *tweets* dapat digunakan untuk menganalisis kelas pekerjaan tertentu.

A. Naïve Bayes

Sebuah model klasifikasi Naïve Bayes terdiri dari satu buah simpul induk, disertai dengan banyak simpul anak sebagai atribut. Model Naïve Bayes didasarkan pada asumsi yang kuat mengenai ketidak-ketergantungan antar atribut [1]. Model Naïve Bayes banyak digunakan dalam pemrosesan citra medis, klasifikasi teks, dan diagnosis kueri. Keunggulan utama dari model klasifikasi Naïve Bayes adalah kemudahan

implementasinya, disertai dengan tingkat keakuratan yang cukup tinggi dalam banyak kasus, bahkan dapat melampaui model regresi logistik yang sering digunakan sebagai suatu standar dalam pembelajaran mesin [2].

Rumus dasar dari model Naïve Bayes dapat dilihat pada Formula (1).

$$P(h/D) = \frac{P(D/h)\ P(h)}{P(D)} \quad \dots (1)$$

Dengan:
- P(h) = probabilitas suatu kejadian h
- P(D) = probabilitas suatu kejadian D
- P(h/D) = probabilitas terjadinya h dengan diberikan kondisi D
- P(D/h) = probabilitas terjadinya D dengan diberikan kondisi h

Algoritma Naïve Bayes sering digunakan oleh banyak orang untuk melakukan penelitian yang menyinggung bidang kajian *machine learning* (pembelajaran mesin), seperti *data mining*, dan *text mining* [4].

B. Ekstraksi Fitur Binomial dan Multinomial

Naïve Bayes *Binomial Classifier* merupakan salah satu cara pengklasifikasian menggunakan algoritma Naïve Bayes dengan cara memasukkan nilai pada variabel himpunan pelatihan hanya dengan angka satu atau nol (*binary*). Angka satu menunjukkan jika sebuah instans memiliki variabel yang ada pada *training dataset* dan angka nol menunjukkan jika sebuah instans tidak memiliki variabel yang ada pada *training dataset*. Hal ini dicontohkan pada Tabel 1.

Tabel 1 Contoh *trainingset Naïve Bayes Binomial*

doc_id	di	kampus	mau	saya	ada	jam
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	0	0	0	0	1	1

Naïve Bayes *Multinomial Classifier* merupakan salah satu cara pengklasifikasian dari algoritma Naïve Bayes dengan cara memperhitungkan jumlah frekuensi kemunculan variabel dalam sebuah instans. Tabel 2 menunjukkan contoh frekuensi kemunculan dari nilai binomial pada Tabel 1.

Tabel 2 Contoh *trainingset Naïve Bayes Binomial*

doc_id	di	kampus	mau	saya	ada	jam
1	5	3	1	2	3	2
2	6	3	2	3	2	4
3	0	0	0	0	1	5

C. Riset Seputar Media Sosial Twitter

Twitter adalah sebuah layanan pengiriman pesan yang membagikan banyak karakter menggunakan alat-alat komunikasi yang sudah ada [4]. Twitter memiliki beberapa kesamaan dengan e-mail, IM (*Instant Message*), SMS (*Short Message Service*), dan semacamnya. Adapun hal-hal yang membuat Twitter unik, adalah sebagai berikut:

1. Pesan yang dikirim atau diterima pada Twitter tidak melebihi dari 140 karakter.
2. Pesan pada Twitter dapat dilihat oleh semua orang, tidak perlu izin dari sang penulis.
3. Pesan dapat diterima atau dikirim dengan banyak mekanisme, seperti lewat *mobile phone*, komputer, situs jaringan, dan lain-lain.

Sejak munculnya media sosial Twitter dan bertambah banyaknya pengguna Internet, banyak penelitian yang dilakukan seputar Twitter. Salah topik penelitian yang sering disinggung adalah *Twitter Sentiment Analysis*. Umumnya, penelitian dengan topik ini bertujuan untuk menentukan apakah *tweets* menunjukkan sesuatu yang positif, negatif atau netral [5, 6].

