

PENERAPAN METODE PROBABILISTIC PARSING SEBAGAI PENDETEKSI POLA KALIMAT HEADLINE COPYWRITING DALAM APLIKASI NATURAL LANGUAGE PROCESSING

Nahar Mardiyantoro

Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, UNSIQ
mardziyant@gmail.com

INFO ARTIKEL

Riwayat Artikel :

Diterima : 9 Desember 2017

Disetujui : 22 Desember 2017

Kata Kunci : Natural Language Processing, Probabilistic Parsing, Copywriting.

ABSTRAK

Dalam ilmu copywriting terdapat formula RTO untuk menyusun headline yang tepat. Sedangkan pada ilmu komputer terdapat pemrosesan bahasa alami atau lebih dikenal dengan Natural Language Processing (NLP) sebagai pengenalan tata bahasa. sehingga pada penelitian ini masalah yang diangkat adalah bagaimana menerapkan metode probabilistic parsing sebagai pendeteksi pola kalimat dalam headline copywriting. Penelitian ini diselesaikan dengan menggunakan metode pengembangan system waterfall dan algoritma yang diterapkan adalah CYK dengan tahapan mengidentifikasi jenis kata dan identifikasi pola kalimat. setelah aplikasi selesai dibangun, kemudian diuji dengan testcase uji blackbox yang menghasilkan 100% navigasi dan fungsi berjalan sesuai dengan rancangan.

ARTICLE INFO

Riwayat Artikel :

Received : December 9, 2017

Accepted : December 16, 2017

Key words: Natural Language Processing, Probabilistic Parsing, Copywriting.

ABSTRACT

In copywriting science there is an RTO formula for preparing appropriate headlines. While in computer science there is natural language processing or better known as Natural Language Processing (NLP) as the introduction of grammar. So in this study the problem raised is how to design and develop NLP applicationais to know the pattern of sentence in the headline copywriting with probabilistic method. This research was completed by using method of waterfall system waterfall and the algorithm applied was CYK with stages identifying word type and identification of sentence pattern. After the application was completed, then tested by blackbox testcase which yield 100% navigation and function goes according to the design.

1. PENDAHULUAN

a. Latar Belakang

Menurut pakar penjualan dewa eka prayoga, *marketing* adalah permainan persepsi. Bagaimana sebuah produk dapat dipersepsikan bagus di mata pasar. Ada produk yang sebenarnya bagus, tapi ketika *marketer* tidak mampu mengkomunikasikannya dengan baik, sehingga persepsi pasar terhadap produk tersebut jadi jelek dan imbasnya produk tidak laku. Adapula produk yang sebenarnya jelek, tapi ketika seorang *marketer* pandai mengkomunikasikannya dengan baik, persepsi pasar terhadap produk tersebut jadi bagus dan efeknya produk laku keras. Proses membuat persepsi melalui tulisan yang membuat pembaca memberikan respon sesuai dengan yang penulis inginkan itulah yang disebut dengan *copywriting*. Pada pembuatan *copywriting* khususnya kalimat *headline* terdapat formula yang dapat dijadikan panduan yaitu formula RTO (*Result-Time Frame-Objection*). Formula tersebut akan membagi bagian *headline* menjadi 3, yaitu bagian pertama *Result* menampilkan hasil akhir yang diharapkan, bagian kedua *Time Frame* menampilkan batas waktu terwujudnya hasil akhir, sedangkan bagian ketiga adalah *Objection* bagian dimana keberatan dimunculkan. Berikut contoh *headline* yang menerapkan *copywriting* formula RTO “Rumah anda akan terjual dalam 90 hari, atau saya beli sendiri”. Pada contoh tersebut, memperlihatkan bahwa pola yang disusun terdiri dari 3 bagian, yaitu : (1) Rumah anda akan terjual masuk dalam bagian *Result*; (2) dalam 90 hari masuk dalam bagian *Time Frame*; (3) saya beli sendiri masuk ke dalam bagian *Objection*. Dari contoh *headline* tersebut, ternyata dalam mengurai menggunakan formula RTO harus diketahui pola kalimatnya, maka itu dibutuhkan proses untuk mengetahui pola kalimat tersebut.

