Jurnal SimanteC
Vol. 8, No. 2 Juni 2020 P-ISSN : 2088-2130
E-ISSN : 2502-4884

KAJIAN LITERATUR MENGENAI KLASIFIKASI BLOG

Husni
Jurusan Teknik Informatika Universitas Trunojoyo
husni@trunojoyo.ac.id

ABSTRAK

Klasifikasi blog merupakan topik kajian baru. Teknik klasifikasi web tradisional tidak dapat diterapkan secara langsung terhadap blog karena sering terjadinya update terhadap isi dan variasi topik pada suatu situs blog. Komponen penyusun blog seperti judul, isi dan komentar, tag (label), penulis, hyperlink, permalink, outlink, tanggal dan jam termasuk obyek yang perlu dilibatkan dalam proses klasifikasi. Tulisan ini mencoba meninjau berbagai pendekatan klasifikasi blog yang hadir sejak 2009. Pada awal kemunculan blog, klasifikasi biner digunakan untuk membedakan blog dari halaman web biasa. Kami fokus pada bagaimana mengkategorikan suatu blog ke dalam daftar topik, genre dan opini (mood dan sentimen) yang telah didefinisikan sebelumnya.Pada klasifikasi topik dan genre, algoritma kNN, Naive Bayes, CFC, SVM dan pendekatan machine learning lainnnya banyak digunakan.Pemanfaatan ontologi topik dan tag dapat meningkatkan akurasi klasifikasi. Pada deteksi opini, pendekatan berbasis lexicon seperti ANEW cenderung lebih banyak digunakan. Opini dari suatu situs blog juga dapat diprediksi berdasarkan opini di sekitar inlink yang menuju situs tersebut. Kajian ini perlu diperluas dan diperdalam, seperti keterlibatan lebih lanjut dari tag, link dan analisis jejaring sosial.

Kata kunci: *Klasifikasi Blog, Analisis Sentimen, Blog Mining*

ABSTRACT

Blog classification is a new study topics. Traditional web classification techniques cannot be applied directly to blogs because of frequent updates to the content and variations of topics on a blog site. The components of the blog, such as title, content and comments, tags (labels), author, hyperlink, permalink, outlink, date and time, including objects that need to be involved in the classification process. This paper tries to review various blog classification approaches that have been present since 2009. At the beginning of the emergence of blogs, binary classification is used to distinguish blogs from ordinary web pages. We focus on how to categorize a blog into a list of topics, genres and opinions (moods and sentiments) that have been previously defined. On the classification of topics and genres, the kNN algorithm, Naive Bayes, CFC, SVM and other machine learning approaches are widely used. topics and tags can improve classification accuracy. In opinion detection, lexicon-based approaches such as ANEW tend to be more widely used. Opinions from a blog site can also be predicted based on opinions around the inlink that goes to that site. This study needs to be broadened and deepened, such as further involvement of tags, links and analysis of social networks.

Keywords: *Blog Classification, Sentiment Analysis, Blog Mining*

PENDAHULUAN

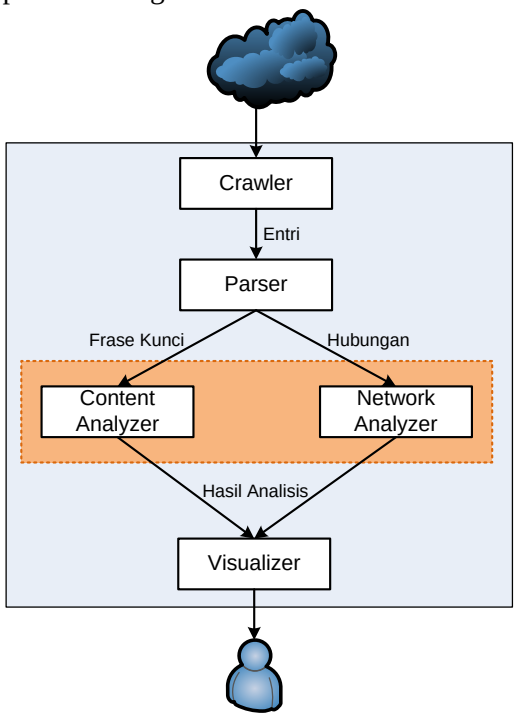
Weblog atau disingkat "blog" adalah halaman web yang sering diubah dimana setiap entri mempunyai tanggal dan jam yang secara kronologis ditampilkan urut terbalik (*descending*). Blogger (orang yang menulis blog) memanfaatkan layanan ini untuk secara bebas mengekspresikan pendapat dan emosinya [1]. Pemanfaatannya semakin meningkat untuk berbagai bidang dan kepentingan. Per 17 Januari 2012, situs repository blog Technorati (technorati.com) menyimpan sebanyak 1.291.160 blog yang dikelompokkan ke dalam sembilan kelas utama, yaitu Entertainment, Business, Sports, Politics, Autos, Technology, Living, Green dan Science [2].

Blogpost bersifat unik, di antaranya mengandung informasi yang terstruktur dan tak-terstruktur, blogger menuliskan blogpost dengan gaya mengalir, naratif tak-terstruktur, bahkan kadang muncul kesalahan ejaan dan tata-bahasa. Blogger juga mungkin menggunakan kata dan tata-bahasa baru untuk mengekspresikan pendapatnya [3]. Menurut Chau dkk. [1] teknik web dan text mining yang ada tidak dapat langsung digunakan karena dua hal, yaitu update yang lebih sering daripada halaman web biasa dan sangat bervariasinya topik yang ditulis pada suatu blog. Pada halaman web tradisional, struktur mining dapat diterapkan terhadap hyperlink antar halaman, tetapi pada blog, hyperlink bukan satu-satunya yang menghubungkan blog-blog. Hubungan juga dapat terwujud melalui komentar dan berlangganan ke blog lain. Tabel 1 memperlihatkan perbedaan antara dokumen teks, halaman web dan blog [4].

Tabel 1. Perbandingan komponen antara blog, web dan teks

Komponen	Blog	Web	Teks
Judul (Title)	•	•	
Isi (Content)	•	•	•
Tag (Label)	•		
Penulis (Author)	•		
URL (hyperlink)	•	•	
Permalink	•		
Outlink	•	•	
Waktu (time)	•		
Tanggal (date)	•		

Berdasarkan komponen pada Tabel 1, secara umum *data mining* terhadap blog dapat dikelompokkan ke dalam lima obyek (dimensi), yaitu isi blog (judul dan isi (*content*)), tag (label atau kategori), penulis (*author* atau *blogger*), link (URL, permalink dan outlink), dan Waktu (tanggal dan jam) [4]. Proses mining terhadap blog terdiri dari beberapa tahapan sebagaimana diperlihatkan oleh framework pada Gambar 1 [1]. Framework ini terdiri dari 5 komponen utama, yaitu Crawler, Parser, Content Analyzer, Network Analyzer dan Visualizer. Cara kerja setiap komponen ini menyerupai framework untuk web mining, karena blog juga termasuk bagian dari web, tetapi dengan beberapa perbedaan signifikan.



