

Penerapan Metode Textrank dalam Rancangan Aplikasi Silogisme Artikel Bahasa Batak

Martha Munthe

Fakultas Ilmu Komputer dan Teknologi Informasi, Program Studi Teknik Informatika, Universitas Budi Darma, Medan, Indonesia

Email: marthamunthe16@gmail.com

Email Penulis Korespondensi: marthamunthe16@gmail.com

Abstrak—Berkomunikasi satu sama lain merupakan salah satu sifat dasar manusia. Komunikasi berfungsi sebagai sarana untuk saling berinteraksi satu sama lain. Manusia terkadang ingin menguasai dari satu bahasa asing. Apalagi jika bahasa itu sedang populer. Manusia pun tidak ingin ketinggalan zaman menguasai bahasa tersebut. Dengan adanya kamus maka manusia akan dapat mempelajari bahasa serta mengetahui bahasa apa saja yang ingin dikuasai. Selama ini jika seseorang ingin mempelajari bahasa asing, salah satu contohnya adalah bahasa Korea, media yang digunakan untuk memperlancar penguasaan kosakatanya adalah melalui kamus. Oleh karena kamus bentuknya masih berupa seperti buku, tidak fleksibel, susah untuk dibawa, mengakibatkan kurang efektif apabila sedang berada di luar atau di jalan. Diperlukan sebuah media baru yang praktis dan efektif seperti media *smartphone* karena *smartphone* mendukung berbagai macam aplikasi seperti aplikasi kamus. Aplikasi kamus di dalam *smartphone* tidak mencari kata secara manual. Tetapi adanya fasilitas pencarian pada aplikasi kamus tersebut mempermudah user mendapatkan kata yang ingin dicarinya. Dengan memanfaatkan salah satu dari algoritma string matching seperti Algoritma *text rank*, maka mempermudah dan mempercepat pencarian kata dalam kamus. Dijelaskan bahwa Algoritma *textrank* mulai mencocokkan pattern pada awal teks. Pencocokan kata dimulai dari posisi kiri hingga akhirnya sampai pada posisi paling kanan, algoritma ini akan mencocokkan karakter per karakter pattern dengan karakter di teks yang bersesuaian, sampai salah satu kondisi berikut dipenuhi yaitu karakter di pattern dan di teks yang dibandingkan tidak cocok (*mismatch*) dan semua karakter di pattern cocok, kemudian algoritma akan memberitahukan penemuan di posisi ini.

Kata Kunci: Text Mining, Silogisme, Bahasa Batak, Metode Textrank

Abstract—Communicating with each other is one of the basic human traits. Communication serves as a means of interacting with one another. Humans sometimes want to master one foreign language. Especially if that language is popular. Humans also don't want to be outdated in mastering the language. dictionary then humans will be able to learn the language and know what language they want to master. So far, if someone wants to learn a foreign language, one example is Korean, the medium used to facilitate vocabulary mastery is through a dictionary. Because the dictionary form is still in the form of like books, inflexible, difficult to carry, resulting in less effective when out or on the road. A practical and effective new media such as *smartphone* media is needed because *smartphones* support various applications such as dictionary applications. The dictionary application on the *smartphone* does not search for words manually. But the search facility in the dictionary application makes it easier for users to get the words they want to look for. one of the string matching algorithms such as the *text rank* algorithm, it makes it easier and faster to search for words in the dictionary. It was explained that the *Textrank* Algorithm begins to match the pattern at the beginning of the text. Word matching starts from the left position until finally reaches the rightmost position, this algorithm will match the character of the pattern character with the characters in the corresponding text, until one of the following conditions is met, namely the character in the pattern and in the text being compared does not match (*mismatch*) and all the characters in the pattern match, then the algorithm will notify the discovery at this position.

Keywords: Text Mining, Syllogism, Batak Language, Textrank Method.