Selain itu, sering pula Twitter digunakan untuk melakukan *profiling* [7]. Informasi tentang seseorang seperti nama, umur, lokasi, dan biodata singkat biasanya sudah tersimpan hampir di setiap media sosial. Akan tetapi, masih banyak juga informasi yang tidak tercantum, seperti kebiasaan sehari-hari, ketertarikan pada politik, dan lain-lain. Penelitian [7] tersebut, melakukan uji coba untuk mendapatkan informasi seseorang secara otomatis dengan cara mengamati bagaimana perilaku *user* melalui interaksinya dengan *user* lain, dan bukan pada gaya bahasa *user* sebagaimana yang diusulkan melalui makalah ini.

Ciri khas lain dari makalah ini, dibandingkan dengan penelitian Twitter lainnnya adalah pengambilan sumber datanya yang berasal dari penduduk di Indonesia yang berarti acuan utamanya adalah bahasa Indonesia, memeriksa kaitan antara profil pengguna (seperti pemeriksaan antara *follower*, dalam satu kategori yang sama), memeriksa *trending topic* yang sedang dibahas pada media sosial Twitter.

3. Rancangan Eksperimen

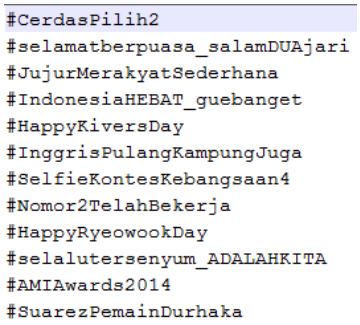
Proses yang dilakukan dalam penelitian ini dibagi menjadi 3 bagian utama, yaitu:
1. Pengambilan data dan pengelompokan data
2. Pengolahan data
3. Pengujian dan evaluasi

A. Pengambilan dan Pengelompokan Data

Sumber data yang digunakan dalam penelitian ini diambil seluruhnya dari media sosial Twitter. Jumlah kelompok yang ditentukan dalam penelitian ini ada empat kelompok, yakni politikus/politisi, artis/aktor, musisi, dan pelajar.

Empat kelompok ini diambil dikarenakan, jika ada suatu topik atau masalah muncul, keempat bidang profesi ini, teramati sebagai kelompok yang paling sering memberikan pendapat mereka. Topik atau masalah yang munculpun seringkali bersangkutan dengan bidang pekerjaan yang dilakukannya. Hal tersebut misalnya terlihat dari kemunculan *trending topic* yang ada pada

media sosial Twitter pada Gambar 1. Setelah ditelusuri dalam profil pengguna Twitter, maka yang paling banyak menulis *tweets* dengan *trending topic* pada Gambar 1 adalah kelompok politisi, musisi, pelajar, dan artis.



Gambar 1 *Trending Topic* pada Media Sosial Twitter

Dalam proses penentuan pengguna dari setiap kelompok, terdapat beberapa kriteria yang harus dipenuhi, yaitu:

1. Pengguna merupakan pengguna yang aktif menulis tweets dengan minimal *tweets* sebanyak 200 dalam empat bulan terakhir (Agustus-Desember 2014).
2. Pengguna berwarga negara Indonesia.

Jumlah pengguna yang berhasil terjaring adalah 39 orang, dengan rincian: kelompok politisi sebanyak 7 pengguna, kelompok artis atau aktor sebanyak 10 pengguna, kelompok musisi sebanyak 11 pengguna, dan kelompok pelajar sebanyak 11 pengguna. Data untuk kelompok politisi, artis dan musisi, adalah nama-nama yang merupakan *public figure*. Khusus untuk kelompok pelajar, data diambil Twitter mahasiswa semester akhir Fakultas Teknologi Informasi Universitas XXX Kota Bandung. Jumlah data yang diambil dari setiap pengguna adalah 200-300 *tweets*. Jumlah data yang terkumpul sekitar 6.800-10.000 *tweets* per pengguna, dengan jumlah total 32.752.

B. Pengolahan Data

Dari proses pengambilan dan pengelompokan data, tahap selanjutnya adalah pengolahan data. Proses yang dilakukan pada saat pengolahan data dalam penelitian ini adalah:

1. Pembersihan isi *tweets* yang diambil pada saat proses pengambilan data agar setiap *tweets* tidak memiliki link, *hashtag*, dan *emoticon*.
2. Pembuatan *training set* yang akan digunakan dalam proses klasifikasi pada aplikasi.

