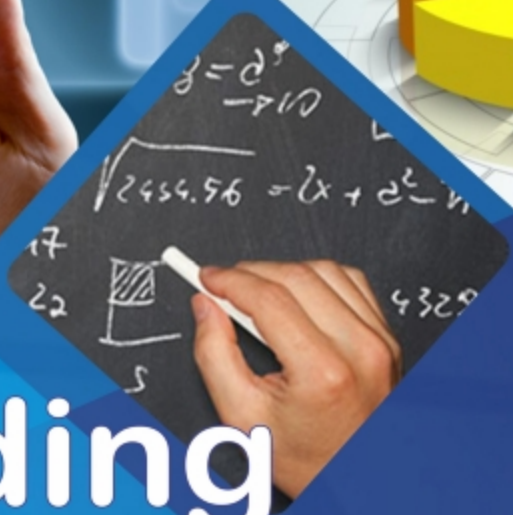
ISBN 978-602-14413-1-2

SEMINAR NASIONAL MATEMATIKA DAN APLIKASINYA

Peranan Matematika dan Sistem Informasi di Era Big Data untuk Menunjang Perkembangan IPTEK di Indonesia

Surabaya, 21 Oktober 2017



Prosiding

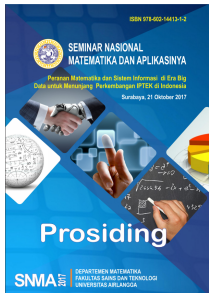
SNMA 2017

DEPARTEMEN MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA

Prosiding SNMA 2017

By On Wednesday, March 06 th, 2019 · *no Comments* · In *Uncategorized*

Prosiding SNMA 2017



NO	MAKALAH	BIDANG
1	KETERBATASAN OPERATOR INTEGRAL FRAKSIONAL PADA RUANG KUASI METRIK TAK HOMOGEN TERBOBOTI	Analisis & Aljabar
2	CRYPTOGRAPHIC RANDOMNESS TESTING PADA ALGORITMA BLOCK CIPHER CAMELLIA MENGGUNAKAN UJI COVERAGE	Analisis & Aljabar
3	KAJIAN OPERATOR ACCRETIVE DAN SIFAT KETERBATASAN PADA RUANG HILBERT	Analisis & Aljabar
4	KESTABILAN MODEL MANGSA PEMANGSA DENGAN FUNGSI RESPON HOLLING TIPE III DAN PENYAKIT PADA PEMANGSA SUPER	Matematika Terapan
5	PENGARUH MAKANAN TAMBAHAN DALAM SISTEM MANGSA–PEMANGSA BEDDINGTON– DEANGELIS	Matematika Terapan

6	EVALUASI SIFAT COMPLETENESS ALGORITMA SNOW 2.0 DENGAN MENGGUNAKAN METODE DIFFUSION TEST	Matematika Terapan
7	PENGEPAKAN LINGKARAN DALAM PERSEGI PANJANG DENGAN METODE ALGORITMA GENETIKA	Matematika Terapan
8	PREDIKSI CUACA MENGGUNAKAN ALGORITMA PARTICLE SWARM OPTIMIZATION-NEURAL NETWORK (PSONN)	Matematika Terapan
9	ANALISIS KONTROL OPTIMAL MODEL MATEMATIKA PENYEBARAN PENYAKIT HIV PADA POPULASI HETEROSEKSUAL	Matematika Terapan
10	STRATEGI OPTIMAL PADA MODEL MATEMATIKA PENYEBARAN PENYAKIT HIV PADA INDUSTRI SEKS KOMERSIAL	Matematika Terapan
11	MODEL PENYEBARAN MIDDLE EAST RESPIRATORY SYNDROME (MERS) DENGAN PENGARUH PENGOBATAN	Matematika Terapan
12	PENYELESAIAN MASALAH DIFUSI PANAS PADA SUATU KABEL PANJANG	Matematika Terapan
13	KESTABILAN MODEL POPULASI SATU MANGSA-DUA PEMANGSA DENGAN PEMANENAN OPTIMAL PADA PEMANGSA	Matematik Terapan
14	ANALISIS STABILITAS PENYEBARAN VIRUS EBOLA PADA MANUSIA	Matematika Terapan
15	ANALISIS MODEL MATEMATIKA PENYEBARAN KOINFEKSI MALARIA-TIFUS	Matematika Terapan
16	FUNGSI BANTU NONPARAMETRIK BARU UNTUK MENYELESAIKAN OPTIMASI GLOBAL	Matematika Terapan
17	ANALISIS KESTABILAN PADA MODEL DUA MANGSA- SATU PEMANGSA DENGAN FUNGSI RESPON HOLLING DAN PEMANENAN	Matematika Terapan
18	LINEARISASI PERSAMAAN PERAMBATAN RETAK UNTUK MENDAPATKAN PARAMETER SIFAT MATERIAL DARI OBSERVASI UJI FATIK	Matematika Terapan
19	ANALISIS SISTEM DINAMIK DAN KENDALI OPTIMAL PADA PENYEBARAN POPULASI ANJING RABIES DI KOTA AMBON	Matematika Terapan
20	KESULITAN BELAJAR GARIS ISTIMEWA DALAM SEGITIGA PADA SISWA BERKEMAMPUAN RENDAH BERDASARKAN TEORI PIAGET	Matematika Pendidikan
21	PEMBELAJARAN PEMECAHAN MASALAH MATEMATIKA DI SEKOLAH DASAR DENGAN MODEL PEMBELAJARAN OSKAR	Matematika Pendidikan
22	PROFIL KOMUNIKASI MATEMATIK TERTULIS CALON GURU MATEMATIKA DENGAN TINGKAT KECEMASAN MATEMATIKA TINGGI DALAM PEMBUKTIAN MATEMATIKA	Matematika Pendidikan
23	ANALISIS KEMAMPUAN MAHASISWA MATEMATIKA DAN PENDIDIKAN MATEMATIKA DALAM MEMAHAMI KONSEP KALKULUS INTEGRAL	Matematika Pendidikan
24	MENINGKATKAN KEMAMPUAN KOMUNIKASI MATEMATIS SISWA SMP MELALUI PEMBELAJARAN KONTEKSTUAL BERBASIS BUDAYA PESISIR	Matematika Pendidikan
25	IMPLEMENTASI BLENDED LEARNING UNTUK MENINGKATKAN KEMANDIRIAN BELAJAR MAHASISWA PADA MATA KULIAH METODE NUMERIK	Matematika Pendidikan
26	PENGEMBANGAN BAHAN AJAR BERBASIS PROJECT-BASED LEARNING BERBANTUAN SCRATCH	Matematika Pendidikan
27	PENGARUH KETERLIBATAN IBU DALAM PEMBELAJARAN MATEMATIKA ANAK USIA DINI DENGAN MEDIA SCRAPMATIK	Matematika Pendidikan
28	PENGEMBANGAN MODEL COLLABORATIVE LEARNING MATEMATIKA BERBASIS MEDIA BLOG MATAKULIAH KALKULUS II	Matematika Pendidikan
29	PENGEMBANGAN MODUL MATEMATIKA DISKRIT UNTUK MENINGKATKAN MULTIPLE INTELLIGENCIES MAHASISWA UNIVERSITAS PGRI ADI BUANA SURABAYA	Matematika Pendidikan
30	PENGEMBANGAN E-MODUL INTERAKTIF BERBASIS CASE (CREATIVE, ACTIVE, SYSTEMATIC, EFFECTIVE) SEBAGAI ALTERNATIF MEDIA PEMBELAJARAN	Matematika Pendidikan

