SENTIPOL: DATASET SENTIMEN KOMENTAR PADA KAMPANYE PEMILU PRESIDEN INDONESIA 2014 DARI FACEBOOK PAGE

Antonius Rachmat[1] , Yuan Lukito[2]
[1]Program Studi Teknik Informatika, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia
anton@ti.ukdw.ac.id
[2]Program Studi Teknik Informatika, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia
yuanlukito@ti.ukdw.ac.id

Abstract

Dataset adalah sekumpulan data yang sudah diverifikasi kebenarannya dan dapat digunakan dalam penelitian sebagai sumber data yang valid. Dataset memiliki banyak jenis, namun dataset teks dan sentimen dalam Bahasa Indonesia masih jarang ditemukan. Penelitian ini akan melakukan pembangunan dataset yang dikumpulkan dari status dan komentar terhadap calon presiden Indonesia pada masa kampanye pemilu tahun 2014 dari Facebook Page. Data status dan komentar diambil dari Facebook menggunakan Facebook API untuk kemudian disimpan dalam basis data lokal. Proses berikutnya dilakukan dengan pemberian label sentimen (positif, negatif, atau netral) untuk setiap data secara crowdsourced labelling menggunakan aplikasi web secara online. Hasil akhir label pada dataset ditentukan secara otomatis menggunakan metode Weighted Majority Voting berdasarkan bobot terbesarnya. Penelitian ini telah menghasilkan dataset sejumlah 3400 komentar dari 68 status dalam format CSV dengan label positif lebih dominan daripada label negatif dan netral sehingga dapat digunakan sebagai data pembelajaran sistem supervised learning lainnya. Berdasarkan hasil validasi dengan tingkat konfiden 95%, dataset ini memiliki validitas sebesar 95.3%.. Dataset ini juga telah dicoba digunakan dalam klasifikasi sentimen analisis menggunakan metode Naïve Bayes dan Support Vector Machine dengan tingkat akurasi masing-masing 82.23% dan 84.82%.

Keywords **:** *dataset politik, pemilu 2014, komentar Facebook, analisis sentimen.*

1. Pendahuluan

Untuk dapat mengklasifikasikan data dengan teknik *data mining* yang bersifat *supervised learning*, dibutuhkan sumber data (dataset) pelatihan yang terpercaya dan valid untuk dapat digunakan sebagai sumber pembelajaran sistem. Berdasarkan kenyataan tersebut, tidak banyak *dataset* yang valid dan bisa diperoleh dengan mudah (Matsubara, Monard, & Prati, 2008), selain karena jarang ada yang membuat, proses pemberian label pada dataset juga memakan waktu yang lama seandainya dilakukan sendirian, bahkan jika dilakukan oleh pakar sekalipun. Dataset yang berasal dari media sosial lebih sulit lagi ditemukan karena sifat datanya yang sangat besar. Walaupun sifat datanya yang sangat besar kini telah muncul beberapa situs maupun penelitian yang berusaha menciptakan dan menyediakan dataset dari media sosial seperti Facebook dan Twitter seperti situs Infochimp Twitter Census (Infochimp, 2010), situs UCIrvine (Markopoulou, et.al., 2011), penelitian yang dilakukan oleh J. McAuley & J. Leskovec (2012) dan Gjorka, et.al. (2011). Beberapa penelitian di atas ternyata telah berhasil menciptakan dataset-dataset yang sudah dipublikasikan dan dapat digunakan oleh peneliti lain untuk melakukan penelitian yang berkaitan dengan *data mining*.

Dataset yang tersedia di Internet tidak banyak yang menggunakan bahasa Indonesia dan sangat jarang yang merupakan dataset bidang politik di Indonesia. Beberapa dataset yang ditemukan tentang Indonesia berupa dataset demografi (DHS, 2013), dan dataset biologi. Penelitian pembuatan dataset politik berbahasa Indonesia diperlukan

untuk menambah tersedianya dataset-dataset berbahasa Indonesia yang berkaitan dengan kondisi di Indonesia.

Metode yang selama ini digunakan dalam pembuatan dataset pelabelan adalah menggunakan tenaga pakar untuk melakukan pelabelan data. Hal ini tentu sangat sulit atau bahkan tidak mungkin dilakukan untuk dataset yang berasal dari media sosial karena jumlah datanya sangat besar yang akan membutuhkan lebih banyak pakar (dalam hal ini manusia) dan waktu yang sangat lama (Hosseini, 2012). Teknik melakukan pelabelan menggunakan banyak pelabel (*crowdsourced*) membutuhkan metode khusus, seperti metode Majority Voting (Hosseini, 2012). Metode tersebut akan digunakan dalam penelitian ini.

Dataset politik yang akan dibangun dalam penelitian ini adalah dataset yang berisi sentimen pengguna media sosial Facebook saat masa Pemilihan Umum (Pemilu) di Indonesia. Pemilu calon presiden pada tahun 2014 yang telah berlangsung tanggal 9 Juli 2014 yang lalu merupakan sebuah peristiwa baru dalam pesta demokrasi di mana hanya menyisakan 2 pasang calon presiden (capres) saja, yaitu pasangan Prabowo Subianto - Hatta Rajasa dan Joko Widodo - Yusuf Kalla. Sejak tanggal 4 Juni hingga tanggal 9 Juli 2014 masyarakat Indonesia melalui berbagai cara dan media berusaha mencari dan menyampaikan dukungan terhadap dua pasang capres melalui Facebook Pages. Pada setiap status dari capres, terdapat ribuan komentar, baik yang mendukung, menolak, netral, maupun yang tidak relevan yang dipostingkan oleh pengguna. Untuk membentuk dataset ini dibutuhkan sebuah aplikasi online yang mampu menjadi alat pelabelan berbasis *crowdsourced.*

Dari permasalahan yang telah dikemukanan sebelumnya didapat beberapa perumusan masalah sebagai berikut: (1). apakah metode *Crowdsourced Labelling* berbasis *Weightened Majority Voting* dapat menghasilkan dataset politik dari media sosial Facebook. (2). apakah dataset yang dihasilkan valid dan dapat digunakan untuk keperluan penelitian klasifikasi sentimen. Tujuan dari penelitian ini adalah menghasilkan suatu dataset politik di Indonesia pada masa kampanye presiden tahun 2014 yang dapat dipergunakan sebagai sumber data valid dalam penelitian *supervised learning* pada sistem cerdas terapan.