1. PENDAHULUAN

Berkomunikasi satu sama lain merupakan salah satu sifat dasar manusia. Komunikasi berfungsi sebagai sarana untuk saling berinteraksi satu sama lain. Manusia terkadang ingin menguasai dari satu bahasa asing. Apalagi jika bahasa itu sedang populer. Manusia pun tidak ingin ketinggalan zaman menguasai bahasa tersebut. Dengan adanya kamus maka manusia akan dapat mempelajari bahasa serta mengetahui bahasa apa saja yang ingin dikuasai. Selama ini jika seseorang ingin mempelajari bahasa asing, salah satu contohnya adalah bahasa Korea, media yang digunakan untuk memperlancar penguasaan kosakatanya adalah melalui kamus. Oleh karena kamus bentuknya masih berupa seperti buku, tidak *fleksibel*, susah untuk dibawa, mengakibatkan kurang efektif apabila sedang berada di luar atau di jalan.

Diperlukan sebuah media baru yang praktis dan *efektif* seperti media *smartphone* karena *smartphone* mendukung berbagai macam aplikasi seperti aplikasi kamus. Aplikasi kamus di dalam *smartphone* tidak mencari kata secara manual. Tetapi adanya fasilitas pencarian pada aplikasi kamus tersebut mempermudah user mendapatkan kata yang ingin dicarinya. Dengan memanfaatkan salah satu dari algoritma string matching seperti Algoritma *text rank*, maka mempermudah dan mempercepat pencarian kata dalam kamus.

Dijelaskan bahwa Algoritma *textrank* mulai mencocokkan pattern pada awal teks. Pencocokan kata dimulai dari posisi kiri hingga akhirnya sampai pada posisi paling kanan, algoritma ini akan mencocokkan karakter per karakter pattern dengan karakter di teks yang bersesuaian, sampai salah satu kondisi berikut dipenuhi yaitu karakter di pattern dan di teks yang dibandingkan tidak cocok (*mismatch*) dan semua karakter di pattern cocok, kemudian algoritma akan memberitahukan penemuan di posisi ini.

2. METODOLOGI PENELITIAN

2.1 TextRank

TextRank merupakan suatu peringkasan atau pemeringkatan teks yang sangat memudahkan bagi setiap orang yang ingin menggunakannya baik dalam dokumen dan sebagainya. Dimana kebutuhan akan informasi di saat ini sangat penting dan kebutuhan akan informasi tersebut dapat diperoleh dari berbagai media, seperti halnya media digital. Untuk dapat mengetahui informasi penting dari suatu dokumen atau teks, pembaca harus banyak meluangkan waktu apalagi ketika sedang membaca kitab suci, pembaca pasti akan banyak menemukan kata yang sering muncul pada setiap ayat [2].

2.2 Silogisme

Merupakan suatu proses penarikan kesimpulan secara *deduktif*. Dalam pengertian umum, silogisme adalah suatu argument *deduktif* yang terdiri dari dua premis dan satu kesimpulan. Kesimpulan itu berhubungan erat sekali dengan premis-premisnya benar maka kesimpulannya juga benar. *Silogisme* disusun dari dua proposisi dan sebuah konklusi (kesimpulan) *silogisme* jenis penalaran deduksi secara tidak langsung. *Silogisme* merupakan penemuan terbesar dari ahli filsafat terkenal yaitu Aristoteles.

3. HASIL DAN PEMBAHASAN

Berkomunikasi satu sama lain merupakan salah satu sifat dasar manusia. Komunikasi berfungsi sebagai sarana untuk saling berinteraksi satu sama lain. Manusia terkadang ingin menguasai dari satu bahasa asing. Apalagi jika bahasa itu sedang populer. Manusia pun tidak ingin ketinggalan zaman menguasai bahasa tersebut. Dengan adanya kamus maka manusia akan dapat mempelajari bahasa serta mengetahui bahasa apa saja yang ingin dikuasai. Selama ini jika seseorang ingin mempelajari bahasa asing, salah satu contohnya adalah bahasa Korea, media yang digunakan untuk memperlancar penguasaan kosakatanya adalah melalui kamus. Oleh karena kamus bentuknya masih berupa seperti buku, tidak *fleksibel*, susah untuk dibawa, mengakibatkan kurang efektif apabila sedang berada di luar atau di jalan.