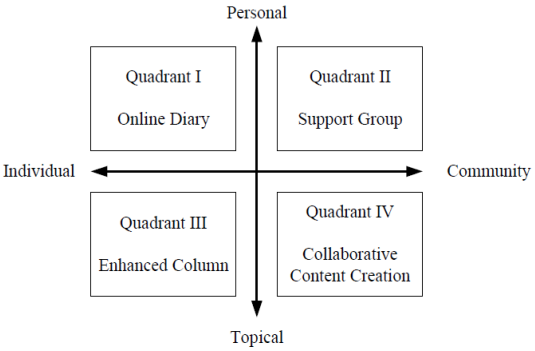
Gambar 1. Tahapan dari proses *data mining* terhadap blog [1]

Fokus dari tulisan ini terdapat pada bagian Analyzer. Blog analyzer bertanggungjawab menganalisis frase

kunci yang telah diekstrak oleh Parser. Analisis dilakukan menggunakan teknik web mining, terutama klasifikasi dan klusterisasi. Analisis terhadap teks atau web biasanya memanfaatkan fitur term sebagai data input dalam proses klasifikasi dan klusterisasi, namun blog analyzer dapat pula mengakses fitur lain seperti komentar dalam *blogpost*, profil *blogger* dan link ke blog lain. Blog anayzer juga mampu menganalisis hubungan jaringan antar *blogger*. Teknik analisis jaringan dapat digunakan untuk mendapatkan *minimum spanning tree* atau partisi graf. Informasi ini berguna untuk mengetahui jarak sosial antar *blogger* dan karakteristik komunitas blog tertentu. Teknik ini juga berguna untuk mengekstrak ciri dari topologi, sentralitas dan komunitas dari suatu jaringan [1].

Lakshmanan dkk. [3] memecah tahapan analisis blog ini menjadi dua bagian lebih detail. Bagian pertama bertugas melakukan *pre-processing* dan *data formatting*, sedangkan bagian kedua berisi algoritma-algoritma *knowledge-discovery*. Strategi *knowledge discovery* dalam blog mencakup pengelompokan, perankingan dan faktorisasi matriks. Pengelompokan data dapat dibagi menjadi dua, yaitu *supervised learning* (klasifikasi) dan *unsupervised learning* (klasterisasi) [5].

Menurut Chang dan Yeh [6], terdapat dua landasan utama klasifikasi blog, yaitu berbasis motif (alasan) dan berbasis isi (*content*). Nardi dkk. [7] membagi blog menjadi lima tipe berdasarkan motivasi blogger, yaitu riwayat hidup, komentar, ekspresi emosi, penjelasan pikiran, dan pertukaran ide. Klasifikasi berbasis isi mengelompokkan blog sesuai dengan subyek tertentu dan melibatkan penyaringan informasi. Krishnamurthy dalam [6] mengidentifikasi empat jenis blog, yaitu buku harian (*diary*) online, kelompok dukungan (*support group*), *enhanced column*, dan pembuatan isi kolaboratif. Pembagian ini diperjelas dengan empat kuadran sebagaimana diperlihatkan pada gambar 2.



Gambar 2. Pembagian blog dalam empat kuadran menurut Krishnamurthy [6]

Menurut Qi dan Davidson [8], penelitian tentang klasifikasi blog dapat dipecah ke dalam tiga kategori, yaitu identifikasi blog, klasifikasi mood dan klasifikasi genre. Kategori pertama bertujuan mengidentifikasi apakah suatu halaman web dapat digolongkan sebagai suatu weblog (klasifikasi biner). Kategori kedua mencakup klasifikasi topik, mood dan sentimen dari suatu *blogpost*. Klasifikasi topik menganalisis isi dari *blogpost* dan memasukkannya ke dalam salah atau lebih daftar topik yang telah disediakan. Koleksi entri blog yang hanya mengandung dua polaritas mood: senang dan sedih (biner), dapat dengan mudah dipisahkan berdasarkan *linguistic content*-nya. Namun pendekatan ini tidak dapat digunakan menentukan mood dari suatu blogpost (atau *blogger*nya) jika terdapat banyak mood yang telah didefinisikan, seperti takut, cemas, marah, dan kecewa. Klasifikasi sentimen bertugas menentukan sentimen dari suatu *blogpost*, biasanya salah satu dari dua (positif, negatif) atau tiga (positif, negatif, netral). Klasifikasi ketiga bertanggungjawab menentukan genre dari situs web blog secara keseluruhan, bukan per blogpost. Suatu blog misalnya terklasifikasi ber-genre (salah satu dari berita, komentar, berkala) atau (salah satu dari buku harian, berita, politik, olah-raga).

Penelitian awal mengenai identifikasi blog dilakukan oleh Nanno et.al [9] yang membuat sistem untuk menghimpun dan mengidentifikasi blog berdasarkan sejumlah heuristik sederhana. Kemudian Elgersma dan Rijke [10] membangun sistem identifikasi yang

melibatkan sejumlah fitur tertentu yang spesifik blog seperti kehadiran komentar dan arsip. Algoritma klasifikasi *machine-learning* umum ternyata mampu mengidentifikasi halaman blog dengan akurasi mencapai 90%. Yu dkk. [11] membangun suatu *classifier* biner untuk mengidentifikasi halaman blog berdasarkan pada struktur dan isinya. Ada 2 cara untuk memisahkan blog dengan halaman web biasa, yaitu melalui identifikasi luas dan sempit. Pada identifikasi luas, *classifier* hanya memanfaatkan informasi link dan fitur kata kunci yang terdapat dalam bagian head. Karena suatu *blogpost* mungkin mengandung kegaduhan (*noise*) seperti banyak informasi link, navigasi, iklan dan login, maka ekstraksi informasi tertentu dari halaman web menjadi lebih sulit. Karena itu digunakan identifikasi sempit. Selain informasi dalam bagian head, identifikasi sempit juga memanfaatkan karakteristik lain, yaitu informasi *blogger*, *blog post*, informasi statistik, kalender pemutakhiran *blog post*, komentar dan kalendar penulisan. Sejumlah halaman blog dijadikan himpunan blog standard. Kemudian halaman web lain dibandingkan dengan himpunan blog standard tersebut dan akan dikategorikan sebagai blog jika nilai kemiripannya di atas ambang yang ditetapkan.

Tulisan ini mencoba untuk meninjau berbagai pendekatan dan teknik klasifikasi untuk mengkategorisasikan topik, genre, mood dan sentimen dari suatu *blogpost* yang melibatkan berbagai komponen penyusunnya, terutama isi, komentar, hyperlink dan tag. Tujuan kajian ini adalah menghadirkan berbagai pendekatan klasifikasi blog yang ada, kelebihan dan kekurangannya, serta contoh aplikasinya sehingga memudahkan peneliti memilih teknik yang tepat untuk masalah yang dihadapi, misalnya dalam pengembangan sistem rekomendasi atau mesin pencarian khusus blog. Kajian ini difokuskan pada naskah ilmiah terkait yang terbit 3 tahun terakhir, sejak 2009. Kajian dilakukan oleh Qi dan Davidson [8] telah merangkum perkembangan klasifikasi blog sebelum tahun 2009.

