Klasterisasi Tweet Terkait Dengan Pemilihan Presiden 2019 Menggunakan Ontology-based Concept Weighting dan DBSCAN

Puput Fajriati Tri Sholekah[1], Anisa Herdiani, S.T., M.T.[2], Ibnu Asror, S.T., M.T.[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]puputfajriati@students.telkomuniversity.ac.id,
[2]anisaherdiani@telkomuniversity.ac.id, [3]iasror@telkomuniversity.ac.id

Abstrak

Informasi yang berada di media sosial twitter berkembang sangat cepat, contohnya seperti *tweet* tentang pemilihan presiden yang berhubungan dengan kedua calon pasang presiden. Topik yang sedang dibicarakan oleh masyarakat mengenai pemilihan presiden di twitter sangat beragam, oleh karena itu diperlukan suatu sistem untuk mengelompokan *tweet* berdasarkan topik pembahasan mengenai pemilihan presiden yang berhubungan dengan kedua calon pasang presiden. Tujuan dilakukan penelitian adalah untuk mengetahui topik apa saja yang dibicarakan oleh masyarakat saat pemilihan presiden, sehingga diperlukan sebuah metode yang dapat mengelompokan *tweet* tersebut dan mengetahui akurasi perfomansi dari *ontology-based concept weighting* dan dbscan. Penelitian ini menggunakan metode *ontology-based concept weighting* yang digunakan untuk menghitung dan menerapkan pengetahuan tentang struktur hierarkis topik dan dbscan untuk mengelompokan *tweet* tersebut. Berdasarkan hasil pengujian, pengelompokan *tweet* menggunakan *ontology-based concept weighting* dan dbscan untuk data pasangan calon nomor urut 1 menghasilkan akurasi sebesar 26.5% dan data pasangan calon nomor urut 2 menghasilkan akurasi sebesar 44.16%.

Kata kunci: ontologi, pemilihan presiden, *tweet, cluster*isasi, dbscan.

Abstract

Information that is on Twitter social media is growing very fast, for example, like tweets about presidential elections related to the two presidential pairs. The topic being discussed by the public regarding the presidential election on Twitter is very diverse, therefore a system is needed to group tweets based on the topic of discussion about presidential elections relating to the two candidates for presidential pairs. The purpose of the research is to find out what topics are discussed by the public during the presidential election, so that a method is needed that can group these tweets and know the performance accuracy of ontology-based concept weighting and dbscan. This study uses ontology-based concept weighting methods that are used to calculate and apply knowledge of topic hierarchical structures and dbscan to group those tweets. Based on the results of the testing, the grouping of tweets using ontology-based concept weighting and dbscan for candidate pair number 1 data produced an accuracy of 26.5% and data on candidate pair number 2 produced an accuracy of 44.16%.

Keywords: ontology, presidential elections, tweet, clustering, dbscan.

1. Pendahuluan

1.1. Latar Belakang

Twitter merupakan salah satu media sosial yang sering digunakan oleh masyarakat Indonesia. Menurut Head of Business Development Twitter South East Asia and Australia, yang dinyatakan pada akun resmi twitter, Indonesia merupakan salah satu negara dengan jumlah pengguna Twitter terbesar di dunia [1]. Berdasarkan hasil survei tersebut, twitter merupakan suatu platform media sosial yang memiliki informasi beragam, salah satu contoh informasi beragam di twitter adalah *tweet* mengenai pemilihan presiden 2019 yang berhubungan dengan kedua calon pasangan presiden. *Tweet* yang beragam tersebut memiliki banyak topik pembahasan dan banyak yang tidak berkaitan dengan pemilihan presiden. Untuk menjadi sebuah topik pembahasan, *tweet* tersebut harus dikelompokkan agar dapat mengetahui topik apa saja yang berkaitan dengan calon pasangan presiden. Selain itu karena kebanyakan *user* twitter membuat *tweet* dengan bahasa yang tidak baku sehingga dapat menyebabkan data yang tidak beraturan terkait dengan pemilihan presiden 2019.