Pada ilmu komputer terdapat pemrosesan bahasa alami atau lebih dikenal dengan *Natural Language Processing* (NLP) sebagai

pengenalan tata bahasa. Penelitian dalam pengolahan bahasa alami telah berlangsung selama beberapa dekade dimulai pada tahun 1940-an. Mesin terjemahan (MT) adalah yang aplikasi pertama berbasis komputer yang berhubungan dengan bahasa alami. Sementara Weaver dan Booth, memulai proyek MT awal tahun 1946 pada terjemahan komputer untuk memecahkan kode musuh selama Perang Dunia II (Karen, 2011).

“ Pola tata bahasa direpresentasikan dalam suatu *tree* dengan menggunakan metode probabilistik (*probabilistic parsing*). Dalam metode *probabilistic parsing*, keberhasilan penguraian elemen pada pemrosesan bahasa alami akan ditentukan oleh algoritma yang digunakan (Jurafsky, 2000). Salah satu yang pernah digunakan dalam penelitian Daniel Jurafsky dan James H. Martin (2000) adalah algoritma CYK (Cocke, Younger, Kasami) yang termasuk pada jenis *parsing bottom-up* dan salah satu kesimpulan dari penelitian tersebut menyebutkan bahwa metode *probabilistic parsing* dengan algoritma CYK efektif dan efisien dalam mengurai elemen tata bahasa (Rosa, 2009). “

Penelitian yang dilakukan Ramon Lefuel dan Brian J. Ross (2004) menggabungkan penguraian probabilistik dengan algoritma genetika. Algoritma genetika digunakan untuk membangkitkan pohon pola tata bahasa dari kalimat masukan. Model yang digunakan pada penelitian ini adalah PCFG. *parse-tree* direpresentasikan dengan kromosom. Fungsi *fitness* yang digunakan adalah perhitungan probabilitas setiap *parse - tree*. Penelitian ini membuktikan bahwa “ algoritma genetika juga dapat digunakan untuk melakukan penguraian (*parsing*) probabilistik pada kalimat walaupun dari segi performansi dianggap masih kurang efisien (Lefuel, 2004). “

“ Penelitian mengenai pengurai yang pernah dilakukan oleh Eugene Charniak . Pengurai (*parser*) yang dibangun oleh Charniak (1997) adalah *parser* untuk bahasa Inggris dan menggunakan *treebank* (kumpulan pohon pola tata bahasa) untuk membangun *system parser* (Rossa, 2009). “

Penelitian Charniak ini sering disebut dengan *parser* (pengurai) menggunakan PCFG yang bersifat leksikal atau dari kamus. Setiap elemen kalimat dipilih berdasarkan grafik untuk menjadi simpul pohon. Algoritma adalah cara kerja dari algoritma *chart parser*.

Parser *bottom-up* adalah parser (pengurai) yang digunakan pada penelitian ini. Menurut Charniak, 1997 “ Setiap *word* pada kalimat akan dianggap sebagai daun pohon, dari setiap daun pohon itu akan disimpulkan apa jenis simpul *parentnya*, demikian terus keatas sampai ditemukan *word head*”.

“ Bonnie (2007) menyebutkan bahwa Algoritma CYK dapat digunakan untuk mengetahui apakah sebuah string dapat diterima oleh suatu *Context Free Grammar* (CFG). Proses pengurai string dengan algoritma CYK memanfaatkan struktur data sebuah *array* 2 dimensi dan merupakan terapan *Dynamic Programming* karena pengolahan pengurai menggunakan hasil pengurai sebelumnya untuk memutuskan apakah pengolahan yang sedang terlaksana *accepted* atau *unaccepted* “.

Algoritma CYK menggunakan tabel dua dimensi untuk mendahulukan penyimpanan keputusan permasalahan yang lebih kecil. Sisi *Dynamic Program* dari algoritma ini terletak pada pembangunan array dua dimensi atau tabel saat mengurai suatu *string*, kemudian ketika penguraian *string* dilakukan dalam iterasi selanjutnya, algoritma ini akan menggunakan larik atau tabel yang telah terbentuk sebelumnya. Dari tabel yang telah terbentuk, string yang diurai dapat diketahui kevalidannya, dengan artian CFG tersebut dapat menghasilkan string tersebut melalui aturan-aturan yang ada. Berikut adalah persyaratan yang dibangun dengan menerapkan CYK : (1) Input: string dengan simbol; (2) Output: tak valid/valid; (3) Struktur data: tabel $n \times n$; (4) Baris dengan indeks nol(0) sampai $n-1$ (atau $1-n$ dengan perubahan).

