PENDETEKSI BAHASA DAERAH PADA TWITTER DENGAN *MACHINE LEARNING*

Ahmad Arif Budiman

Jurusan Teknik Informatika Fakultas Teknologi Industri

Universitas Islam Indonesia

Email: 11523262@students.uii.ac.id

Abstrak

Indonesia merupakan negara yang memiliki jumlah bahasa etnis terbesar kedua di dunia. Menurut Ethnologue, lembaga bahasa di dunia, Indonesia memiliki 707 bahasa daerah. Negara dengan jumlah bahasa etnis paling banyak adalah Papua Nugini, dengan jumlah 839 bahasa etnis. Namun, tidak sedikit dari bahasa daerah di Indonesia yang terancam, bahkan sudah punah. Beberapa diantara penyebab punahnya bahasa daerah adalah faktor urbanisasi, perkawinan antar etnis, penggunaan bahasa daerah dalam pendidikan, serta bahasa dominan dalam suatu wilayah multibahasa yang berdampingan.

Salah satu wadah yang memungkinkan untuk menjadi media pendekatan kepada generasi muda saat ini adalah melalui sosial media. Hal ini disebabkan oleh maraknya penggunaan sosial media oleh seluruh kalangan di Indonesia khususnya generasi muda. Sehingga diperlukan suatu wadah untuk mengakses sosial media yang bisa menampilkan tulisan dari pengguna yang ada di sosial media tersebut dan menampilkan bahasa daerah mana yang digunakan dalam penulisan tersebut.

Penelitian dilakukan dengan menggunakan data dari Twitter, salah satu sosial media yang berbasis teks*, yang mana hanya mende*teksi *bahasa dari negara tertentu dan belum bisa menentukan bahasa dari daerah-daerah yang spesifik seperti bahasa daerah di Indonesia.*

Kata kunci: Twitter, klasifikasi, sosial media, bahasa daerah, model.

I. PENDAHULUAN

Bahasa menurut (Departemen Pendidikan Nasional; Pusat Bahasa (Indonesia), 2008) adalah sistem lambang bunyi berartikulasi yang bersifat sewenang-wenang dan konvensional yang dipakai sebagai alat komunikasi untuk melahirkan perasaan dan pikiran. Bahasa juga bisa diartikan perkataan-perkataan yang dipakai oleh suatu bangsa (suku bangsa, negara, daerah, dan sebagainya). Kata tersebut terkadang digunakan untuk mengacu pada kode, sandi dan bentuk lain dari sistem komunikasi yang dibentuk secara artifisial seperti yang digunakan pada pemrograman komputer. Makna bahasa dalam hal ini adalah suatu sistem isyarat untuk menyandikan dan menerjemahkan informasi. Dalam konteks Indonesia, Indonesia merupakan negara yang memiliki jumlah bahasa etnis terbesar kedua di dunia. Menurut ethnologue (lembaga bahasa di dunia), Indonesia memiliki 707 bahasa daerah. Negara dengan jumlah bahasa etnis paling banyak adalah Papua Nugini, dengan jumlah 839 bahasa etnis (Patji, 2016).

Pengolahan Bahasa Alami (PBA) atau Natural Language Processing (NLP) adalah salah satu cabang dalam Artificial Intelligence (AI) yang memungkinkan komputer seolah-olah mengetahui bahasa manusia. Text Classification (TC) adalah salah satu cabang dalam NLP, selain Speech Regocnition, Named Entity, dan sebagainya. Sebagaimana sebutannya, TC dapat digunakan untuk melakukan klasifikasi teks, biasanya dengan indikator atau aturan-aturan tertentu untuk masing-

Kondisi seperti ini di level tertentu dapat menjadi pemicu atau alasan terjadinya distorsi pengetahuan tentang bahasa dan serta merta mengancam kelestariannya, yang mana perlu dihindari dengan upaya-upaya tertentu. Salah satu upaya untuk mempertahankan kelestarian bahasa adalah dengan membuat suatu aplikasi yang bertujuan untuk mengedukasi atau memberikan pengetahuan tentang bahasa. Kemungkinan percampuran bahasa dan kesalahan penafsiran bisa dijembatani dengan suatu aplikasi yang memungkinkan penggunanya untuk mengetahui kategori dari bahasa-bahasa yang digunakan.

