PENGKLASIFIKASIAN TOPIK DAN ANALISIS SENTIMEN DALAM MEDIA SOSIAL

Tegar Heru Susilo [1)] **Siti Rochimah** [2)]
1) Program Studi/Jurusan Sistem Informasi, STIKOM Surabaya, email: tegar@stikom.edu
2) Institut Teknologi Sepuluh Nopember Surabaya, email: siti@if.its.ac.id

Abstract: Social media has the capability to increase student's potencies (Sturgeon, 2009) measured by intellectual, social, and performance level. This potency is affected by informal relation between lecturer and students. In the other hand, lecturer has the responsibilities for teaching, research, and public service (Peraturan Pemerintah no 37 tahun 2009). In relation to teaching, lecturer is responsible for guiding students. Guidance has meant to direct students to learn and have a good behavior by encouraging and providing examples. Therefore, lecturer should know, and understand the emotions his students have in order to provide appropriate treatment. These emotions can be seen from the student's statuses in social media.
In this research, an application is proposed. This application has the ability to retrieve information about student statuses in social media, doing topic classification using SVM (Yu, 2011) between academic and non-academic label, and doing sentiment analysis using Maximum Entropy (Soria, 2010) between positive and negative emotions. Testing was conducted in a form of dataset testing using learn and classify approach for testing SVM and MaxEnt classification result. In the dataset testing for SVM, the result shows an accuracy rate of 93%. While in the dataset testing for MaxEnt, the result shows an accuracy rate of 70% for positive document and 53% for negative document. Improved accuracy of sentiment analysis is obtained from the use of word-shape feature in the learning process.

Keywords: Social Media, Academic Performace, SVM, Maximum Entropy.

Media sosial seperti Facebook, Twitter, LinkedIn, YouTube telah mengubah cara orang dalam berinteraksi. Media sosial telah menjadi bagian dari kehidupan mereka. Media sosial telah menjadi identitas mereka dalam bersosialisasi, tidak hanya bagi kalangan sendiri, tetapi juga untuk masyarakat yang lebih luas, dunia.

Dalam sistem akademis, dosen yang efektif adalah mereka yang menjalin hubungan informal dengan mahasiswanya (Sturgeon, 2009). Dikatakan juga oleh Sturgeon (Sturgeon, 2009) bahwa interaksi dosen dan mahasiswa mampu memberikan dampak yang luar biasa bagi intelektual dan tingkat sosial, serta ada hubungan tidak langsung antara dosen yang menggunakan Facebook dengan performa akademik. Namun sesuai dengan peraturan pemerintah (Pemerintah Republik Indonesia, 2009), beban kerja dosen diatur dalam tri dharma perguruan tinggi yaitu pengajaran, penelitian, dan pengabdian masyarakat. Sehingga ada celah antara penelitian akademisi dan peraturan pemerintah mengenai interaksi yang terjadi antara dosen sebagai pengajar dan mahasiswa. Celah inilah yang menjadi latar belakang utama dalam penelitian ini yaitu dengan membangun sebuah aplikasi yang dapat memonitoring perkembangan anak wali/didik dalam media sosial.

Dalam media sosial, setiap status yang diunggah oleh pengguna, tidak semuanya bermakna akademis. Sehingga butuh pengklasifikasian dokumen untuk dapat membedakan status bertopik akademis dari yang non-akademis. Status-status ini juga mempunyai sentimen dari penulisnya. Dengan melakukan analisis sentimen, dapat diketahui informasi tentang emosi pengguna.

Dalam penelitian ini, diusulkan sebuah aplikasi Student Status Retrieval (S.Star, seterusnya dalam makalah ini disebut S.Star) yang mampu mengklasifikasikan topik status, dan mampu melakukan analisis sentimen terhadap status-status tersebut. Fungsi utama S.Star adalah sebagai alat bantu bagi dosen untuk **memonitoring** perkembangan mahasiswa. Masukan dalam S.Star merupakan sebuah trigger yang memicu sistem untuk bekerja secara otomatis mencari informasi mahasiswa yang menjadi anak wali dan anak didik dosen yang bersangkutan, didalam media sosial. Dengan menggunakan informasi tersebut, sistem melakukan pengklasifikasian topik dan analisis sentimen.

KAJIAN PUSTAKA

Klasifikasi

Klasifikasi merupakan sebuah cara untuk memilah obyek kedalam satu atau beberapa kategori yang telah ditentukan (Tan, 2005). Dalam sebuah dokumen, klasifikasi digunakan untuk memilah dokumen kedalam class, yang telah ditentukan, sesuai dengan kontennya (Kamruzzaman, 2007). Tujuan dari klasifikasi ini adalah untuk meminimalkan usaha (effort) yang dikeluarkan oleh organisasi untuk mengelola dokumen, dan bahkan mencari informasi dari dokumen tersebut.