Berdasar uraian diatas, maka peneliti sangat tertarik menjadikan Aplikasi NLP dengan metode *probabilistic parsing* menggunakan algoritma CYK untuk mengetahui pola kalimat dalam *copywriting* sebagai tema penelitian ini.

b. Rumusan Masalah

Bagaimana menerapkan metode probabilistic parsing pada aplikasi NLP untuk mengetahui pola kalimat dalam copywriting?

c. Batasan Masalah

Pada penelitian ini jumlah kata tidak lebih dari 15 kata tanpa singkatan ber bahasa Indonesia.

d. Tujuan

Menerapkan metode *probabilistic parsing* pada aplikasi NLP untuk mengetahui pola kalimat dalam *copywriting*.

e. Manfaat

Kepada Masyarakat, penelitian ini menghasilkan produk aplikasi yang dapat digunakan untuk mengetahui apakah headline yang dibuat untuk promosi sudah memenuhi formula *copywriting* RTO atau belum.

Kepada Ilmu Pengetahuan dan Teknologi, penelitian ini memberikan kontribusi berupa khasanah pengetahuan dalam hal penerapan ilmu komputer dan kecerdasan buatan dalam bidang promosi dan *marketing*.

2. METODE

a. Objek Penelitian

Objek pada penelitian ini adalah kalimat *copywriting*. Kalimat *copywriting* yang digunakan adalah dalam bahasa Indonesia tanpa singkatan dengan batas maksimal 15 kata.

b. Materi Penelitian

Hardware yang digunakan adalah Notebook dengan spesifikasi *processor* core I5.

Software yang digunakan adalah system operasi windows 7 ultimate, web editor adobe dreamweaver, xampp.

c. Metode Pengumpulan Data

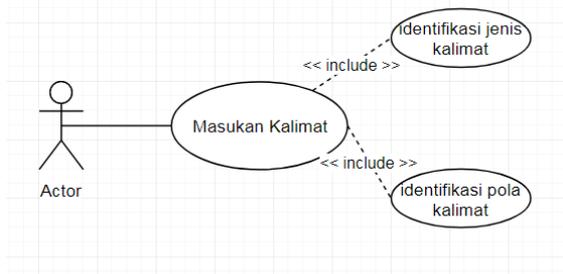
Pengumpulan data pada penelitian ini, peneliti menerapkan metode observasi terhadap kalimat *copywriting* yang lazim digunakan. Sedangkan untuk *parsing* kalimat, proses

steaming dan penelitian-penelitian terkait pemrosesan temu kembali informasi bahasa khususnya Indonesia dengan algoritma CYK parser menggunakan metode studi literatur.

d. Metode Pengembangan Sistem

Metode yang digunakan dalam pengembangan aplikasi NLP ini adalah waterfall. Dimana dengan menggunakan metode ini, aplikasi harus sampai ke tahap akhir terlebih dahulu baru setelah itu dilakukan evaluasi. Tahap ini sangat cocok untuk sebuah penelitian karena modelnya adalah *Linear Sequential*.

Business Process Identification : Use Case

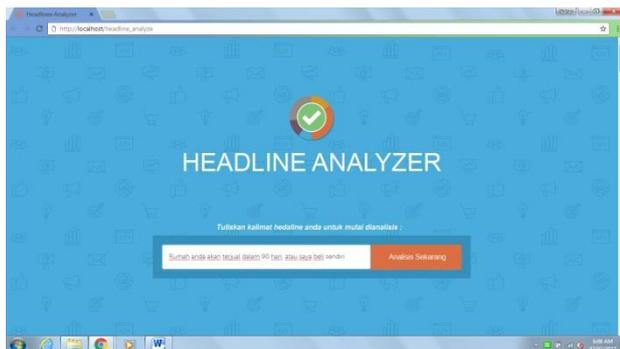


Gambar 2.1 Use Case

3. HASIL DAN PEMBAHASAN

3.1 Pengembangan Sistem

Aplikasi ini telah dijalankan menggunakan dua *browser* yaitu Google Chrome dan Firefox. Seperti gambar 3.1 dibawah, setelah pengguna mengisikan alamat pada Address Bar, akan tampil halaman depan dari aplikasi.



