Fidelity : Jurnal Teknik Elektro
Vol. 3, No. 3, September 2021, page 56-58

p-ISSN: 2686-3650
e-ISSN: 2686-3642

Review Klasifikasi Berita Online

Fatekhul Muiz
R&D
PT. Cogindo DayaBersama
Sukabumi, Indonesia
fatekhul.muiz@cogindo.co.id

***Abstrak*—Data mining adalah subjek studi dan eksperimen. Dalam transaksi manual, mengklasifikasikan berita online selalu menjadi kendala. Pekerjaan yang disajikan di sini bertujuan untuk mengembangkan algoritma baru untuk mengkategorikan struktur dalam dari informasi rahasia yang penting. Dalam kasus saat ini, pekerjaan hanya dilakukan untuk mengidentifikasi kluster luar sistem, tetapi tidak ada pekerjaan yang dilakukan pada kumpulan dalam kumpulan data. Dalam tugas yang diusulkan ini, pengelompokan internal akan dibuat untuk setiap bidang dari sistem yang diusulkan, seperti olahraga dan hiburan.**

Kata Kunci—Model Markov Tersembunyi (HMM), Mesin Vektor Dukungan (SVM), K Mean, CART

I. Pengantar

Penambangan data mengekstraksi informasi menarik dari sejumlah besar data dalam basis data, gudang, atau tempat penyimpanan informasi lainnya, seperti pola, asosiasi, perubahan, anomali, dan struktur yang mengesankan. Penambangan data baru-baru ini mendapat banyak perhatian di industri informasi, berkat ketersediaan data dalam jumlah besar dalam bentuk elektronik dan kebutuhan mendesak untuk mengubah data itu menjadi informasi dan pengetahuan yang berguna untuk berbagai aplikasi, termasuk analisis pasar dan dukungan keputusan manajemen bisnis. . Meskipun beberapa akademisi menganggap penambangan data sebagai langkah mendasar dalam penemuan pengetahuan, ini biasanya digunakan sebagai sinonim untuk penemuan pengetahuan basis data. Urutan berulang dari langkah-langkah berikut membentuk proses penemuan pengetahuan[9]:

- Pembersihan data digunakan untuk membersihkan data yang berisik, salah, hilang, atau tidak relevan.

- Integrasi data, di mana beberapa sumber data yang berbeda digabungkan menjadi satu.

- Pemilihan data, yang melibatkan pengambilan data dari database yang relevan dengan kegiatan analisis.

- Transformasi data melibatkan melakukan operasi agregat pada data untuk mengubah atau menggabungkannya ke dalam bentuk yang cocok untuk penambangan.

- Data mining adalah prosedur penting yang menggunakan cara canggih untuk mengekstrak pola data.

- Evaluasi pola mengidentifikasi metode yang paling menarik untuk mengungkapkan pengetahuan berdasarkan beberapa ukuran minat.

- Presentasi pengetahuan, di mana pemahaman yang digali disajikan kepada pengguna menggunakan pendekatan visualisasi dan representasi pengetahuan.

II. Klasifikasi Teks

Sebagian besar data disimpan dalam teks, seperti email, halaman web, artikel surat kabar, laporan riset pasar, surat keluhan pelanggan, dan laporan yang dibuat secara internal, karena surat kabar online mencakup berbagai topik, termasuk nasional, internasional, politik, keuangan, olahraga, dan hiburan [4]. Klasifikasi teks adalah aspek penting lain dari penambangan teks [5]. Pengetahuan ahli tentang mengklasifikasikan dokumen di bawah set kategori yang ditentukan digunakan untuk mengklasifikasikan teks. Kumpulan dokumen pelatihan yang sudah diberi label dengan kelas digunakan untuk memulai klasifikasi data mining. Ada dua jenis klasifikasi teks: single label dan multi label[5]. Sebuah dokumen label tunggal hanya dapat dimiliki oleh satu kelas, tetapi dokumen multi-label dapat dimiliki oleh beberapa kategori. Sebagian besar database teks menyimpan data semi-terstruktur, yang merupakan materi yang tidak sepenuhnya tidak terstruktur atau sepenuhnya terstruktur. Sebuah dokumen, misalnya, mungkin menyertakan beberapa bidang terstruktur seperti judul, penulis, tanggal publikasi, dan kategori, tetapi mungkin juga memiliki beberapa komponen teks tidak terstruktur seperti abstrak dan konten.