1) Pembersihan Isi Tweet

Setiap data yang diambil pada saat proses pengambilan data belum tentu murni hanya berisi teks saja. Satu *tweets* dapat terdiri atas teks, *link*, *hashtag*, dan *emoticon*. *Tweet* atau data yang dibutuhkan dalam penelitian ini harus murni berisi teks saja. Maka dari itu, setiap *tweets* yang diambil harus dibersihkan terlebih dahulu.

Gambar 2 Contoh *Tweets* yang Tidak Bersih

Gambar 3 merupakan salah satu contoh tweets yang tidak bersih. *Tweet* ini harus dibersihkan agar murni berisi teks saja. Setelah melakukan pembersihan, maka isi *tweets* akan menjadi: "`Girls day out!`".

2) Pembuatan Training Set

Setelah membersihkan semua *tweets* yang diambil, maka tahapan selanjutnya adalah pembuatan *training set* yang akan digunakan untuk menjalankan fungsi klasifikasi. Pembentukan *training set* dan proses klasifikasi menggunakan pendekatan binomial dan multinomial sebagaimana disampaikan dalam bagian kajian literatur.

Alasan utama dari penggunaan algoritma Naïve Bayes adalah digunakannya data-data statistik sebagai bahan acuan, dan konsep Naïve Bayes sudah dikenal tingkat keakuratannya yang tinggi dalam hal menganalisis perhitungan-perhitungan statistik.

C. Pengujian dan Evaluasi

Evaluasi dilakukan dengan memasukkan sebuah *tweets* yang ingin diuji, lalu melakukan klasifikasi berdasarkan model profesi yang sudah dibuat. Hasil dari pengujian ini diharapkan sesuai dengan informasi kebenaran yang didapat pada dunia nyata dan memiliki tingkat keakuratan yang dapat dipercaya.

Setiap model pekerjaan yang dibentuk oleh aplikasi diuji pula melalui *cross-validation* (pengujian silang) dan validasi kesalahan dengan data pengujian baru, yaitu dari *tweets* bulan Januari 2015. Adapun yang hendak diuji dengan data baru ini adalah kecenderungan munculnya kesalahan tipe satu, yaitu kondisi dimana sebuah instans atau pernyataan bernilai benar tetapi hasil dari pengujian yang dilakukan terhadap pernyataan atau instans tersebut bernilai salah.

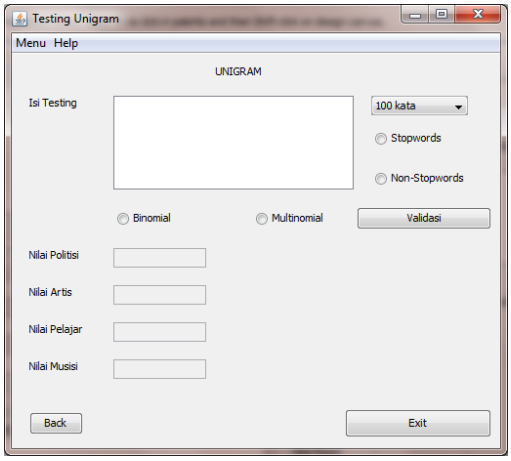
Kombinasi dari eksperimen yang dilakukan dibagi menjadi empat skenario besar, yaitu:

1. Penggunaan Naïve Bayes Binomial dan Multinomial;
2. Penggunaan *word stopping*[1];
3. Penggunaan jumlah kata yang dipakai dalam pembentukan model; dan
4. Penggunaan proses pemecahan kata secara *n*-gram (*uni-*, *bi-*, dan *trigram*), yaitu: satu kata tunggal, dua dan tiga kata secara berurutan.

[1] http://lucene.apache.org/core/4_6_0/analyzers-common/org/apache/lucene/analysis/id/IndonesianAnalyzer.html

Untuk mempermudah jalannya skenario evaluasi di atas, maka dibuatlah tampilan antarmuka sehingga nilai parameter pengujian di atas dapat dengan fleksibel dilakukan. Gambar 3 merupakan implementasi antarmuka untuk form Pengambilan Data. Form ini berfungsi untuk mengambil data (*tweets*) yang paling baru dari daftar pengguna yang ada pada *text area*.