	GEOMETRI TRANSFORMASI UNTUK MENDUKUNG KEMANDIRIAN BELAJAR DAN KOMPETENSI MAHASISWA	
31	MODEL PROSES TITIK BERTANDA TERINDEKS WAKTU PADA DATA GEMPA BUMI DI PANTAI SELATAN JAWA	Statistika
32	PEMODELAN DAN PEMETAAN FAKTOR UNMET NEED KB DI JAWA TIMUR SEBAGAI PERENCANAAN MENCEGAH LEDAKAN PENDUDUK DENGAN REGRESI LOGISTIK BINER	Statistika
33	FUNGSI GOODNESS OF FIT DALAM KRITERIA PENALIZED SPLINE PADA ESTIMASI REGRESI NONPARAMETRIK BIRESPON UNTUK DATA LONGITUDINAL	Statistika
34	PEMODELAN FUNGSI RELIABILITAS DISTRIBUSI GUMBEL TERSENSOR TYPE III DENGAN PENDEKATAN METODE BOOTSTRAP	Statistika
35	ANALISIS ARIMA BOX JENKINS UNTUK PERAMALAN JUMLAH KUNJUNGAN WISATAWAN MANCANEGARA DI INDONESIA	Statistika
36	ANALISIS DISKRIMINAN UNTUK STATUS KUALITAS KELAYAKAN RUMAH TINGGAL DI KAB. KUPANG	Statistika
37	KONSTRUKSI DAN ESTIMASI MATRIKS KOVARIANSI DALAM MODEL REGRESI NONPARAMETRIK MULTIRESPON BERDASARKAN ESTIMATOR SMOOTHING SPLINE UNTUK BEBERAPA KASUS UKURAN SAMPEL	Statistika
38	ESTIMASI FUNGSI REGRESI NONPARAMETRIK MULTIRESPON MENGGUNAKAN REPRODUCING KERNEL HILBERT SPACE BERDASARKAN ESTIMATOR SMOOTHING SPLINE	Statistika
39	ANALISIS PETA KENDALI DEWMA (DOUBLE EXPONENTIALLY WEIGHTED MOVING AVERAGE) DALAM PENGENDALIAN KUALITAS PRODUKSI FJLB (FINGER JOINT LAMINATING BOARD) PADA PT. INHUTANI 1 GRESIK	Statistika
40	STRUCTURAL EQUATION MODELING – PARTIAL LEAST SQUARE UNTUK PEMODELAN INDEKS PEMBANGUNAN KESEHATAN MASYARAKAT (IPKM) DI JAWA TIMUR	Statistika
41	ANALISA KEANDALAN PADA PERALATAN UNIT PENGGILINGAN AKHIR SEMEN UNTUK MENENTUKAN JADWAL PERAWATAN MESIN (STUDI KASUS PT SEMEN INDONESIA PERSERO TBK.)	Statistika
42	PENGELOMPOKKAN DESA DI KABUPATEN SORONG PROVINSI PAPUA BARAT TAHUN 2016 BERDASARKAN STATUS KETERTINGGALAN	Statistika
43	ANALISIS PENGARUH INDIKATOR K3 TERHADAP KEPUASAN KARYAWAN PT. TELKOM WITEL SURABAYA	Statistika
44	REGRESI LOGISTIK UNTUK PEMODELAN INDEKS PEMBANGUNAN KESEHATAN MASYARAKAT DI PULAU KALIMANTAN	Statistika
45	ANALISIS SURVIVAL LAJU INDEKS KINERJA DOSEN STKIP PGRI TULUNGAGUNG DENGAN MODEL REGRESI COX	Statistika
46	PENDEKATAN MODEL EKONOMETRIKA UNTUK MEMPREDIKSI INDEKS SAHAM SYARIAH INDONESIA	Statistika
47	PENGGUNAAN MODEL GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSCEDASTICITY (P,Q) UNTUK PERAMALAN HARGA DAGING AYAM BROILER DI PROVINSI JAWA TIMUR	Statistika
48	PEMODELAN KADAR GULA DARAH DAN TEKANAN DARAH PADA REMAJA PENDERITA DIABETES MELITUS TIPE II DENGAN PENDEKATAN REGRESI NONPARAMETRIK BIRESPON BERDASARKAN ESTIMATOR SPLINE	Statistika
49	ANALISIS UJI STATISTIK BERBASIS KORELASI PADA ALGORITMA SNOW 2.0	Sistem Informasi
50	SISTEM INTEGRASI PENDISTRIBUSIAN OBAT PADA INSTALASI FARMASI BERBASIS SUPPLY CHAIN MANAGEMEN	Sistem Informasi
51	TEXT MINING PADA MEDIA SOSIAL TWITTER STUDI KASUS: MASA TENANG PILKADA DKI 2017 PUTARAN 2	Sistem Informasi
52	MONITORING KESEHATAN MENGGUNAKAN COMPILER ARDUINO & MODUL WIFI ESP8266 UNTUK KOMUNITAS PASIEN HIPERTENSI	Sistem Informasi
53	IMPLEMENTASI K-MEANS CLUSTERING UNTUK PEMETAAN DESA DAN KELURAHAN DI KABUPATEN BANGKALAN BERDASARKAN CONTRACEPTIVE PREVALENCE RATE DAN TINGKAT PENDIDIKAN	Sistem Informasi
54	PENDEKATAN NUMERIK FUNGSI GAMMA UNTUK PERHITUNGAN LEVY FLIGHT PADA ALGORITMA CUCKOO SEARCH	Sistem Informasi
55	PENYIRAM TANAMAN MEDIA OTOMATIS BERBASIS TELEPON SELULER PINTAR dan JARINGAN SENSOR FUZZY TANPA KABEL	Sistem Informasi
56	IDENTIFIKASI POLA PENYAKIT ANAK DIBAWAH 5 TAHUN (BALITA) DENGAN MENGGUNAKAN ALGORITMA APRIORI	Sistem Informasi
57	SISTEM INFORMASI BANK SOAL TEKNIK INFORMATIKA POLITEKNIK NEGERI TANAH LAUT	Sistem Informasi
58	PROTOTYPE SISTEM MONITORING DAN PENGENDALIAN PINTU AIR OTOMATIS SEBAGAI PERINGATAN DINI BAHAYA BANJIR BERBASIS INTERNET OF THINGS	Sistem Informasi
59	PERANCANGAN SISTEM INFORMASI MONITORING DISTRIBUSI LOGISTIK BANTUAN BENCANA (MDB) BERBASIS FRAMEWORK CODEIGNITER	Sistem Informasi
60	SISTEM PENDUKUNG KEPUTUSAN PENENTUAN LOKASI PEMBUKAAN KANTOR BARU DENGAN METODE FUZZY TOPSIS (STUDI KASUS : PT BANK PEMBANGUNAN DAERAH JAWA TIMUR)	Sistem Informasi
61	CASE BASED REASONING MENENTUKAN KELOMPOK UKT (STUDI UNIVERSITAS SEMBILANBELAS NOVEMBER KOLAKA)	Sistem Informasi
62	ANALISIS KUALITAS LAYANAN WEBSITE PUSAT PENERIMAAN MAHASISWA BARU UNIVERSITAS AIRLANGGA BERDASARKAN PERSEPSI PENGGUNA MENGGUNAKAN METODE WEBQUAL 4.0 DAN IMPORTANCE PERFORMANCE ANALYSIS (IPA)	Sistem Informasi
63	PENERAPAN CLUSTERING K-MEANS PADA CUSTOMER SEGMENTATION BERBASIS RECENCY FREQUENCY MONETARY (RFM) (STUDI KASUS : PT. SINAR KENCANA INTERMODA SURABAYA)	Sistem Informasi
64	DESAIN ACADEMIC BUSINESS INTELLIGENCE UNTUK AKREDITASI: STUDI KASUS UNIVERSITAS TRUNOJOYO	Sistem Informasi

