MEMANFAATKAN *BIG DATA* UNTUK MENDETEKSI EMOSI

Aprianti Putri Sujana
Teknik Komputer Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung (STEI ITB)
e-mail: putrisujana@students.itb.ac.id

ABSTRAK

Layanan sosial media merupakan penyedia sumber daya yang menyediakan data yang cukup besar. Data yang cukup besar ini kemudian dapat dimanfaatkan dengan berbagai kebutuhan. Kebutuhan yang digunakan untuk berbagai tujuan. Tujuan tersebut dapat berupa pengolahan data untuk memonitoring pengguna.

Dengan berkembangnya sosial media, akan tersimpan banyak data yang akan terus menerus bertambah. Setiap pertambahan pengguna, bertambah pula lokasi data yang tersimpan. Apabila pengguna menikmati layanan dari social media tersebut, ditambah jika layanan tersebut menambah benefit, maka semakin besar pula data yang akan tersimpan pada server sosial media tersebut. Data yang berukuran raksasa ini dapat diproses dan dimanfaatkan.

Paper ini akan menjelaskan tentang pemanfaatkan data yang tersimpan di microblogging twitter untuk mendeteksi emosi melalui hastag. Dengan memanfaatkan hastag, akan dibuat dataset tersendiri. Menggunakan algoritma Naïve Bayes dan Liblinear.

Kata kunci: Big Data, Naïve Bayes

1. PENDAHULUAN

Emosi bersifat umum dan penting untuk semua aspek kehidupan kita. Ini mempengaruhi keputusan dalam hubungan social. Membentuk perilaku kita sehari-hari, bahkan kenangan kita. Dengan pertumbuhan teknologi yang cepet kita dapat mengekspresikan emosi tersebut dan mempublikasikannya dengan microblog, posting blog dan forum diskusi. Sehingga banyak dikembangkan alat otomatis untuk menganalisis emosi seseorang yang dapat dinyatakan dengan teks.

Mengidentifikasi emosi diekspresikan dalam teks sangat menantang untuk setidaknya dua alasan. Pertama, emosi bisa implisit dan dipicu oleh peristiwa atau situasi tertentu. Teks menggambarkan suatu peristiwa atau situasi yang menyebabkan emosi bisa tanpa kata-kata secara eksplisit.

Sebagian besar penelitian identifikasi emosi saat ini bergantung pada data training dijelaskan secara manual [1], [2]. Penjelasan data manual oleh para ahli memakan waktu lebih lama. Selain itu, berbeda dengan tugas-tugas penjelasan lain seperti entitas atau deteksi topik, menentukan emosi dalam teks cenderung subyektif dan bervariasi, dan karenanya, kurang dapat diandalkan. Akibatnya, sebagian besar dataset emosi yang ada relatif kecil, dari urutan ribuan entri, yang gagal untuk menyediakan cakupan yang komprehensif dari peristiwa emosi pemicu dan situasi.

Meskipun ada kekurangan data berlabel cukup untuk penelitian emosi , banyak layanan sosial media telah memasuki era data yang besar. *Twitter*, layanan microblogging populer , menyediakan lebih dari 340 juta tweet per hari pada berbagai topik , dan menjadi bagian penting dari itu adalah tentang apa yang terjadi dalam kehidupan kita sehari-hari dinyatakan menggunakan *hashtags* emosi. Misalnya, " berangkat ke rumah sakit #sedih", dalam tweet pengguna menambahkan catatan tweet dengan hashtag #sedih untuk mengekspresikan kegelisahan emosi .

2. DASAR TEORI

Big Data

Big data adalah data berukuran besar yang volumenya akan terus bertambah, terdiri dari berbagai jenis atau varietas data, terbentuk secara terus menerus dengan kecepatan tertentu dan harus diproses dengan kecepatan tertentu pula.

Big data dapat juga didefinisikan data yang sudah sangat sulit untuk dikoleksi, disimpan dan dikelola maupun dianalisa dengan menggunakan system database yang biasa karena volumenya yang terus berlipat.

Dari segi teknologi, akan bermunculan akan pentingnya kemampuan untuk memproses big data. Semenjak itu, teknik akses dan penyimpanan data KVS (*Key-Value Store*) dan teknik komputasi parallel yang disebut *MapReduce*.