Tahap Persiapan

Data masukan untuk proses klasifikasi adalah kumpulan rekaman. Setiap rekaman, atau yang lebih dikenal sebagai dokumen, dikarakterisasikan dengan sebuah tuple (x,y), dimana x adalah kumpulan perangkat atribut, dan y adalah *class*. Kumpulan perangkat atribut terdiri dari properti, atau fitur berkesinambungan yang menentukan *class*. Sehingga dari penjelasan ini dapat dikatakan bahwa klasifikasi merupakan proses pembelajaran sebuah fungsi target (*target function*) f yang memetakan setiap kumpulan perangkat atribut x kedalam salah satu *class-label* y. Fungsi target ini dikenal sebagai model klasifikasi.

Pendekatan dan Penyelesaian

Teknik klasifikasi adalah sebuah pendekatan sistemik untuk membangun model klasifikasi dari dataset masukan. Setiap teknik menggunakan sebuah algoritma pembelajaran untuk mengidentifikasikan sebuah model yang paling cocok dalam menghubungkan kumpulan perangkat atribut dengan *class*-nya. Model yang dibuat oleh algoritma pembelajaran harus cocok dengan data masukan dan secara tepat memprediksi *class-label* dari rekaman yang belum pernah dilihat.

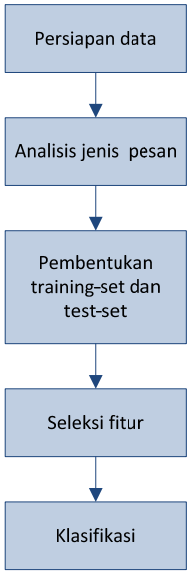
Sedangkan evaluasi terhadap performa model klasifikasi didasarkan pada jumlah pengujian rekaman yang diprediksikan secara tepat atau tidak tepat (*miss-classification*; salah klasifikasi) oleh model tersebut.

Dalam penelitian ini, klasifikasi topik difokuskan pada bagaimana memilah dokumen, yang dalam hal ini adalah status dinding mahasiswa dalam media sosial, kedalam dua buah *class-label* yaitu label akademis, dan non-akademis. Sedangkan untuk kumpulan perangkat atribut dibentuk dari kumpulan fitur yang diperoleh dari hasil tokenisasi kalimat, penghapusan *stop-word*, dan pembobotan kata dalam status dinding mahasiswa.

Support Vector Machine

Klasifikasi SVM digunakan secara luas di bidang bio-informatika karena tingkat keakuratannya yang tinggi, kemampuannya dalam menghadapi data berdimensi tinggi, dan fleksibilitasnya dalam pemodelan sumber data yang beragam.

Yu dalam penelitiannya menggunakan SVM untuk melakukan pengklasifikasian pesan bisnis pada Facebook (Yu, 2011). Penelitian ini mencoba memisahkan dua jenis pesan bisnis yaitu *direct marketing message* dan *communication message*. Metodologi yang dipakai untuk klasifikasi topik dapat dilihat pada Gambar 1.



Gambar 1. Metodologi dalam Klasifikasi Topik (Yu, 2011)

Tahap Persiapan

Karena pengklasifikasian ini mencoba untuk mengklasifikasikan topik dalam status dinding, Yu menggunakan *Bag-of-Words* dalam SVM dengan mengubah dokumen dari *string* kedalam fitur-fitur representatif serta menghilangkan *stop-word*. Setiap kata menjadi fitur, dan jumlah kata tersebut muncul dalam dokumen merupakan nilai fitur.

Karena banyaknya fitur yang mungkin terbentuk, maka perlu adanya fungsi seleksi fitur untuk meningkatkan keakuratan generalisasi dan menghindari *overfitting*. Untuk mencapai apa yang dimaksud, disarankan menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF) untuk melakukan seleksi fitur (Yu, 2011; Joachim, 1999).