65	PENERAPAN ALGORITMA APRIORI UNTUK TRANSAKSI PENJUALAN OBAT PADA APOTEK AZKA	Sistem Informasi
66	APLIKASI SPK UNTUK REKOMENDASI SISTEM E-LEARNING MENGGUNAKAN ADAPTIVE INTERVAL TRIANGULAR FUZZY NUMBER	Sistem Informasi

Leave a Reply

*Your email address will not be published. Required fields are marked **

Comment

*Name **

*Email **

Website

Save my name, email, and website in this browser for the next time I comment.

Post Comment

KULTAS SAINS DAN TEKNOLOGI
ɔgram Studi S1 Matematika

mpus Merr C, Jl. Dr. Ir. H. Soekarno
ılyorejo Surabaya – 60115
p.(031)5936501, 5924617 Fax.(031)5936502
ıail: admin@fst.unair.ac.id

TEXT MINING PADA MEDIA SOSIAL TWITTER STUDI KASUS: MASA TENANG PILKADA DKI 2017 PUTARAN 2

Alfian Futuhul Hadi[1)], Dimas Bagus C. W.[2)], Moh. Hasan[3)]

[1)3)] *Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Jember*
Jln. Kalimantan 37, Jember 68121

[1)]afhadi@unej.ac.id

[3)]moh.hasan@unej.ac.id

[2)]*rakyatkecil99@gmail.com*

Abstract— **Sebanyak 20.000 data *tweet* diambil pada 15-19 April 2017 selama pelaksanan Pilkada DKI Putaran 2. Kemudian data tersebut direduksi dengan menggunakan langkah *preprocessing* serta menghapus data dengan nilai TD-TDF yang rendah. Kemudian sentimen diberikan kepada data dengan menghitung jumlah kata positif dan negatif yang telah definisikan oleh peneliti berdasarkan observasi terhadap beberapa sampel data yang diambil secara acak. Kami menemukan bahwa terdapat ledakan *tweet* bersentimen negatif pada hari kedua masa tenang. Sedangkan ledakan selanjutnya terjadi pada hari ketiga, namun pada sentimen positif. Temuan kami yang lain yaitu "ahok" selalu mendapatkan sentimen negatif lebih tinggi dan sentimen positif lebih rendah dari pada *tweet* "anies". Hasil temuan ini ternyata sejalan dengan hasil pemilihan, yaitu kemenangan pada pasangan Anies-Sandi. Selain itu pada metode-metode *unsupervised learning* kami menemukan bahwa metode k-*means* tidak dapat memberikan hasil yang merata pada setiap kelompoknya. Sebaliknya, hasil luaran dari pemodelan topik (*topic modeling – Latent Dirichlet Allocation*) lebih merata. Selain itu hasil pengelompokkan dari metode k-*means* dan *topic modeling* pada data tanggal 18 April 2017 memiliki nilai SSE (k-*means*) dan *marginal likelihood (topic model*) yang lebih baik dari pada data tanggal lainnya. Hal ini disebabkan oleh karena keragaman data yang terdapat pada tanggal 18 April 2017 tersebut lebih rendah.**

Keywords— ***text mining*, pemilihan gubernur jakarta, twitter, sentimen analisis, hasil pilkada**

I. PENDAHULUAN

Text Mining telah digunakan secara luas dalam menggali informasi yang tersembunyi pada data yang sangat banyak. Salah satu media yang menyimpan data teks yaitu berupa opini masyarakat terhadap suatu peristiwa adalah Twitter. Twitter merupakan salah satu media sosial yang paling sering digunakan. Bahkan pada tahun 2015 lalu Indonesia telah menjadi salah satu negara yang memiliki pengguna aktif Twitter terbesar didunia dengan jumlah *tweets* nya yang mencapai 500 juta tiap harinya.

Pilkada DKI merupakan salah satu pilkada yang mendapat sorotan di Indonesia. Hal ini disebabkan oleh posisi DKI Jakarta sebagai ibu kota negara yang menjadi contoh oleh daerah-daerah lain serta sebagai jantung perpolitikan Indonesia. Tak heran apabila prosesnya menjadi pembicaraan baik dilingkungan sekitar hingga di media sosial. Oleh karenanya media sosial populer seperti Twitter menjadi salah satu sarana bagi pasangan calon untuk meningkatkan popularitasnya.

Serunya proses Pilkada DKI serta perannya media sosial pada proses tersebut membuat peneliti tertarik untuk mengetahui apa saja yang dilakukan pengguna media sosial pada saat pilkada tersebut. Sehingga pada penelitian ini, peneliti menggunakan data *tweets* pada media Twitter terkait dengan Pilkada DKI Putaran 2 untuk menjadi objek penelitian dalam mengaplikasikan metode-metode dalam *text mining* seperti analisis sentimen, pengelompokkan dan visualisasi data.