Berdasarkan paparan mengenai fakta, masalah, metode dalam pengolahan teks, serta solusi yang diajukan dalam menjembatani masalah, peneliti berusaha membuat suatu sistem untuk deteksi bahasa (dalam hal ini bahasa Indonesia, Jawa, dan Sunda) menggunakan gabungan dari pendekatan yang dipaparkan sebelumnya, serta sumber data yang berasal dari media sosial Twitter.

II. PENELITIAN TERKAIT

Hidayatullah (2014) dalam tesisnya melakukan suatu klasifikasi sentimen dan kategori tweet menggunakan data yang didapat melalui Twitter. Data tersebut dikenakan suatu metode pre-processing dan processing antara lain : NBC, negation detection, laplace smoothing, dan TF-IDF.

Groot (2012) melalukan suatu penelitian mengklasifikasikan tweet, menggunakan data Twitter, yang dikenai suatu rangkaian metode antara lain :support vector machine, dan naïve bayes. Sedangkan Romelta (2012) dengan sumber data dan metode yang sama melakukan suatu penelitian untuk mendapat opini pengguna smartphone. Adapun Saraswati (2011) melakukan suatu teks mining dan sentimen terhadap data dari review film serta menggunakan metode yang sama dengan Groot (2012) dan Romelta (2012).

Go (2009) melakukan penelitian dengan mengklasifikasikan tweet dalam hal sentimen dengan judul . Dalam penelitian ini menggunakan NBC yang dikombinasikan dengan metode lain, yaitu Support Vector Machine dan Maximum Entropy. Kemudian dalam pengujian atau evaluasis modelnya menggunakan Experimental Set-up.

Zulfa (2017) melakukan penelitian tentang analisis sentimen pada data Twitter dengan mengguanakn Deep Belief Network. Penelitian ini berjudul "Sentimen Analisis Tweet Berbahasa Indonesia dengan Deep Belief Network".

III. METODE PENELITIAN

1.1 Analisis Kebutuhan

Dalam penelitian ini pastinya membutuhkan beberapa hal yang harus dipersiapkkan sebelum memulai untuk meneliti. Kebutuhan dalam pembuatan penelitian ini meliputi dua hal, yaitu kebutuhan penelitian dan kebutuhan sistem.

Pada penelitian ini memiliki dua hal yang dibutuhkan yaitu perangkat keras dan lunak. Adapun kebutuhan perangkat keras seperti komputer untuk melakukan pengkodean serta pembutan laporan penelitian, buku sebagai bahan acuan dalam pengerjaan penelitian, dan media penyimpanan eksternal (*flashdisk*) untuk memudahkan mobilitas membawa data.

Kebutuhan perangkat lunak yang dibutuhkan dalam penelitian ini berupa *text editor* untuk membuka dan merubah isi *file* text yang dalam penelitian ini menggunakan Sublime, Python IDE sebagai dasar untuk menjalankan perintah menggunakan bahasa pemrograman Python, Microsoft Word digunakan dalam pembuatan laporan, aplikasi Excell digunakan dalam pembuatan *dataset*.

Sedangkan pada kebutuhan sistem, untuk pembuatan penlitian membutuhkan data *tweet* yang berupa teks menggunakan bahasa yang akan diteliti pada penelitian ini.

1.2 Data dan Sumber Data

Data yang digunakan dalam penelitian ini adalah data *tweet* yang didapatkan dari aplikasi Twitter. Data yang dikumpulkan berupa data teks berbahasa Indonesia, Jawa, dan Sunda.

Adapun jumlah akun untuk masing-masing bahasa yaitu :

Bahasa Indonesia : 4 akun

Bahasa Jawa : 6 akun

Bahasa Sunda : 4 akun

Akun dipilih jika terindikasi menggunakan bahasa-bahasa yang dimaksud, indikasinya dilakukan secara manual.Terhadap akun-akun yang telah terindikasi menggunakan bahasa-bahasa yang dimaksud, dilakukan suatu proses pengambilan data (*retrieve*), prosesnya sebagai berikut :

- Masuk apps.twitter.com
- Mendaftarkan akun twitter sebagai *developer*
- Mendaftarkan nama aplikasi atau projek yang akan dibuat.
- Mendapat *public* dan *secret key*
- Menggunakan *public* dan *secret key* untuk *retrieve*data dari akun-akun terpilih.

