EKSTRAKSI KATA KUNCI DENGAN KEYPHRASE EXTRACTION ALGORITHM PADA DOKUMEN SOCIAL MEDIA BERBAHASA INDONESIA

Pratama Adhi Guna[1], Ema Rachmawati[2], Arie Ardiyanti Suryani[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Salah satu cara untuk mengetahui isi dari satu dokumen teks adalah dengan melakukan ekstraksi kata kunci pada dokumen tersebut. Penggunaan ekstraksi kata kunci pada social media sangat menguntungkan karena perkembangan social media yang pesat. Ekstraksi kata kunci pada dokumen social media berbeda karena memiliki teks yang relatif sangat pendek.

Metode untuk menemukan kata kunci salah satunya adalah Keyphrase Extraction Algorithm (KEA). KEA membentuk model klasifikasi lalu menghitung nilai fitur dari setiap kandidat kata kunci untuk diketahui nilai probabilitasnya. Data yang digunakan pada penelitian ini adalah data dokumen teks Twitter. Eksperimen menunjukkan bahwa penggunaan metode ini sangat efektif untuk ekstraksi kata kunci pada dokumen social media.

Kata Kunci : ekstraksi kata kunci, KEA, social media

Abstract

One way to know the contents of the text documents is extracting the keywords in the document. The use of keyword extraction in social media is very profitable because of rapid development of social media. Keyword extraction in social media documents is different because it has a relatively very short text.

One of methods to find keyword is Keyphrase Extraction Algorithm (KEA). KEA makes a classification model then calculates the feature value of each keywords candidate to find probability value. The data used in this study are Twitter text documents. Experiment showed that the use of this method is highly effective for keywords extraction in social media documents.

Keywords : keyword extraction, KEA, social media

1. Pendahuluan

1.1 Latar Belakang

Dalam pemahaman isi dari suatu teks, kata kunci dapat membantu para pembaca untuk memahami teks tanpa membaca keseluruhan isinya. Kata kunci tersebut dapat menggambarkan konten yang sedang dibicarakan [15]. Menemukan kata kunci secara manual sangat sering digunakan, tetapi hal itu akan sulit jika dokumen berjumlah banyak hal tersebut akan memakan waktu yang lama. Agar kata kunci pada dokumen yang berjumlah banyak dapat diperoleh dengan cepat dibutuhkan ekstraksi kata kunci secara otomatis.

Penggunaan ekstraksi kata kunci otomatis pada dokumen teks *social media* adalah contoh penggunaan yang sangat menguntungkan karena perkembangan dari *social media* yang begitu pesat. Saat ini sudah banyak social media yang bermunculan seperti Facebook, Twitter, Plurk, Linkedin, dan lain sebagainya yang sudah memiliki jutaan pengguna di seluruh dunia. Ekstraksi kata kunci pada dokumen *social media* ini dapat digunakan untuk *advertising*, *search*, dan *content filtering* [9].

Tidak seperti dokumen tradisional, teks pada *social media* biasanya sangat pendek dan cenderung tidak formal [9]. Dengan teks yang seperti itu, ekstraksi kata kunci pada dokumen *social media* memiliki perbedaan dengan ekstraksi kata kunci pada umumnya.

Ada banyak metode yang dapat digunakan dalam ekstraksi kata kunci. Salah satu metode yang dapat digunakan adalah KEA (*Keyphrase Extraction Algorithm*). Pada paper yang diacu menyatakan bahwa KEA merupakan metode yang sederhana dan efektif dalam penggunaannya [15]. KEA hanya menggunakan dua *feature* yang sederhana dalam penerapannya pada dokumen teks baik yang panjang ataupun pendek. Oleh karena itu, ekstraksi kata kunci pada dokumen *social media* yang memiliki teks pendek ini menggunakan KEA.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang penulis uraikan sebelumnya, permasalahan yang akan diteliti dalam tugas akhir ini adalah sebagai berikut.

1. Bagaimana melakukan ekstraksi kata kunci pada dokumen *social media* yang bentuk teks nya relatif pendek.
2. Bagaimana mengetahui keakuratan KEA untuk menghasilkan kata kunci pada dokumen *social media* berbahasa Indonesia.

Dan adapun batasan masalah untuk tugas akhir ini adalah:

1. *Social media* yang digunakan adalah Twitter
2. Data yang dianalisis berupa *plain text* dari Twitter tidak temasuk foto, video, *link web*, dan lokasi
3. Data akan dianalisis secara *offline learning*

1.3 Tujuan

Tujuan yang diharapkan oleh penulis dalam pengerjaan tugas akhir ini adalah:

1. menentukan kata kunci pada dokumen *social media* menggunakan KEA
2. mengevaluasi akurasi dari hasil ekstraksi kata kunci menggunakan KEA

1.4 Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelesaikan tugas akhir ini adalah sebagai berikut.