Salah satu tujuan penelitian ini ialah untuk menggali informasi dari media sosial Twiter yang dapat digunakan sebagai gambaran tentang apa saja yang terjadi selama proses pilkada (khususnya pada masa tenang). Selain itu, peneliti menduga bahwa ada keterkaitan antara media sosial dengan hasil dari pencoblosan.

II. METODE PENELITIAN

A. Data Penelitian

Data yang digunakan dalam penelitian ini adalah data tweets dengan menggunakan kata kunci "ahok" dan "anies". pengambilan data dilakukan selama 5 hari yaitu mulai tanggal 15 April 2017 sampai dengan 19 April 2017. Secara keseluruhan jumlah data yang digunakan ditunjukkan oleh Tabel 1 berikut:

TABEL 1. JUMLAH DATA PENELITIAN

Tanggal	Kata Kunci	Jumlah data *Tweets*
15 April 2017	Anies	20.000
	Ahok	20.000
16 April 2017	Anies	20.000
	Ahok	20.000
17 April 2017	Anies	20.000
	Ahok	20.000
18 April 2017	Anies	20.000
	Ahok	20.000
19 April 2017	Anies	20.000
	Ahok	20.000

B. Text Preprocessing

Data teks yang diperoleh masih berupa data mentah yang memiliki banyak kata yang tidak lengkap, *noise* (simbol, *link*, *e-mail*, dsb.) ataupun data yang *inconsistent*. *Text preprocessing* bertujuan untuk membersihkan data-data yang tidak diperlukan serta menyeragamkan kata-kata yang memiliki arti sama agar proses *mining* lebih akurat. Sehingga tahap-tahap pada *text preprocessing* diperlukan agar data olah yang diperoleh menjadi lebih baik. Proses ini dapat dikatakan sebagai kunci kualitas data serta *output* yang dihasilkan nanti.

C. Pembobotan Term

Pembobotan term bertujuan untuk memberikan sebuah nilai pada sebuah term berdasarkan tingkat kepentingan term tersebut didalam sekumpulan dokumen masukan. Pada penelitian ini akan digunakan metode *Term Frequency – Inverse Term Frequence* (TF-IDF) sebagai proses pembobotan, yaitu dengan cara mencari representasi dari tiap-tiap dokumen dari sekumpulan data training dan akan dibentuk menjadi vektor. [13] merumuskannya sebagai berikut:

$$w(t,d)_{ij} = \frac{tf(t_i, d_j)}{\sum t_{ij}} \times idf \quad (1)$$

$$idf = \log_2\left(\frac{N}{df}\right) \quad (2)$$

dimana:

$tf(t,d)_{ij}$ = kemunculan kata t_i pada dokumen d_j
N = jumlah dokumen keseluruhan
df = jumlah dokumen yang memiliki kata t_i
$\sum t_{ij}$ = jumlah kata pada dokumen d_j

D. Supervised Learning

Supervised learning merupakan metode yang digunakan untuk memperoleh pembelajaran pada suatu data yang telah memiliki variabel masukan dan variabel luaran (label) untuk melakukan prediksi terhadap data yang hanya memiliki variabel masukan saja. Seperti yang dijelaskan oleh [5] bahwa pembelajaran yang diperoleh dari metode *supervised learning* nantinya akan digunakan untuk melakukan prediksi pada *unseen data*. Disamping itu *supervised learning* adalah metodologi yang paling penting pada mesin pembelajaran (*machine learning*) dan juga memiliki *central importance* dalam memproses data multimedia.

1. Pelabelan: Dokumen yang terdapat pada internet merupakan data yang tidak terlabeli (unsupervised data). Sehingga untuk dapat diproses dengan menggunakan supervised learning, diperlukan metode untuk melabeli data-data tersebut. Kendala yang muncul selanjutnya ialah terlalu banyaknyadata yang harus diberi pelabelan. Oleh karena itu [8] mencoba memberikan label pada dokumen dengan cara menghitung banyak teks yang terkandung pada setiap dokumen. Proses pelabelan dilakukan sebagai berikut [8]:

1. Tentukan kata-kata yang memiliki arti positif juga negatif
2. Hitung jumlah kata positif dan negatif pada dokumen
3. Jika jumlah kata positif > jumlah kata negatif, maka label sentimennya adalah positif (label 1)
4. Jika jumlah kata positif < jumlah kata negatif, maka label sentimennya adalah negatif (label -1)
5. Jika jumlah kata positif = jumlah kata negatif, maka skor sentimennya adalah netral (label 0)

2. Klasifikasi Naïve Bayes: Metode Klasifikasi Naïve Bayes merupakan salah satu supervised learning untuk memberikan klasifikasi terhadap dokumen teks [10]. Bayangkan bahwa suatu dokumen digambarkan dari angka kelas dokumen yang mana dapat dimodelkan sebagai himpunan dari kata-kata dimana peluang (independen) kata yang ke-i dari suatu dokumen muncul pada dokumen dari kelas C dapat ditulis sebagai:

$$p(w_i|C)$$

Maka probabilitas bahwa diberikan dokumen D mengandung semua kata w_i, dari kelas C adalah

$$p(D|C) = \prod_i p(w_i|C)$$

Dalam pengklasifikasian teks kita menandai (*tokenize*) dokumen untuk dapat melakukan klasifikasi *in its appropriate class*. Menggunakan aturan pengambilan keputusan dengan "*Max a Posterior Probability*" maka diperoleh pengklasifikasi sebagai berikut [10]:

$$c_{MAP} = \arg\max_{c \in C}\big(P(c|d)\big) = \arg\max_{c \in C}\left(P(c)\prod_{1 \le i \le n} P(w_i|c)\right) \quad (3)$$

dengan:

d = dokumen
w_i = kata-kata ke-*i* dalam dokumen

c = himpunan dari penggunaan kelas klasifikasi
$P(c|d)$ = *conditional probability* dari kelas c yang diberikan dokumen
$P(c)$ = probabilitas prior dari kelas c
$P(w_i|c)$ = *conditional probability* dari kataw_i diberikan kelas c

Dapat dilihat bahwa untuk dapat menyimpulkan dokumen tersebut masuk dalam kelas tertentu, harus melakukan estimasi *the product of the probability* dari setiap kata pada dokumen pada kelas khusus (*likelihood*) setelah itu kita kalikan lagi dengan peluang pada kelas khusus (*prior*).

Unsupervised Learning

Unsupervised learning merupakan mesin pembelajaran yang khusus digunakan pada data yang hanya memiliki variabel *output* saja. Mesin pembelajaran ini digunakan untuk menemukan pola tersembunyi pada data yang harapannya dapat digunakan untuk menggambarkan karakteristik data. Tidak ada kesimpulan benar ataupun salah pada hasil (*output*) dari *unsupervised learning*. Oleh karena itu, peneliti harus memahami kajian teori pada topik tertentu yang menjadi objek *unsupervised learning*. Sehingga dengan begitu, peneliti dapat memberikan gambaran yang ideal terhadap *output* dari metode ini.