Gambar 3 Desain Antarmuka Pengambilan Data



Gambar 4 Desain Antarmuka Pengujian *Unigram*

Gambar 4 merupakan implementasi antarmuka untuk form *Testing Unigram*. Pengguna memasukkan *tweets* yang hendak diuji, maka nilai untuk masing-masing pekerjaan kemudian ditampilkan pada *textbox* yang telah disediakan.

4. Hasil Penelitian dan Evaluasi

A. Evaluasi Hasil Pengujian Silang

Pengujian ini bertujuan untuk menguji tingkat kestabilan model yang dibuat berdasarkan kombinasi-kombinasi dari rencana pengujian untuk keempat bidang profesi yang diujikan. Hasil pengujian silang yang dilakukan dengan metode Naïve Bayes *Binomial dan Multinomial* dapat dilihat pada Tabel 3 dan 4. Dari hasil-hasil tersebut, maka didapatkan beberapa hasil rata-rata berdasarkan kombinasi-kombinasi yang dilakukan:

1. Rata-rata hasil pengujian dengan metode Naïve Bayes Binomial: 0.51.
2. Rata-rata hasil pengujian dengan metode Naïve Bayes Multinomial: 0.65.
3. Rata-rata hasil pengujian dengan fitur *stopping*: 0.58.
4. Rata-rata hasil pengujian tanpa fitur *stopping*: 0.57.
5. Rata-rata hasil pengujian dengan pemecahan kata secara *unigram*: 0.73.
6. Rata-rata hasil pengujian dengan pemecahan kata secara *bigram*: 0.51.
7. Rata-rata hasil pengujian dengan pemecahan kata secara *trigram*: 0.50.

Tabel 3 Akurasi Hasil 5-Pengujian Silang dengan Metode Naïve Bayes *Binomial*

	Binomial					
	Stopwords			***Nonstopwords***		
	1 gram	**2 gram**	**3 gram**	**1 gram**	**2 gram**	**3 gram**
100 kata	0.63	0.51	0.51	0.62	0.51	0.51
500 kata	0.51	0.51	0.51	0.51	0.51	0.51
10% kata	0.51	0.51	0.49	0.49	0.49	0.49
50% kata	0.51	0.51	0.49	0.49	0.51	0.49

Berdasarkan hasil uji validasi silang pada Tabel 3 dan 4, serta perhitungan nilai rata-ratanya, beberapa analisis dapat dilakukan sebagai berikut:

- Performa akurasi model multinomial lebih tinggi jika dibandingkan model binomial secara signifikan dengan nilai p=0.05. Hal ini menunjukkan bahwa frekuensi kemunculan kata dalam sebuah model profesi sangat menentukan hasil prediksi.

Tabel 4 Akurasi Hasil 5-Pengujian Silang dengan Metode Naïve Bayes *Multinomial*

	Multinomial					
	Stopwords			***Nonstopwords***		
	1 gram	**2 gram**	**3 gram**	**1 gram**	**2 gram**	**3 gram**
100 kata	0.86	0.52	0.49	0.82	0.49	0.49
500 kata	0.94	0.49	0.53	0.94	0.49	0.49
10% kata	0.96	0.55	0.53	0.96	0.52	0.49
50% kata	0.94	0.55	0.49	0.96	0.52	0.48

- Penghilangan *stopwords* tidak mempengaruhi rata-rata akurasi dalam setiap model profesi. Hal ini menunjukkan bahwa ada kecenderungan bahwa setiap profesi memiliki kata-kata khas yang muncul secara dominan, namun belum tentu hal tersebut menunjukkan gaya bahasa (urutan kata) yang dominan dalam sebuah model profesi.
- Pengujian dengan variasi jumlah kata pada saat pembentukan model, menunjukkan bahwa penggunaan 10% dari keseluruhan kata yang muncul dalam koleksi, telah cukup untuk menghasilkan stabilitas performa.

- Hasil pengujian menunjukkan bahwa *tweets* yang muncul dalam setiap model profesi tidak menunjukkan adanya gaya bahasa yang dominan. Hal tersebut ditunjukkan melalui nilai rata-rata akurasi yang selalu lebih tinggi pada pengujian *unigram*, yaitu pada data-data pengujian yang hanya menggunakan kata tunggal.

