

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

Antonius Rachmat dan Yuan Lukito

KNASTIK 2016 Universitas Kristen Duta Wacana
19 November 2016

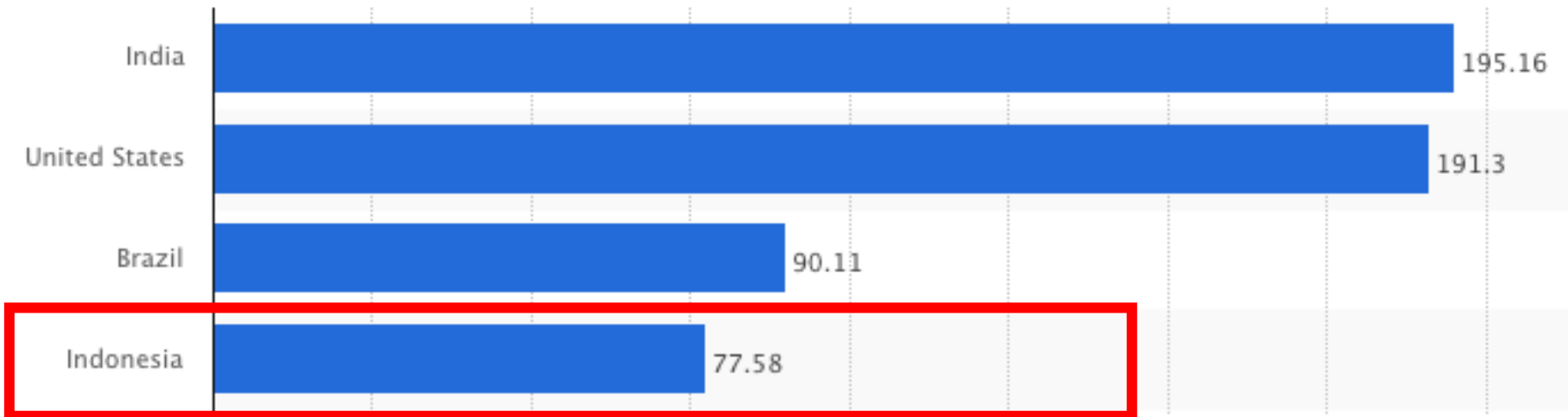


- Pengguna Internet Indonesia nomor 6 di dunia
 - https://kominfo.go.id/content/detail/4286/pengguna-internet-indonesia-nomor-enam-dunia/0/sorotan_media

Top 25 Countries, Ranked by Internet Users, 2013-2018 millions

	2013	2014	2015	2016	2017	2018
1. China*	620.7	643.6	669.8	700.1	736.2	777.0
2. US**	246.0	252.9	259.3	264.9	269.7	274.1
3. India	167.2	215.6	252.3	283.8	313.8	346.3
4. Brazil	99.2	107.7	113.7	119.8	123.3	125.9
5. Japan	100.0	102.1	103.6	104.5	105.0	105.4
6. Indonesia	72.8	83.7	93.4	102.8	112.6	123.0

- Indonesia merupakan pengguna Facebook terbesar ke-4 di dunia (77,58 juta)
 - <https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/>



- Selebriti, politikus maupun tokoh publik lainnya menggunakan Facebook, dengan beberapa macam tujuan:
 - Menyampaikan visi/misi/pandangan/program kegiatan
 - Berinteraksi langsung dengan penggemar/pendukungnya
 - Meningkatkan popularitas
 - **Mengetahui respon masyarakat (tingkat dukungan)**

- Penelitian mengenai **Sentimen Analisis** sudah banyak dilakukan
 - Sumber data dari Facebook, Twitter, situs berita online maupun sumber-sumber lainnya di Internet
- Dataset berbahasa Indonesia
 - **Masih jarang**, umumnya peneliti membangun dataset sendiri
 - Antar penelitian menggunakan dataset yang **berbeda-beda**, sehingga sulit dibandingkan
 - Membutuhkan **waktu** untuk pengumpulan dan pemrosesan awal data



Prabowo Subianto

August 17 · 🌐

Dirgahayu Kemerdekaan Republik Indonesia ke-71.

Sahabatku dimanapun engkau berada. Banyak kekuatan besar ingin Indonesia selalu lemah, mereka ingin rakyat Indonesia selalu miskin. Ingin Indonesia menjadi negara kacung, mereka tidak suka Indonesia punya pemerintahan yang bersih. Oleh karena itu mereka dengan segala cara berusaha untuk menghentikan perjuangan kita.

Saya menghimbau kepada seluruh rakyat Indonesia yang peduli akan masa depan bangsa kita. Masa depan anak kita, cu... [See More](#)



👍 Like 💬 Comment ➦ Share



👍❤️😱 Tjondro Purnomo and 139K others

[Top Comments](#) ▾

139.000 Response
(Like, Love, ...)

8695 Komentor

- Jumlah komentar yang **sangat banyak** membutuhkan banyak waktu untuk **standarisasi** dan **pelabelan**
 - **Standarisasi**: penyeragaman format, penghapusan karakter-karakter yang tidak bermakna, serta beberapa tahap pemrosesan awal data teks yang umum digunakan
 - **Pelabelan**: setiap komentar harus diberi **label sentimen** (negatif, netral atau positif)
- Membutuhkan “*ahli bahasa*” untuk memberi label
 - Biaya relatif mahal
 - Membutuhkan waktu lama