1. K-Means: Metode klasifikasi k-means merupakan metode yang paling sering digunakan. Hal tersebut disebabkan karena metode ini adalah metode yang cukup sederhana, sehingga dapat dipahami dengan mudah. K-means bertujuan untuk mengelompokan data berdasarkan kemiripan data satu dengan yang lainnya. Berikut adalah algoritma dasar dari k-means[9]:

1. Pertama-tama tentukan titik k sebagai pusat kelompok yang mungkin terbentuk.
2. Menentukan secara acak posisi k yang akan menjadi *centroid*(C_i).
3. Menghitung jarak semua titik (data) ke k *centroids* tersebut lalu memilih jarak terdekat dari masing-masing titik ke k *centroid*. Sehingga setiap titik yang jaraknya paling dekat dengan *centroid* menjadi sebuah kelompok.
4. Menghitung ulang *centroid* dari masing-masing kelompok, hingga *centroid* tidak berubah dengan menggunakan persamaan berikut.

$$c_i = \frac{1}{m_i}\sum_{x \in C_i} x \tag{7}$$

dimana:
c_i = centroid baru
m_i = jumlah x yang masuk pada centroid C_i

Prosedur penentuan titik k dilakukan secara acak. Setelah itu, menghitung kembali jarak dari masing-masing titik ke k titik pusat (*centroid*) dengan menggunakan metode jarak Euclid.

$$d(x_i, x_{i+1}) = \sqrt{\sum_{i=1}^{n_m}(x_i - x_{i+1})^2} \tag{8}$$

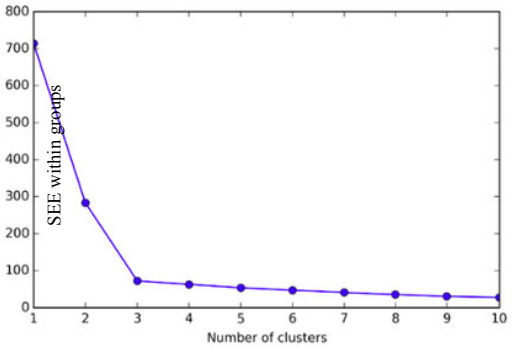
dimana:
x_i = nilai pengamatan obyek ke$-i$ dengan $i = 1, 2, \cdots, n_m$
x_{i+1} = nilai pengamatan obyek ke$-(i+1)$
n_m = banyak obyek pada klaster ke$-m$

setelah itu proses kembali ketahap 3 hingga nilai *centroid* tidak berubah.

Namun dalam menentukan k kluster haruslah optimal, karena k ditentukan secara manual sehingga sangat memungkinkan apabila k kluster yang dipilih terlalu sedikit ataupun terlalu banyak. Salah satu metode yang dapat digunakan untuk membantu menentukan jumlah k kluster adalah metode *elbow*. Metode ini bekerja dengan menghitung nilai SSE (*Sum Square of Error*) dari setiap kluster yang terbentuk sebagai berikut [1]:

$$\text{SSE} = \sum_{i=1}^{k}\sum_{x \in C_i} dist(c_i, x)^2 \tag{9}$$

dengan *dist* adalah fungsi untuk menghitung jarak dengan menggunakan metode jarak Euclid. Setelah itu nilai tersebut digambarkan pada grafik 2 dimensi (x, y) dimana x adalah jumlah kluster dan y adalah SSE masing-masing kluster. Berikut adalah contoh grafik dari metode *elbow*:



Gambar 2 Contoh Grafik *Elbow* Nilai SSE

Pada Gambar 2 diatas maka dapat diputuskan bahwa penggunaan 3 kluster sudah mencapai optimal. Hal tersebut terlihat dari nilai SSE yang mengalami penurunan drastis terjadi sampai pada k-means dengan k = 3 dan setelah itu penurun SSE yang terjadi tidak begitu drastis.

2. Topic Modeling (Latent Dirichlet Allocation): *Topic modeling* bertujuan untuk menemukan topik dari masing-masing dokumen. *Topic modeling* adalah metode statistik yang melakukan analisa kata-kata dari teks asli untuk menemukan tema yang terdapat pada teks tersebut, bagaiman tema tersebut berhubungan satu dengan lainnya dan

bagaimana mereka berubah setiap waktu. Salah satu metode *topic modeling* adalah dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). LDA mengasumsikan bahwa topik telah terspesifikasi sebelum setiap data terbentuk. Andaikan D adalah kumpulan dokumen d, proses pembentukan dokumen berdasarka LDA model adalah sebagai berikut[6]:

1. Topik telah terspesifikasi dan telah memiliki distribusi sebagai berikut:
   $\varphi^{(k)} \sim Dirichlet(\beta), \quad$ untuk $k = 1, \dots, K$
2. Selanjutnya untuk setiap dokumen, kita membentuk kata dengan 2 tahap:
   a. Pilih secara acak sebuah distribusi topik untuk dokumen *d*.
      $\theta_d \sim Dirichlet(\alpha), \quad d \in D$
   b. Setiap kata pada dokumen
      i. Pilih secara acak sebuah topik pada distribusi topik.
      $z_i \sim Discrete(\theta_d)$
      ii. Secara acak pilih sebuah kata dari topik yang berkoresponden pada berdistribusi kosa kata.
      $w_i \sim Discrete\left(\varphi^{(z_i)}\right)$

dimana:

K = angka dari topik laten pada kumpulan dokumen
$\varphi^{(k)}$ = distribusi probabilitas diskrit pada kosa kata yang merepresentasi distribusi topik ke-*k*
θ_d = distribusi dokumen ke-*d* dari topik yang tersedia
z_i = indeks topik pada kata ke-*i*
w_i = kata ke-*i*
α, β = *hyperparameters* untuk distribusi Dirichlet

Proses pembuatan dokumen diatas menghasilkan distribusi gabungan sebagai berikut :

$$p(\mathrm{w}, \mathrm{z}, \theta, \varphi | \alpha, \beta) = p(\varphi|\beta)p(\theta|\alpha)p(\mathrm{z}|\theta)p(\mathrm{w}|\varphi_z) \quad (1)$$

LDA tidak hanya digunakan untuk melakukan pendeteksian topik saja, namun LDA juga digunakan sebagai salah satu *tools* untuk melakukan analisis *Business Intelligence* pada Bank, yaitu untuk mengetahui hubungan antara kebijakan tertentu dengan trend yang dihasilkan [12]. Jumlah topik yang optimum pada model LDA dapat ditentukan dengan menghitung nilai *Harmonic Mean* dari *loglikehood* hasil iterasi pada masing-masing model [14].