III. Proses Klasifikasi Teks

Bagian berikut membahas langkah-langkah klasifikasi teks [5].

a. Pengumpulan Dokumen Kumpulkan berbagai macam dokumen, seperti dokumen HTML, PDF, dan Word, pada tahap ini.

b. Persiapan Dalam hal ini, dokumen teks akan diubah menjadi format kata yang jelas, dengan fitur-fitur seperti Tokenization, stemming words, dan delete stop words. Sebuah dokumen ditangani sebagai string dan kemudian dipartisi ke dalam daftar token di Tokenization. Hapus stopword seperti "the", "a", dan "and" saat menghapus stop word. Stemming kata menggabungkan kata-kata yang beragam menjadi satu bentuk kanonik.

c. Pengindeksan: Ini membuat indeks untuk setiap dokumen agar mudah diidentifikasi.

d. memilih fitur Seleksi fitur adalah tahap kritis dalam klasifikasi teks setelah prapemrosesan dan pengindeksan. Ide utama di balik pemilihan fitur adalah untuk mengambil subset dari fitur dokumen asli dan menggunakannya untuk membuat dokumen baru. Hal ini dilakukan dengan

mempertahankan kata-kata dengan skor tertinggi berdasarkan penilaian nilai kata yang ditentukan.

e. Identifikasi dan klasifikasi Ini mengkategorikan dokumen ke dalam kelompok yang telah ditentukan. Makalah dapat diklasifikasikan menggunakan pendekatan terawasi atau tidak terawasi. Klasifikasi terbimbing adalah ketika label kelas dari setiap dokumen diketahui. Kategorisasi tanpa pengawasan terjadi ketika label kelas dokumen tidak diketahui.

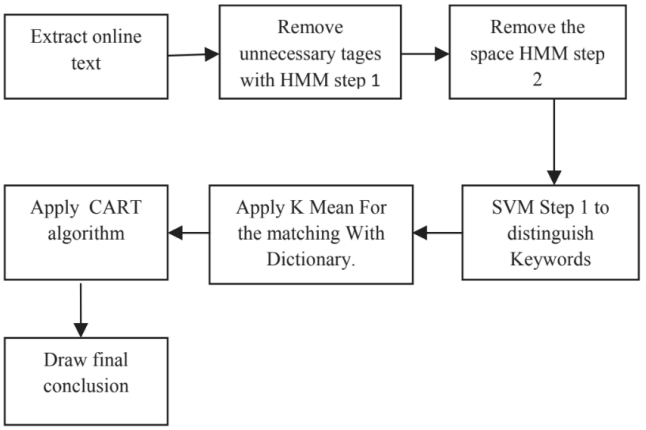
f. Evaluasi Kinerja Ini adalah langkah terakhir dalam proses klasifikasi teks. Dalam hal ini, kinerja diukur secara eksperimental daripada analitis. Presisi dan recall adalah dua contoh pengukuran yang telah digunakan.

IV. Pekerjaan Terkait Dengan Ini

Mengklasifikasikan dokumen teks ke dalam kumpulan jenis yang telah ditentukan adalah kategorisasi teks [1]. Penelitian ini membahas tiga metode untuk membuat meta classifier untuk meningkatkan akurasi klasifikasi. Konsep utamanya adalah mempelajari pengklasifikasi meta yang akan secara optimal memilih pengklasifikasi komponen terbaik untuk setiap titik data. Hasil menunjukkan bahwa menggabungkan pengklasifikasi meningkatkan akurasi klasifikasi secara signifikan dan bahwa teknik klasifikasi meta mengungguli pengklasifikasi individu. [2] memberikan pengenalan dasar SVM dan berbagai aplikasi SVM untuk tantangan pengenalan pola. Deteksi dan pengenalan wajah, deteksi dan pengenalan objek, pengenalan huruf dan angka tulisan tangan, pengambilan informasi dan gambar, prediksi, dan lebih banyak aplikasi menggunakan SVM. Dua pendekatan untuk mengembangkan pengklasifikasi dokumen teks berbasis Teori Nave Bayes dan mengintegrasikannya ke dalam pengklasifikasi meta meningkatkan akurasi klasifikasi[3]. Untuk kategori Olahraga, Keuangan, dan Politik, pengklasifikasi berita cerdas dirancang dan diuji dengan berita online dari web[4]. Strategi inovatif yang menggabungkan dua algoritme yang kuat, Model Markov Tersembunyi dan Mesin Vektor Dukungan, menghasilkan hasil yang sangat baik dalam klasifikasi berita online. Kata kunci dari materi surat kabar online diekstraksi oleh sistem intelijen dan diklasifikasikan ke dalam kategori yang telah ditentukan. Klasifikasi surat kabar online dibagi menjadi tiga tahap: (1) preprocessing teks, (2) ekstraksi fitur berbasis HMM, dan (3) ekstraksi fitur. (3) Klasifikasi SVM—sebuah gambaran singkat dari beberapa skema representasi teks[5]. Pendekatan klasifikasi yang ada dievaluasi dan dikontraskan berdasarkan berbagai parameter klasifikasi, implementasi algoritma, dan kompleksitas waktu klasifikasi. Tergantung pada data yang dikumpulkan, algoritma yang berbeda berperilaku berbeda. Sebagian besar data disimpan sebagai teks. Penambangan teks ini dianggap memiliki nilai komersial yang signifikan. Setiap sumber informasi dapat digunakan untuk memperoleh pengetahuan. Namun, teks yang tidak terstruktur tetap menjadi sumber pengetahuan yang paling tersedia. Penelitian ini memperkenalkan klasifikasi teks, memberikan pengenalan pengklasifikasi, dan membandingkan beberapa pengklasifikasi yang ada berdasarkan kompleksitas waktu, prinsip, dan kinerja. Kategorisasi teks otomatis adalah aktivitas pembelajaran mesin semi-diawasi di mana dokumen yang diberikan secara otomatis ditetapkan ke kategori tertentu berdasarkan konten tekstual dan fitur yang diekstraksi [6]. Klasifikasi teks otomatis memiliki aplikasi dunia nyata dalam manajemen konten, penggalian opini, analisis tinjauan produk, dan solusi survei yang ada untuk masalah signifikan seperti teks tidak terstruktur, menangani banyak atribut, dan memilih teknik pembelajaran mesin yang tepat untuk aplikasi klasifikasi teks. Membangun model prediksi[7] yang dapat mengatur berita saat naik atau turun digunakan untuk mengklasifikasikan berita keuangan berdasarkan isi berita yang relevan. Mekanisme peramalan disajikan dalam penelitian ini. Untuk menyesuaikan klasifikasi, gunakan metode klasifikasi SVM[8]. Pengguna dapat menentukan kategori mereka menggunakan beberapa kata kunci untuk kueri pencarian dalam klasifikasi yang dipersonalisasi. Pengkategori mengumpulkan teks positif dan negatif, yang diperlukan untuk membangun pengklasifikasi. [9] Makalah ini membahas penambangan data dan kegunaannya yang berbeda dan siklus hidup penambangan data, penemuan pengetahuan, dan berbagai pendekatan penambangan data.