B. Evaluasi Hasil Pengujian dengan Data baru

Untuk mengujicoba model profesi pada himpunan data yang lebih luas, maka semua data *tweets* pada hasil evaluasi bagian A di atas digunakan sebagai model. Himpunan data *tweets* dari setiap obyek penelitian pada bulan Januari 2015 dijadikan sebagai data pengujian.

Tabel 5 Hasil Pengujian Himpunan Data baru

Unigram	0.4982
Bigram	0.4969
Trigram	0.4981

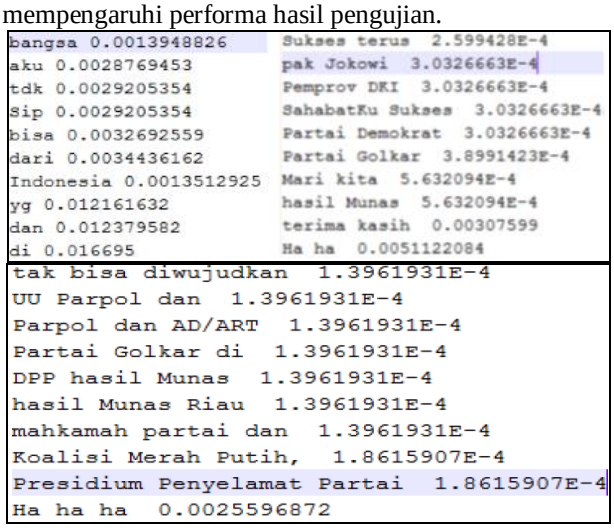
Kombinasi eksperimen yang digunakan mengikuti hasil pada bagian A, yaitu: dengan model multinomial, tanpa menggunakan stopwords, menggunakan hanya 10% dari kata dalam koleksi, dan kemunculan *unigram*. Hasil akurasi dari eksperimen ini dapat dilihat pada Tabel 5. Hal menarik yang dapat dianalisis dari hasil pada Tabel 5 adalah performa akurasi menurun secara signifikan dibandingkan dengan hasil validasi silang. Dari pengamatan yang dilakukan lebih jauh, terungkap bahwa model yang dibentuk dengan kata-kata melalui *tweets* bulan Agustus – Desember 2014, ternyata memiliki nuansa yang sangat jauh berbeda. Mayoritas kata-kata pada model pelatihan cenderung mengarah pada peristiwa besar di Indonesia saat itu, yaitu masa pemilihan Presiden, sedangkan mayoritas kata-kata pada *tweets* pengujian di bulan Januari 2015, berisi mayoritas ucapan tahun baru, dan bencana alam yang terjadi di penghujung dan awal tahun 2015. Hal ini berlaku untuk semua profesi yang dijadikan sebagai obyek penelitian.

Hasil ini hendak menunjukkan bahwa Twitter sebagai salah satu media sosial yang paling cepat berevolusi secara konten. Fenomena-fenomena yang terjadi di berbagai pelosok daerah dan dunia dapat menjadi pemicu konten kicauan yang dimasukkan oleh para pengguna Twitter.

C. Evaluasi DaftarKata per Model Profesi

Dalam bagian ini disampaikan kata-kata dominan disertasi pembobotannya untuk setiap model profesi. Evaluasi ini hendak menunjukkan kemunculan dominasi kata-kata dalam setiap bidang profesi. Daftar kata-kata yang dimunculkan dalam evaluasi ini adalah 10 besar kata (*uni-*, *bi-*, dan *trigram*) yang muncul dalam setiap model profesi, tanpa memperhitungkan dampak jumlah *stopwords*, dengan menggunakan nilai pembobotan TF-IDF. Gambar 5-8 memberikan daftar kata-kata tersebut beserta pembobotannya untuk setiap profesi yang diamati.

Secara garis besar, daftar kata-kata dari bidang-bidang profesi tersebut memiliki variasi yang sangat besar. Setiap jenis profesi memiliki kecenderungan kata-kata yang berpengaruh, misalnya: kata ‘partai’ dalam kelas politisi atau ‘tiket’ dalam kelas musisi. Salah satu hal yang cukup mencolok dari hasil ekstraksi kata-kata ini adalah banyaknya penggunaan kata-kata asing (terutama bahasa Inggris), juga munculnya dan kata-kata yang tidak baku dan tidak konsisten (singkatan, akronim, dll) dalam setiap model profesi. Misalnya kata ‘yang’ disingkat menjadi ‘yg’. Atau kata ‘tidak’, dituliskan sebagai ‘gak’. Berdasarkan observasi, hal-hal ini sangat mempengaruhi performa hasil pengujian.