3. Visualisasi Data (Word Cloud): Terdapat beberapa cara untuk memudahkan user dalam menyimpulkan hingga menggambarkan karakteristik hingga korelasi data. Diantaranya yang paling sederhana ialah memvisualisasikannya kedalam plot 2 dimensi ataupun 3 dimensi, seperti halnya biplot, boxplot, dan yang biasa digunakan dalam analisis teks atau text mining ialah word cloud. *Boostlabs.com* menjelaskan bahwa wordcloud (juga dikenal sebagai text cloud atau tag cloud) bekerja dengan cara yang sederhana. Kata-kata yang spesifik dimunculkan dengan kondisi semakin banyak teks yang muncul dalam suatu tambang data, maka kata tersebut semakin besar dan tebal. Gambar berikut ini adalah salah satu contoh dari *wordcloud*.

III. HASIL DAN PEMBAHASAN

Pemberian Sentimen

Pemberian sentimen dilakukan dengan membandingkan banyak kata positif dan negatif dalam data. Apabila jumlah kata positif pada suatu dokumen lebih banyak daripada kata negatifnya, maka dokumen tersebut bersentimen positif. Begitu pula sebaliknya. Oleh karena itu daftar kata-kata negatif dan positif menjadi kunci kesuksesan dalam memberikan sentimen pada data. Selama penelitian, peneliti menemukan beberapa kata-kata khusus yang digunakan pengguna Twitter untuk menunjukan keberpihakannya pada salah satu pasangan calon. Pada kata-kata khusus tersebut, peneliti memberikan skor 2 kali lipat daripada kata-kata bersentimen lainnya. Tabel 2 berikut adalah hasil pemberian sentimen pada dokumen *tweets*:

TABEL 2. HASIL PEMBERIAN SENTIMEN

Tanggal	Kata Kunci	Twitter		
		Negatif	netral	positif
15 April 2017	Anies	4.372	9.863	5.765
	Ahok	6.422	7.097	6.481
16 April 2017	Anies	6.937	6.649	6.414
	Ahok	7.504	5.887	6.609
17 April 2017	Anies	8.733	6.664	4.603
	Ahok	12.844	4.350	2.806
18 April 2017	Anies	1.843	3.576	14.581
	Ahok	3.712	4.279	12.009
19 April 2017	Anies	2.613	5.048	12.339
	Ahok	6.956	6.513	6.531

Analisis Sentimen Naive Bayes

Metode Klasifikasi Naive Bayes umunya digunakan untuk melakukan prediksi sentimen yang muncul pada data yang masih belum memiliki sentimen. Namun untuk dapat memberikan sentimen, metode ini terlebih dahulu memerlukan data awal (*data training*) yang digunakan untuk melakukan pembelajaran. Hasil pembelajaran tersebut lalu digunakan untuk melakukan prediksi sentimen yang muncul pada data. Peneliti telah menyiapkan 5 skenario yang digunakan untuk metode Naive Bayes sebagai berikut:

1. Skenario 1, melakukan pembelajaran dan prediksi pada data olah.

2. Skenario 2, pembelajaran dilakukan pada 33% dari data kemudian melakukan prediksi pada 33% data yang lain.
3. Skenario 3, pembelajaran dilakukan pada 33% dari jumlah data pada hari sebelumnya dan digunakan untuk menduga sentimen pada 33% data di hari setelahnya.
4. Skenario 4, pembelajaran dilakukan pada gabungan dari 2500 data pada tanggal 15–18 April 2017, untuk menduga data tanggal 19 April 2017
5. Skenario 5, pembelajaran dilakukan pada setiap hari untuk menduga sentimen yang muncul pada tanggal 19 April 2017

Rata-rata hasil dari pengujian metode Naive Bayes pada kelima skenario tersebut disajikan pada Tabel 2.

TABEL 3. KEAKURATAN METODE NAIVE BAYES

Skenario	Keakuratan
Skenario 1	81,84%
Skenario 2	79,62%
Skenario 3	48,47%
Skenario 4	46,45%
Skenario 5	42,78%

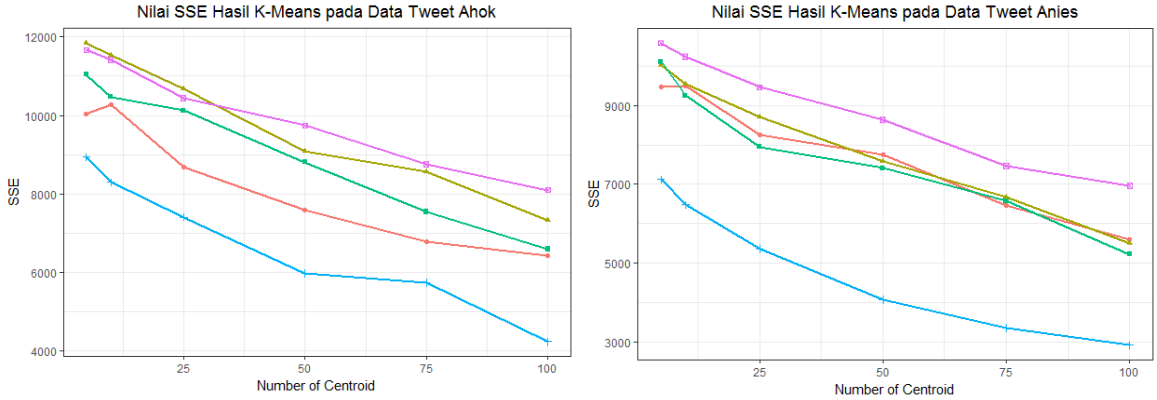
K-Means

Peneliti menentukan sejumlah *k centroid* yang digunakan untuk mengelompokkan data *tweets* yaitu 5, 10, 25, 50, 75 dan 100 *centroid*. Hasilnya adalah hampir secara keseluruhan data *tweets* tidak dapat dikelompokkan dengan baik oleh K-Means walaupun ada beberapa data yang telah memiliki *k centroid* yang optimum saat nilainya 75 seperti yang ditunjukkan oleh plot *elbow* nilai SSE oleh Gambar 4.

Hasil pengelompokkan K-Means selalu cenderung ada 1 kelompok yang memiliki anggota yang sangat dominan seperti yang ditunjukkan oleh Tabel 4 berikut:

TABEL 5. JUMLAH ANGGOTA KLUSTER SAAT 5 CENTROID

Kluster	Tweet
1	925
2	9.787
3	90
4	143
5	184



(a) (b)

Gambar 4. Plot Besar *Sum Square Errors* Masing-masing Kluster pada Data Ahok (a) dan Anies (b)

Topic Modeling

Jumlah topik yang ditentukan untuk pemodelan *Topic Modeling* sama seperti jumlah *centroid* pada K-Means yaitu 5, 10, 25, 50, 75, dan 100 topik. Selanjutnya penentuan topik optimun dilakukan dengan milihat nilai *harmonic mean* yang paling besar dari masing-masing topik. Agar lebih mudah, nilai *harmonic mean* di plot kedalam grafik 2 dimensi seperti yang ditunjukan oleh Gambar 5.