Data didapat dalam bentuk JSON, yaitu suatu format data ringan yang digunakan dalam pertukaran data, mudah dibaca dan ditulis oleh manusia, serta mudah diterjemah dan dibuat oleh komputer (JSON, n.d.).

1.3 Gambaran Sistem

Gambaran sistem adalah suatu perancangan yang menggambarkan proses sejak data diambil sampai menghasilkan keluaran/*output*.

Terdapat 3 subsistem yang masing-masing dijabarkan dalam paparan berikut ini :

1.3.1 Pengambilan Data.

Seperti yang telah dipaparkan pada subbab 3.2 tentang data dan sumber data, data diambil melalui Twitter API menggunakan *key* yang telah didapatkan setelah mendaftarkan akun dan sistem yang akan dibuat. Setelah data didapatkan dalam format JSON, dilakukan suatu fungsi *decode*atau pengubahan data yang berformat JSON menjadi sebuah objek, sehingga isi data bisa diambil. Dari banyak atribut yang ada pada objek, maka diambil data '*text'* yang merupakan isi dari *tweet* pengguna akun yang diambil datanya.

Karena penelitan ini akan membahas pendeteksian bahasa daerah yang digunakan pada teks, maka ada beberapa hal yang harus diperlakukan terlebih dahulu kepada teks yaitu*data cleaning. Data cleaning* adalah proses untuk menghilangkan nama akun, tanda baca, dan url pada teks. Dan proses ini dilakukan menggunakan program, sehingga

dilakukan secara otomatis sebelum menyimpan hasil *decode* ke dalam bentuk txt.

Setelah data tersimpan kedalam format txt, maka dilakukan *datacleaning* lagi untuk menghapus dokumen-dokumen yang terindikasi kosong dikarenakan sebelumnya teks hanya berisi url saja, namun para proses penghapusan dilakukan secara manual.Sampai pada tahapan ini, ada tiga buah *file* txt yang berisi sejumlah dokumen teksyang merepresentasikan masing-masing bahasa.

Sebelum masuk ke tahapan *pre-processing*, data dari 3 *file* digabung ke dalam format .xlsx dengan isi setiap barisnya meruppakan perwakilan dari setiap dokumen, dan pada kolom pertama berisi teks yang sudah dibersihkan pada proses sebelumnya, dan pada kolom kedua berisikan label atau bahasa yang digunakan dalam teks tersebut. Label pada dokumen tersebut diisi oleh peneliti sebagai basis untuk nantinya dilakukan *pre-processing* dan *training.* Label juga bisa disebut sebagai kelas dari teks tersebut apabila dipandang dasi segi klasifikasi.

1.3.2 *Pre-processing*

Pada subsistem ini, ada beberapa proses yang dilakukan secara sekuensial untuk mendapatkan *data training*, yaitu tokenisasi atau membuat data yang berupa teks menjadi kumpulan *array* kata, misal:

Kalimat "saya sedang belajar"

Dirubah menjadi
['saya'],['sedang],['belajar']

Setelah semua kata sudah ter-tokenisasi, maka langkah selanjutnya adalah membuat *word vector*. *Word vector* atau dalam Bahasa Indonesia bisa disebut vektor kata, yaitu membuat kalimat yang sudah menjadi kumpulan *array* menjadi suatu matriks, yang mana setiap baris matriks tersebut mewakili baris dokumen, sedangkan kolom pada matriks mewakili seluruh kata yang ada di seluruh teks yang ada, bisa digambarkan seperti berikut:

Kumpulan kalimat seperti berikut (['saya'],['sedang],['belajar']),

(['saya'],['ingin],['bermain]),

(['kamu],['teman],['saya])

Akan dirubah menjadi matriks seperti ini:

Tabel 3.1 Tabel contoh *word vector*

saya	sedang	belajar	ingin	bermain	teman	kamu
1	1	1	0	0	0	0
1	0	0	1	1	0	0
1	0	0	0	0	1	1