Berdasarkan masalah di atas, maka dibutuhkan sebuah sistem pengelompokan *tweet* untuk membedakan topik pembahasan mengenai pemilihan presiden yang terkait dengan kedua calon pasangan presiden. Sistem pengelompokan ini menggunakan metode dbscan yang dikombinasikan dengan *ontology-based concept weighting*. Topik pembahasan dalam penelitian ini ditandai dengan angka dan data *tweet* dikelompokkan berdasarkan kata yang sering muncul. Alasan menggunakan *ontology-based concept weighting* adalah untuk menerapkan pengetahuan tentang struktur hierarkis topik, jadi setiap topik yang berada pada tingkat hierarki yang sama, memiliki kesetaraan. Kelebihan dari *ontology* adalah membantu menemukan dokumen dengan kata-kata

secara sintaksis berbeda tetapi memiliki makna yang sama [2]. Sedangkan, alasan menggunakan algoritma dbscan karena berdasarkan bentuk data yang tidak beraturan algoritma ini dapat menghasilkan *cluster* yang lebih akurat dibandingkan dengan *partitional clustering*. Kelebihan dari algoritma dbscan adalah dapat digunakan untuk mengelompokan data dalam jumlah besar dan dapat menangani *noise/outlier* [3]. Setalah melakukan pengelompokan *tweet*, dilakukan perhitungan akurasi yang dinyatakan dalam bentuk *confusion matrix* dengan menggunakan evaluasi *F1 Score.*

1.2. Topik dan Batasan Masalah

Berdasarkan latas belakang, maka rumusan masalah dalam penelitian ini adalah mengelompokan *tweet* terkait pemilihan presiden yang berhubungan dengan calon pasangan presiden, serta mengukur tingkat akurasi pengelompokan *tweet* menggunakan *ontology-based concept weighting* dan dbscan dengan menggunakan *F1 Score*.

Pada penelitian ini memiliki batasan masalah yang digunakan untuk membatasi penelitian agar berjalan sesuai dengan yang diharapkan. Batasan masalah pada penelitian ini adalah *tweet* yang digunakan adalah *tweet* yang berbahasa Indonesia dan tidak menamai *cluster* yang dihasilkan oleh algoritma dbscan.

1.3. Tujuan

Tujuan dari penelitian ini adalah untuk mengelompokan *tweet* menggunakan *ontology-based concept weighting* dan dbscan yang dapat menjadikan ke dalam beberapa *cluster* sehingga menjadi sebuah informasi *tweet* tersebut termasuk ke dalam topik mana dan menunjukan performansi pengelompokan *tweet* menggunakan *ontology-based concept weighting* dan dbscan.

1.4. Organisasi Tulisan

Laporan penelitian ini terdiri dari 4 bagian yaitu bagian 2 menjelaskan mengenai teori-teori yang berhubungan dengan penelitian yang dilakukan. Bagian 3 menjelaskan perancangan sistem Klasterisasi *Tweet* Terkait Dengan Pemilihan Presiden 2019 Menggunakan *Ontology-based Concept Weighting* dan DBSCAN. Pada bagian 4 dari penelitian ini adalah bagian evaluasi yang menjelaskan tentang hasil pengujian dan analisis hasil pengujian yang telah dilakukan. Sedangkan bagian 5 menjelaskan mengenai kesimpulan dan saran untuk mengembangkan sistem ini.

2. Studi Terkait

2.1. Twitter

Twitter adalah layanan jejaringan sosial dan situs *microblogging* yang didirikan oleh Jack Dorsey pada tahun 2006 memungkinkan untuk pengguna mengirim teks dengan batasan 140 karakter saja akan tetapi pada tahun 2017 batasan tersebut ditambah hingga 280 karakter [4]. Twitter ini hanya dapat menulis *tweet* untuk pengguna twitter saja sedangkan yang bukan pengguna twitter hanya dapat membaca *tweet* saja. Twitter memiliki beberapa fitur diantaranya adalah fitur *reply tweet* yang digunakan untuk membalas *tweet* yang dibuat oleh pengguna lain fitur *retweet* yang berfungsi untuk membagikan *tweet* pengguna orang lain, fitur *like tweet* digunakan untuk menyimpan *tweet* pengguna lain yang disukai, fitur *following* berfungsi untuk mengikuti pengguna lain dan fitur *follower* yang digunakan untuk meliha pengguna lain yang mengikuti pengguna lain juga [5].