Analisis Sentimen

Zabin dan Jefferies, dalam Pang (Pang, 2008), memberikan catatan tentang terminologi mengenai analisis sentimen:

> *"'The beginning of wisdom is the definition of terms,' wrote Socrates. The aphorism is highly applicable when it comes to the world of social media monitoring and analysis, where any semblance of universal agreement on terminology is altogether lacking. Today, vendors, practitioners, and the media alike call this still-nascent arena everything from 'brand monitoring,' 'buzz monitoring' and 'online anthropology,' to 'market influence analytics,' 'conversation mining' and 'online consumer intelligence'. . . . In the end, the term 'social media monitoring and analysis' is itself a verbal crutch. It is placeholder [sic], to be used until something better (and shorter) takes hold in the English language to describe the topic of this report."*

Kutipan ini menyoroti permasalahan yang muncul dalam percobaan untuk mendefinisikan era baru dengan pemilihan kata "*social media monitoring and analysis*". Beberapa istilah telah banyak dipakai dalam ruang lingkup ini, beberapa diantaranya adalah frase *opinion mining*, *sentiment analysis*, dan/atau *subjectivity analysis*. Frase *review mining* juga merupakan salah satu istilah pada ruang lingkup ini yang bertujuan memberikan kemungkinan bagi komputer untuk mengenali dan mengekspresikan emosi. Sedangkan *subjectivity analysis* merupakan pengakuan bahasa berorientasi opini dalam rangka untuk membedakannya dari bahasa obyektif.

Istilah *opinion mining* dalam makalah oleh Dave, dalam Pang (Pang, 2008), menjelaskan bahwa istilah ini berhubungan dengan pencarian Web atau temu kembali informasi. Sejarah mengenai analisis sentimen (sentiment analysis) mempunyai kesamaan dengan *opinion mining* dalam beberapa hal. Istilah sentimen digunakan dalam referensi untuk analisis otomatis untuk mengevaluasi teks dan penelusuran pertimbangan prediktif oleh Das and Chen, dalam Pang (Pang, 2008). Beberapa penelitian lain dalam Pang (Pang, 2008) juga menggunakan istilah ini untuk hal yang sama yang menggunakan Natural Language Processing (NLP). Banyak dari penelitian tersebut yang menyebutkan analisis sentimen fokus pada pengaplikasian khusus dari pengklasifikasian ulasan (*review*) menggunakan polaritas (positif atau negatif). Namun, banyak penelitian saat ini yang menafsirkan istilah analisis sentimen keranah yang lebih luas dalam hal perlakuan komputasi terhadap opini, sentimen, dan subyektifitas dalam teks.

Maximum Entropy

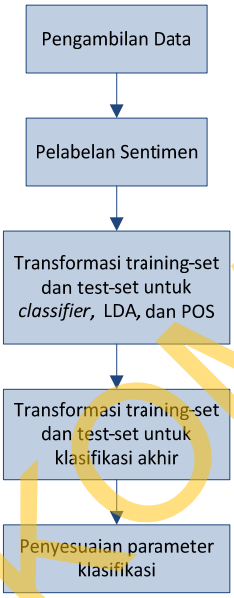
Maximum Entropy adalah teknik umum yang digunakan untuk mengestimasi probabilitas distribusi data (Nigam, 1999). Dikatakan pada teknik ini, bahwa ketika tidak ada yang diketahui, maka distribusi diusahakan untuk *uniform*, yaitu mempunyai maximum entropy. Dalam klasifikasi teks, Maximum Entropy mengestimasi distribusi label dalam dokumen. Dokumen direpresentasikan oleh seperangkat fitur penghitung kata. Dalam kasus yang diambil oleh Nigam, penggunaan Maximum Entropy dapat mengurangi kesalahan klasifikasi sampai dengan 40% dibandingkan dengan Naïve Bayes.

Penggunaan Maximum Entropy juga dapat digunakan untuk menganalisis sentimen dari status

dalam Facebook dengan menambahkan fitur Part-of-Speech (POS) tagging (Soria, 2010). Dikatakan bahwa sebenarnya Maximum Entropy digunakan untuk melatih dataset dengan *corpus* yang telah didefinisikan sebelumnya tentang kata positif/negatif. Metodologi yang dipakai untuk analisis sentimen dapat dilihat pada Gambar 2.

Representasi Fitur

Label pada status dinding disesuaikan dengan kesetaraan *emoticon*, yang didefinisikan pada artikel dalam Wikipedia, List of Emoticon (Soria, 2010). Penyaringan dilakukan pada 14 kategori untuk kemudian dikategorikan kembali kedalam dua *class-label* sentimen yaitu sentimen positif dan sentimen negatif. Dari *emoticon* ini, didapatkan sentimen mutlak sebagai alat bantu dalam pembelajaran dan klasifikasi.