Bagian selanjutnya dari tulisan ini akan menjelaskan mengenai Tinjauan Pustaka yang terkait erat dengan klasifikasi blog (bagian 2). Pada bagian 3, akan diuraikan pendekatan untuk klasifikasi topik dan genre dari blogpost (biasanya berdasarkan isi blog). Bagian 4 akan membahas teknik-teknik klasifikasi mood dan sentimen (berdasarkan motif). Kesimpulan dan *future work* dari tulisan ini disampaikan pada bagian 5.

TINJAUAN PUSTAKA

Qi dan Davidson [8] telah melakukan kajian awal mengenai klasifikasi blog dan merangkum pendakatan yang hadir sebelum 2009. Setelah dapat membedakan halaman blog dari web biasa, titik berat penelitian adalah bagaimana menentukan topik atau genre dari blog. Lex [14,15] mengklasifikasi blog memanfaatkan pengetahuan dari domain surat kabar (*cross-domain blog classification*). Hagiwara dkk. [19] mengklasifikasi blog memanfaatkan label dari topik pada tingkatan situs blog, sedangkan Tsai [4] mengkategorisasikan blog berdasarkan tag yang dibuat oleh blogger. Subramaniyaswamy [22] membangun suatu ontologi topik dan digunakan pada tahapan klasifikasi blog. Wiegand [23] mengklafikasi blog pada tingkatan kalimat. *Supervised learning* seperti kNN, Naive Bayes, CFC, SVM dan modifikasinya digunakan dalam proses klasifikasi.

Deteksi opini juga menjadi fokus dari kajian blog mining. Tinjauan yang baik mengenai deteksi opini ditulis oleh Pang [30] pada tahun 2008. Demartini [26] membangun suatu model yang secara otomatis memperkirakan opini publik terkait pemilihan presiden amerika. Missen [25] mengajukan framework deteksi opini berbasis jejaring sosial. Nguyen [31] dan Jung [32] mengklasifikasi mood memanfaatkan daftar lexicon bernama ANEW.Keshtkar dan Inkpen [34] memperkenalkan suatu

pendekatan hirarkis untuk klasifikasi mood. Liu dkk. [35] melakukan klasifikasi biner halaman blolg memanfaatkan lexicon subyektif. Kemudian Link [36] membedakan kalimat kunci dan trivial dalam proses klasifikasi sentimen dari blog. Kale dkk [39], Martineuw dan Hurst [40] dan Leskovec dkk. [41] telah mengklasifikasikan link-link ke dalam sentimen positif atau negatif. Ishino dkk. [42] mengklasifikasikan sentimen dari suatu blog berdasarkan opini terhadap link yang menuju (inlink) ke blog tersebut. Khan dkk [43] memanfaatkan konsep semantik dalam penentuan sentimen, sedangkan Li dkk [44] melibatkan pandangan *personal* dan *impersonal* dan klasifikasi sentimen suatu halaman blog.

1. Klasifikasi Topik dan Genre

Halaman-halaman web sebagian besar dibedakan berdasarkan topik (misal: politik, olahraga, gaya hidup dan ekonomi) dan genre (misal: blog, homepage dan e-commerce) [12]. Topik dan genre menyediakan cara untuk menggambarkan sifat dari teks yang memungkinkan penempatan dokumen ke suatu kelas. Genre dicirikan oleh kriteria eksternal seperti pembaca yang dijadikan sasaran dan tujuan komunikatif. Topik adalah mengenai teks itu sendiri. Suatu *blogpost* mengenai "naga" dapat dikatakan bergenre fiksi daripada berita [13]. Topik merupakan inti bahasan dari teks, sedangkan genre didefinisikan oleh maksud komunikatif yang disebarkan. Kadang sulit menentukan batas yang jelas antara topik dan genre. Karena itu, bagian ini membahas teknik klasifikasi topik dan genre secara bersamaan.

Lex dkk. [14, 15] mengklasifikasikan blog-blog berita (*news*) ke dalam salah satu dari lima kategori dari surat kabar Jerman dan Austria, yaitu politik, ekonomi, olah raga, budaya dan sains. Bagian penting dari penelitian ini adalah menggunakan berita yang telah terkelompok dari surat kabar untuk membentuk *classifier* (fase pelatihan), kemudian menguji *classifier* memanfaatkan data blog (fase pengujian). Algoritma klasifikasi yang digunakan adalah *Class Feature Centroid* (CFC). Pada pendekatan ini, setiap kelas diwakili oleh serangkaian term yang paling mewakili kelas tersebut, sehingga akurasinya sangat tergantung pada kualitas ***centroid*** dari masing-masing kelas. Dalam penentuan *centroid* kelas, term-term dibobot memanfaatkan distribusi *inner-class* dan *inter-class*. Detail dari CFC dapat dilihat dalam [16, 17]. Evaluasi dilakukan dengan membandingkan CFC dengan algoritma k-*Nearest Neighbor* (kNN) dan *Support Vector Machine* (SVM) yang termasuk algoritma klasifiksi teks standard dengan kinerja terbaik [18]. Hasil pengujian memperlihatkan bahwa akurasi dari CFC bersaing dengan kNN dan SVM. CFC unggul dalam waktu komputasi dan pemanfaatan ruang memory terutama untuk fase pengujian.

Hagiwara dkk. [19] membangun suatu sistem klasifikasi blog memanfaatkan label dari topik pada tingkatan situs blog. Metode klasifikasi yang digunakan adalah Naive Bayes yang dipadukan dengan algoritma EM. Nigam dkk. [20, 21] menyatakan bahwa kombinasi demikian dapat memberikan kinerja lebih baik dalam klasifikasi teks. Pengujian terhadap 634 blog (75.161 blogpost) memperlihatkankan bahwa akurasi dari classifer ini lebih baik daripada classifier Naive Bayes tanpa algoritma EM.

Sesungguhnya blog telah terkategorisasi melalui tag atau label yang diberikan oleh *blogger* untuk setiap *blog post*. Namun, tag ini bersifat subyektif dan tidak konsisten untuk setiap blog [4]. Tag "apple" dapat mengacu ke dokumen yang berisi informasi terkait buah-buahan atau perusahaan teknologi. Informasi tag juga tidak sejalan dengan domain lain yang telah mempunyai kategori umum, seperti pada surat kabar. Selain itu, kosa kata tag dari setiap blog berbeda dan berubah secara dinamis. Karena itu, folksonomi tidak dapat dimanfaatkan langsung untuk klasifikasi blog [15]. Tag merupakan fitur yang lebih efektif dalam klasifikasi dari pada judul dan deskripsi blog. Tag dapat

pula dijadikan sebagai obyek klasifikasi, yaitu bagaimana memprediksi rangkaian tag yang paling mewakili untuk suatu topik tertentu [8].

Tsai [4] mengelompokkan tag-tag pada blog ber-genre "Security" menggunakan model *tag-topic* yang merupakan pengembangan dari *Latent Dirichlet Allocation* (LDA). Setiap tag direpresentasikan oleh distribusi peluang terhadap topik, sedangkan setiap topik direpresentasikan sebagai suatu distribusi peluang terhadap term untuk topik tersebut. Pendekatan ini menyelesaikan masalah untuk mendapatkan tag dan term yang paling mewakili. Rangkaian tag sangat mungkin berisi *noise* karena tag dibuat oleh masing-masing blogger. Karena itu diperlukan teknik reduksi dimensi yang baik untuk menghilangkan tag-tag yang tidak perlu dan mengganggu.