2. Tinjauan Pustaka

2.1. Media Sosial

Media Sosial adalah suatu wahana yang diciptakan berbasis komputer, biasanya berbasis Internet yang memungkinkan para penggunanya untuk menciptakan, membagikan, bertukar informasi, ide, baik dokumen, foto, atau video di dalam suatu komunitas virtual dalam jaringan. Menurut Kaplan & Haenlein (2010), sosial media didefinisikan sebagai "*a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content*".

Media sosial berkembang sangat pesat berkat teknologi pendukungnya seperti Internet, dan web 2.0 & 3.0. Perkembangan tersebut membuat media sosial dapat dibagi-bagi berdasarkan bentuknya seperti *blog, wiki, social network, podcast, video on demand, social bookmarking*, dan berbagai macam lainnya. Facebook merupakan salah satu contoh media sosial, sedangkan Twitter merupakan contoh *blog (microblogging).* Pada prinsipnya media sosial mengandalkan unsur partisipasi aktif dari para penggunanya.

Menurut KOMINFO (2014), pengguna Internet Indonesia mencapai 82 juta orang (peringkat ke-8 dunia) dimana 95% nya menggunakan media sosial. Dari jumlah pengguna internet tersebut, 80% diantaranya adalah remaja, dengan pengguna Facebook di peringkat ke-4 besar dunia. Dengan jumlah pengguna yang sangat besar, Facebook (1.11 triliun pengguna (Statistic Brain (2014)) dan Twitter (645.750.000 pengguna (Statistic Brain (2014)) memiliki banyak sekali data yang secara implisit dapat digali lebih lanjut dengan berbagai metode *data mining.* Facebook memiliki data yang beraneka ragam bentuknya, mulai data data teks, data gambar / foto, data suara bahkan data dalam bentuk video. Dari berbagai jenis data tersebut, data teks pada Facebook yang perlu dianalisis adalah data status dan komentar yang dibuat oleh para penggunanya.

2.2. Dataset

Dataset adalah kumpulan data yang berelasi / berkaitan satu dengan lainnya dalam satu kesatuan

yang biasanya bersifat spesifik terhadap suatu kasus tertentu, misalnya dataset medikal, dataset komentar pengguna Twitter terhadap layanan Apple, dataset curah hujan selama 1 tahun, dataset pergerakan harga emas selama tahun tertentu, dan lain sebagainya. Dataset bidang medis bahkan biasanya bersifat sangat banyak, kompleks, heterogen, dan hierarkis (Hosseinkhah, Ashktorab, Veen, & Owrang, 2009). Dataset dapat direpresentasikan dalam berbagai bentuk misalnya bentuk tabel dalam basis data, bentuk matriks, bentuk teks, bentuk *Comma Separated Value* (CSV) dan sebagainya. Dataset dapat dikumpulkan dan dibentuk oleh seseorang, sekelompok atau bahkan suatu organisasi dan dipublikasikan secara online seperti misalnya situs DataHub (www.datahub.io) yang berisi berbagai macam dataset.

Dataset dipergunakan sebagai referensi data yang valid untuk suatu penelitian selanjutnya, misalnya untuk referensi data dalam pembelajaran sistem cerdas (sistem pengenalan pola, machine learning, dan lain-lain), atau juga sebagai referensi data dalam pengujian sistem otomatis seperti misalnya pada sistem klasifikasi, klasterisasi, dan sentimen analisis. Dataset yang baik memiliki ciri memiliki data yang lengkap, selalu *up to date*, bersifat konsisten dalam representasi datanya, jumlah variabelnya jelas, tidak mengandung *noise*, menarik, dan mudah dimengerti (Hosseinkhah, Ashktorab, Veen, & Owrang, 2009)

2.3. Pelabelan Data

Dalam proses pembuatan dataset untuk sistem klasifikasi, diperlukan mekanisme bagaimana agar dataset yang telah dikumpulkan memiliki label kelas yang benar. Pada kenyataanya dataset yang sudah dilabeli sangat sedikit dan sulit dicari (Matsubara, Monard, & Prati, 2008). Proses pelabelan dataset akan mudah jika datanya berjumlah sedikit dan tidak terlalu besar, namun akan sangat membutuhkan waktu yang sangat lama bahkan tidak mungkin dikerjakan sendiri jika dataset berjumlah sangat besar. Proses pelabelan data dapat dilakukan secara manual sendirian atau dikerjakan bersama-sama oleh beberapa bahkan puluhan hingga ratusan orang menggunakan teknik *crowdsourced labelling*.

2.3.1 *Crowdsourced Labelling*

Proses pemberian label pada data berukuran besar menjadi sulit atau bahkan tidak mungkin dilaksanakan dalam waktu yang singkat. Pelabelan dapat dilakukan dalam waktu lebih singkat jika jumlah pelabel semakin banyak. Salah satu teknik yang memanfaatkan banyak pelabel untuk memberi label pada data berukuran besar adalah *Crowdsourced Labelling*. Teknik pemberian label ini memiliki kemungkinan untuk menghasilkan kualitas label yang bervariasi. Tetapi dengan semakin banyak pelabel, maka kemungkinan untuk mendapatkan kualitas label yang lebih baik akan lebih besar.