Gambar 2. Metodologi dalam Analisis Sentimen (Soria, 2010)

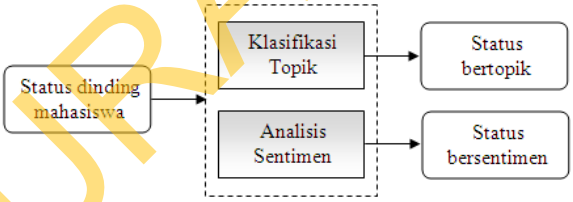
Fitur didapatkan dengan melakukan POS-tagging terhadap dokumen. Berbeda dengan penggunaan bahasa Inggris dalam status media sosial yang mengikuti struktur kalimat baku, penggunaan bahasa Indonesia dalam penyampaian sentimen yang ditulis oleh mahasiswa dalam media sosial lebih disesuaikan dengan struktur yang berkembang di lingkungan mereka masing-masing. Sehingga tidak hanya struktur kalimat, tetapi juga penggunaan tanda baca terkadang tidak sesuai dengan maksud dari kalimat. Namun hal ini tidak dibahas dalam penelitian ini.

PEMBAHASAN

Model Pengembangan

Model ini mengkolaborasikan kedua metode klasifikasi untuk mendapatkan status dinding mahasiswa yang lebih informatif, yaitu bertopik dan bersentimen.

Sesuai dengan Gambar 3, masukan dari sistem adalah status dinding mahasiswa. Dari status ini, dilakukan dua macam klasifikasi yang masing-masing memberikan output yang berbeda, antara lain:



Gambar 3. Model Pengembangan

1. Klasifikasi topik

   Seluruh status yang telah diambil, diklasifikasi menjadi dua label class yaitu label akademis dan label non-akademis. Proses yang terjadi dalam klasifikasi ini antara lain:

   a. Pembobotan dan seleksi fitur menggunakan TF-IDF. Fitur dibentuk berdasarkan kata dan disimpan didalam Bag-of-Words.
   b. Pembelajaran SVM untuk membuat model klasifikasi SVM.
   c. Klasifikasi menggunakan model klasifikasi SVM.

   Output yang diberikan oleh klasifikasi ini adalah status yang sudah mempunyai label topik.

2. Analisis sentiment

   Status yang sama, dianalisis sentimennya untuk menentukan bagaimana emosi pengguna dalam status tersebut. Proses yang terjadi dalam analisis ini antara lain:

a. Penilaian sentimen berdasarkan emoticon.
b. POS-Tag
c. Pembelajaran Maximum Entropy untuk membuat model klasifikasi MaxEnt.
d. Klasifikasi menggunakan model klasifikasi MaxEnt.

Output yang diberikan oleh klasifikasi ini adalah status yang sudah mempunyai label sentimen.

Pembelajaran untuk Klasifikasi Topik

Data Mentah

Data mentah didapatkan dari status dinding mahasiswa, dari dua media sosial yaitu Facebook dan Twitter. Pengambilan status dilakukan pada bulan Februari 2013 sampai dengan 1 Mei 2013. Data yang dikumpulkan berjumlah 3021 data. Masing-masing data ini disebut dengan dokumen.

Pembobotan TF-IDF

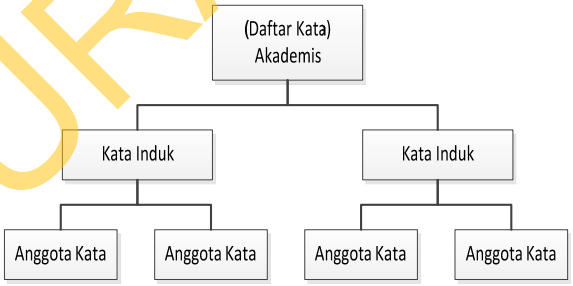
Sebelum pembobotan dilakukan, dilakukan tokenisasi untuk setiap dokumen dan menghilangkan stop-word. Setelah itu, untuk semua dokumen dilakukan pembobotan fitur berdasarkan hasil tokenisasi. Sehingga didapatkan bobot fitur dalam seluruh dokumen.

Pembentukan BoW dan Transformasi Fitur

Bag-of-Words dibentuk dari fitur hasil proses tokenisasi dengan menghilangkan fitur yang sama. Dari 3021 dokumen terbentuk 9246 fitur. Bag-of-Words ini digunakan sebagai acuan dalam transformasi fitur dari bentuk string menjadi bentuk integer, untuk meningkatkan performa proses klasifikasi, lalu diurutkan. Contoh: fitur `tugas:0.0543892701` menjadi `829:0.0543892701`.