Data Mining

Kemajuan dalam pengumpulan data dan teknologi penyimpanan yang cepat memungkinkan organisasi menghimpun jumlah data yang sangat luas. Alat dan teknik analisis data yang tradisional tidak dapat digunakan untuk mengekstrak informasi dari data yang sangat besar. Untuk itu diperlukan suatu metode baru yang dapat menjawab kebutuhan tersebut.

Data mining merupakan teknologi yang menggabungkan metode analisis tradisional dengan algortima yang canggih untuk memproses data dengan volume besar.

Data mining atau Knowledge Discovery in Databases (KDD) adalah pengambilan informasi yang tersembunyi, dimana informasi tersebut sebelumnya tidak dikenal dan berpotensi bermanfaat. Proses ini meliputi sejumlah pendekatan teknis yang berbeda, seperti *clustering*, *data summarization*, *learning classification rules*.

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

Salah satu tuntuntan dari data mining ketika diterapkan pada data berskala besar adalah diperlukan metodologi sistematis tidak hanya ketika melakukan analisa saja tetapi juga ketika mempersiapkan data dan juga melakukan interperstasi dari hasilnya sehingga dapat menjadi aksi ataupun keputusan yang bermanfaat.

Text Mining

Menurut Feldman, R. dan Sanger, J. "*text mining* adalah sebuah proses pengetahuan intensif dimana pengguna berinteraksi dan bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis" (2007, hlm. 1). *Text mining* mencoba untuk mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi dari suatu pola menarik. Sumbner data berupa sekumpulan dokumen dan pola menarik yang tidak ditemukan dalam bentuk database record, tetapi dalam data text yang tidak terstruktur.

3. PEMBAHASAN

Dengan menggunakan 131 hastags emosi sebagai kata kunci dan mengumpulkan 5 juta tweet untuk 7 kategori emosi (sukacita, kesedihan, kemarahan, cinta, rasa takut , rasa syukur, kejutan) antara 10 November 2011 dan 22 Desember , 2011 (lihat Tabel II). Dengan menggunakan algoritma Multinomial Naïve Bayes (MNB)

3.1 Mengumpulkan Data Emosi

Pada bagian ini, menjelaskan bagaimana secara otomatis membuat sebuah dataset emosi berlabel dari Twitter. Kami pertama kali mengumpulkan 7 set kata-kata emosi selama 7 emosi yang berbeda (misalnya, kata "mengganggu" untuk marah emosi) dari psikologi literatur yang ada [12], dan kemudian dimanfaatkan Twitter API streaming untuk mengumpulkan tweet yang memiliki salah satu dari kata-kata emosi ini dalam bentuk dari hashtag (misalnya, #menjengkelkan).

Setiap tweet yang dikumpulkan secara otomatis berlabel dengan satu emosi sesuai dengan emosi hashtag nya, dan hashtag sendiri dihapus dari tweet. Sebagai contoh, dari tweet yang masuk ". Aku benci ketika ibuku membandingkan saya ke teman-teman saya #menjengkelkan", diperoleh contoh data berikut: "Aku benci ketika ibuku membandingkan saya ke teman-temanku" diberi label dengan kemarahan, karena mengandung "#menjengkelkan" hashtag.

Sumber dari kata-kata emosi adalah Shaver's dkk dikutip dari prototype psikologi [12] , di mana para penulis mengatur emosi menjadi sebuah hirarki di mana lapisan pertama berisi enam emosi dasar yaitu, (kasih, sukacita, kejutan, kemarahan, kesedihan, dan ketakutan) dan lapisan kedua berisi 25 emosi sekunder yang subkategori dari enam emosi dasar . Setiap emosi sekunder memiliki daftar kata-kata emosi . Selanjutnya memperluas daftar kata-kata emosi dengan memasukkan varian leksikal mereka , misalnya menambahkan "mengejutkan" dan "terkejut" untuk "surprise" . Selain itu, menghapus kata-kata ambigu . Untuk setiap emosi dasar menggunakan kata-kata emosi yang sesuai dengan emosi sekunder ketika mengumpulkan tweet . Selain yang disebutkan di atas enam emosi dasar, menambahkan satu lagi emosi dasar, syukur, yang tidak tercakup oleh [12]. Tabel I menunjukkan tujuh emosi , hashtags sampel emosi , contoh tweet dan jumlah tweet di masing-masing kategori setelah penyaringan yang relevan .