Gambar 3.1. Halaman Depan

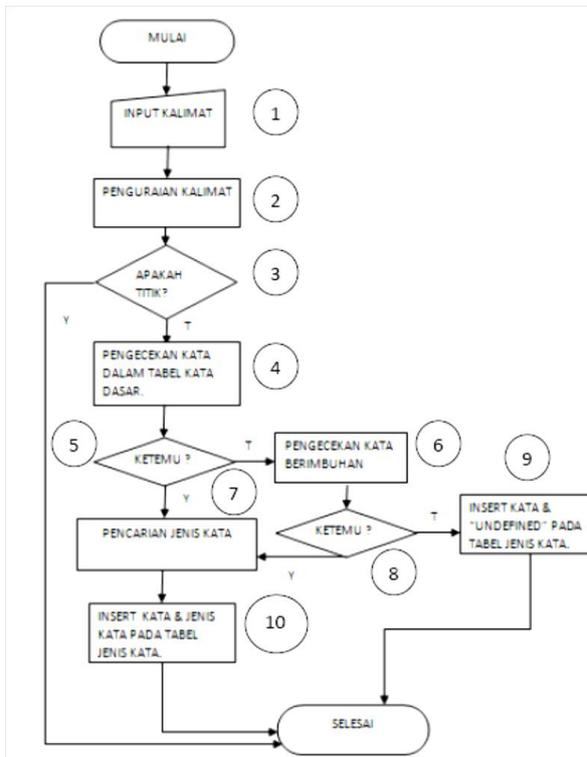
Gambar 3.1 memperlihatkan antar muka pengguna yang sederhana, dengan komposisi satu *text field* dan satu tombol untuk mulai melakukan analisis kalimat di tengah layar. Tampilan tersebut mengacu pada prinsip rancangan antarmuka beberapa diantaranya adalah *Task Compability*, *Work Flow Compability*, *Simplicity* dan *Responsiveness*.

Pada tampilan tersebut, pengguna akan memasukan kalimat *headline* pada *textfield* kemudian dilanjutkan dengan menekan tombol Analisis Sekarang. Proses yang terjadi ketika tombol ditekan adalah mengidentifikasi jenis kata.