Diperlukan sebuah media baru yang praktis dan *efektif* seperti media *smartphone* karena *smartphone* mendukung berbagai macam aplikasi seperti aplikasi kamus. Aplikasi kamus di dalam *smartphone* tidak mencari kata secara manual. Tetapi adanya fasilitas pencarian pada aplikasi kamus tersebut mempermudah user mendapatkan kata yang ingin dicarinya. Dengan memanfaatkan salah satu dari algoritma string matching seperti Algoritma *textrank*, maka mempermudah dan mempercepat pencarian kata dalam kamus.

Dijelaskan bahwa Algoritma *textrank* mulai mencocokkan pattern pada awal teks. Pencocokan kata dimulai dari posisi kiri hingga akhirnya sampai pada posisi paling kanan, algoritma ini akan mencocokkan karakter per karakter pattern dengan karakter di teks yang bersesuaian, sampai salah satu kondisi berikut dipenuhi yaitu karakter di pattern dan di teks yang dibandingkan tidak cocok (*mismatch*) dan semua karakter di pattern cocok, kemudian algoritma akan memberitahukan penemuan di posisi ini.

Penelitian ini mempunyai beberapa tahapan yaitu, *Text processing* (Tokenisasi, pembuangan, *stopword*, dan proses *stemming*), pembobotan kata (*term*), dan mengelompokkannya dengan menghitung nilai similaritas *termnya* dengan data yang ada pada *data resource*.

3.1 Penerapan Algoritma TextRank

Pada bagian ini sinopsis yang sudah diinput akan melalui proses persiapan dan seleksi dokumen dimana tahapan ini bertujuan untuk mempersiapkan text menjadi data yang akan mengalami pengolahan menjadi data untuk proses penentuan genre. Pada bagian ini terdapat beberapa tahapan yang akan dilakukan yaitu: tokenisasi (*tokenization*), pembuangan *stopword* (*stopword removal*) dan yang terakhir proses *stemming*.

1. Tokenisasi

Sebelum kata dipisahkan dari kalimat, terlebih dahulu dibersihkan dari tanda baca, tag html, dan angka. Proses ini dilakukan sebelum proses tokenisasi dilakukan agar bisa memperkecil hasil dari tokenisasi tersebut. Dalam proses tokenisasi dokumen yang akan dibaca berupa teks yang kemudian akan dilakukan proses pemotongan *string* input berdasarkan setiap kata yang menyusunnya. Secara umum setiap kata akan terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenisasi mengandalkan karakter spasi pada dokumen teks tersebut untuk melakukan pemisahan kata.

2. Pembuangan Stopword

Sebelum tahap pembuangan *stopword* dilakukan terlebih dahulu dilakukan normalisasi dengan mengubah semua huruf kapital menjadi huruf kecil. Proses pembuangan *stopword* merupakan proses pembuangan *term* yang tidak memiliki arti atau relevan. *Term* tersebut di peroleh setelah tahap tokenisasi dilakukan, kemudian dicek ke dalam daftar *stopword*, jika kata tersebut masuk ke dalam daftar *stopword* maka kata tersebut tidak akan diproses lebih lanjut. Sedangkan jika sebuah kata tidak termasuk ke dalam daftar *stopword* maka kata tersebut akan masuk ke proses berikutnya.

Stopword list diperoleh dari daftar *stopword* dari tala dengan memproses data set ke dalam daftar *stopword* tala, setiap kata yang ada di dalam data set akan dicek ke dalam daftar *stopword* tala, jika kata tersebut ada di dalam daftar *stopword* tala maka kata tersebut akan dibuang dan tidak diproses lagi ke tahap selanjutnya. Tabel *Stopword list* yang diperoleh dapat dilihat pada tabel dan kata yang ada di dalam daftar *stopword* tala disebut sebagai hasil filtering dari

proses *stopword* yang akan diproses ke tahap selanjutnya. Tabel hasil filtering dari proses *stopword* dapat dilihat pada tabel 1.