Teknik penyaringan dikembangkan pada set tersebut dari 400 tweet, penyaringan tersebut berupa hanya mengambil tweet dengan emosi hastag diakhir, karena jika hastag tidak diakhir kecil kemumngkinan adalah emosi penulis.

Kemudian penyaringan berupa tweet yang memiliki kurang dari lima kata, karena tidak dapat ditarik kesimpulan bahwa tidak dapat menyimpulkan sebuah emosi.

Penyaringan selanjutnya menghapus tweet yang berisi url. Karena tweet yang mengandung url kecil kemungkanan merupakan luapan emosi penulis. Sejumlah besar tweet yang berisi url hanyalah sebuah informasi yang disampaikan oleh penulis.

Setelah menerapkan penyaringan pada semua tweet akhirnya memperoleh koleksi 2.488.982 tweet. Distribusi tweets per emosi diringkas dalam Tabel I.

TABEL I
Klasifikasi Emosi Hasil Penyaringan

Emosi	Hastag	# tweet
Kegembiraan	excited, happy,	706.182

	elated, proud (36)	
Kesedihan	Sadness, sorrow, unhappy, depressing, (36)	616.471
Kemarahan	irritating, annoyed, frustrate, fury (23)	574.170
Rasa cinta	affection, lovin, loving, fondness (7)	301.759
Ketakutan	fear, panic, fright, worry, scare (22)	135.154
Rasa Syukur	thankfulness, thankful (2)	131.340
Terkejut	surprised, astonished, unexpected (5)	23.906
TOTAL	131	2.488.982

4. PENGUJIAN DAN ANALISA

Dari 2.488.982 tweet pada Tabel II , secara acak sampel 250.000 tweet sebagai dataset uji Te, selain itu secara acak sampel 247.798 tweet sebagai dataset pengembangan untuk tuning algoritma, dan menggunakan 1.991.184 tweet tersisa (dilambangkan sebagai Tr) untuk training data.

Analisa dilakukan dengan data dibagi menjadi delapan Tr subset (dilambangkan sebagai Tr1 , Tr2 , ... , TR8), masing-masing terdiri 248.898 tweet . Tr1 digunakan untuk menjelajahi fitur yang efektif , dan semua delapan subset yang digunakan untuk pengujian.

Data preprocessing : *lower-cased*, mengganti user yang ditautkan (misalnya, @ladygaga) dengan @user menjadi anonim mengganti tanda baca yang diulang lebih dari dua kali dengan dua huruf yang sama / tanda baca (misalnya , cooool → keren , →) ; ! ! ! dinormalisasi beberapa sering digunakan ekspresi formal (misalnya , *'ll* → *will* , *dnt* → do not) , dan menanggalkan simbol hash (#besok → besok) .

Klasifikasi data menggunakan LIBLINEAR [5] dan Multinomial Naïve Bayes (MNB) [18], karena mereka sangat efisien bahkan untuk menangani jutaan tweet . Dengan mengimplementasi Weka ini [7] untuk MNB . Dan menggunakan regresi logistik untuk cabang LIBLINEAR dan nilai-nilai default untuk semua parameter di kedua pengklasifikasi .

Kinerja keseluruhan *classifier* dapat dihitung dengan : $accuracy = \frac{\#\ tweet\ dengan\ label}{\#\ semua\ tweet\ pada\ dataset}$

Untuk dataset dilambangkan dengan E adalah tweet dengan emosi, E' tweet dengan klasifikasi dari emosi setelah dilakukan penyaringan. Kemudian dapat dihitung *precision* dari emosi :

$$pre(e) = \frac{|E \cap E'|}{|E'|}$$

Precision adalah tingkat ketepatan hasil klasifikasi terhadap suatu kejadian. Dan menghitung *recall* dari emosi :

$$rec(e) = \frac{|E \cap E'|}{|E|}$$

Recall adalah tingkat keberhasilan suatu kejadian dari seluruh kejadian yang harusnya dikenali.

Kami meneliti efek meningkatkan ukuran dataset pelatihan pada keakuratan LIBLINEAR dan MNB pengklasifikasi . Karena kebanyakan identifikasi emosi yang masih ada [2] dilakukan pada dataset ribu kalimat , kami berharap untuk mendapatkan wawasan baru dan manfaat menggunakan data pelatihan yang besar .