Setelah sudah berubah menjadi vektor kata, maka selanjutnya adalah memberikan pembobotan terhadap setiap kata pada setiap kalimat atau dokumen menggunakan Unigram dan TF-IDF menggunakan rumus yang bisa dilihat pada bab sebelumnya. Sehingga data bisa ditampilkan bisa seperti pada contoh berikut (bukan hasil perhitungan sebenarnya):

Tabel 3.2 Tabel contoh *word vector* dengan bobot

Saya	sedang	belajar	ingin	bermain	teman	kamu
0.7546	0.7546	0.7546	0.0	0.0	0.0	0.0
0.7546	0.0	0.0	0.7546	0.7546	0.0	0.0
0.7546	0.0	0.0	0.0	0.0	0.7546	0.7546

Setelah data sudah terbentuk seperti contoh pada tabel diatas, maka *dataset* sudah siap untuk digunakan dalam training menggunakan perhitungan pada naive bayes.

1.3.3 Pembentukan Model

Setelah didapatkan data training yang sudah dikenai bobot, maka data tersebut di-*training*berdasarkan aturan pada Naive Bayes sehingga menghasilkan sebuah model yang bisa digunakan untuk pengklasifikasian kalimat selanjutnya. Model sendiri berisikan info seperti jumlah setiap kelas yang ada dalam *dataset,* nilai probabilitas dari setiap kata di setiap dokumen dan semua data yang dibutuhkan dalam perhitunan untuk prediksi kelas apabila ada kalimat baru yang nantinya dimasukkan.

1.3.4 Uji Model

Uji model dilakukan untuk mengetahui kinerja model. Nilainya didapat dengan menghitung*accuration*, *precision*, dan *recall* (tiga pendekatan perhitungan dalam PEM). Untuk menghitung nilai itu, maka dibutuhkan kalimat yang sudah diketahui terlebih dahulu kelasnya. Oleh karena itu dilakukan pengecekan silang antara setiap data dalam dataset, untuk menyilangkannya data akan dibagi menjadi dua bagian, bagian pertama (*data training*) sebagi data training, yaitu data yang dijadikan basis seperti yang dijelaskan sebelumnya, sedangkan bagian kedua (*data test*) dianggap sebagai teks baru yang belum diketahui kelasnya (kelasnya disembunyikan terlebih dahulu).

Setelah *data test* diujikan terhadap data training, maka akan menghasilkan daftar kelas-kelas dari *data test,* sebut sata prediksi kelas. Kemudian prediksi kelas dibandingkan dengan kelas yang sebenarnya dari data *test* yang disembunyikan sebelumnya. Sehingga dapat dilihat dan dihitung nilai *accuration*, *precision*, dan *recall* menggunakan cara-cara yang sudah dijelaskan pada bab sebelumnya.

IV. HASIL DAN PEMBAHASAN

1.4 Hasil Pengolahan *Dataset*

Sesuai dengan tahapan penelitian yang dibahas di bab sebelumnya, hal yang dilakukan pertama kali adalah pengambilan data untuk dijadikan *dataset. Dataset* diambil dari Twitter menggunakan Twitter API menggunakan *key* yang disediakan oleh Twitter.

Dengan menggunakan *API key* tersebut, maka dapat dilakukan pengambilan data *tweet.* Pengambilan data *tweet* pada penelitian ini menggunakan nama akun sebagai pemicu. Data yang dihasilkan berbentuk JSON. Data ini adalah data mentah yang berisi seluruh informasi mengenai aktivitas suatu akun Twitter, antara lain : *tweet*, tanggal *tweet* dibuat, ada *retweet/replay* atau tidak dan berapa jumlah serta apa isi replay tersebut, serta informasi-informasi lain yang di-*generate* oleh *generator* API Twitter

Dari data JSON yang didapat, hanya diambil teks *tweet* dari pengguna. *Tweet* tersebut kemudian langsung di sesuai bahasa mayoritas yang digunakan dalam tweet tersebut : label Indonesia untuk tweet bahasa Indonesia, Jawa untuk tweet bahasa Jawa, begitu pula untuk bahasa Sunda. Selain mengambil teks *tweet* saja dari rangkaian data JSON serta melabelinya sesuai bahasa, dilakukan juga beberapa aktivitas antara lain :

- mengubah teks menjadi bentuk *lowercase*-nya, kemudian
- menghilangkan url dan tanda baca
- menyimpan *file* ke ekstensi txt

Sebab menggunakan 3 bahasa, sampai pada tahap ini adalah 3 buah file .txt yang merepresentasikan dokumen-dokumen teks sesuai bahasanya (Indonesia, Sunda, Jawa).