Welinder dan Perona (2010) menggunakan *Crowdsourced Labelling* untuk memberi label pada sekumpulan data gambar memanfaatkan layanan Amazon Mechanical Turk. Beberapa faktor yang diteliti antara lain kualitas label dan besarnya biaya yang diperlukan.

2.3.2 *Weightening Majority Voting*

Metode pelabelan data yang selama sering ini digunakan adalah menggunakan tenaga pakar tertentu untuk melakukan pelabelan. Hal ini tentu sangat sulit atau bahkan tidak mungkin dilakukan untuk *dataset* yang berasal dari media sosial karena jumlah datanya sangat besar, maka akan membutuhkan lebih banyak pakar (dalam hal ini manusia) dan waktu yang sangat lama (Hosseini, 2012). Teknik melakukan pelabelan menggunakan banyak pelabel membutuhkan metode khusus, seperti metode *Majority Voting*, metode *Expectation Maximization* (Hosseini, 2012) Dari berbagai metode tersebut metode yang sederhana dan banyak digunakan adalah metode *Majority Voting*. Metode ini menggunakan konsep pengambilan keputusan hasil vote yang diperoleh dari jumlah terbesar dari masing-masing pilihan vote yang ada. Metode ini menggunakan rumus (1) dan (2) berikut (asumsi menggunakan 3 pelabel) (James, 1998):

$$C(X) = \text{mode}\{h_1(X), h_2(X), h_3(X)\} \quad (1)$$

Dimana:
C(X) : class/label X
h1(X) : hasil vote pelabel 1
h2 (X) : hasil vote pelabel 2
h3 (X) : hasil vote pelabel 3

Dalam perkembangannya metode Majority Vote dapat diberi bobot menjadi Majority Vote berbobot sebagai berikut:

Masukkan data ke dalam suatu label jika : (Stefano, Cioppa, & Marcelli, 2002)

$$\sum_{i=0}^{N} w_{ik} * \delta_{ik} = \max_{1 \le j \le N} \sum_{i=0}^{N} w_{ij} * \delta_{ij} \quad (2)$$

Dengan

$$\delta_{ik} = \begin{cases} 1 \; if \; E \; gives \; the \; class \; k \\ 0 \; otherwise \end{cases}$$

Dimana W_{ik} adalah bobot dan d_{ik} adalah label tertentu.

3. Hasil dan Pembahasan

3.1. Format Dataset

Dataset yang dihasilkan akan dipublikasikan secara terbuka sehingga dapat digunakan sebagai sumber referensi penelitian-penelitian selanjutnya. Publikasi dataset akan dilakukan dalam format file teks dan file CSV (*Comma Separated Value*), yang atribut dan isi datanya dapat dilihat pada Tabel 1 berikut ini:

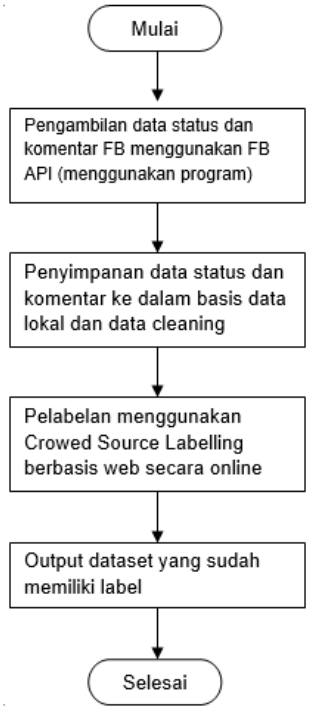
Tabel 1**.** Atribut dan Contoh Isi Dataset

Id_Status	Id_Komentar	Id_User_Komen	Komentar	Label_Sentimen
23383061178_10152076252911179	10152076252516179_10152076253441179	794948677210695	"Prabowo insya Allah jadi presiden 2014 :), yang like berarti setuju. hehe"	positif

Id_status adalah nomor unik dari dari setiap status yang dikeluarkan. Id_komentar adalah nomor unik dari suatu komentar yang dikeluarkan pada waktu tertentu oleh pengguna FB. Id_user_komen berupa nomor unik dari pengguna FB yang membuat komentar yang diberikan pada status tersebut, komentar adalah teks komentar yang diberikan, sedangkan label sentimen adalah hasil pelabelan yang sudah dihitung menggunakan *Weighted Majority Voting* berbasis *Crowdsourced Labelling.*

3.2. Metode Penelitian

Metode penelitian yang dilakukan pada penelitian ini dirangkum dalam Gambar 1. Dari gambar dapat dijelaskan bahwa tahap pertama adalah pembuatan program untuk mengunduh data status dan komentar dari Facebook Page yang dilakukan oleh peneliti. Tahap kedua adalah pengambilan data menggunakan program yang dibuat sehingga terkumpul data yang siap diberi label menggunakan program yang telah dihasilkan pada tahap pertama. Tahap ketiga adalah pelabelan data menggunakan aplikasi online berbasis *Crowdsourced Labelling* menggunakan metode *Weighted Majority Voting* yang dilakukan oleh pelabel secara acak. Tahap terakhir adalah melakukan validasi dataset dan menghasilkan dataset final yang telah terlabeli yang dilakukan oleh peneliti dan menghasilkan dataset final. Secara detail tahapan penelitian dijelaskan pada sub bab berikutnya.