Tabel 1. Stopword List

BAHASA INDONESIA	BAHASA BATAK
Tunggu	Peitte
Siapa	Ise
Kenapa	Maua
Lihat	Bereng
Makan	Mangan
Berjalan	Mardalan
Kemana	Tudia
Mengapa	Boasa
Teringat	Taringot
Sayang	Holong
Jangan	Unang

3. Pembobotan kata dengan TF-IDF

Proses pembobotan kata adalah proses pemberian nilai atau bobot ke sebuah kata berdasarkan kemunculannya pada suatu dokumen teks (Baeza-Yates et al, 1999). Pada proses sebelumnya atau proses *Text Processing* akan didapat kumpulan kata atau *term* yang akan dipresentasikan kedalam sebuah *terms* vector. *Term* vector suatu dokumen teks adalah *tuple* bobot semua *term* pada a. Nilai bobot sebuah *term* inilah yang nantinya akan mempresentasikan dokumen teks. Pada penelitian ini proses pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF).

Term Frequency-Inverse Document Frequency (TF-IDF) adalah suatu metode pembobotan kata dengan menghitung nilai TF dan juga menghitung kemunculan sebuah kata pada dokumen teks. Pada pembobotan ini, jika kemunculan *term* pada sebuah dokumen teks tinggi dan kemunculan *term* tersebut pada dokumen teks yang lain rendah, maka bobotnya akan semakin besar. Sedangkan jika kemunculan *term* pada dokumen teks lain tinggi, maka bobotnya akan semakin kecil. Tujuan penghitungan IDF adalah untuk mencari kata-kata yang benar-benar mempresentasikan dokumen teks pada suatu koleksi. Metode ini merupakan metode paling baik dalam perolehan informasi (Khodraet al, 2005). Adapun rumus dari TF-IDF dapat dilihat pada persamaan berikut (Salton, 1983).

$$tfidf(i,j) = tf(i,j) \times \log\left(\frac{N}{df(j)}\right) \quad (1)$$

Dengan $tf(i,j)$ adalah frekuensi kemunculan *term* j pada dokumen teks di $\in D^*$, dimana $i = 1,2,3,\dots,N$, $df(j)$ adalah frekuensi dokumen yang mengandung *term* j dari semua koleksi dokumen, dan N adalah jumlah seluruh dokumen yang ada di koleksi dokumen. Berdasarkan rumus diatas berapapun besarnya nilai $tf(i,j)$, apabila $N = df(j)$ maka akan didapatkan hasil 0 (nol) untuk perhitungan *idf*. Untuk itu dapat ditambahkan nilai 1 pada sisi *idf*, sehingga perhitungan untuk pembobotan dapat dilihat padarumus (2) berikut :

$$tfidf(i,j) = tf(i,j) \times \left(\log\left(\frac{N}{df(j)}\right) + 1\right) \quad (2)$$

Pada penelitian ini, proses penentuan genre dari komik dilakukan setelah melakukan pembobotan kata. Proses penentuan genre dilakukan dengan memetakan kata pada komik ke daftar kata yang mewakili dari genre yang ada di dalam *data resource*, kemudian dihitung nilai kemiripan kata yang di dapat dari proses TF-IDF dengan menjumlahkan bobot setiap kata yang sama dengan daftar kata pada *data resource*, dan kemudian akan di klasifikasikan tepat ke salah satu genre yang memiliki nilai kemiripan tertinggi atau terbesar. Dalam tahap ini beberapa proses untuk menggunakan metode TextRank dapat dilakukan beberapa langkah - langkah yaitu :

1. Tahap pertama

Melakukan *Text cleaning* yang bertujuan untuk menghilangkan setiap tanda baca yang tidak penting pada kata bahasa batak yang akan diringkas untuk mengetahui setiap kata yang sering muncul, tanda baca yang dimaksud ialah seperti tanda petik dua, petik satu dan lainnya kecuali tanda titik.

2. Tahap kedua

Setelah tanda baca yang tidak penting dihilangkan maka dilakukan *Case folding* yang bertujuan untuk menormalkan huruf atau mengubah setiap huruf besar menjadi huruf kecil.

3. Tahap ketiga

Selanjutnya dilakukan tahap *Tokenizing* yang bertujuan untuk membentuk satu-satu kalimat menjadi kalimat yang lebih jelas dan tepat.