Subramaniyaswamy [22] mengusulkan suatu pendekatan untuk meningkatkan akurasi klasifikasi blog berbasis ontologi topik menggunakan *Support Vector Machine* (SVM). Ontologi topik menghubungkan satu topik dengan lainnya melalui hubungan semantik. Topik disajikan sebagai *node* dan relasi topik sebagai *edge*. Ontologi topik dimaksudkan untuk mengenali dokumen-dokumen yang terkait untuk setiap topik. Isi blog diasosiasikan dengan sehimpunan kata kunci ontologi topik yang telah didefinisikan. Tag dan isi blog digunakan sebagai input. Kinerja pendekatan ini memberikan akurasi lebih baik daripada algoritma Naive Bayes.

Wiegand [23] mengevaluasi beberapa metode untuk klasifikasi polaritas terkait-topik pada tingkatan kalimat. Hasilnya menyatakan bahwa *classifier* polaritas berbasis klasifikasi teks *bag-of-words* (BoW) sederhana memberikan hasil yang kurang baik. Kinerja lebih baik dapat dicapai oleh classifier yang berbasis pada pencarian *lexicon*. Informasi polaritas yang di-*encode*-kan dalam *lexicon* polaritas lebih bersifat independen. Kinerja optimal dari *classifier* jenis ini dicapai saat suatu himpunan kecil dari fitur polaritas linguistik kecil digunakan dalam kombinasi dengan fitur jarak.

2. Klasifikasi Opini

Cakupan topik yang didiskusikan di dalam blog sangat luas, mulai dari tema santai sampai dengan masalah berat, termasuk perbedaan pandangan politik seperti pendapat mengenai aborsi, pemilihan umum dan imigrasi. Blog telah menjadi pembangkit informasi yang dapat digunakan untuk menggali opini dan trend [24] dan menjawab kebutuhan informasi opini bagi pengguna. Blogosphere merupakan fokus dari penelitian deteksi opini atau sentimen. Sebagai contoh, suatu partai politik memerlukan jawaban atas pertanyaan "apa pendapat media, *newsgroup*, *chat room*, dan blog mengenai kebijakan terbaru partainya?" atau suatu vendor komputer tablet memerlukan jawaban "mengapa produk tipe tertentu jarang dibeli oleh pelanggan?" [25]. Demartini [26] membangun suatu model yang secara otomatis memperkirakan opini publik dari blogosphere dengan menggali dan mengumpulkan informasi yang diekstrak dari blog terkait dengan kandidat presiden Amerika: Obama dan McCain.

Secara umum terdapat dua pendekatan umum dalam deteksi opini, yaitu teknik berbasis lexicon [27] dan berbasis *machine learning* [28]. Pendekatan berbasis *lexicon* menggunakan *lexicon* opini untuk menentukan apakah suatu dokumen bersifat opini (*opinionated*) atau tidak. Pendekatan *machine learning* menggunakan himpunan data anotasi untuk membangkitkan suatu model yang kemudian dapat digunakan untuk mengevaluasi data uji [29].

Missen [25] mengusulkan suatu *framework* deteksi opini mengikuti fitur berbasis jejaring sosial. *Blogosphere* dilihat sebagai suatu graf berarah dimana setiap *node* (yaitu blog) dihubungkan ke *node* lain melalu koneksi pertemanan atau koneksi relevansi. *Framework* ini bersandar pada dua faktor utama, yaitu:

1. Menggunakan koneksi antar blog (atau *blog post*) yang memberikan

banyak kepentingan dalam jejaring sosial. Ini disebut pendekatan jejaring sosial murni.

2. Memanfaatkan kemungkinan memandang setiap *node* dalam konteks berbeda. Ada dua konteks yang paling berpengaruh, yaitu: konteks tingkat Network dan tingkat Topik.

Mood

Mood merupakan suatu kondisi pikiran seperti senang, sedih dan marah. Klasifikasi mood berbasis teks termasuk dalam masalah *opinion and sentiment mining* [30]. Klasifikasi mood menghadirkan tantangan tambahan diluar kategorisasi teks standard. Proses kognitif kompleks dari formulasi mood membuatnya bergantung pada konteks sosial khusus dari pengguna, asosiasi *idiosyncratic* dari mood dan kosakata, sintaks dan gaya bahasa atau genre dari teks. Pada blog, ini diperlihatkan oleh blogger dalam mengekspresikan gaya berbeda, teks relatif pendek dan bahasa informal seperti jargon, singkatan dan error tata-bahasa. Teknik pemilihan fitur dalam *machine learning* mahal secara komputasi, bersandar pada data berlabel untuk mempelajari fitur-fitur diskriminatif; tetapi koleksi blog tumbuh cepat dan berlanjut, mengharuskan adanya himpunan fitur yang bekerja tanpa memerlukan tahapan pemilihan fitur terbimbing untuk klasifikasi mood. Salah satu solusinya adalah menggunakan *affective norm for English words* (ANEW) yang merupakan irisan antara psikologi dan linguistik dalam proses klasifikasi mood [31]. Daftar ANEW berisi 1.034 *term* unik dengan score *valence* afektif (*unpleasant ~ pleasant*), *arousal* (*calm ~ excited*), dan dominansi (*submissive ~ dominance*). Daftar ini dapat digunakan untuk mengidentifikasi jenis mood berdasarkan pada analisis *lexical* dengan memetakan *term* dalam teks ke *term* dalam daftar tersebut [32].

Jung dkk. [32] telah menggunakan ANEW bersama dengan ConceptNet untuk membangun suatu sistem klasifikasi mood yang tidak melibatkan panjang dan gaya penulisan. ConceptNet adalah suatu basis pengetahuan *commonsense* yang tersedia bebas dan merupakan salah satu toolkit pemrosesan bahasa alami yang mendukung banyak aplikasi tugas penalasan tekstual, termasuk peng-intisari-an topik, *affect-sensing*, pembuatan analogi dan inferensi berorientasi konteks lainnya. Pendekatan ini mengambil fitur unik teks blog dan menghitung statistik sederhana seperti *term frequency*, n-gram, dan *point-wise mutual information* (PMI) untuk metode klasifikasi SVM. Transisi mood dalam teks blog ditangkap dengan mengembangkan segmentasi tingkatan paragraf berbasis analisis aliran mood menggunakan operasi GuessMood dari ConceptNet dan modul sensing afektif berbasis ANEW. Sistem ini mengklasifikasi teks blog ke dalam salah satu dari 4 mood: senang, sedih, marah atau takut.

Emosi dapat direpresentasikan dalam pendekatan dimensi dan diskret. Pada pendekatan dimensi, status dari emosi dikodekan sebagai kombinasi beberapa faktor seperti valensi dan arousal (penimbulan). Sedangkan pada pendekatan diskret, setiap emosi mempunyai *coincidence* unik dari pengalaman, psikologi dan perilaku. Nguyen dkk. [31] menggunakan basis dimensi untuk memperkirakan emosi dalam blog.