Pembentukan Dataset

Untuk setiap dokumen, dilakukan pembobotan topik secara manual. Topik dibagi kedalam dua *class*, yaitu akademis dengan nilai +1, dan non-akademis dengan nilai -1. Penentuan topik ini mengacu pada *academic-word-list* (AWL) yang dibangun oleh Averil Coxhead dari Victoria University. AWL ini terdiri dari sepuluh sublist dengan sekitar 3000 kata yang membentuk 570 keluarga kata (*word-families*). Struktur AWL dapat dilihat pada Gambar 4. AWL ini dibentuk dari *corpus* yang terdiri dari 3.500.000 kata yang diambil dari empat fakultas termasuk salah satunya adalah jurusan Computer Science. Seluruh kata ini tersebar kedalam 414 teks (dokumen). Dengan adanya AWL, menurut Averil Coxhead, dapat digunakan oleh guru dan pembelajar untuk mempelajari kata-kata yang sangat dibutuhkan dalam pembelajaran akademis di tingkat pendidikan tinggi (*tertiary level study*).



Gambar 4. Struktur AWL

Dalam penelitian ini, kata-kata dalam AWL ditranslasikan kedalam Bahasa Indonesia menggunakan kamus Bahasa Inggris - Bahasa Indonesia Online yang dibentuk oleh STANDS4 LLC, yang merupakan penyedia layanan referensi online, yaitu www.kamus.net. Dalam AWL, *corpus* dibentuk dalam keluarga kata sehingga tidak menghilangkan imbuhan. Oleh karena itu, hasil translasi tidak diubah meskipun terdapat imbuhan. Namun, kata yang mempunyai translasi yang sama dengan kata yang lain, akan diabaikan agar tidak dimasukkan kedalam AWL hasil translasi.

Mengacu pada AWL Bahasa Indonesia ini, dilakukan klasifikasi topik pada status. Pengklasifikasian ini dilakukan oleh dosen. Dari hasil klasifikasi tersebut, dibentuk data dalam dataset. Bentuk

data ini berupa gabungan dari bobot topik hasil klasifikasi oleh dosen, diikuti oleh fitur yang membentuk data tersebut. Bobot dan fitur dipisahkan oleh spasi. Sebagai contoh:

```
-1 1:0.43 3:0.12 9184:0.2
```

mempunyai bobot topik `-1` (non-akademis) dan dibentuk dari fitur `1`, `3`, dan `9184` dengan masing-masing bobot fiturnya. Dataset ditulis dengan urutan bobot topik +1 (akademis) terlebih dahulu, baru bobot topik -1 (non-akademis).

Pembelajaran

Pembelajaran dilakukan dengan menggunakan SVMLight yang merupakan implementasi algoritma SVM dalam bentuk C. Untuk pembelajaran, aturan yang dipakai adalah aturan standar yang telah dispesifikasikan oleh SVMLight.

Pembelajaran untuk Analisis Sentimen

Data mentah

Data mentah untuk analisis sentimen adalah sama dengan data mentah untuk klasifikasi topik.

Pembobotan Sentimen berdasarkan Emoticon

Emoticon mengacu pada daftar *emoticon* yang diambil Wikipedia. Selain dari Wikipedia, daftar emoticon juga dibentuk berdasarkan *emoticon* yang ada didalam dokumen. Namun tidak semua *emoticon*, karena permasalahan kompleksitasnya. Dari daftar *emoticon* yang terbentuk, dibuat dua kategori sebagai *class-label*, yaitu sentimen positif dan sentimen negatif.

Dari daftar *emoticon*, dilakukan pencarian *emoticon* disetiap dokumen dan dilakukan penyimpanan sentimen berdasarkan *emoticon* tersebut. Dokumen yang memiliki *emoticon*, dihilangkan *emoticon*-nya. Dokumen tersebut menjadi dataset dalam proses pembelajaran.

POS-Tagging

POS-Tagging dilakukan dengan bantuan Pebahasa. Masukan dari aplikasi ini adalah dokumen yang akan di POS-tag. Keluaran dari aplikasi ini adalah dokumen dengan hasil POS-tag di masing-masing fiturnya.

Pembentukan Dataset

Untuk setiap dokumen, dilakukan pembobotan sentimen secara manual. Sentimen dibagi kedalam dua *class*, yaitu positif dengan nilai +1, dan negatif dengan nilai -1. Bentuk data dalam dataset ini adalah gabungan dari bobot topik diikuti dengan fitur yang dipisahkan dengan tab. Sebagai contoh:

```
-1    Tegar/NN sedang/RB makan/VBT
      malam/NN
```

mempunyai bobot sentimen `-1` (sentimen negatif) dan dibentuk dari fitur "`Tegar sedang makan malam`" dengan masing-masing POS-Tag-nya.