V. Penelitian yang Diusulkan

Diagram alir penelitian ditunjukkan di bawah ini. Ini menggambarkan proses klasifikasi berita.



Gambar 1. Proses Klasifikasi Berita

Kami menggunakan dua model dalam penelitian ini: HMM dan SVM. Hidden Markov Model (HMM) digunakan untuk ekstraksi teks, sedangkan untuk klasifikasi, Support Vector Machine (SVM) digunakan. Berikan alamat URL surat kabar online terlebih dahulu, dan kemudian teks akan muncul, diikuti dengan proses yang berbeda untuk menghapus elemen HTML dan spasi yang berlebihan. K adalah singkatan dari metode yang digunakan untuk menghasilkan cluster, dan CART adalah singkatan dari algoritma yang akan digunakan untuk menggambarkannya dalam bentuk hierarki.

VI. KESIMPULAN

Implementasi yang sukses dari penggalian berita dari portal internet untuk pemrosesan tambahan telah tercapai. Selain itu, kelompok kategori yang berbeda telah dikembangkan sehingga kombinasi lebih lanjut dari HMM DAN SVM dapat diterapkan untuk meningkatkan efisiensi.

Referensi

[1] Daniel I. Morariu, Lucian N. Vintan, and Volker Tresp,"Meta-Classification using SVM Classifiers for Text Documents,"World Academy of Science, Engineering and Technology 21 2006.

[2] Hyeran Byun1 and Seong-Whan Lee2,"Applications of Support Vector Machines for Pattern Recognition: A Survey,"SVM 2002, LNCS 2388, pp. 213-236, 2002.

[3] D. Morariu, R. Cre ţulescu and L. Vin ţan,"Improving a SVM Meta-classifier for Text Documents by using Naïve-Bayes,"Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844.

[4] Krishnalal G, S Babu Rengarajan, K G Srinivasagan ,"A new text mining approach based on HMM -SVM for web news classification,"International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 19,2010.

[5] Vandana Korde, C Namrata Mahender,"Text classification and classifier a survey," International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[6] Mita K. Dalal, Mukesh A.Zaveri,"Automatic text classification,"International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011.

[7] Rama Bharath Kumar, Bangari Shravan Kumar, Chandragiri Shiva Sai Prasad,"Financial news classification using SVM",International Journal of Scientific and Research Publications, Volume 2, Issue 3, March 2012 .

[8] Chee-Hong Chan Aixin Sun Ee-Peng Lim,"Automated Online News Classifcation with Personalization,"4th international conference on asian digital libraries, Dec 2001.

[9] Mr. S. P. Deshpande , Dr. V. M. Thakare,"data mining system and applications: a review,"International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.