Gambar 5 Daftar Kata (*uni-*, *bi-*, dan *trigram*) Politisi

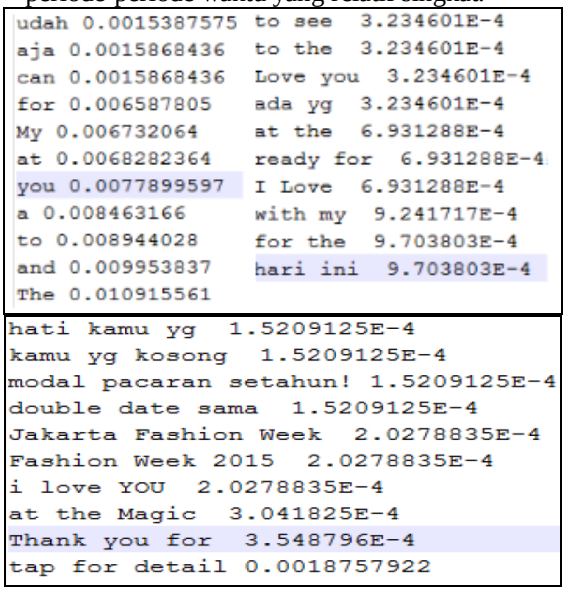
Dalam model profesi politisi (Gambar 5), kata-kata yang seringkali muncul terkait dengan partai politik, kebangsaan dan tokoh-tokoh berpengaruh dalam bidang politik. Dalam model artis (Gambar 6), kata-kata dominan banyak dipengaruhi oleh penyelenggaraan *event-event*, seperti: konser atau *fashion show*. Di samping itu muncul juga gaya-gaya bahasa ‘keartisan’, yang seringkali didominasi oleh campuran antara bahasa Inggris dan Indonesia, seperti misalnya urutan *trigram*: ‘*double date* sama’. Hal yang mirip juga terjadi dalam model musisi pada Gambar 7.

Dalam model pelajar (Gambar 8), kemunculan kata-kata banyak dipengaruhi oleh pengerjaan tugas yang sedang dilakukan dalam kurun waktu selama pengambilan data. Misalnya, kemunculan banyaknya kata-kata berbahasa Inggris terjadi karena pada saat pengambilan data dari Twitter, sedang dilangsungkan tutorial penulisan makalah berbahasa Inggris bagi mahasiswa tingkat akhir. Muncul pula frasa ‘Pemimpin yg baik’, hal ini didukung oleh kenyataan pada saat pengambilan data sedang dilangsungkan sesi perkuliahan kepemimpinan.

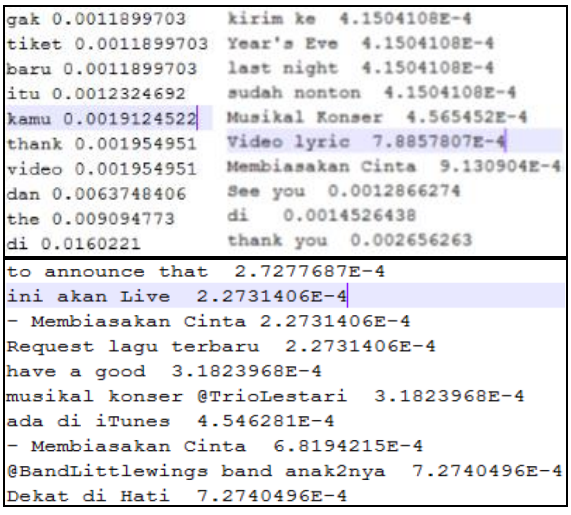
5. Kesimpulan dan Saran

Mengacu pada hasil-hasil dan observasi terhadap kata-kata yang berpengaruh dalam setiap model, dapat ditarik kesimpulan sebagai berikut:

1. Performa model multinomial lebih akurat dibandingkan dengan model binomial untuk melakukan prediksi profesi berdasarkan kemunculan kata-kata.
2. Model prediksi profesi dengan menggunakan kata-kata dari tweets memiliki performa yang baik saat validasi silang, namun menurun drastis dengan data baru yang berbeda periode pengambilan. Hal tersebut menunjukkan perlunya pembentukan model secara berkesinambungan (*incremental*) dalam periode-periode waktu yang relatif singkat.