4. Tahap keempat

Setelah dilakukannya *Tokenizing* maka selanjutnya dilakukan tahap *Stop word removal* dimana akan dilakukannya pemilihan kata pada kata bahasa batak yang tidak penting atau tidak perlu sehingga lebih mudah untuk menemukan kata yang sering muncul.

5. Tahap kelima

Setelah semua tahap dilakukan maka dilanjutkan dengan membuat sebuah *vertex* yang berupa kolom dan baris yang berisi setiap kalimat isi bahasa batak, *vertex* ini bertujuan untuk menghitung jumlah kata yang sama atau sering muncul pada setiap kolom dan baris pada *vertex*. Jumlah *vertex* ditentukan seberapa banyak kalimat atau paragraph yang ada pada text bahasa batak. Setelah semua kata yang terdapat pada *vertex* diketahui jumlahnya maka dilakukanlah penjumlahan dengan menggunakan rumus yang ada atau yang sudah ditentukan, dengan demikian jika hasil sudah diketahui maka dapat dilakukan sebuah peringkatan dimana, hasil yang lebih tinggi dan rendah dapat dikelompokkan dengan menggunakan atau membuat sebuah grafik / graf untuk mengetahui setiap jumlah tertinggi dan terendah. Setelah sudah dilakukannya peringkatan dan sudah mengetahui setiap jumlah kata tertinggi dan terendah maka selanjutnya dilakukanlah penomoran dengan menggunakan Silogisme, penomoran dilakukan pada setiap kata yang terbanyak berupa kata yang sama atau pun kata yang sering muncul.

Setelah yang sudah diketahui untuk melakukan beberapa tahapan ini atau tahap representasi graf dilakukan dengan menggunakan rumus perhitungan :

$$\text{Similarity } (Si, Sj) = \frac{|\{wk | wk \sum Si \& wk \sum Sj\}|}{\log(|Si|) + \log(|Sj|)}$$

$$\text{Similarity } (Si, Sj) = \frac{1}{\log(14) + \log(10)} \quad 1/14 = 0.07$$

Keterangan :

Wk = Jumlah kata (term) yang sama antara kalimat Si dan Sj.

Si = Panjang kalimat Si.

Sj = Panjang kalimat Sj.

Dengan tahap – tahap inilah akan dilakukannya proses untuk mempermudah pencarian dan penomoran pada hasil kata yang sering muncul dan yang sudah dihitung.

4. KESIMPULAN

Dari pembahasan pada bab sebelumnya penulis dapat membuat beberapa kesimpulan yang merupakan inti sari dari penelitian ini. aplikasi yang dirancang dapat menjadi suatu acuan untuk memecahkan permasalahan dalam memahami aplikasi tersebut. Adapun kesimpulan yang diperoleh sebagai berikut Pada jumlah kata dari artikel yang dibahas *Silogisme dalam Bahasa Batak* dengan menggunakan metode *Textrank*, menjelaskan cara membuat Silogisme dengan menggunakan aplikasi. Menerapkan metode *TextRank* untuk menyusun *Silogisme dalam Bhasa Batak* yang menjelaskan tahapan yang dilakukan dalam penerapan metode *Textrank* untuk membuat Silogisme dari artikel berbahasa batak.

REFERENCES

- [1] Hartman & Rokitta, "Plant Propagation and Principles Practice," 2013.
- [2] Mihalchea, "Bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing," 2004.
- [3] Arikunto, "Ilmu Sebuah Pengantar Populer," 2005.
- [4] uli kozok, "Warisan Leluhur Sastra Lama dan Aksara Batak," 1999.
- [5] kerap Gorys, "Sebuah Pengantar Kemahiran Bahasa," 2004.
- [6] Rosa A.S. M.Shalahuddin, *Rekayasa Perangkat Lunak*. Bandung, 2015.
- [7] Yuni Sugiarti, *Analisis dan Perancangan UML*. 2013.
- [8] H. wibowo Raditya, "Visual Basic Database (Indonesia Edition)," 2010.