Tahapan penting sebelum klasifikasi atas pemilihan fitur. Fitur dapat ditampilkan dalam bentuk *bag of word*. Skema pembobotan TF, DF, dan TF.IDF sering digunakan untuk meningkatkan kekuatan diskriminatif dari fitur atau term. Pemilihan fitur juga dapat berbasis interaksi *Term-Class* yang menangkap ketergantungan antara term dan label kelas yang berkorespondensi. Ada tiga metode pemilihan umum dalam kelompok ini, yaitu *information gain* (IG), *mutual information* (MI) dan CHI [33]. IG mengambil information gain (diukur dalam bit) kapan suatu *term* hadir atau absen; MI mengukut mutual information antara suatu term dan kabel kelas; dan

CHI mengukur ketergantungan antara suatu *term* dan label kelas dengan membandingkan terhadap satu derajat dari distribusi statistik X^2 bebas. Pada analisis sentimen, beberapa emosi menggunakan himpunan *lexicon* yang nilai subyektifnya telah ditentukan seperti ANEW. Kata-kata dalam ANEW dinilai dalam istilah *valence*, *arousal* dan dominansi yang disampaikannya [31].

Komponen linguistik seperti pemanfaatan khusus dari *adverb*, *adjective* atau *verb* dapat menjadi indikator kuat bagi penyimpulan mood [30]. Nguyen dkk. [31] menjalankan suatu tagger *part-of-speech* bernama SS-Tagger untuk mengenali semua term yang dapat ditag sebagai verb, adjactive dan adverb. Akurasi yang diperoleh *reasonable*. Tiga basis pembobotan term (TF, DF, TF.IDF) dan tiga metode pemilihan berbasis interaksi Term-Class (IG, MI, CHI) diberlakukan dalam ujicoba yang dievalusi dengan 10-*fold cross validation*. Metode pemilihan fitur ini diterapkan terhadap semua term (unigram) dan subset dari term yang ditag dengan POS spesifik. Hasil penelitiannya memperlihatkan bahwa skema pemilihan fitur yang memberikan akurasi dan F-Score (berbasis precision dan recall) tertinggi pada Naive Bayes Classifier (NBC) adalah IG (77 - 79%), sebagaimana telah dinyatakan dalam [33]. Skema MI dan CHI tidak tepat jika digunakan untuk klasifikasi mood, bahkan tidak muncul dalam 10 hasil tertinggi. Pembobotan TF dan DF memberikan hasil lebih baik dari pada TF.IDF untuk semua kasus unigram, berlawanan dengan laporan pada kasus text mining. Terkait dengan komponen linguistik, kombinasi *adjactive*, *verb* dan *adverb* mendominasi 10 hasil tertinggi (75 – 76%) dan memberikan kinerja hampir sama dengan pendekatan yang menggunakan semua term. Tanpa perlu tahapan pemilihan fitur terbimbing, fitur ANEW memberikan hasil baik (71%) meskipun tidak tertinggi.

Keshtkar dan Inkpen [34] memperkenalkan suatu pendekatan hirarkis untuk klasifikasi mood. Berbagai label mood dilibatkan dan pendekatan ini bersifat fleksibel sehingga himpunan mood dapat diubah. Klasifikasi hirarkis digunakan atas pertimbangan adanya kelas-kelas (label mood) yang secara alami terkelompok dalam suatu hirarki. Ini membuat classifier mempelajari *coarse-grained distinctions* terlebih dahulu dan mempelajari *more fine grained* pada tingkatan berikutnya dalam hirarki. Ujicoba yang dilakukan menyatakan bahwa pendekatan ini lebih baik daripada klasifikasi standard (*flat*) ke dalam 132 mood. Penelitian ini juga menyatakan bahwa meskipun corpus blog mengandung banyak kata dalam berbagai domain, sistem dapat memilih fitur yang mengarah ke klasifikasi akurat dari emosi dan mood yang diekspresikan. Dimulai dengan semua kata sebagai fitur, kemudian menggunakan teknik pemilihan fitur dilakukan reduksi terhadap ruang fitur. Fitur berorientasi sentimen dan *emoticon* ditambahkan. Perlakuan ini ternyata secara signifikan meningkatkan akurasi klasifikasi.

Sentimen

Analisis sentimen adalah tugas kategorisasi teks yang fokus pada pengenalan dan pengklasifikasian teks yang mengandung opini ke arah suatu subyek yang diberikan [35]. Kajian terhadap sentimen pada blog menjadi hotspot riset dengan prospek aplikasi yang luas [36].Masalah kunci dalam analisis sentimen adalah menentukan polaritas dari suatu dokumen, condong ke arah positif (*thumb up*) atau negatif (*thumb down*). Pada klasifikasi teks berbasis topik, akurasi tinggi dapat dicapai karena kelompok topik biasanya dapat dipisahkan dengan baik satu dengan lainnya, berangkat dari fakta bahwa pemanfaatan kata membedakan dengan sangat antara dua dokumen yang secara topik berbeda. Banyak review membingungkan secara sentimen karena berbagai alasan. Pernyataan obyektif disisip dengan pernyataan subyektif dapat mengacaukan metode pembelajaran dan pernyataan subyektif dengan sentimen yang berkonflik lebih lanjut menyulitkan tugas klasifikasi.

Liu dkk. [35] menyelidiki dan dan mencoba meningkatkan kinerja klasifikasi polaritas biner (positif vs. negatif) dalam blog menggunakan dua metode, yaitu (1) mengintregrasikan analisis relevansi topik untuk mengerjakan klasifikasi polaritas spesifik topik dan (2) mengadopsi metode adaptif dengan menggabungkan banyak hipotesis classifier dari beberapa domain sebagai fitur. Dalam klasifikasi, digunakan himpunan *lexicon subjective* yang terdiri dari 2034 kata positif dan 4145 kata negatif. Terdapat 4 jenis fitur yang dilibatkan, yaitu fitur lexical (LF), polarized lexical (PL), bigram polarized (BP) dan transition word (T). Saat diterapkan maximum entropy classifier dan evaluasi menggunakan 10-*fold cross validation*, hasil klasifikasi polaritas memperlihatkan bahwa kinerja terhadap blog (71 – 72.5) lebih buruk daripada terhadap data review (82 – 84%). Kemungkinan besar ini disebabkan oleh variasi besar dalam term dari content dan style dalam blog.

Klasifikasi polaritas blog ditingkatkan dengan memanfaatkan konteks blog relevan topik. Konteks relevan dapat berupa semua kata dalam kalimat yang relevan dengan suatu topik (termasuk kalimat sebelum dan setelahnya) atau kata-kata content yang hanya terdiri dari noun, verb, adjactive dan adverb. Hasil evaluasi memperlihatkan adanya peningkatan akurasi terutama klasifikasi yang hanya melibatkan kata-kata content, yaitu mencapai lebih dari 75%. Ini ditingkatkan lagi memanfaatkan dua metode adaptif. Pertama.dalam setiap 10-*fold cross validation*, dilakukan penyatuan data training blog (90%) dengan semua data review dari 5 domain. Kedua, memperbesar fitur dengan hipotesis yang diperoleh dari classifier yang ditraining menggunakan data domain lain, yaitu dengan mentraining 5 classifier menggunakan data dari 5 domain review dan kemudian mengencodekan hipotesis dari classifier tersebut sebagai fitur bagi data training dari blog. Dari dua pendekatan ini, hanya pendekatan kedua yang memberikan peningkatan akurasi (77%).