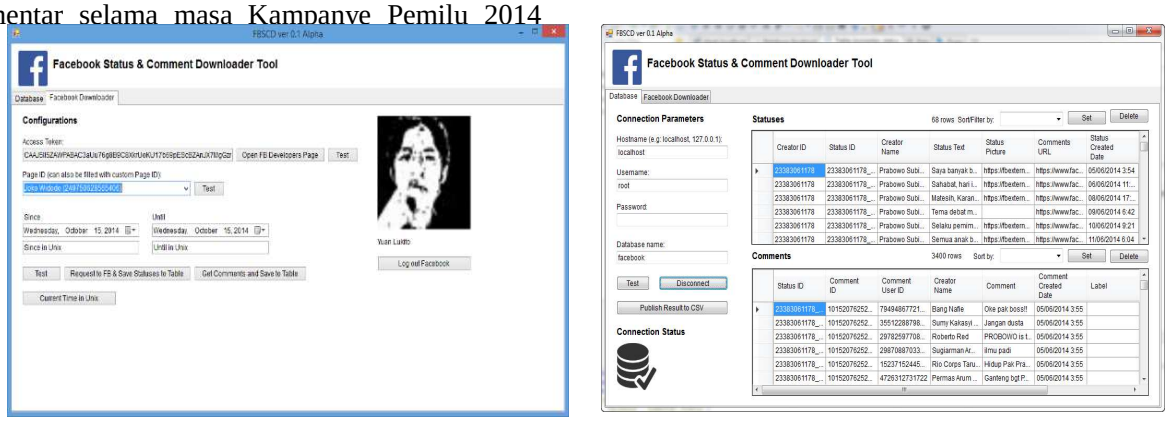
Gambar 1. Metode Penelitian.

3.2.1 Pengambilan Data

Pengambilan data pada penelitian ini dilakukan langsung dari Facebook Page Calon Presiden Prabowo Subianto dan Calon Presiden Joko Widodo oleh peneliti menggunakan bantuan program yang dikembangkan sendiri berbasis Facebook API. Facebook Page Prabowo Subianto diambil dari alamat: https://www.facebook.com/PrabowoSubianto dan Facebook Page Joko Widodo diambil dari alamat: https://www.facebook.com/Jokowi/ Dari kedua Facebook Page tersebut diambil Status dan Komentar selama masa Kampanye Pemilu 2014 (Juni 2014 - 9 Juli 2014). Proses pengambilannya dilakukan menggunakan Facebook Developer Key dan Facebook API yang dapat diperoleh secara gratis. Tahap pengambilan data merupakan tahap pertama penelitian seperti pada gambar 1 di atas dilakukan sebagai sebagai berikut:

1. Pengambilan data ke Facebook Page Prabowo Subianto menggunakan Query Facebook Graph Explorer: 249750628565406/posts?fields=id,message,full_picture,link,actions,created_time&limit=250&until=1404579599&since=1402074001
2. Pengambilan data ke Facebook Page Joko Widodo menggunakan Query Facebook Explorer: 23383061178/posts?fields=id,message,full_picture,link,actions,created_time&limit=250&until=1404579599&since=1402074001
3. Query untuk mengambil 50 komentar pertama tiap Page adalah: 23383061178_10152167482826179?fields=id,message,comments.limit(50){id,message,created_time,from}&limit=250
4. Pengembangan aplikasi pengambil data otamatis terdiri dari modul : login, get_statuses, get_comments, dan save to database.

Setelah aplikasi sudah selesai menyimpan data ke dalam basis data, sistem tersebut juga dapat menghasilkan output ke dalam file .CSV. Aplikasi downloader yang telah dikembangkan dapat dilihat pada Gambar 2.



Gambar 2. Aplikasi Downloader Facebook Page

3.3. Implementasi Preprocessing Data

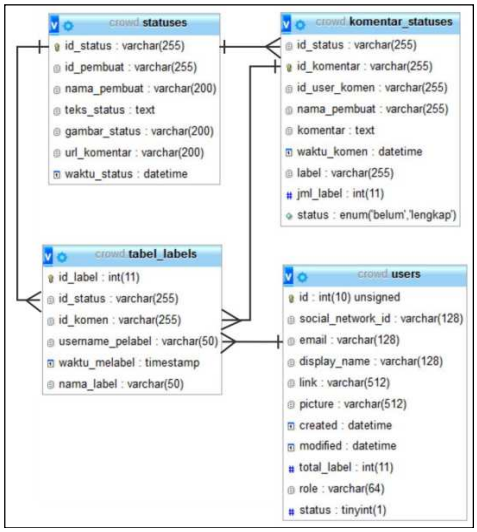
Tahap setelah pengembangan aplikasi *downloader* pengambil data status dan komentar dari Facebook Page adalah tahap implementasi preprocessing data hasil status dan komentar tersebut. Tahapan preprocessing ini dilakukan terhadap 263 status dan 65487 komentar. Tahap ini melakukan beberapa hal sebagai berikut:

1. Tahap penghapusan komentar kosong dan komentar yang hanya mengandung 1 huruf, dari 65487 komentar tersisa 58720 komentar.
2. Dari data tersebut dilakukan pengambilan data 1 status per hari dari masing-masing calon presiden Prabowo dan Jokowi dalam masa kampanye Pemilu 2014, sehingga diperoleh 68 status dan 15781 komentar. Dari data tersebut Prabowo Subianto tidak memposting status tanggal 7 Juni dan tgl 15 Juni 2014.
3. Pembuatan tabel-tabel pada basis data sebagai alat penyimpanan data untuk proses berikutnya. Tabel yang dibuat adalah statuses dan komentar.
4. Sesuai dengan skenario penelitian akan diambil 50 komentar secara acak dari 68 status menggunakan *stored procedure* pada MySQL. Dari *stored procedure* tersebut akhirnya dihasilkan 3400 komentar, sehingga hasil akhirnya adalah 68 status dan 3400 komentar.