Proses Pembelajaran

Pembelajaran dilakukan dengan menggunakan Stanford-Classifier yang mengimplementasikan algoritma klasifikasi Maximum Entropy dalam bentuk Java (stanford-classifier.jar).

Fitur yang dipakai dalam melakukan klasifikasi mengikuti apa yang telah didefinisikan oleh Soria (Soria, 2010). Fitur-fitur dalam pengklasifikasian ini merupakan properti dari Stanford-Classifier. Fitur-fitur ini merupakan fitur dasar, antara lain:

1. *useSplitWords*: membuat fitur dari kata yang dipisahkan berdasarkan Regex. Nama fitur dalam klasifikasi adalah SW-str.
2. *useSplitWordPairs*: membuat fitur dari kata yang saling berdekatan. Nama fitur dalam klasifikasi adalah SWP-str1-str2.
3. *useSplitFirstLastWords*: membuat fitur dari kata pertama dan kata terakhir dalam dokumen. Nama dalam klasifikasi adalah SFW-str, SLW-str.

4. *useSplitPrefixSuffixNGrams*: membuat fitur dari prefiks dan suffiks setelah dipisah dari kata utama menggunakan Regex.

Pengujian Klasifikasi Topik

Langkah pertama adalah pembelajaran dataset yang dibangun menggunakan AWL. Dari 3021 dokumen yang diklasifikasikan, terbentuk 2643 dokumen berlabel non-akademis, dan 378 dokumen berlabel akademis. Sesuai dengan Krejcie dan Morgan (Krejcie, Morgan, 1970), untuk jumlah data kurang dari 3500, ukuran sample adalah 341 data dengan asumsi standar galat 5%. Oleh karena itu, dari dokumen ini dipilih 200 dokumen akademis, dan 200 dokumen non-akademis. Sehingga total dokumen yang dipakai untuk *sample* adalah 400 dokumen untuk menjadi dataset pembelajaran.

Pembelajaran SVMLight dijalankan menggunakan `svm_learn.exe` yang membaca dataset pembelajaran dengan parameter pembelajaran *default*. Dari hasil pembelajaran didapatkan tingkat akurasi 97,89%, tingkat galat 16% dan tingkat *recall* 69,50% dengan 385 *support-vector*. Hasil ini disimpan menjadi model klasifikasi. Dari model ini, dilakukan klasifikasi terhadap 100 dokumen acak dengan komposisi 50 dokumen akademis dan 50 dokumen non-akademis yang diambil dari data mentah, selain dataset pembelajaran.

Langkah selanjutnya adalah klasifikasi SVMLight yang dijalankan menggunakan `svm_classify.exe` yang membaca model klasifikasi dan dataset pengujian. Dari hasil klasifikasi, didapatkan tingkat akurasi 93%, tingkat precision 95,74% dan tingkat recall 90%.

Pada dokumen *"kenapa seh "deadline" ini ngejar-ngejar aku terus? padahal udah jelas-jelas aku gak suka ama dia... :(:(:("* terjadi salah-klasifikasi, yang seharusnya positif menjadi negatif. Dari dokumen ini, ada enam fitur yang dihasilkan oleh TF-IDF yaitu "*padahal*", "*ama*", "*udah*", "*deadline*", "*seh*", dan "*ngejar*". Dari model klasifikasi yang dihasilkan, fitur "*padahal*", "*udah*", dan "*seh*" mempunyai kedekatan dengan label non-akademis. Sedangkan fitur lainnya tidak pernah dilatih didalam model. Sedangkan dalam klasifikasi manual, "*deadline*" mempunyai porsi terbesar dalam menentukan label dokumen menjadi positif.

Dokumen lain seperti "*menikmati masa2 karantina. :)*" terjadi salah-klasifikasi yang seharusnya negatif menjadi positif. Dalam banyak dokumen untuk pembelajaran, fitur "*menikmati*" dan "*masa*" lebih banyak digunakan pada dokumen berlabel akademis. Sedangkan fitur "*karantina*" belum dilatih didalam model. Sehingga dokumen ini dianggap positif oleh *classifier*.