- Membangun dataset SENTIPOL (Sentimen Politik) berdasarkan data komentar pada masa kampanye PEMILU presiden Indonesia tahun 2014
- Dataset komentar tersebut sudah dilabeli dengan benar dan akurat

- Pengumpulan Data

- Mengambil komentar dari setiap post yang dibuat oleh kandidat presiden Indonesia pada PEMILU 2014 selama periode kampanye (4 Juni – 9 Juli 2014)
- Didapatkan 68 status
- Dari masing-masing status diambil 50 komentar pertama, sehingga didapatkan 3400 komentar secara keseluruhan

- Pengembangan Crowdsourced Labelling berbasis Web
 - Dapat diakses di <http://ti.ukdw.ac.id/~crowd>
 - Menggunakan mahasiswa (setelah melalui tes awal) sebagai pelabel
 - Pilihan label: negatif, netral dan positif
 - Setiap komentar dilabeli oleh 5 orang mahasiswa yang berbeda
 - Label akhir dihitung menggunakan metode Weighted Majority Voting

[lihat daftar status](#)

Joko Widodo

Sungguh merupakan kebanggaan tersendiri bagi saya melihat apa yang diharapkan generasi muda bangsa ini terhadap pemimpinnya dan juga terhadap masa depan negeri ini.

Komentar:

Alvin Waruwu

Jokowi yg tepat memimpin bangsa INI

Sudah dilabeli **5x**Positif **5**Netral **0**Negatif **0**Status: **lengkap**Label Final: **positif**

Ok Nurman

matap ini pak,saya tetap memili bapak...

Sudah dilabeli **5x**Positif **5**Netral **0**Negatif **0**Status: **lengkap**Label Final: **positif**

Renal Waisama

Pemimpin, Penggerak perubahan. Spirit anak Muda, tp sarat pengalaman, dan punya integritas... Harapan saya, Pa Jokowi bs membuat peruhan di bidang Hukum... Basmi para Koruptor. Bravo Jokowi.

Sudah dilabeli **5x**Positif **5**Netral **0**Negatif **0**Status: **lengkap**Label Final: **positif**

Weighted Majority Voting

- Majority Voting
 - Hasil akhir ditentukan berdasarkan hasil vote dari beberapa pelabel yang berbeda

Negatif	Netral	Positif
3	1	1

Negatif	Netral	Positif
2	1	2

Weighted Majority Voting

- Weighted Majority Voting
 - Setiap vote diberi bobot (negatif = -1, netral = 0 dan positif = +1)

Negatif	Netral	Positif
2	1	2

$$\text{Hasil Akhir} = 2 * (-1) + 1 * (0) + 2 * (+1) = 0$$

Negatif	Netral	Positif
2	0	3

$$\text{Hasil Akhir} = 2 * (-1) + 0 * (0) + 3 * (+1) = 1$$

Renal Waisama

Pemimpin, Penggerak perubahan. Spirit anak Muda, tp sarat pengalaman, dan punya integritas... Harapan saya, Pa Jokowi bs membuat peruhan di bidang Hukum... Basmi para Koruptor. Bravo Jokowi.

Sudah dilabeli 5x

Positif 5

Netral 0

Negatif 0

Status: lengkap

Label Final: positif

Positif

nico.raynaldo@ti.ukdw.ac.id

kent.wijaya@ti.ukdw.ac.id

tabitha.banu@ti.ukdw.ac.id

rico.manurung@ti.ukdw.ac.id

untung.putra@ti.ukdw.ac.id

- Validasi label dataset
 - Dilakukan secara manual oleh peneliti
 - Diambil 10% (340 komentar) secara acak
 - Didapatkan 95,3% sudah sesuai
- Validasi juga dilakukan dengan uji coba pengujian dengan Naïve Bayes dan Support Vector Machine (SVM)
 - Naïve Bayes: akurasi 83,32%
 - SVM: akurasi 84,82%

- Publikasi dataset
 - Dapat diunduh dalam format CSV di:
<https://ti.ukdw.ac.id/~crowd/>
 - Dalam proses unggah ke UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)

- Sulit mencari pelabel yang memenuhi syarat
 - Lulus tes awal pelabelan dengan nilai minimal 80% (20 dari 25)
 - Solusi: perekrutan terbuka dan dilakukan dalam beberapa tahap
- Jumlah dataset yang cukup besar menyebabkan web crowdsourced menjadi lambat
 - Solusi: pindahkan ke hosting yang lebih baik dan membagi pelabelan menjadi beberapa tahap

- Penelitian telah menghasilkan dataset sentimen politik (SENTIPOL) dengan sistem Crowdsourced-labelling dan metode Weighted Majority Voting
- Metode Crowdsourced-labelling berbasis Web dapat membantu proses pelabelan data dalam jumlah besar walaupun ada beberapa kendala di bagian sumber daya hostingnya
- Dataset yang dihasilkan sudah diuji validitasnya dengan tingkat validitas 95,3% dan sudah diuji untuk sentimen analisis dengan metode Naïve Bayes dan SVM

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

Terima kasih

Antonius Rachmat dan Yuan Lukito



KNASTIK 2016 Universitas Kristen Duta Wacana

19 November 2016

NOVASI
E-HEALTH DAN BIOMEDIKA
UNTUK INDONESIA