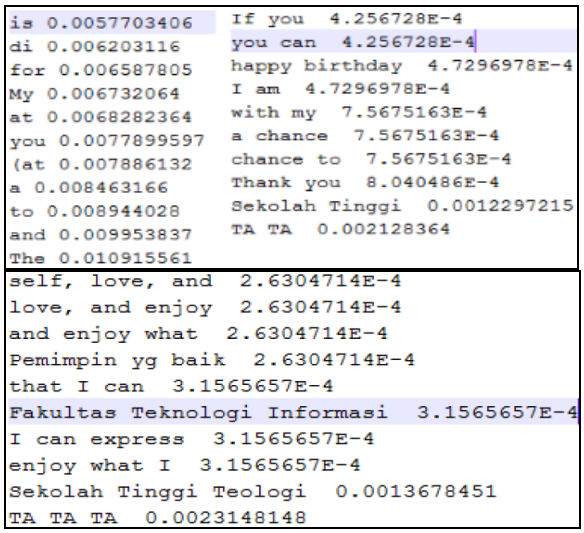


Gambar 6 Daftar Kata (*uni-*, *bi-*, dan *trigram*) Artis



Gambar 7 Daftar Kata (*uni-*, *bi-*, dan *trigram*) Musisi

3. Kata-kata berpengaruh dalam setiap model profesi tidak menunjukkan gaya bahasa (keterurutan kata-kata) tertentu. Dengan demikian, tantangannya sangat besar untuk dapat menghasilkan model yang akurat untuk prediksi profesi. Kumpulan kata-kata dalam *tweets* cenderung lebih mengarah pada fenomena-fenomena yang terjadi dalam kurun waktu tertentu.



Gambar 8 Daftar Kata (*uni-*, *bi-*, dan *trigram*) Pelajar

Mengacu pada keterbatasan model profesi yang telah dibentuk ini, maka disarankan hal-hal sebagai berikut:

1. Pemanfaatan heuristik atau model untuk memformalisasi penggunaan singkatan dan penulisan kata-kata yang terkadang tidak konsisten pada media sosial Twitter.
2. Mendeteksi perbedaan kata bahasa asing dan bahasa Indonesia, sehingga dapat fokus pada kata-kata yang berpengaruh dalam bahasa Indonesia saja.

Daftar Pustaka

[1] Bui, A. A., & Taira, R. K. (2010). *Medical Imaging Informatics*. London: Springer Science-Business Media, LLC.

[2] Mitchell, T. (2015). *Generative and Discrimintative Classifiers: Naïve Bayes and Logistic Regression*. Available online: https://www.cs.cmu.edu/~tom/mlbook/Nbayes LogReg.pdf. Access: November 2015.

[3] Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, July). *Transferring naive bayes classifiers for text classification*. In Proceedings of the National Conference on Artificial Intelligence (Vol. 22, No. 1, p. 540). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.

[4] Reilly, T. O., & Milstein, S. (2011). *The Twitter Book*. Sebastopol: O'Reilly Media, Inc.

[5] Pak, A., & Paroubek, P. (2010, May). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. In LREC (Vol. 10, pp. 1320-1326).

[6] Go, A., Huang, L., & Bhayani, R. (2009). *Twitter sentiment analysis*. Entropy, 17.

[7] Pennacchiotti, M., & Popescu, A. M. (2011). *A Machine Learning Approach to Twitter User Classification*. ICWSM, 11, 281-288.

Biodata Penulis

Hapnes Toba, memperoleh gelar S2 teknik informatika dari TU Delft (2002), dan S3 ilmu komputer dari Universitas Indonesia (2015). Saat ini menjadi tenaga pengajar di Universitas Kristen Maranatha Bandung.

William Stefanus, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika Universitas Kristen Maranatha Bandung (2015). Saat ini bekerja sebagai pengembang perangkat lunak di salah satu perusahaan swasta nasional di Jakarta.