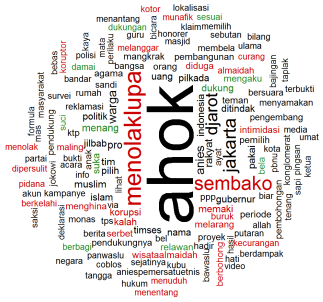
Hasilnya ialah rata-rata model optimum saat menggunakan 75 topik. Selain itu, *Topic Modeling* berhasil memberikan pengelompokan yang seragam pada topik yang terbentuk. Setiap kelompok yang dibentuk oleh *Topic Modeling* memiliki keanggotaan yang cukup merata seperti yang ditunjukan oleh Tabel 5.

TABEL 6. JUMLAH TWEET PADA MASING-MASING TOPIK

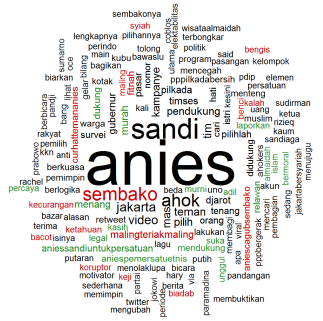
Topik	Tweet
1	2.124
2	1.621
3	1.855
4	1.125
5	1.285

Keterkaitan dengan Lembaga Survei dan Hasil KPU Pilkada DKI Putaran 2

Media sosial yang bersifat bebas telah menjadi ajang untuk saling mendukung atau bahkan menjatuhkan popularitas pasangan calon dengan menggunakan isu-isu ataupun kejadian-kejadian tertentu. Seperti yang terjadi pada Ahok dan Anies pada tanggal 17 April 2017 yang ditunjukkan oleh gambar berikut:



Gambar 6. *Word cloud* Data Ahok tanggal 17 April 2017

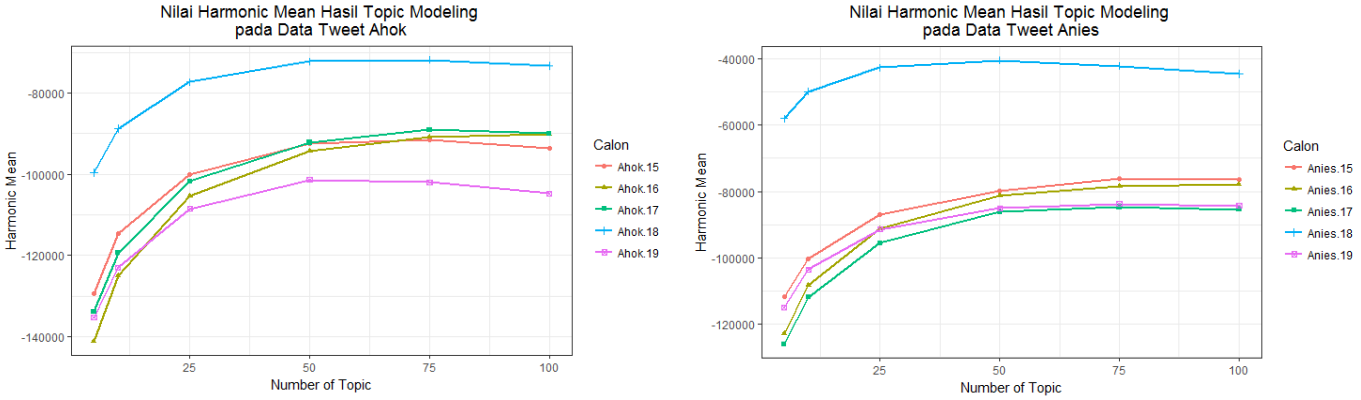


Gambar 7. *Word cloud* Data Anies tanggal 17 April 2017

Gambar 6 dan Gambar 7 menunjukan bahwa isu-isu negatif yang diberikan terhadap kedua pasangan calon. Namun frekuensi serta banyaknya sentimen negatif yang diterima oleh Ahok terlihat lebih banyak daripada yang diterima oleh Anis.

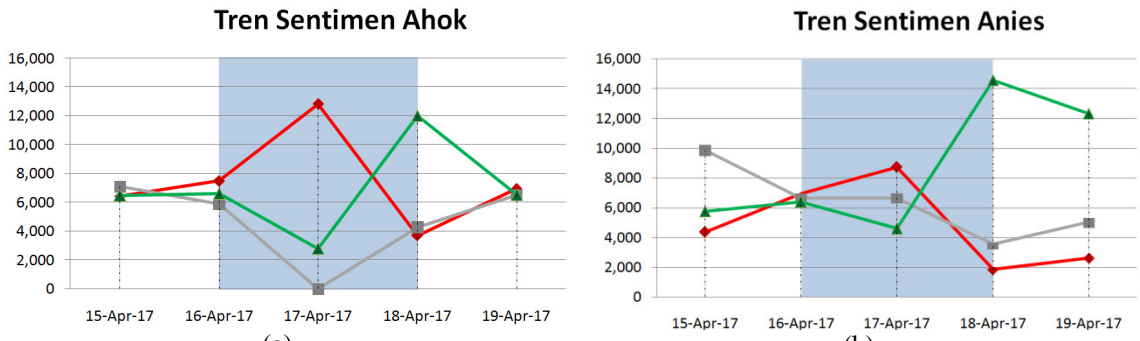
Namun yang terjadi pada hari selanjutnya justru berkebalikan. Tren sentimen positif dari kedua pasangan calon sama-sama mendominasi (Gambar 8) dan meningkat drastis (Gambar 9) daripada hari sebelumnya. Selain itu juga ditemukan bahwa pada tanggal 18 April 2017 tersebut masyarakat telah menentukan pasangan calon mana yang akan dipilihnya pada saat pencoblosan esok hari.

Hasil dari analisis sentimen yang dilakukan linier dengan hasil yang diperoleh dari beberapa lembaga survei yang telah mengumumkan hasil survei mereka. Gambar 10 berikut diperoleh dari Ridwan (2017) pada Pojoksatu.id menunjukan bahwa 6 dari 7 lembaga survei menyatakan Anies lebih unggul daripada Ahok. Begitu pula hasil sementara perhitungan yang dilakukan oleh KPU pada *website* resminya yang menunjukkan bahwa Anies memperoleh suara sebesar 57,95% sedangkan Ahok hanya berhasil mengumpulkan suara sebesar 42,05% (Kpu.go.id, 2017).