Pengujian Analisis Sentimen

Langkah pertama adalah pembelajaran dataset yang dibangun menggunakan data hasil analisis sentimen berdasarkan *emoticon* karena ke-mutlak-annya. Dari seluruh 3021 dokumen, ditemukan 870 dokumen ber-*emoticon* dengan komposisi 351 dokumen bersentimen negatif dan 519 dokumen bersentimen positif. Dataset dibentuk menggunakan komposisi 200 dokumen positif dan 200 dokumen negatif sehingga total dokumen yang dipakai adalah 400 dokumen. Pembelajaran Stanford-Classifier dijalankan menggunakan `java.exe –jar stanford-classifier.jar` yang membaca file *property* yang berisi fitur standar dalam pembentukan model klasifikasi. Dalam file *property* ini dituliskan juga dataset yang dipakai untuk pembelajaran dan untuk pengujian. Hasil dari pembelajaran berupa model klasifikasi. Untuk proses klasifikasi, dataset dibentuk dari 40 dokumen hasil analisis psikolog pendidikan untuk melihat tingkat akurasi.

Pengujian dilakukan dalam beberapa skenario untuk membandingkan (1) fitur dasar dengan (2) fitur dasar dan bentuk kata (*word-shape*). Urutan skenario pengujian dapat dilihat pada Tabel 1. Dari pengujian berdasarkan skenario, diperoleh hasil klasifikasi seperti pada Tabel 2.

Tabel 1: Skenario Pengujian MaxEnt

No.	Bentuk skenario, pengujian dataset dengan-
1	Fitur dasar
2	Fitur dasar dan fitur word-shape “dan1”
3	Fitur dasar dan fitur word-shape “dan2”
4	Fitur dasar dan fitur word-shape “chris1”
5	Fitur dasar dan fitur word-shape “chris2”
6	Fitur dasar dan fitur word-shape “chris4”

Tabel 2: Hasil Skenario Pengujian MaxEnt

Skenario ke-	Nilai F1 POSITIF	Nilai F1 NEGATIF
1	0.625	0.400
2	0.625	0.400
3	0.708	0.533
4	0.694	0.483
5	0.680	0.429
6	0.667	0.370

Dalam bentuk fitur dasarnya, klasifikasi memberikan nilai F1 positif 0.625 dan F1 negatif 0.400. Pada dokumen “*Butuh ketenangan hati dan pikiran*” terjadi salah klasifikasi dari positif menjadi negatif. Dari hasil klasifikasi dengan fitur dasar, ditemukan:

	+1	-1
1-SW-hati/NN	-0.03	0.03

Dalam dokumen pembelajaran, kata “hati” digunakan sebagai NN dalam banyak dokumen positif. Namun penekanan kata hati sebagai kata negatif mempunyai nilai terbesar dalam klasifikasi. Sehingga hal ini menjadikan dokumen ini 53% bersentimen negatif. Namun hasil klasifikasi benar ketika klasifikasi dilakukan menggunakan *word-shape* “dan2”.

Fitur *word-shape* “dan2” mengukur bobot fitur menggunakan komposisi huruf kecil, huruf besar, angka, kata dengan campuran huruf besar dan kecil, kata dengan tanda baca, dan ekivalensi kelas kata yang memiliki bentuk yang sama dengan panjang 3 karakter atau kurang. Poin terbesar dari fitur ini adalah komposisi yang terbentuk yaitu penggunaan kata dengan campuran huruf besar dan kecil yang mempunyai nilai positif. Sesuai dengan pembicaraan yang dilakukan dengan psikolog pendidikan, seseorang menulis sesuatu dengan benar ketika mereka dalam keadaan nyaman, rileks, dan tanpa tekanan. Dengan penggunaan *word-shape* ini, dokumen 52% bersentimen positif. Hasil klasifikasi dengan fitur *word-shape* “dan2” dapat dilihat pada Tabel 3.

Tabel 3: Hasil Pengujian MaxEnt Menggunakan Fitur *Word-Shape* “dan2”

	+1	-1
CLASS	0.02	-0.02
1-SSHAPE-WT-Xx/X	0.32	-0.32
…	…	…
1-SW-dan/CC	0.14	-0.14
…	…	…
1-SW-hati/NN	-0.01	0.01
1-SSHAPE-WT-x/X	-0.25	0.25
…	…	…
1-SW-pikiran/NN	-0.17	0.17
…	…	…
Prob:	0.47	0.52

Namun dari hasil pengujian menggunakan fitur *word-shape* “chris4” yang menggunakan bentuk kata dengan campuran huruf besar dan kecil dengan panjang karakter lebih panjang dari fitur “dan2”, hasil klasifikasi menurun baik untuk penilaian dokumen positif maupun penilaian dokumen negatif. Setelah dilakukan analisis terhadap dokumen pembelajaran, kata dengan panjang karakter > 3, umumnya dipakai untuk kalimat-kalimat negatif dengan penggunaan kosakata yang salah seperti kata “*setaaann*”, “*cacaaad*“, dan lain-lain.