1. Studi Litelatur

   Mempelajari literatur-literatur meliputi *crawling tweet using Twitter API*, *text mining*, dan ekstraksi frasa kunci terutama dengan metode KEA.

2. Pengumpulan Data

   Mengumpulkan twit untuk dianalisis.

3. Perancangan Sistem

   Melakukan analisis kebutuhan yang diperlukan dalam sistem serta merancang sistem yang sesuai dengan identifikasi kebutuhan tersebut.

4. Implementasi

   Melakukan implementasi berdasarkan rancangan yang telah dibuat sebelumnya

5. Analisis Hasil

   Analisis akurasi terhadap hasil pengujian yang didapat berdasarkan parameter uji.

6. Pengambilan Kesimpulan dan Pembuatan Laporan

   Membuat kesimpulan dari analisa yang telah dilakukan dan membuat laporan tugas akhir yang mendokumentasikan tahap-tahap penelitian.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan pengujian yang dilakukan dalam Tugas Akhir ini, dapat disimpulkan beberapa poin sebagai berikut.

1. Keyphrase Extraction Algorithm (KEA) dapat digunakan untuk menentukan kata kunci pada dokumen *social media* berbahasa Indonesia baik satu kata atau lebih dari satu kata (frasa).
2. Nilai *Keyword Matches* sangat bergantung pada jumlah kata maksimum dalam satu frasa dan pada penelitian ini nilai yang memberikan akurasi paling tinggi adalah 3.
3. Nilai *Keyword Matches* sangat bergantung pada jumlah maksimum kata kunci keluaran sistem dan pada penelitian ini semakin besar maksimum kata kunci semakin besar akurasi yang dihasilkan.

5.2 Saran

Berdasarkan pengujian yang dilakukan di atas, sebaiknya perlu diperhatikan beberapa poin sebagai berikut.

1. Sistem dapat dikembangkan menjadi sistem *online*, sehingga pengambilan dan ekstraksi kata kunci dapat dilakukan dengan cepat.
2. Implementasi KEA pada topik tertentu dapat menggunakan *thesaurus* agar akurasinya lebih baik.

6. Daftar Pustaka

[1] Choy, Murphy. (2012). Twitter Topic Extraction using Principal Component Analysis.Singapore

[2] Domingos, Pedro dan Michael Pazzani. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. University of California, Irvine.

[3] Ercan, Gonenc. (2006). AUTOMATED TEXT SUMMARIZATION AND KEYPHRASE EXTRACTION. Master Thesis. Bilkent University.

[4] Fayyad, U.M. and Irani, K.B. (1993) “Multi-interval discretization of continuous-valued attributes for classification learning.” Proc IJCAI’93, 1022-1027.

[5] Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.G. (1999) "Domain-specific keyphrase extraction." Proc. 16th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668-673.

[6] Hearst, Marti. (2003). What is Text Mining?.SIMS, UC Berkeley.

[7] Hulth, Anette. (2004). Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction. Doctoral Dissertation. Department of Computer and Systems Sciences. Stockholm University.

[8] Kwak, Haewon., Changhyun Lee., Hosung Park. and Sue Moon. (2010). What si Twitter, a Social Network bor a News Media?. Department of Computer Science, KAIST. Korea

[9] Li, Zhenhui, Ding Zhou, Yun-Fang Juan, Jiawei Han. (2010). Keyword Extraction for Social Snippets.

[10] O'Connor, Brendan, Michel Krieger, David Ahn. (2010). TweetMotif: Exploratory Search and Topic Summarization for Twitter. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[11] Pochampally, Ravali dan Vasudeva Varma. (2011). User context as a source of topic retrieval in Twitter. Search and Information Extraction Lab, LTRC. IIIT Hyderabad, India.

[12] Suhardi. 2013. *Dasar-dasar Ilmu Sintaksis*. Yogjakarta: Ar-Ruzz Media.

[13] Tala, Fadillah Z. 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation, Universiteit van Amsterdam. Netherlands.

[14] Turney, P. 1999. Learning to Extract Keyphrases from Text. Canada

[15] Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. (2000) "KEA: Practical automatic keyphrase extraction." Working Paper 00/5, Department of Computer Science, The University of Waikato.