Pendekatan yang hampir sama diteliti oleh Lin dkk. [36]. Berlandaskan pada anggapan umum bahwa tidak semua bagian dokumen bernilai informatif sama untuk menyimpulkan polaritas dari dokumen, maka diberikan nilai berbeda antara kalimat kunci dan trivial, dan ini diyakini dapat meningkatkan kinerja klasifikasi sentimen. Kalimat kunci diekstrak dan kemudian memasukkan kalimat tersebut dalam klasifikasi sentimen *supervised* dan *semi-supervised*. Perbedaan dua klasifikasi ini dapat dilihat dalam [37].

Pada tahapan ekstraksi kalimat, diambil tiga atribut, yaitu atribut sentimen, posisi dan kata-kata khusus. Atribut sentimen membedakan apakah suatu kalimat mengandung perasaan subyektif. Atribut posisi menjamin bahwa kalimat pada awal dan akhir (dari paragraf atau dokumen) mempunyai peluang lebih tinggi daripada bagian tengah. Atribut kata khusus meningkatkan bobot dari kalimat yang mengandung kata khusus, seperti “overall”. Nilai akhir dari setiap kalimat adalah penjumlahan terbobot dari score tiga atribut dan kalimat dengan nilai tertinggi dianggap sebagai kalimat kunci.

Selanjutnya data training dibagi ke dalam dua bagian, yaitu kalimat kunci dan trivial. Ruang fitur dari kalimat kunci biasanya lebih kecil daripada ruang kalimat trivial karena kalimat kunci hanya terdiri dari satu kalimat sedangkan kalimat trivial terdiri dari banyak kalimat. Kalimat kunci biasanya bersifat sumatif dan kalimat trivial bersifat deskriptif, sehingga ekspresi dalam kalimat kunci menjadi kurang variatif tetapi lebih diskriminatif daripada kalimat trivial.

Dalam klasifikasi sentimen supervised, dilakukan adopsi metode kombinasi classifier. Tiga classifier basis dilatih: f1 dan f2 dilatih menggunakan dataset dari kalimat kunci dan kalimat trivial, dan f3 dilatih menggunakan data training lengkap. Setiap classifier basis mengeluarkan tidak hanya label kelas tetapi juga beberapa jenis ukuran kepercayaan seperti peluang posterior dari

contoh testing yang masuk ke setiap kelas. Berikutnya, label kelas dari contoh testing ditentukan dengan kombinasi dari f1, f2 dan f3.

Jika suatu dokumen mempunyai kalimat kunci, *classifier* kalimat kunci lebih dipercaya dengan keputusannya. Tetapi tidak setiap dokumen mempunyai kalimat kunci. Pada pembelajaran semi-supervised digunakan suatu algoritma bernama co-training dengan menggabungkan data tak berlabel dalam klasifikasi sentimen. Saat suatu contoh dapat diberikan secara confident oleh classifier kalimat kunci, tidak menjamin bahwa akan mudah diklasifikasi oleh classifier kalimat trivial. Karena itu, classifier kalimat trivial mengambil informasi berguna untuk meningkatkan dirinya dan sebaliknya. Perbedaan dari setiap classifier mengakibatkan algoritma *co-training* menjadi *applicable*.

Algoritma klasifikasi Naive Bayes tanpa pemilihan dan reduksi fitur digunakan pada klasifikasi supervised. Akurasi terbaik diperoleh saat diterapkan klasifikasi kombinasi terhadap classifier kalimat kunci. Algoritma Transductive SVM digunakan sebagai pembanding pada klasifikasi semi-supervised. Classifier kombinasi (Naive Bayes) yang melibatkan co-training memberikan akurasi paling tinggi [36].

Tingkat kepentingan (*authoritative*) informasi dari suatu halaman web dapat diketahui berdasarkan jumlah link yang menuju halaman web tersebut [38]. Tetapi algoritma ini tidak mencerminkan sentimen penulis (author) mengenai situs yang dihubungkan. Ini memungkinkan blog yang disalahgunakan akan diranking tinggi oleh mesin pencarian. Penelitian awal mengenai klasifikasi sentimen berdasarkan link dilakukan oleh Kale dkk. [39] yang mengusulkan suatu metode untuk mengklasifikasi link-link dalam suatu blog dalam kategori positif atau negatif. Himpunan *lexicon* positif dan negatif dicocokkan dengan kata yang diambil sebelum dan setelah link untuk mendapatkan polaritas. Martineau dan Hurst [40] menerapkan pendekatan *machine-learning* untuk klasifikasi link dari beberapa sudut pandang menggunakankata-kata yang muncul dalam konteks sitasi URL sebagai fitur. Leskovec dkk. [41] menghimpunan data set dari Epinions, Slashdot dan Wikipedia untuk mengklasifikasi blog dalam suatu social network dan melibatkan terori keseimbangan dan status dalam psikologi sosial.

Ishino dkk. [42] mengusulkan suatu metode untuk mengklasifikasi sentimen dari suatu blog dengan meninjau sentimen dari link yang menuju ke blog tersebut. Langkah pertama yang dilakukan adalah mengekstrak letak teks link (citing area) dalam suatu blog. Kemudian polaritas link diklasifikasi menggunakan informasi di dalam citing area tersebut. Informasi ini berguna untuk mengidentifikasi blog-blog otoritatif dalam blogosphere secara efisien, karena blog yang dihubungkan secara positif dari banyak blog lain dianggap lebih reliable.Berdasarkan pengamatan, dari 840 link yang dihimpun, ternyata hanya 5 link yang bersentimen negatif. Karena itu polaritas link dibagi menjadi 2, yaitu positif atau lainnya. Kata-kata yang diperoleh dari sekitar link dibandingkan dengan suatu himpunan lexicon yang berisi kata-kata yang berorientasi positif. Hasil penelitian menunjukkan bahwa pendekatan ini efektif.