SIMPULAN

1. Kolaborasi metode Support Vector Machine (SVM) dengan pembobotan fitur Term Frequency – Inverse Document Frequency (TF-IDF), dapat dipakai untuk melakukan klasifikasi topik dalam bahasa Indonesia dengan tingkat akurasi 93%.
2. Kolaborasi metode Maximum Entropy (MaxEnt) dengan fitur *word-shape* “dan2” serta POS-tag menggunakan Hidden Markov Model (Wicaksono, 2010), dapat dipakai untuk melakukan analisis sentimen dalam bahasa Indonesia dengan tingkat akurasi 70% untuk dokumen positif dan 53% untuk dokumen negatif.

PENELITIAN SELANJUTNYA

1. Penambahan fungsi untuk menormalisasi kecacatan kata (tidak sesuai dengan kosakata) menggunakan *spelling corrector*.

2. Koleksi data untuk klasifikasi diperbanyak dengan mengambil komentar dan *liked/disliked*.
3. Analisis terhadap histori status pengguna media sosial untuk melihat perilaku pengguna dalam media sosial.
4. Pembuatan Knowledge Management System (KMS), yang mengolah pengalaman dan pengetahuan dari pakar, untuk mendukung keputusan perlakuan kepada mahasiswa tertentu.

RUJUKAN

Alba, A., Bhagwan, V., Grandison, T., (2008), *Accessing The Deep Web: When Good Ideas Go Bad*, IBM, California.

Ben-Hur, A., Weston, J., (2008), *A User's Guide to Support Vector Machines*, Colorado State University.

Choudhury, M., et.al, (2007), *How Difficult is it to Develop a Perfect Spell-Checker? A Cross-linguistic Analysis through Complex Network Approach*, Department of Computer Science and Engineering, IIT Kharagpur.

Dumais, S., et.all, (1998), *Inductive Learning Algorithms and Representations for Text Categorization*, Microsoft Research.

Gruchawka, S., (2005), *Using the Deep Web: A How-To Guide for IT Professional*, TechDeepWeb.com

Joachim, T., (1999), *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, University of Dortmund.

Joachims, T., et.al, (1999), *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, MIT-Press.

Kamruzzaman, SM., 2007, *Text Classification using Artificial Intelligence*, University of Rajshah, Bangladesh.

Kridalaksana, H., 2001, *Kamus Linguistik*, Gramedia Pustaka Utama, Jakarta.

Kumar, S., Sanaman, G., Rai, N., (2008), *Federated Search: New Option for Libraries in the Digital Era*, International CALIBER.

Mehra, N., Khandelwal, S., Patel, P., (2002), *Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews*, Stanford University.

Nigam, K., Lafferty, J., McCallum, A., (1999), *Using Maximum Entropy for Text Classification*, Carnegie Mellon University.

Orenstein, B., (2000), *QuickStudy: Application Programming Interface (API)*, Online, http://www.computerworld.com, diakses 29 September 2011.

Pang, B., Lee, L., 2008, *Opinion Mining and Sentiment Analysis*, Journal of Foundations and Trends ® in Information Retrieval.

PCMag Encyclopedia, *API Definition*, Online, http://www.pcmag.com, diakses 29 September 2011.

Pemerintah Republik Indonesia, (2009), *Peraturan Pemerintah Republik Indonesia No 37 Tahun 2009 Tentang Dosen*, Jakarta.

Rajaraman, A., (2009), *Kosmix: Exploring the Deep Web using Taxonomies and Categorization*, Kosmix Corporation, California.

Rubinger, B., Bultan, T., (2010), *Contracting the Facebook API*, University of California.

Soria, S., Akhter, JK., (2010), *Sentiment Analysis: Facebook Status Message*, Stanford University.

Sturgeon, M., Walker, C., (2009), *Faculty on Facebook: Confirm or Deny?*, 14th Annual Instructional Technology Conference, Middle Tennesse State University, Tennesse.

Supratiknya, A., 1993, *Psikologi Kepribadian 3: Teori-Teori Sifat dan Behavioristik*, Kanisius, Yogyakarta

Tan, P., et.al., 2005, *Introduction to Data Mining, Addison Wesley*, Boston.

Tang, L., Liu, H., (2010), *Towards Predicting Collective Behaviour via Social Dimension Extraction*, Arizona State University, Arizona.