3.4. Tahap Pengembangan Sistem Pelabelan Data

Setelah data siap digunakan, tahap ketiga seperti pada gambar 1 adalah melakukan pengembangan sistem pelabelan data *crowdsourcing* berbasis web menggunakan metode Majority Voting. Pengembangan sistem ini telah dilakukan melalui penelitian "Implementasi Crowdsourced Labelling Berbasis Web Menggunakan Metode Weighted Majority Voting" (Rachmat & Lukito, 2015). Hal-hal yang berkaitan dalam konfigurasi dan pengembangan sistem pelabelan dapat dijelaskan sebagai berikut:

1. Pengembangan skema basis data yang digunakan dalam sistem pelabelan online dapat dilihat pada Gambar 3.



Gambar 3. Skema Basis Data Sistem Crowdsourced Labelling (Rachmat & Lukito, 2015).

Sistem pelabelan ini dapat diakses pada alamat http://ti.ukdw.ac.id/~crowd. Implementasi halaman muka sistem berupa sebuah halaman web yang menerima login menggunakan akun Google Mail dalam hal ini akun email @ti.ukdw.ac.id.

2. Tahap berikutnya adalah pelabelan data dimana pelabel yang telah login akan mendapat tampilan dari 1 buah status dari salah satu calon presiden secara acak dan 1 komentarnya dari pengguna Facebook. Data yang diambil merupakan data komentar dan status yang memang belum dilabeli sebanyak 5 kali oleh 5 pengguna yang berbeda dan belum dilabeli oleh pelabel yang bersangkutan. Tampilan aplikasi web saat menampilkan status dan komentar dapat dilihat pada Gambar 4 berikut.

Gambar 4. Aplikasi Online Crowdsourced Labelling Status dan Komentar(Rachmat & Lukito, 2015)

Pada halaman web seperti pada gambar 4 di atas pengguna dapat memberi label, dengan memilih salah satu, yaitu positif, netral, atau negatif. Jika pengguna telah memilih salah satu label, maka sistem akan langsung melakukan validasi label dengan cara:

- Jika status dan komentar yang diberi label tersebut ternyata sudah mencapai 5 label dari 5 pelabel yang berbeda, maka sistem akan menampilkan informasi bahwa pelabelan gagal karena data sudah diberi label oleh pengguna lainnya.
- Jika status dan komentar yang diberi label memang belum diberi label oleh 5 pelabel yang berbeda maka sistem akan memberikan informasi pelabelan berhasil dan data label akan disimpan di dalam basis data yang telah dibuat sebelumnya, dan kemudian sistem akan merandom lagi untuk menampilkan data komentar baru lainnya kepada pengguna.

3. Hasil akhir pengembangan sistem pelabelan online ini adalah tahap pengujian sistem menggunakan data 300 komentar yang berasal dari 15 status Facebook Page masing-masing calon presiden, dimana dari masing-masing status tersebut diambil 10 komentar secara random sehingga total diperoleh 15 status x 10 komentar = 150 x 2 page = 300 komentar. Setelah tahap uji coba selesai dilakukan maka semua data yang telah dilabeli dapat diekspor dalam format .CSV.

3.5. Tahap Pelabelan Data Secara *Crowdsourced* Menggunakan *Weighted Majority Voting*

Setelah data siap digunakan, tahap ketiga seperti pada gambar 1 adalah melakukan pengembangan sistem pelabelan data *crowdsourcing* berbasis web menggunakan metode Majority Voting. Pengembangan sistem ini telah dilakukan melalui penelitian "Implementasi Crowdsourced Labelling Berbasis Web Menggunakan Metode Weighted Majority Voting" (Rachmat & Lukito, 2015). Hal-hal yang berkaitan dalam konfigurasi dan pengembangan sistem pelabelan dapat dijelaskan sebagai berikut:

1. Tahap 1 (3500 data komentar).
2. Tahap 2 (3500 data komentar).
3. Tahap 3 (3500 data komentar).
4. Tahap 4 (3250 data komentar).
5. Tahap 5 (3250 data komentar).

Pelabel yang akan digunakan pada tahap pelabelan ini menggunakan kriteria sebagai berikut: mahasiswa UKDW dan lulus seleksi awal pelabelan dalam bentuk pengisian soal-soal seleksi pre-test rekrutmen secara online pada alamat http://goo.gl/forms/HvTNKkLeB927414u1. Yang terpilih menjadi pelabel tetap adalah calon pelabel yang memiliki tingkat kesalahan menjawab lebih kecil dari 5 soal dari 25 soal pelabelan komentar yang diberikan.

Pada tahap pelabelan ini dapat diperoleh data sebagai berikut:

1. Tahap 1 7 April – 25 April 2016. Dikerjakan oleh 8 pelabel.
2. Tahap 2 28 April – 13 Mei 2016. Dikerjakan oleh 7 pelabel.
3. Tahap 3 18 Mei – 30 Mei 2016. Dikerjakan oleh 7 pelabel.
4. Tahap 4 6 Juni – 20 Juni 2016. Dikerjakan oleh 6 pelabel.
5. Tahap 5 15 September – 1 Oktober 2016. Dikerjakan oleh 7 pelabel.

Dari tahap pertama hingga kelima, terkumpul calon pelabel sebanyak 102 pelamar, dan total pelabel yang terpilih adalah 33 orang. Setelah selesai tahap kelima pada tanggal 1 Oktober maka terkumpulah 3400 data komentar yang telah terlabel secara final.