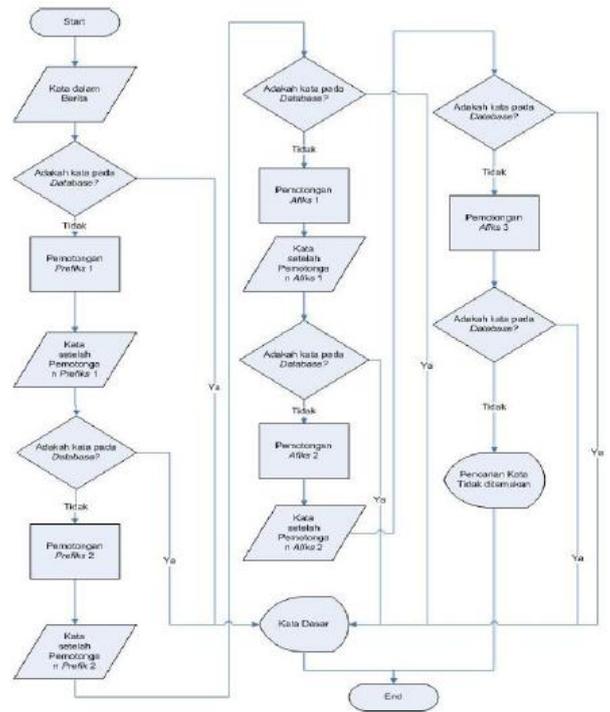
3.1.1 Identifikasi Jenis Kalimat

Dalam mengidentifikasi jenis kata dalam suatu kalimat yang telah diinputkan, diambil data jenis kata pada tabel kata dasar yaitu nomina, adjective, adverb, verbam interjeksi, preposisi, konjungsi, numeralia, pronomina dan sebagai terminator adalah token atau tanda baca titik. Namun, yang menjadi kendala adalah apabila terdapat kata berimbuhan karena pada tabel kata dasar hanya memuat kata dasar. Untuk mengatasi hal tersebut, maka harus dibentuk tiga tabel untuk memuat data imbuhan. Tiga tabel tersebut adalah tabel prefiks untuk awalan, tabel konfiks untuk awalan dan akhiran, dan tabel postfiks untuk akhiran. Dalam mengenali kata berimbuhan, dilakukan proses stemming yaitu pemangkasan kata sehingga dapat diketahui kata dasarnya. Dalam proses stemming algoritma yang digunakan adalah algoritma Nazief & Adriani.

Lebih jelasnya tentang proses identifikasi jenis kata, bisa dilihat pada alur diagram gambar 3.2.



Gambar 3.2 Flowchart Proses Identifikasi Jenis Kata



Gambar 3.3 Flowchart proses steaming algoritma Nazief & Adriani

Keterangan:

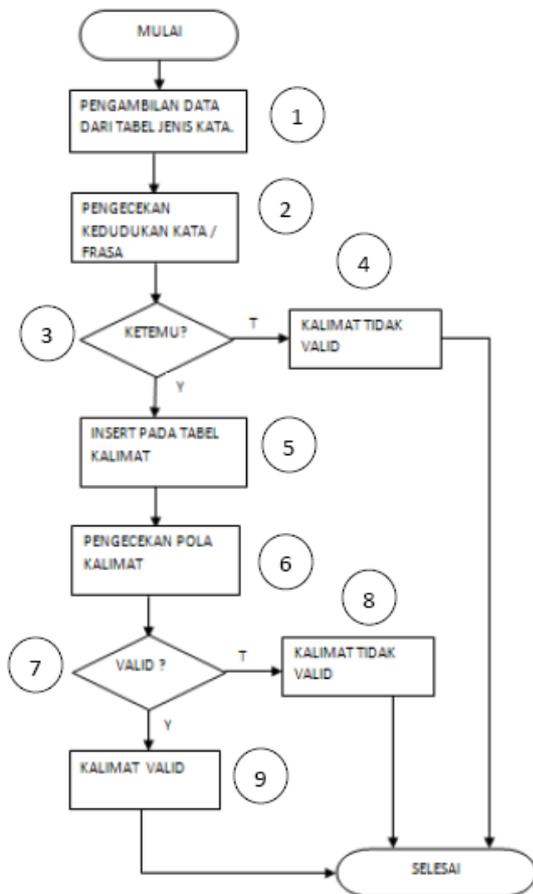
1. Pengguna memasukkan kalimat dalam bahasa Indonesia yang berupa kalimat pernyataan.
2. Mengurai kalimat menjadi kata-kata.
3. Periksa apakah sudah terdapat titik atau belum, apabila ditemukan titik langsung selesai.
4. Apabila belum ditemukan titik maka dilakukan pemeriksaan kata pada tabel kata dasar.
5. Apabila kata diketemukan pada tabel kata dasar maka dilakukan pencarian jenis kata sesuai dengan kata yang ditemukan, jika tidak ketemu maka dilakukan pengecekan kata berimbuhan.
6. Apabila kata tidak ada pada tabel kata dasar maka dilakukan proses stemming untuk dicari kata dasarnya menggunakan algoritma Nazief & Adriani. Adapun proses stemming dari algoritma tersebut adalah sebagai berikut:

7. Jika kata ditemukan maka dilakukan pencarian jenis kata sesuai dengan kata yang ditemukan.
8. Cek apakah kata tersebut merupakan kata berimbuhan, apabila iya maka dilakukan proses nomor 7.
9. Jika tidak ditemukan maka insert kata dan jenis kata sebagai undefined pada tabel jenis kata.
10. Setelah dilakukan pencarian jenis kata dan ditemukan maka insert kata dan jenis kata pada tabel jenis kata.