Data yang disimpan ke ekstensi txt kemudian di cek secara manual oleh peneliti untuk menghilangkan data-data yang salah *tag* sampai data yang tidak ada teksnya karena kecenderungan ada campuran bahasa dan adanya kemungkinan hanya memposting url, sehingga teks yang disimpan tidak sesuai dengan tag bahasa yang seharusnya bahkan tidak ada teks sama sekali.

Setelah dilakukan pembersihan secara manual, maka semua data digabungkan ke dalam satu *file* berekstensi xlsx. Data inilah yang nantinya digunakan dalam proses pembentukan *data training*.

1.5 Hasil Pembentukan Model

Setelah terbentuknya *file* yang akan dijadikan *dataset,* maka selanjutnya data tersebut akan dibentuk menjadi sebuah model klasifikasi. Namun sebelum membentuk model, ada beberapa tahapan yang harus dilakukan agar terbentuknya suatu model yang baik. Yang pertama dilakukan adalah membaca *file* xlsx dan kemudian dilakukan tokenisasi terhadap seluruh dokumen dalam *file* tersebut.

Data yang sebelumnya berupa kalimat dipisah menjadi kata-kata yang nantinya akan diproses untuk menjadi *word vector.*Ketika dilakukan proses tokenisai, data label disimpan kedalam variabel yang berbeda, namun dengan urutan dokumen yang sama agar tidak terjadi kesalahan dalam membandingkan *training* dan *test* nantinya.

Data yang sudah menjadi *word vector* kemudian dihitung menggunakan rumus N-Gram dan TFIDF sehingga menghasilkan *word vector* dengan nilai yang sudah terbobot.

Setelah terbentuknya data yang terbobot, maka kemudian data di *training* dengan Algoritma Naive Bayes, apabila tidak ada error atau kesalahan maka terbentuklah model. Kalimat yang akan dicoba ujikan kepada model harus dirubah terlebih dahulu ke dalam bentuk *word vector* sesuai dengan kata yang ada di model. Kemudian baru lah dihitung probabilitas antara vektor data test dengan vector model.

Dari perhitungan probabilitas antar kalimat terhadap setiap kelas, maka barulah bisa mendapatkan hasil jekas prediksi dari data yang dimasukkan. Setelah semua proses itu dilakukan, barulah bisa

menghitung performa dari algoritma yang digunakan.

1.6 Performa Algoritma

Untuk mengetahui performa dari Algoritma Naive Bayes, maka dilakukan pengujian evaluasi terhadap sistem. Pengujian sistem menggunakan *cross-validation*sebanyak 10 kali dan kemudian dihitung nilai *accuracy, precision,* dan *recall*.

Untuk nilai *accuracy*, model ini menghasilkan nilai secara berurutan sebagai berikut:

Accuracy: 0.84848485; 0.73282443; 0.75572519; 0.77099237; 0.83846154; 0.72307692;

0.80769231; 0.77692308; 0.82307692; 0.78294574;

Sehingga dapat dilihat, nilai terbaik dari 10 kali perhitungan adalah **0.84615385** dan rata-rata dari semua nilai *accuracy* adalah sebesar**0.78602**.

Seperti pada perhitungan *accuracy,* pada proses ini juga membutuhkan parameter berupa label dan pred. Label berisikan kelas asli dari data yang diujikan seperti pada penjelasan *accuracy,* sedangkan pred adalah kelas hasil prediksi kalimat menggunakan model. Sehingga bisa menghasilkan nilai yang tertera pada tabel berikut:

Tabel 4.1 Tabel nilai *precision* dan *recall*

	Precision	***Recall***
Indonesia	1.00	0.98
Jawa	0.98	1.00
Sunda	1.00	0.99
Rata-rata	0.99	0.99

Dari hasil diatas dapat dilihat bahwa nilai *accuracy* bernilai kurang lebih **0.8** dan nilai *precision* dan *recall* **0.99**. Maka,nilai itu bisa dianggap nilai yang baik dalam pembuatan model.

V. KESIMPULAN

Berdasarkan rumusan masalah yang sudah dibahas pada bab pertama pada penelitian ini, yakni bagaimana mengetahui bahasa daerah apa yang digunakan dalam penulisan tweet, maka dapat disimpulkan bahwa bahasa daerah dapat diketahui (dideteksi) menggunakan metode klasifikasi pada pengolahan bahasa alami. Pendeteksi bisa mengetahui bahasa yang digunakan menggunakan sistem yang dibuat dari *dataset* yang telah diolah terlebih dahulu dengan beberapa tahapan.

Tahapan pembuatan sistem dimulai dengan pengambilan data JSON menggunakan Twitter API dan kemudian data teks dari data yang didapat diambil dan dibersihkan dari tanda baca, url, dan nama akun yang tertulis pada teks tersebut. Setelah bersih dari tanda baca, data dirubah ke bentuk *word vector* yang kemudian diberikan pembobotan menggunakan Unigram dan TF-IDF. Setelah terbentuk matriks kata dengan bobot, kemudian diolah menggunakan Naive Bayes sehingga menghasilkan model yang bisa digunakan untuk mendeteksi bahasa.

Dengan terbentuknya model, maka dapat digunakan dalam memenuhi tujuan dibuatnya penelitian, yaitu terbentuknya model pendeteksi bahasa daerah secara otomatis. Sehingga dapat memberikan banyak manfaat yang beberapa diantaranya:

- Menambah wawasan tentang bahasa daerah.
- Mengetahui bahasa daerah yang sering digunakan di sosial media.
- Mengenalkan bahasa daerah kepada generasi muda secara tidak langsung melestarikan bahasa daerah

VI.DAFTAR PUSTAKA

DBpedia.*About: Tokenisasi*. Retrieved from DBpedia: http://id.dbpedia.org/page/Tokenisasi

Departemen Pendidikan Nasional; Pusat Bahasa (Indonesia). (2008). *Kamus besar bahasa Indonesia Pusat Bahasa.* Jakarta: Gramedia Pustaka Utama.

Hidayatullah, A. F. (2014). ANALISIS SENTIMEN DAN KLASIFIKASI KATEGORI TERHADAP TOKOH PUBLIK PADA TWITTER. *Seminar Nasional Informatika 2014 (semnasIF 2014).* Yogyakarta.

Kaswidjanti, W., Aribowo, A. S., & Wicaksono, C. B. (2014). Implementasi Fuzzy Inference System Metode Tsukamoto Pada Pengambilan Keputusan Pemberian Kredit Pemilikan rumah. *Telematika*, 137-146.

KOMINFO. (2013, 11 07). *Berita Kementrian*. Retrieved from KOMINFO: https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker

Kumar, E. (2011). *NATURAL LANGUAGE PROCESSING.* New Delhi: LK. Intrnational Publishing House Pvt, Ltd.

Kusumadewi, S., & Purnomo, H. (2004). *Aplikasi logika fuzzy untuk mendukung keputusan.* Yogyakarta: Graha ilmu.

Mayangningsih, Siswanto, & Mesterjon. (2013). Metode Logika Fuzzy Tsukamoto Dalam Sistem Pengambilan Keputusan Penerimaan Beasiswa. *Jurnal Media Infotama*, 140-165.

Patji, A. R. (2016, Agustus 05). *lipimedia*. Retrieved from Lembaga Ilmu Pengetahuan Indonesia (LIPI): http://lipi.go.id/lipimedia/139-bahasa-daerah-di-indonesia-terancam-punah/15938

Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

Sholihin, M., Fuad, N., & Khamiliyah, N. (2013). Sistem Pendukung Keputusan Penentuan Warga Penerima Jamkesmas Dengan Metode Fuzzy Tsukamoto. *Jurnal Teknika*, 501-505.

Zhang, J. (2008). *Visualization for Information Retrieval.* Heidelberg: Spinger-Verlag.