3.6. Kendala yang Dihadapi

Kendala-kendala yang dihadapi adalah tidak mudahnya menemukan pelabel yang mampu menyelesaikan soal-soal seleksi pre-test pelabelan dan memperoleh batas skor tertentu sebagai syarat menjadi pelabel selanjutnya. Kendala berikutnya adalah masalah sistem, ketika dataset semakin banyak maka pelabelan menjadi lambat karena sistem mendapat beban pencarian SQL ke database yang diharuskan mencari data yang belum dilabeli. Proses tersebut membutuhkan waktu yang cukup lama sehingga sistem terkesan lambat. Hal ini dapat dikurangi dengan cara mengunggah dataset secara perlahan-lahan tidak secara keseluruhan sekaligus.

3.7. Validasi Dataset

Setelah seluruh dataset dihasilkan dalam bentuk CSV, maka dataset tersebut dilakukan validasi oleh pakar (dalam hal ini penulis) untuk memastikan bahwa dataset yang dihasilkan memiliki label yang benar dan dapat dipertanggungjawabkan. Statistik data label sebelum validasi adalah 2940 positif, 337 negatif, dan 123 netral. Validasi dataset dilakukan dengan cara mengambil sampel random data dari dataset sejumlah 10% (340 data) yang terdiri dari label positif, negatif, dan netral. Dataset terebut kemudian divalidasi dengan memeriksa secara manual apakah setiap data memiliki label yang benar atau tidak, kemudian dihitung tingkat persentase kebenarannya dengan rumus jumlah benar dibagi jumlah seluruh data (340 buah) dikali 100%. Dari 340 data tersebut dihasilkan 95.3% data benar / valid, sehingga dapat dikatakan bahwa dataset ini dapat digunakan sebagai sumber data pelatihan *supervised learning* lainnya. Dataset yang dihasilkan bersifat final, tidak akan bertambah lagi karena dataset ini diambil hanya masa kampanye pemilu 2014 saja. Contoh data dan metode validasi dataset dapat dilihat pada Tabel 2 berikut.

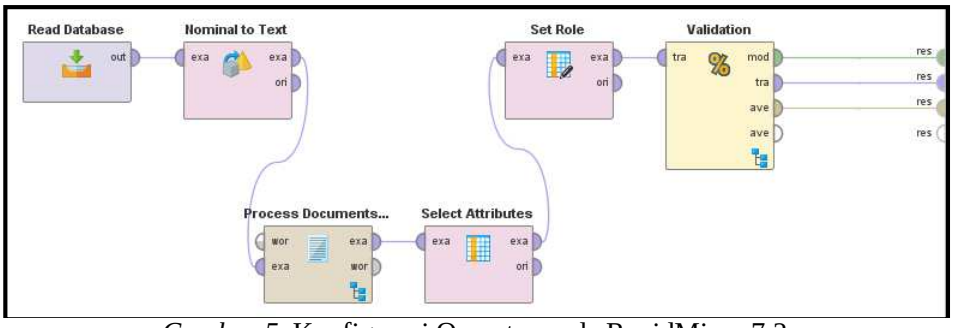
Tabel 2**.** Contoh Data dan Metode Validasi Dataset

No.	Status	Komentar	Label Sistem	Label Manual	Validasi
1.	Sahabat, hari ini saya kembali berkeliling Jawa Timur. Saya teringat kata-kata seorang pemimpin yang baik, pemimpin yang berhasil yaitu bapak Muhammad Noer (Cak Noer), gubernur Jawa Timur dari tahun 1967 sampai 1978 yang sangat legendaris dan sangat dicintai oleh rakyat Jawa Timur. Pada suatu saat ia mengatakan kepada saya, ?seorang pemimpin berhasil kalau rakyat biasa bisa tersenyum?. Dalam bahasa Jawa, ?yen wong cilik iso gumuyu maka pemimpin rakyat itu berhasil?. Itulah mimpi saya. Suatu saat rakyat Indonesia bisa saling menyapa dengan damai apapun suku, agama dan rasnya. Kita semua adalah manusia, kita semua adalah hamba Tuhan. Kita semua punya hak hidup di alam semesta ini. Kita semua bersaudara, kenapa kita harus bertikai?	KAMPANYE HARI PERTAMA | TIPS MENCOBLOS PILPRES 9 JULI 2014 1. Jika anda menyukai Prabowo, langsung coblos gambar Prabowo. 2. Jika anda membenci Prabowo,coblos gambar Prabowo sekuat tenaga, dengan kekuatan penuh, agar kebencian anda terpuaskan. 3. Jika anda menyukai hatta rajasa,silakan coblos gambar hatta rajasa. 4. Jika anda menyukai Prabowo tapi tidak menyukai hatta rajassa,coblos gambar Prabowo saja. Jangan gambar hatta rajasa 5. Jika anda menyukai hatta rajasa,tapi tidak menyukai Prabowo,coblos gambar hatta rajasa nya saja. Gambar Prabowo pandengin aja, atau kalau perlu pelototin; ?Ndesit Lu!? 6. Jika anda menyukai Jokowi dan tidak menyukai Prabowo, coblos Prabowo dengan semangat, agar kejengkelan anda terlampiaskan dan kesukaan anda pada Prabowo terpenuhi. 7. Jika anda menyukai Hatta Radjasa namun tidak menyukai Prabowo, coblos Prabowo saja, karena memang benar pendapat anda; Hatta lebih pantas jadi capres daripada Prabowo.	Positif	Positif	Valid