3.1.2 Identifikasi Pola Kalimat

Setelah jenis kata telah terproses pengidentifikasiannya dilakukan maka proses selanjutnya adalah pengenalan pola kalimat. Pada proses pengenalan pola kalimat inilah metode *probabilistic parsing* digunakan dengan menerapkan algoritma CYK. Pada tahapan ini, sudah terdapat tabel yang berisi kedudukan kata atau frasa dari jenis kata yang ada. Jadi proses yang pertama dilakukan adalah pengambilan data dari tabel jenis kata dan setelah itu dilakukan pengecekan kedudukan kata/frasa pada tabel kedudukan. Setelah

dilakukan pengecekan kedudukan maka diperiksa apakah pola kalimat tersebut valid atau tidak dengan memeriksa pola kalimat pada tabel pola kalimat. Untuk lebih jelasnya tentang proses identifikasi pola kalimat, bisa dilihat pada alur diagram gambar 3.4

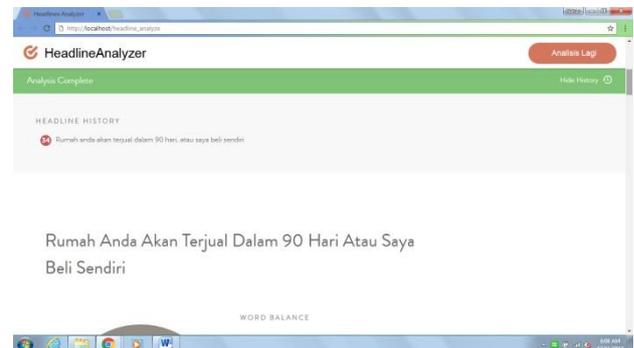


Gambar 3.4 Alur Diagram Pemrosesan Pola Kalimat

Keterangan:

1. Pengambilan data kata dan jenis kata dari tabel jenis kata.
2. Pemeriksaan kedudukan kata atau frasa dari data kata dan jenis kata pada tabel kedudukan.
3. Jika ketemu maka masukan pada tabel kalimat, jika tidak ketemu maka kalimat tidak valid.
4. Jika kedudukan kata atau frasa tidak ditemukan pada tabel kedudukan, maka kalimat tidak valid.
5. Apabila kedudukan kata atau frase ditemukan pada tabel kedudukan, maka masukan pada tabel kalimat.

6. Pemeriksaan kalimat pada tabel pola kalimat.
7. Jika ditemukan pola kalimat tersebut maka valid, jika tidak ditemukan maka kalimat tersebut tidak valid.



Gambar 3.5 Tampilan Hasil Analisis Copywriting

3.2 Pengujian NLP

Pada pengujian, hasil rancang bangun *ditestcase*. Pengujian tersebut meliputi fungsi, proses, dan hasil. Metode uji black box akan digunakan pada *test* ini, yaitu pengujian dengan melihat input dan output yang akan di dihasilkan. Pada test case ini, akan dilakukan pengujian terhadap 20 kalimat *copywriting*.

Hasil dari uji *blackbox* dinyatakan 100% navigasi dan fungsi berjalan dengan baik atau sesuai dengan rancangan.

4. PENUTUP

4.1. Simpulan

Metode *probabilistic parsing* dan algoritma CYK berhasil diterapkan pada aplikasi NLP untuk mengetahui pola kalimat dalam *copywriting*.

Aplikasi ini mampu mengenali pola kalimat dalam *copywriting* formula RTO.

4.2. Saran

Perluasan kalimat dengan menggunakan tanda baca selain titik seperti koma, titik dua, tanda Tanya, tanda seru.

5. DAFTAR PUSTAKA

- Karen Sparck Jones. 2011. *Natural language processing historical review*. Computer Laboratory, University of Cambridge.
- Rosa. 2009. *Penguraian Bahasa Indonesia dengan Menggunakan Pengurai Collins*.

- Tesis Magister Informatika. Institut Teknologi Bandung.
- Bonnie Webber. 2007. "*Chart Parsing: The CYK Algorithm*". Lecture of School of Informatics University of Edinburgh.
- Lefuel, Ramon dan Brian J. Ross. 2004. *Parsing Probabilistic Context Free Languages with Multiple-Objective Genetic Algorithms*. Technical Report. Brock University.
- Jurafsky, James H, Daniel dan Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, New Jersey.
- Charniak, Eugene. 1997. *Statistical Parsing with a Context-free Grammar and Word Statistics*. American Association for Artificial Intelligence: AAAI Press