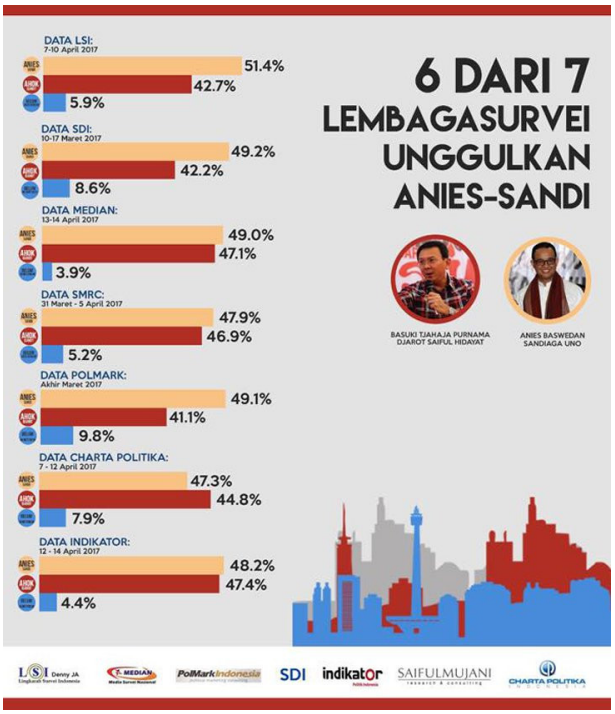
Gambar 5. Plot nilai *Marginal Likelihood* terhadap Jumlah Topik pada Data Ahok (a) dan Anies (b)

Gambar 8. *Word cloud* Data Ahok (a) dan Anies (b) tanggal 18 April 2017



(a) (b)

Gambar 9. Tren Sentimen Data Ahok (a) dan Anies (b)



Gambar 10. Hasil Survei 7 Lembaga Survei (Pojoksatu.id, 2017)

Dari hasil yang telah diperoleh dalam sentimen analisis dapat ditarik suatu keterkaitan dengan hasil Pilkada DKI yaitu,

1. Besarnya sentimen positif yang diperoleh Anies pada media sosial linier dengan hasil survei lembaga survei dan juga *quick count* yang dilakukan KPU yaitu lebih unggul daripada Ahok.
2. Sentimen negatif yang muncul pada pasangan Ahok-Djarot selalu konsisten lebih tinggi daripada sentimen negatif pasangan Anies-Sandi. Hal ini juga bisa jadi telah mempengaruhi keputusan masyarakat dalam menentukan pilihannya.
3. Sentimen netral yang diperoleh Anies tidak pernah lebih sedikit daripada sentimen negatifnya kecuali pada tanggal 17 April 2017. Pada tanggal tersebut baik pasangan Ahok-Djarot dan Anies-Sandi sama-sama mendapatkan sentimen negatif yang sangat tinggi daripada hari-hari yang lainnya.

IV. PENUTUP

Metode text mining dapat digunakan untuk menemukan informasi yang tersembunyi pada data. Selain itu dengan menggunakan metode yang tepat akan membantu peneliti dalam mengamati serta mengambil informasi yang sesuai dengan harapan peneliti. Berikut ini adalah beberapa kesimpulan yang didapat dari penelitian ini,

1. Metode Naive Bayes dalam melakukan prediksi sentimen pada data tweet “anies” dan “ahok” di hari yang sama memiliki tingkat keakuratan rata-rata sebesar 80%. Namun apabila hasil pembelajaran yang dilakukan hari ini digunakan untuk memprediksi hari esok ataupun hari lain, metode Naive Bayes mengalami penurunan keakuratan yang cukup besar yaitu hanya mampu memprediksi dengan keakuratan rata-rata 50% kebawah.

2. *Association Rules* dapat memberikan informasi yang lebih jelas dengan melihat besar nilai *Lift* serta *Confidence* pada suatu kata dengan kata-kata yang lain. Setiap kata yang memiliki besar lift yang lebih dari 1 terhadap kata lain mengindikasikan bahwa kedua kata tersebut saling mendukung kemunculannya. Dengan informasi tersebut, peneliti dapat lebih mudah menyimpulkan apa yang sedang menjadi pembahasan pada data tersebut.
3. Pengelompokan data berdasarkan metode K-Means cenderung memerlukan jumlah centroid yang besar. Namun besarnya jumlah centroid tersebut masih memiliki pola yang sama yaitu ada 1 kelompok yang memiliki anggota yang dominan daripada kelompok-kelompok yang lain.
4. *Topic Modeling* berhasil memberikan pengelompokan yang seragam pada topik yang terbentuk. Setiap kelompok yang dibentuk oleh *Topic Modeling* memiliki keanggotaan yang cukup merata. Selain itu, sebagaian besar data tweet berhasil dikelompokan secara optimum oleh *Topic Modeling* dengan menggunakan 75 topik. Dengan begitu peneliti lebih mudah dalam mengeksplorasi topik yang muncul dibandingkan dengan pengelompokan K-Means yang cenderung memilik 1 kelompok yang keanggotaannya mendominasi.
5. Visualisasi dengan menggunakan Word Cloud dapat digunakan untuk meringkas informasi pada data. Namun metode ini menjadi tidak efektif apabila tidak ada kata yang mendominasi pada data tweet.
6. Terdapat dugaan-dugaan bahwa tren yang terjadi pada media sosial Twitter memiliki keterkaitan atau bahkan mempengaruhi hasil pilkada. Hal ini ditunjukkan dari tren sentimen yang muncul pada Twitter linier dengan hasil pilkada, baik itu pada sentimen yang positif maupun sentimen yang negatif.

DAFTAR PUSTAKA

Bholowalia, P., dan A. Kumar. 2014. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN

Cunningham, P., Cord, M., Delany, S. J. 2008. Supervised Learning. Journal of Cognitive Technologies pp 21-49.

Darling, W. M. 2011. A Theoretical and Pratical Implementation Tutorial on Topic Modeling and Gibbs Sampling.

Khushboo, N., Swati , T., Vekariya, K., Shailendra, M. 2012. Mining of Sentences Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm. int. J. Computer Technology & Applications, Volume 3, pp. 987991

Kumar, V., Steinbach, M., Tan, P-N. 2006. Introduction to Data Mining.

Li, Y. H., Jain, A. K. 1998. Classification of Text Document. The Computer Journal, vol. 41, no. 8.

Moro, S., Cortez, P., Rita, P.2014. Business intelligence in banking: A literature analysis from 2002 to 2013. Elsevier.

Salton, G., Buckley, C. 1988. *Term-weighting approaches in automatic text retrieval*. *Information Processing and Management*, **24**/5, 513--523.

Steyvers, M., Griffiths, T. L. 2004. *Finding Scientific Topics*. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS).