KLASIFIKASI BAHASA DAERAH DENGAN N-GRAM MODELING LANGUAGE PADA TWITTER

DISUSUN OLEH:

KHAIRANI

11551202703

JURUSAN TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UIN SUSKA RIAU

2018

KATA PENGHANTAR

Assalamu'alaikum Warohmatullohi Wabarokatuh. Segala puji dan rasa syukur hanya penulis panjatkan ke hadirat Allah subhanahu wa ta'ala, yang telah melimpahkan segala kemudahannya hingga akhirnya penulis mampu menyelesaikan laporan besar ini tepat waktu. Laporan Tugas Besar ini disusun untuk memenuhi sebagian persyaratan memperoleh nilai di Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau. Dalam pelaksanaan pembuatan laporan besar yang didalamnya termasuk kegiatan pembuatan laporan ini, penulis mendapat banyak bantuan dari berbagai pihak. Tanpa bantuan Allah subhanahuwa ta'ala melalui tangan mereka niscaya laporan besar ini tidak akan berjalan dengan lancar. Untuk itu dalam beberapa lembar kertas yang mungkin tiada berarti ini penulis sampaikan rasa hormat dan menghaturkan rasa terima kasih kepada:

1. Bapak Prof. Dr. H. Akhmad Muhahiddin, MA selaku Rektor Universitas Islam Negeri Sultan Syarif Kasim Riau.
2. Bapak Dr. Hartono, M.Pd selaku Dekan Fakultas Sains dan Teknologi Universitas Negeri Sultan Syarif Kasim Riau.
3. Bapak M. Irsyad, MT selaku Ketua Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau.
4. Ibu Yusra,ST selaku Dosen dalam Mata Kuliah Natural Language Processing Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau.
5. Orang tua yang tetap memberikan semangat kepada penulis untuk menyelesaikan laporan besar ini.
6. Teman-teman yang telah mendorong penulis untuk tetap memberikan semangat kepeda penulis ketika mental penulis jatuh.
7. Terimakasih kepada semua pihak yang membantu dalam penyelesaikan laporan besar ini untuk bidang studi Natural Language Processing yang ada di Universitas Islam Negeri Sultan Syarif Kasim Riau.

Bukan lagi rahasia, sebuah karya selalu disertai kekurangannya, oleh sebab itu penulis memohon kelapangan hati pembaca sekalian untuk menerima kekurangan yang ada dalam laporan besar ini. Semoga bermanfaat untuk kita semua. Wassalamu'alaykum Warohmatullohi Wabarokatuh.

Pekanbaru, 28 Juni 2018

Penyusun

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar belakang

Manusia adalah sebuah entitas dalam kehidupan sosial, oleh karena itu manusia disebut juga sebagai makhluk sosial. Salah satu media yang digunakan untuk bersosialisasi adalah bahasa. Bahasa adalah sebuah sistem yang dibentuk oleh sejumlah komponen yang berpola secara tetap dan dapat di kaidahkan. Salah satu karakteristik bahasa adalah keberagaman suku bahasa karena bahasa itu digunakan oleh penutur yang heterogen yang mempunyai latar belakang sosial dan kebiasaan yang berbeda. Di Indonesia sendiri memiliki berbagai macam bahasa yang disebut sebagai bahasa daerah. Bahasa daerah digunakan hanya untuk berkomunikasi antar sesama daerah saja. Bahasa daerah hanya dimengerti oleh sesama daerah saja maka dari itu dibutuhkan sebuah klasifikasi data.

Klasifikasi bahasa daerah merupakan permasalahan yang mendasar dan penting. Setiap bahasa daerah memiliki makna yang terkandung di dalam bahasa tersebut, yang merupakan inti dari suatu bahasa yang kompleks dan jumlah kata yang sangat banyak. Oleh karena itu, permasalahan ini merupakan masalah yang cukup kompleks dikarenakan penggunaan kata yang tergolong tidak sedikit, sehingga perlu mengetahui inti dari makna setiap bahasa daerah tersebut. Salah satu dari beberapa metode yang bisa digunakan dalam tujuan untuk mengklasifikasi bahasa daerah adalah menggunakan metode N-gram yang menyangkut *Unigram*, *Bigram*, dan *Trigram*.

Metode N-gram menggunakan frekuensi yang menunjukkan kemunculan apakah teks tersebut menggunakan bahasa daerah tertentu. Salah satu keunggulan menggunakan N-gram adalah bahwa N-gram tidak akan terlalu sensitif terhadap

kesalahan penulis yang terdapat pada suatu teks tersebut. (Ahmad Hanafi et all., 2009)

Penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem klasifikasi bahasa daerah yaitu bahasa Minang, Bahasa Jawa, dan Bahasa Sunda. Tujuan dari penelitian ini adalah untuk mempermudah menentukan bahasa daerah pada sebuah media sosial yaitu Twitter dan juga bertujuan untuk menentukan metode N-gram manakah yang terbaik untuk menentukan bahasa daerah tersebut.

1.2 Rumusan Masalah

Masalah yang diteliti adalah sebagai berikut:

1. Bagaimana mendeteksi bahasa daerah pada chattingan di twitter menggunakan metode N-gram.
2. Bagaimana akurasi dari sistem yang akan dibuat dalam melakukan pendeteksi bahasa daerah.

1.3 Batasan Masalah

Batasan masalah yang digunakan dalam menyelesaikan permasalahan klasifikasi bahasa daerah dengan N-gram, yaitu sebagai berikut:

1. Bahasa yang digunakan sebagai sampel acuan dalam proses identifikasi bahasa yaitu bahasa yang ada pada chattingan di media sosial yaitu twitter.
2. Bahasa yang digunakan hanya berupa huruf.
3. Tanda baca dan angka akan diabaikan.
4. Banyaknya bahasa daerah yang digunakan hanya tiga bahasa (Bahasa Minang, Bahasa Melayu, dan Bahasa Jawa.

1.4 Tujuan Penelitian

Melalui penelitian ini tujuan yang ingin dicapai penulis adalah membuat sistem pendeteksi bahasa daerah menggunakan metode N-gram.

BAB II
LANDASAN TEORI

2.1 Bahasa Daerah

Bahasa adalah alat atau sarana untuk berkomunikasi antar manusia yang berupa sistem pertukaran informasi dengan lambang bunyi yang dihasilkan dari alat ucap manusia. Bahasa disusun dari ribuan kata yang masing-masing memiliki makna berbeda. Di Indonesia sendiri kaya akan bahasa dan adat istiadat. Indonesia memiliki ribuan pulau dan kebanyakan pulau tersebut memiliki bahasa tersendiri. Bahasa setiap daerah di Indonesia di sebut dengan bahasa daerah.

Bahasa daerah adalah bahasa yang digunakan dalam satu wilayah di sebuah negara kebangsaan dan digunakan dalam berkomunikasi sehari-hari pada suatu daerah. Bahasa daerah sering diperumpamakan sebagai jati diri daerah tersebut. Dari sekian banyak bahasa yang ada di Indonesia tetapi penelitian ini hanya meneliti tiga bahasa daerah yaitu bahasa Minang, Bahasa Jawa, dan Bahasa Sunda.

2.1.1 **Bahasa Minang**

Bahasa Minang adalah bahasa dari suku Minangkabau yang berasal dari Sumatera Barat. Bahasa Minang termasuk bahasa dari rumpun bahasa Melayu. Bahasa Minang bisa ditemukan dibanyak tempat seperti di Sumatera Barat, pantai barat Aceh, barat Riau, pantai barat Sumatera Utara, dan wilayah-wilayah lain di sekitar Sumatera Barat. Bahasa Minang sangat kental dengan dialek Melayu, namun bahasa Minang bukanlah bahasa melayu melainkan salah satu cabang dari bahasa dialek melayu itu sendiri.

2.1.2 **Bahasa Jawa**

Bahasa Jawa adalah bahasa tradisional yang dituturkan oleh suku Jawa yang banyak tinggal di Jawa Tengah, Daerah Istimewa Yogyakarta, Jawa Timur, dan beberapa daerah Jawa Barat. Bahasa Jawa adalah bahasa daerah yang paling banyak penyebarannya di Indonesia maupun di luar Indonesia karena persebaran penduduk suku Jawa sangat luas hampir seluruh Indonesia.

2.1.3 **Bahasa Sunda**

Bahasa Sunda adalah bahasa tradisional dari suku Sunda yang populasinya berada di Jawa Barat, Jakarta, Banten, dan sedikit di wilayah Jawa Barat. Bahasa Sunda termasuk di dalam rumpun bahasa Melayu Polinesia dan merupakan salah satu cabang dari bahasa Austronesia.

2.2 Twitter

Twitter adalah sebuah media sosial dan layanan *microblogging* yang mengizinkan penggunanya untuk mengirimkan pesan *realtime*. Pesan ini populer dengan sebutan *tweet. Tweet* adalah sebuah pesan pendek dengan panjang karakter yang dibatasi hanya sampai 140 karakter, karena keterbatasan karakter yang bisa dituliskan sebuah *tweet* seringkali mengandung singkatan. Twitter memang dibuat sebagai layanan berbasis mobile yang didesain sesuai dengan batasan karakter pada sebuah pesan teks seperti SMS dan sampai saat ini twitter bisa digunakan pada setiap *telephone* genggam yang memiliki kemampuan untuk mengirim dan menerima pesan (The Twitter Government and Election Team, 2014).

Twitter diciptakan untuk menjadi tempat saling bertukar informasi dengan orang yang berada di seluruh dunia. Menggunakan twitter bisa mengikuti tren terbaru, berita, dan informasi dari penjuru dunia. Ketika penggunanya mengirimkan *tweet*, pesan tersebut bersifat publik dan bisa diakses oleh siapapun. Bahkan, orang yang mengikuti *follow* orang tersebut bisa melihat *tweet* yang kita kirim pada halaman lini. (Nurrun Muchammad Shiddieqy Hadna et all., 2016)

Berikut ini adalah istilah yang dikenal pada twitter :

1. *Mention (@)*

   Mention menyebut atau memanggil pengguna twitter lain dalam sebuah *tweet.*

2. *Hashtag* (#)

   Hashtag digunakan untuk menandai sebuah topik pembicaraan di twitter dan juga untuk meningkatan vasibilitas *tweet* pengguna.

3. *Emoticon*

   Emoticon adalah ekespresi wajah yang dipresentasikan dengan kombinasi antara huruf, tanda baca, dan angka. Pengguna menggunakan emoticon adalah untuk mengekpresikan mood yang sedang mereka rasakan.

4. *Trending Topic*

   Trending Topic adalah kumpulan dari topik pembicaraan di twitter.

Sebuah fakta menunjukkan bahwa media twitter memiliki pengaruh yang signifikan terhadap pengurangan ketidakpastian informasi (Nila Nur Apriliani et all., 2015). Bahwa pada dasarnya sebagian orang banyak membuat *tweet* berupa opini atau berita hoax.

2.3 Metode N-gram

N-gram adalah potongan N-karakter yang diambil dari suatu string. Mendapatkan N-gram yang utuh ditempuh dengan menambahkan blank pada awal dan akhir string, misalnya suatu string "TEKS" setelah ditambah aal dan akhir dengan "_" sebagai pengganti blank akan didapat N-gram sebagai berikut:

Unigram :T,E,K,S

Bigram : _T,TE,EK,KS, dan T

Trigram : _TE,TEK,EKS,KS_ dan T

Dapat disimpulkan bahwa untuk string berukuran n akan dimiliki n unigram dan n+1 bigram, n+1 trigram, dan seterusnya. Penggunaan N-gram untuk *matching* kata memiliki keuntungan sehingga dapat diterapkan pada *recovery* pada input karakter ASCII yang terkena noise, interpretasi kode pos, *information retrieval* dan berbagai aplikasi dalam pemrosesan bahasa alami.

Keuntungan N-gram dalam *matching* string adalah berdasarkan karakteristik N-gram sebagai bagian dari suatu string sehingga kesalahan pada sebagian string hanya akan berakibat perbedaan pada sebagian N-gram. Jika, N-gram dari dua string dibandingkan kemudian kita menghitung cacah N-gram yang sama dari dua string

tersebut maka akan didapatkan nilai similaritas atau kemiripan dua string tersebut yang bersifat resistan terhadap kesalahan tekstual.

Kemiripan antara JOKO dan JOKI, disini ada perbedaan satu huruf maka dapat diukur derajat kesamaan dengan cara menghitung berapa buah N-gram yang diambil dari dua kata tersebut yang bernilai sama, yaitu:

JOKO : _J,JO,OK,KO,O_

JOKI : _J,JO.OK,KI,,I_

Disini terdapat tiga persamaan

Sementara antara kata JOKO dengan JONI, disini ada perbedaan dua huruf, nilai kesamaannya yaitu:

JOKO : _J,JO,OK,KO,O_

JONI : _J,JO,ON,NI,O_

Disini terdapat dua persamaan.

Sehingga dapat disimpulkan bahwa JOKO-JOKI-lah yang ada kemiripan atau kesamaan karena JOKO-JOKI mendapatkan lebih banyak persamaan yaitu tiga persamaan sedangkan JOKO-JONI hanya memiliki dua persamaan.

Keunggulan menggunakan N-gram adalah bahwa N-gram tidak akan terlalu sensitif terhadap keslahan penulisan yang terdapat pada suatu teks *chattingan* tersebut. Karakteristik N-gram yaitu dapat berfungsi dengan baik walaupun terdapat kesalahan tekstual dan dapat berjalan secara efisien, membutuhkan penyimpanan yang sederhana dan waktu proses yang cepat (Shoffan Azizurahman, 2011).

2.3.1 **Klasifikasi Bahasa dengan N-gram**

Penggunaan N-gram untuk mengklasifikasi bahasa didasarkan pada anggapan bahwa pola sebaran N-gram dari suatu bahasa bersifat unik karena ini terkait dengan frekuensi penggunaan huruf atau pasangan huruf baik itu vokal atau konsonan suatu

bahasa yang umumnya berbeda dengan bahasa yang lain. Unigram yang jika dihitung frekuensinya adalah frekuensi kemunculan huruf dalam teks bahasa tertentu yang akan berbeda.

BAB III
METODE PENELITIAN

3.1 Identifikasi Masalah

Masalah yang dihadapi dalam tugas saya ini adalah bagaimana cara mendeteksi bahasa daerah yang begitu banyak dengan metode N-gram. Apakah aplikasi ini nantinya bisa berjalan dan digunakan untuk mengklasifikasikan bahasa daerah pada media sosial yaitu twitter.

3.2 Pengumpulan Data

Pengumpulan data merupakan sebuah cara bagaimana saya mendapatkan bahan-bahan yang diperlukan agar aplikasi ini berjalan. Pengumpulan data dilakukan dengan mendeteksi langsung apakah bahasa daerah yang digunakan oleh orang yang mengirimkan *tweet* pada twitter.

3.3 Studi Pustaka

Studi Pustaka yang saya lakukan adalah dengan membaca dan memahami dari teori-teori jurnal yang saya temui dari internet dan juga dari buku-buku yang saya temui di pustaka walaupun buku matakuliah lain. Adapun beberapa jurnal yang kami temui sampai saat ini.

No	Nama Pengarang	Bahasan Penelitian
1	Ahmad Hanafi, Rimba Whidiana, Retno Novi Dayawati	Pengenalan Bahasa Suku Bangsa Berbasis Teks Menggunakan Metode N-gram
2	Nurrun Muchammad Shiddieqy Hadna, Paulus Insap Santosa, Wing Wahyu Winarni	Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter
3.	Nila Nur Apriliani, Yuliani Rahma Putri, Dini Salmiyah Fithrah Ali	Pengaruh Penggunaan Media Twitter Terhadap Pengurangan Ketidakpastian Informasi
4	Shoffa Azizurahman, Yanuar Firdaus A.w, Ari Ardiyanti Suryani	Analisis dan Implementasi Metode N-gram Pada Informasi Retrieval

DAFTAR PUSTAKA

Hanafi.Ahmad, Rimba.Whidiana, D.N.Retno, (2009), Pengenalan Bahasa Suku Bangsa Berbasis Teks Menggunakan Metode N-gram.

H.S.M.Nurrun, S.I.Paulus, W.W.Wing, (2016), Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter.

A.N.Nilla, P.R.Yuliani, A.F.S.Dini, (2015), Pengaruh Penggunaan Media Twitter Terhadap Pengurangan Ketidakpastian Informasi.

Azizurahman.Shoffa, A.W.F.Yanuar, S.A.Ari, (2011), Analisis dan Implementasi Metode N-gram Pada Informasi Retrieval.