		8. Jika anda tidak menyukai keempatnya, coblos gambar Prabowo saja berkali-kali, asal tetap dalam frame, untuk mewakili ketidaksukaan anda terhadap keempat tokoh ini, atau nafsu golput anda terpuaskan. :p =))			
2	Tema debat malam ini, pukul 20.00 sampai 22.00 WIB adalah pembangunan demokrasi, pemerintahan yang bersih dan kepastian hukum.	jangan memaki,mengolok olok dan menghujat ya....	Positif	Netral	Tidak Valid
3.	Tema debat malam ini, pukul 20.00 sampai 22.00 WIB adalah pembangunan demokrasi, pemerintahan yang bersih dan kepastian hukum. Saya mohon doa' dan masukan dari sahabat Facebook, agar acara malam ini dapat berlangsung dengan baik, serta bermanfaat untuk demokrasi kita. Terima kasih.	Sholat belum pak,,, sholat dulu ya, och iya slamat kemaren kmpanye disolo sampe ngundang SERA mlh ga main karna ga ada yg datang. saya sebagai warga solo ketawa sampe jingkrak" karna lucu kasian penyanyinya kkkkk wkekwkwkwkwkekw JOKOWI YESSSSS, SHOLAT NO 1, PRESIDEN NO 2..!!!	Negatif	Negatif	Valid

3.8. Pengujian Dataset

Dataset yang telah divalidasi kemudian dilakukan pengujian untuk diuji coba digunakan sebagai data training klasifikasi sentimen menggunakan dua buah metode yaitu Naive Bayes dan Support Vector Machine menggunakan tool RapidMiner 7.2. Konfigurasi RapidMiner 7.2 yang digunakan dapat dilihat pada Gambar 5 sebagai berikut:



Gambar 5. Konfigurasi Operator pada RapidMiner 7.2

Kemudian hasil pengujian dengan kedua metode adalah sebagai berikut:

1. Hasil Pengujian menggunakan Metode Naive Bayes:

```
Accuracy: 83.32% +/- 1.35% (mikro: 83.32%)
ConfusionMatrix:
True: positif    negatif  netral
positif:2801     296       112
negatif:102      29        8
netral:37        12        3
Classification_error: 16.68% +/- 1.35% (mikro: 16.68%)
```

2. Hasil Pengujian menggunakan Metode Support Vector Machine:

```
Accuracy: 84.82% +/- 2.31% (mikro: 84.82%)
ConfusionMatrix:
True:positif     negatif     netral
positif:2865     316          110
negatif:56       15           9
netral:19        6            4
classification_error: 15.18% +/- 2.31% (mikro: 15.18%)
```

Dari hasil pengujian menggunakan dua metode berbeda, data komentar pada dataset yang telah dibangun mampu menghasilkan tingkat akurasi yang cukup tinggi yaitu di atas 80%. Memang metode Support Vector Machine memang menghasilkan keakuratan lebih tinggai dibandingkan dengan metode Naive Bayes namun tidak terlalu besar. Terlihat pula tingkat error pada klasifikasi mencapai sekitar 15% untuk kedua metode. Karena dataset tersebut sudah terbukti dapat digunakan pada pembelajaran dan pengujian klasifikasi sentimen dengan tingkat keakuratan di atas 80%, maka dataset tersebut dapat digunakan sebagai bahan data pembelajaran sistem *supervised learning* lainnya.

4. Penutup

4.1. Kesimpulan

Dari penelitian yang telah dilakukan, diperoleh kesimpulan sebagai berikut:

1. Penelitian ini telah berhasil menyelesaikan pelabelan data komentar berdasarkan status calon presiden pada masa kampanye Pemilu 2014 menggunakan metode crowdsourced labelling berbasis weightened majority voting secara online. Dataset yang dihasilkan telah dapat dipublikasikan menggunakan format CSV sehingga dapat digunakan secara luas oleh peneliti lain di bidang *supervised learning*.
2. Metode *crowdsourced labelling* berbasis weighted majority voting secara online sangat tepat digunakan dalam pengembangan dataset yang memiliki jumlah data yang sangat besar karena dapat dilakukan oleh banyak orang, tidak terikat oleh waktu, dan dapat dilakukan dari manapun menggunakan Internet.
3. Metode *weightened majority voting* telah mampu mengotomatisasi keputusan pengambilan label akhir dari tiap data komentar yang dilabeli oleh banyak pelabel sekaligus.
4. Validasi dataset yang telah dilakukan menghasilkan validitas data sebesar 95.3%.
5. Dataset telah diuji coba digunakan sebagai data pelatihan klasifikasi sentimen menggunakan metode Naive Bayes dan Support Vector Machine dengan kekuratan masing-masing adalah 83.32% dan 84.82%.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Ristek DIKTI atas bantuan dana dan dukungan penelitian yang telah diberikan melalui kontrak Skema Penelitian Dosen Pemula No. 009/HB-LIT/III/2016, tahun pendanaan 2016

Daftar Pustaka

DHS. (2013, September 25). Indonesia: Standard DHS. Diakses 28 Agustus 2014, dari http://dhsprogram.com/what-we-do/survey/survey-display-357.cfm

Geertzen, J. (2012). Inter-Rater Agreement with multiple raters and variables. Diakses 14 August 2014 dari https://mlnl.net/jg/software/ira/

Gjoka, M., T. Butts, C., Kurant, M., & Markopoulou, A. (2011). Multigraph Sampling of Online Social Networks. *COMMUNICATIONS,29*(1893), 1905-1905. Diakses 28 Agustus 2014, dari http://minasgjoka.com/papers/jsac11_multigraph_sampling.pdf

Hosseini, M., J. Cox, I., Mili ć –Frayling, N., Kazai, G., & Vinay, V. (2012). On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. *ECIR*. Diakses 14 Agustus 2014 dari http://www0.cs.ucl.ac.uk/staff/ingemar/Content/papers/2012/ECIR2012.pdf

Hosseinkhah, F., Ashktorab, H., Veen, R., & Owrang O., M. M. (2009). Challenges in Data Mining on Medical Databases. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology, Second Edition* (pp. 502-511). Hershey, PA: . doi:10.4018/978-1-60566-026-4.ch083

Infochimps. (2010). Twitter Census. Diakses 28 Agustus 2014, dari http://www.infochimps.com/collections/twitter-census

James, G. (1998). *Majority Vote Classifiers: Theory and Applicaitons* (Doctoral dissertation, Stanford). Retrieved from http://web.stanford.edu/~hastie/THESES/gareth_james.pdf

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media.