Analisis sentimen dapat dilakukan pada tingkatan dokumen, kalimat dan kata/fitur [35]. Khan dkk. [43] mencoba mengklasifikasi kalimat dari review dan blog dengan memanfaatkan score semantik dari kalimat subyektif yang diekstrak dari SentiWordNet [Esuli] untuk menghitung polaritas kalimat. Pada tahap awal, dokumen dipercah menjadi kalimat-kalimat dan disimpan dalam bentuk *Bag of Sentences* (BoS). Kemudian *noise* di dalam kalimat dihilangkan menggunakan pembetulan ejaan, simbol dan karakter khusus dikonversi ke ekspresi teks, setiap kata darikalimat sitag menggunan *Part of Speech* (POS) dan posisi dari setiap kata dalam kalimat disimpan. Selanjutnya dibangun suatu kamus holistik (vektor

fitur) dari fitur-fitur penting dengan posisinya dalam kalimat. Kalimat-kalimat dikelompokkan ke dalam obyektif dan subyektif menggunakan pendekatan machine-learning dan lexical. Kamus lexical digunakan sebagai basis pengetahuan untuk memeriksa polaritas dari kalimat subyektif, apakah positif, negatif atau netral. Terakhir dilakukan pemeriksaan dan update polaritas menggunakan struktur kalimat dan fitur kontekstual dari setiap term dalam kalimat. Ujicoba yang dilakukan memperlihatkan metode ini mencapai akurasi 97.8% pada tingkat feedback dan 86.6% pada tingkat kalimat, melampaui metode tingkat kata dan machine-learning.

Li dkk [44] mengajukan model dua pandangan, yaitu pandangan personal dan impersonal untuk klasifikasi sentimen. Pandangan personal (terkait langsung orang tertentu) terdiri dari kalimat subyektif yang subyeknya merupakan seseorang, sedangkan pandangan impersonal (tidak mengenai seseorang) terdiri dari kalimat obyektif yang subyeknya bukan seseorang. Pandangan demikian secara lexical bersifat isyarat dan harus diperoleh tanpa adanya data yang diberikan label sebelumnya. Pendekatan unsupervised learning digunakan untuk mendapatkannya. Metode kombinasi dan algoritma co-training digunakan dengan klasifikasi sentiment supervised dan semi-supervised. Evaluasi terhadap tinjauan produk dari 8 domain memperlihatkan bahwa pendekatan ini secara signifikan meningkatkan kinerja, baik pada klasifikasi supervised maupun semi-supervised.

KESIMPULAN

Kami telah melakukan studi literatur mengenai berbagai pendekatan klasifikasi blog sekitar 3 tahun terakhir. Framework blog mining sengaja diberikan di awal tulisan agar dapat diketahui kerangka lengkap suatu sistem *knowldge discovery* dari blog. Beberapa penyesuaian harus dilakukan agar teknik klasifikasi web tradisional dapat diterapkan terhadap blog. Komponen penyusun blog seperti judul, isi dan komentar, tag (label), penulis, hyperlink, permalink, outlink, tanggal dan jam dapat digunakan untuk meningkatkan akurasi klasifikasi. Pada awal hadirnya blog, klasifikasi biner digunakan untuk membedakan blog dari halaman web biasa. Blog-blog dapat diketegorisasikan ke dalam kelas-kelas topik (termasuk genre) dan opini (mood dan sentimen) yang telah didefinisikan. Pada klasifikasi topik dan genre, algoritma kNN, Naive Bayes, CFC, SVM dan pendekatan *machine learning* lainnnya memberikan kinerja yang baik. Pemanfaatan ontologi topik dan tag terbukti dapat meningkatkan akurasi klasifikasi. Pada deteksi opini, pendekatan berbasis lexicon seperti ANEW cenderung lebih banyak digunakan, dan sering dikombinasikan dengan teknik *machine* learning. Opini dari suatu situs blog juga dapat diprediksi berdasarkan opini di sekitar teks link (*inlink*) yang menuju situs tersebut. Pada waktu mendatang, kajian akan diperluas dan diperdalam, seperti keterlibatan lebih lanjut dari tag, link dan analisis jejaring sosial.

DAFTAR PUSTAKA

[1] Michael Chau, Porsche Lam, Boby Shiu, Jennifer Xu, Jinwei Cao (2009) **A Blog Mining Framework**, IT Pro January/February 2009, IEEE Computer Society

[2] Technorati (2011) **State of the Blogosphere 2011**, URL: http://technorati.com/blogging/article/state-of-the-blogosphere-2011-introduction/page-2/

[3] Geetika T. Lakshmanan, Marten A. Oberhofer (2010) **Knowledge Discovery in the Blogosphere Approaches and Challenges**, IEEE Internet Computing

[4] Flora S. Tsai (2011) **A Tag-Topic Model for Blog Mining**, Expert System with Applications (ESwA) Journal, Vol. 38, Page 5330 – 5335

[5] Jiawei Han, Micheline Kamber, Jian Pei (2012) **Data Mining**

Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers

[6] Hae-Ching Chang, Kao-chi, Yeh (2008) Clarifying The Difficulties And Management Of Blogging, Journal of Information, Technology, and Society (JITAS), Vol. 8 No.2, URL: jitas.im.cpu.edu.tw/2008-2/1.pdf

[7] Bonnie A. Nardi, Diane J. Schiano, Michelle Gumbrecht (2004) Blogging As Social Activity, Or, Would You Let 900 Million People Read Your Diary?, dalam Proceedings of Conference On Computer Supported Cooperative Work: 222-231, ACM, URL: http://home.comcast.net /~diane.schiano/CSCW04.Blog.pdf

[8] Xiaoguang Qi, Brian D. Davidson (2009) Web page Classification: Features and Algorithms, ACM Computing Survey, Vol. 41, No. 2, Article 12, URL: http://www.cse.lehigh.edu/~xiq204 /pubs /classification-survey/LU-CSE-07-010.pdf

[9] Tomoyuki Nanno, Yasuhiro Suzuki, Toshiaki Fujiki, Manabu Okumura (2004) Automatically Collecting, Monitoring, And Mining Japanese Weblogs, dalam Proceedings Of The 13th International World Wide Web Conference On Alternate Track Papers & Posters (WWW Alt.): 320-321, ACM, URL: www.iw3c2.org/WWW2004/docs/2p320.pdf

[10] Erik Elgersma, Maarten De Rijke (2005) Learning To Recognize Blogs: A Preliminary Exploration, EACL Workshop: New Text—Wikis And Blogs And Other Dynamic Text Sources, URL: http://www.Sics.Se/Jussi/Newtext/Working_Notes/05_Elgersma_ Derijke.Pdf

[11] Feng Yu, Dequan Zheng, Tiejun Zhao, Xiao Cheng (2008) Structure and Content Based Blog Pages Identification, dalam Fifth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE

[12] Ioannis Kanaris & Efstathios Stamatatos (2009) Learning To Recognize Webpage Genres, Information Processing and Management Journal, Volume 45 Issue 5, URL: http://www.icsd.aegean.gr/lecturers/stamatatos/papers/IPM2009%20preprint.pdf

[13] Philipp Petrenz (2009) Assessing Approaches To Genre Classification, M.Sc. Thesis, School Of Informatics University Of Edinburgh, URL: http://www.inf.ed.ac.uk/publications/thesis/online /IM090692.pdf

[14] Alisabeth Lex, Christin Seifert, Michael Granitzer, Andreas Juffinger (2009) Automatic Blog Classification: A Cross Domain Approach, dalam Proceedings of IADIS International Conference WWW/Internet 2009, URL: http://www.iadisportal.org/digital-library/automated-blog-classification-a-cross-domain-approach