TABEL II
Liblinear dengan Data Training

Emosi	Precision (%)	Recall (%)
Joy 28.5%	67.6	72.1
Sadness 24.6%	62.6	64.7
Anger 23.0%	69.8	71.5
Love 12.1%	58.1	51.5
Fear 5.6%	59.7	43.9
Thankfulness 5.3%	66.6	57.1
Surprise 1.0%	44.7	13.9

Tabel II menunjukkan kinerja LIBLINEAR classifier (dengan semua tweet di Tr) pada masing-masing kategori emosi. Terdapat tiga emosi yang paling populer *joy, sadness* dan *anger,* yang merupakan 76,1% dari seluruh tweet , *classifier* mencapai *precision* lebih dari 62% dan *recall* lebih dari 66% untuk masing-masing dari tiga emosi. Penurunan kinerja dapat dilihat pada tiga emosi kurang popular yaitu *love, fear,* dan *thankfulness*

yang terdiri dari 23,0 % dari semua tweet dalam dataset. *Precision* dari tiga kategori emosi yang relatif tinggi (dengan *precision* terendah 58,1%) dibandingkan dengan *recall*, tetapi karena ingat rendah untuk masing-masing emosi. Untuk minoritas emosi yang tersisa yaitu *surprise*, dengan hanya 1,0 % dari seluruh tweet, classifier mendapatkan *precison* terendah begitu pula dengan *recall* dari data pelatihan.

5. KESIMPULAN

Kesimpulan yang dapat diambil adalah :

1. Hastag dapat diidentifikasi menjadi sebuah emosi yang secara otomatis ditulis oleh penulis ini lebih akurat dibandingkan metode pendeteksian emosi dengan text.
2. Cara ini sangat cepat mengingat data yang dikumpulkan lebih banyak.
3. data training yang lebih besar akan menyebabkan akurasi yang lebih tinggi untuk identifikasi emosi karena dapat memberikan cakupan yang komprehensif dari momen emosional dalam hidup kita sehari-hari .

6. DAFTAR PUSTAKA

[1.] C. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of HLT and EMNLP*. ACL, 2005, pp. 579–586.

[2.] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in *Proceedings of IJCNLP*, 2008, pp. 296–302.

[3.] P. Chesley, B. Vincent, L. Xu, and R. K. Srihari, "Using verbs and adjectives to automatically classify blog sentiment," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 27–29.

[4.] M. D. Choudhury, S. Counts, and M. Gamon, "Not all moods are created equal! exploring human emotional states in social media," in *Proceedings of ICWSM*, 2012.

[5.] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[6.] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of- speech tagging for twitter: annotation, features, and experiments," in *Proceedings of HLT:short papers*, ser. HLT '11. Stroudsburg, PA, USA: ACL, 2011, pp. 42–47.

[7.] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[8.] G. Mishne, "Experiments with mood classification in blog posts," in *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

[9.] S. Mohammad, "#emotional tweets," in *Proceedings of the Sixth Inter- national Workshop on Semantic Evaluation*. ACL, 7-8 June 2012, pp. 246–255.

[10.] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: Novel rule-based approach to affect sensing from text," *Natural Language Engineering*, vol. 17, no. 1, pp. 95–135, 2011.

[11.] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classi- fication using machine learning techniques," in *Proceedings of EMNLP*. ACL, 2002, pp. 79–86.

[12.] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowl- edge: Further exploration of a prototype approach." *Journal of person- ality and social psychology*, vol. 52, no. 6, pp. 1061–1086, 1987.

[13.] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 2008, pp. 1556–1560.

[14.] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of LREC*, vol. 4. Citeseer, 2004, pp. 1083– 1086.

[15.] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, ser. SemEval '07, 2007, pp. 70–74.

[16.] R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion classification using massive examples extracted from the web," in *Proceedings of COLING*. ACL, 2008, pp. 881–888.

[17.] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of HLT and EMNLP*. ACL, 2005, pp. 347–354.

[18.] A. Witten, E. Frank, and M. Hall, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[19.] C. Yang, K. Lin, and H. Chen, "Emotion classification using web blog corpora," in *IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 2007, pp. 275–278.

[20.] W. Wang, Lu Chen, K. Thirunarayan, A. P. Sheth, "*Harnessing Twitter 'Big Data' for Automatic Emotion Identification*". *in IEEE/ASE International Congference on Social Computing and International Conference on Privacy, Security, Risk, and Trust.* IEEE, 2012 pp. 587 – 592