Business Horizons, *53*(1), 59-68. doi:10.1016/j.bushor.2009.09.003

KOMINFO. (2014). Kemkominfo: Pengguna Internet di Indonesia Capai 82 Juta. Website Resmi Kementerian Komunikasi dan Informatika RI. Retrieved 12 October 2016, from https://kominfo.go.id/content/detail/3980/kemkominfo-pengguna-internet-di-indonesia-capai-82-juta/0/berita_satker

Markopoulou, A., T. Butts, C., Kurant, M., Gjoka, M., Cici, B., & Wang, Y. (2011). Sampling Online Social Networks. Diakses 28 Agustus 2014, dari http://odysseas.calit2.uci.edu/doku.php/public:online_social_networks#facebook_social_graph_-_breadth_first_search

Matsubara, E. T., Monard, M. C., & Prati, R. C. (2008). Exploring Unclassified Texts Using Multiview Semisupervised Learning. In H. A. do Prado & E. Ferneda (Eds.), *Emerging Technologies of Text Mining* (pp. 139-161). Hershey, PA: Information Science Reference. Diakses 6 Februari 2015darihttp://go.galegroup.com/ps/i.do?id=GALE%7CCX2657900017&v=2.1&u=idpnri&it=r&p=GVRL&sw=w&asid=e1743a115ec1f3ce6f8caf50fd6bbe9d

McAuley, J., & Lescovec, J. (2012).Learning to Discover Social Circles in Ego Networks. *NIPS*. Stanford, USA. Diakses 28 Agustus 2014 dari https://snap.stanford.edu/data/egonets-Facebook.html

Oyama, S., Baba, Y., Sakurai, Y., & Kashima, H. (2013). Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2554-2560. Diakses 14 August 2014 dari http://ijcai.org/papers13/Papers/IJCAI13-376.pdf

Rachmat, A. & Lukito, Y. (2015). Implementasi Crowdsourced Labelling Berbasis Web Menggunakan Metode Weighted Majority Voting. *Ultima Infosys*, *6*(2).

Statistic Brain. (2014). Facebook Statistics. Diakses 28 Agustus 2014 dari http://www.statisticbrain.com/facebook-statistics/

Statistic Brain. (2014, July 11). Twitter Statistics. Diakses 28 Agustus 2014 dari http://www.statisticbrain.com/twitter-statistics/

Stefano, C. D., Cioppa, A. D., & Marcelli, A. (2002). An Adaptive Weighted Majority Vote Rule for Combining Multiple Classifiers. *Proceedings on Pattern Recognition. 16th International Conference*, *Vol2*, p. 192 – 195. Doi:10.1109/ICPR.2002.1048270. Diakses 5 September 2014 dari http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1048270

Biodata Penulis

Antonius Rachmat, memperoleh gelar S1 di Universitas Kristen Duta Wacana Yogyakarta. Memperoleh gelar S2 di Universitas Gadjah Mada Yogyakarta. Saat ini menjadi pengajar di Prodi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana Yogyakarta.

Yuan Lukito, memperoleh gelar S1 di Universitas Gadjah Mada Yogyakarta. Memperoleh gelar S2 di Universitas Gadjah Mada Yogyakarta. Saat ini menjadi pengajar di Prodi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana Yogyakarta

BERITA ACARA PELAKSANAAN HASIL SEMINAR SESI PARALEL KNASTIK 2016

Judul : SENTIPOL: Dataset Sentimen Komentar Pada Kampanye PEMILU Presiden Indonesia 2014 dari Facebook Page

Pemakalah : Antonius Rachmat, Yuan Lukito

Moderator : Gloria Virginia, S.Kom., MAI, Ph.D

Notulis : Abadi

Peserta : 11 orang di ruang : D.3.2

Tanya Jawab :

Pertanyaan (Oleh sdr. Alz):

1. Bagaimana cara memasukkan atau menyeleksi semua data yang masuk ? Bahasa gaul bagaimana cara melabelkannya ?
2. Apa yang dilihat pada *Naïve Bayes* disini ? Lalu bagaimana cara mengklasifikasinya ?

Jawaban :

1. Data yang akan dimasukkan terlebih dahulu dibersihkan atau dikonversi dengan thesaurus tanpa merubah makna sama sekali. Setiap data dilabeli dengan cara manual per 5 mahasiswa dengan memperhatikan makna kalimatnya.
2. Yang dilihat dalam naïve bayes adalah frekuensi kata dokumen positif, negative dan netral. Cara melakukan klasifikasi menggunakan tools.

Masukkan : Mungkin ada pengecekan di satu kalimat apakah ada kemiripan dengan tempat lainnya, namun bisa menjadi topik penelitian yang baru.

Masukan Seminar :

Metodologi penelitian sudah dijelaskan dengan baik.

Yogyakarta, 19 November 2016

Moderator Kelas	Penyaji Makalah
Gloria Virginia, S.Kom., MAI, Ph.D.	Yuan Lukito