[15] Alisabeth Lex, Christin Seifert, Michael Granitzer, Andreas Juffinger (2010) Efficient Cross-Domain Classification of Weblogs, International Journal of Intelligeny Computing Research (IJICR), Vol 1, Issue 1/2, URL: http://infonomics-society.org/IJICR/Efficient%20Cross_Domain%20Classification%20of%20Weblogs.pdf

[16] Verayuth Lertnattee, Thanaruk Theeramunkong (2004) Effect Of Term Distributions On Centroid-Based Text Categorization, Information Sciences - Informatics and Computer Science Journal, Volume 158 Issue 1, URL: http://ccc.inaoep.mx/~villasen/index_archivos/cursoTL/articulos/Lertnattee-EffectOfTermDistributionsOn

Centroid-based TextCategorization.pdf

[17] Hu Guan, Jingyu Zhou, Minyi Guo (2009) A Class-Feature-Centroid Classifier for Text Categorization, WWW 2009, April 20–24, Madrid, Spain, ACM, URL: www2009.eprints.org/21/1/p201.pdf

[18] Fabrizio Sebastiani (2002) Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1: 1-47, URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf

[19] Ken Hagiwara, Hiroya Takamura, Manabu Okumura (2010) Constructing Blog Entry Classifiers Using Blog-Level Topic Labels, Proceedings of Asia Information Retrieval Symposium (AIRS) 2010, Springer

[20] Andrew McCallum, Kamal Nigam (1998) A Comparison of Event Models for Naive Bayes Text Classification. Proceedings of AAAI 1998 Workshop on Learning for Text Cetegorization: 41-48, 1998, URL: http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf

[21] Kamal Nigam, Andrew McCallum, Sebastian Thrun, Tom Mitchell (2000) Text Classication from Labeled and Unlabeled Documents using EM, Machine Learning Journal: 1-34, Vol 39 Issue 2-3, URL: www.kamalnigam.com/papers/emcat-mlj99.pdf

[22] Subramaniyaswamy, V, S. Chenthur Pandia (2012) An Improved Approach for Topic Ontology Based Categorization of Blogs Using Support Vector Machine, Journal of Computer Science 8 (2): 251-258, URL: http://thescipub.com/pdf/10.3844/jcssp.2012.251.258

[23] Michael Wiegand, Dietrich Klakow (2009) Topic-Related Polarity Classification of Blog Sentences, Proceeding EPIA '09 Proceedings of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence, URL: www.lsv.uni-saarland.de/epia.pdf

[24] Macdonald, C.; Santos, R. L.; Ounis, I.; and Soboroff, I. (2010) Blog track research at TREC. SIGIR Forum 44:58–75, URL: http://www.sigir.org/forum/2010J/2010j-sigirforum-macdonald.pdf

[25] Malik Muhammad Saad Missen, Guillaume Cabanac , Mohand Boughanem (2010) Opinion Detection in Blogs: What is still Missing?, International Conference on Advances in Social Networks Analysis and Mining, IEEE, URL: http://acadmedia.wku.edu/Zhuhadar/nikhile/ASONAM-2010/ASONAM-47.pdf

[26] Gianluca Demartini, Stefan Siersdorfer, Sergiu Chelaru,Wolfgang Nejdl (2011) Analyzing Political Trends in the Blogosphere, Fifth International AAAI Conference on Weblogs and Social Media, URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2838/3244

[27] B. Ernsting, W. Weerkamp, and M. de Rijke (2007) The University of Amsterdam at the TREC 2007 Blog Track, URL: http://staff.science.uva.nl/~mdr/Publications/Files/trec2007-wn-blog.pdf

[28] S. Gerani, M. Carman, and F. Crestani (2009) Investigating Learning Approaches for Blog Post Opinion Retrieval, ECIR 2009, URL: bradipo.net/mark/papers/gerani_ecir2009.pdf

[29] Andrea Esuli, Fabrizio Sebastiani (2006) SentiWordNet: A Publicly Available Lexical Resource For Opinion Mining, LREC-06: ELRA, URL: gandalf.aksis.uib.no/lrec2006/pdf/384_pdf.pdf

[30] Bo Pang, Lillian Lee (2008) Opinion Mining And Sentiment Analysis, Foundations And Trends In Information Retrieval, Vol. 2, No 1-2: 1–135, URL: www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[31] Thin Nguyen, Dinh Phung, Brett Adams, Truyen Tran, Svetha Venkatesh (2010) Classification and Pattern Discovery of Mood in Weblogs, PAKDD 2010, Springer

[32] Yuchul Jung, Hogun Park, Sung Hyon Myaeng (2006) A Hybrid Mood Classification Approach for Blog Text, PRICAI 2006: 1099 - 1103, Springer

[33] Yiming Yang, Jan O. Pedersen (1997) A Comparative Study On Feature Selection In Text Categorization. Proceedings. of ICML, pp. 412–420, URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.9956

[34] Fazel Keshtkar, Diana Inkpen (2011) A Hierarchical Approach To Mood Classification In Blogs, Natural Language Engineering 18 (1): 61–81, Cambridge University Press

[35] Feifan Liu, Dong Wang, Bin Li, Yang Liu (2010) Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods, Human Language Technology: The 2010 Annual Conference of the North American Chapter of the ACL, 309-312, URL: www.aclweb.org/anthology/N10-1042

[36] Zheng Lin, Songbo Tan, Xueqi Cheng (2011) Using Key Sentence to Improve Sentiment Classification, Proceddings of Advanced Information Retrieval Systems (AIRS) 2011, Springer

[37] Shoushan Li, Zhongqing Wang, Guodong Zhou, Sophia Yat Mei Lee (2011) Semi-Supervised Learning for Imbalanced Sentiment Classification, Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, URL: ijcai.org/papers11/Papers/IJCAI11-306.pdf

[38] Sergey Brin, Larry Page (1998) The Anatomy of A Large-Scale Hypertext Web Search Engine. Computer Networks and ISDN System Archive 30 (1-7), 107-117

[39] Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java, Tim Finin, Anupam Joshi (2007) Modelling Trust and Influence in the Blogosphere Using Link Polarity. International Conference on Weblogs and Social Media, URL: http://ebiquity.umbc.edu/_file_directory_/papers/364.pdf

[40] Justin Martineau, Matthew Hurst (2008) Blog Link Classification. International Conference on Weblogs and Social Media, URL: http://ebiquity.umbc.edu/_file_directory_/papers/510.pdf

[41] Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg (2010) Predicting Positive and Negative Links in Online Social Networks, 10th WWW, URL: cs.stanford.edu/~jure/pubs/signs-www10.pdf

[42] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa (2011) Automatic Classification of Link Polarity in Blog Entries, Proceedings of Asia Information Retrieval Symposium (AIRS) 2011, Springer

[43] Aurangzeb Khan, Baharum Baharudin, Khairullah Khan (2011) Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews, Trends in Applied Science Research 6 (10); 1141-1157, URL: eprints.utp.edu.my/6435/1/207-627-1-PB.pdf

[44] Shoushan Li, Chu-Ren Huang, Guodong Zhou, Sophia Yat Mei Lee (2010) Employing Personal/Impersonal Views in Supervised and Semi-supervised

Sentiment Classification, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 414–423, 2010, URL: www.aclweb.org/anthology/P10-1043