2.2. *Ontology*

Ontology adalah representasi formal dari suatu knowledge dan dapat mendeskripsikan sebuah domain dengan membaginya ke dalam beberapa konsep dan mendeskripsikan relasi yang ada [6]. Ontology dapat membantu menemukan dokumen dengan kata-kata secara sintaktis berbeda akan tetapi memiliki makna yang sama. Terdapat 4 komponen yaitu komponen *concepts* yang merupakan kumpulan entitas dalam domain, komponen *relations* adalah interaksi antara konsep atau properti konsep, komponen *instances* merupakan contoh konkret dari konsep dalam domain, dan komponen *axioms* adalah aturan eksplisit untuk membatasi konsep penggunaan [7]. Sedangkan untuk kontruksi dari ontologi terdapat 2 cara yaitu kontruksi *bottom-up* merupakan konstruksi yang mendefinisikan konsep dari yang lebih spesifik ke konsep yang paling umum, dan kontruksi *top-down* adalah konstruksi yang mendefinisikan konsep dari yang paling umum kemudian konsep selanjutnya yang lebih spesifik [7]. Pada penelitian ini akan menggunakan kontruksi *top-down*.

2.3. *DBSCAN*

Density-Based Spatial Clustering of Applications with Noise adalah sebuah algoritma yang menumbuhkan area-area yang memiliki kepadatan cukup tinggi ke dalam *cluster-cluster* dan menemukan *cluster-cluster* dalam bentuk sembarang dalam suatu database spatial yang memuat noise (Sander *et* al., 1998). Algoritma ini medefinisikasi *cluster* sebagai himpunan maksimum dari titik-titik kepadatan yang terkoneksi (*density-connected*) [8]. Semua objek yang tidak masuk ke dalam cluster manapun dianggap sebagai *noise*. Dbscan memerlukan 2 parameter yaitu MinPts digunakan untuk menentukan minimal banyak *items* dalam suatu *cluster* dan Eps sebagai

nilai yang digunakan untuk jarak antar items yang menjadi dasar pembentukan *neighborhood* dari suatu titik item. Kelebihan dari algoritma dbscan adalah [9]:

a. Dapat menghasilkan *cluster* yang lebih akurat dari bentuk data yang tidak beraturan jika dibandingkan dengan *partitional clustering*
b. Dapat menangani *noise/outlier*
c. Baik untuk data dalam jumlah besar

Kekurangan dari algoritma dbscan adalah [9]:

a. Kurang maksimal pada *dataset* berdimensi tinggi
b. Tidak dapat membentuk fungsi kepadatan yang sesungguhnya, akan tetapi lebih ke arah poin-poin kepadatan yang saling berhubungan dan membentuk *graf*
c. Memiliki permasalahan saat identifikasi *cluster* dari kepadatan yang bervariasi

2.4. *Ontology-based Concept Weighting*

Pada tahap ini dilakukan implementasikan penghitungan bobot berdasarkan ontologi. Saat merancang metode menghitung bobot, sistem ini memiliki beberapa asumsi diantaranya adalah [10]:

1. Lebih sering kata-kata muncul di dokumen, lebih mungkin itu adalah kata-kata yang khas;
2. Panjang kata-kata juga akan mempengaruhi pentingnya kata-kata. Rupanya, satu konsep dalam ontologi terkait dengan konsep lain dalam ontologi domain tersebut. Itu juga berarti bahwa hubungan antara dua konsep dapat ditentukan dengan menggunakan panjang dari dua jalur penghubung konsep ini (jarak topologi) dalam konsep kisi.
3. Jika probabilitas satu kata tinggi, maka kata itu akan mendapatkan berat tambaha n;
4. Satu kata mungkin dapat menjadi kata yang khas bahkan jika itu tidak muncul di dokumen.

Kombinasi yang lebih ketat di atas menggambarkan empat asumsi mengarah ke struktur pembobotan dengan aspek ontologi. Makalah ini mempertimbangkan frekuensi, panjang, area spesifik dan skor dari konsep ketika menghitung berat, menggunakan fungsi dengan nilai berat sebagai berikut [10]

$$W = len * Freq * Correlation\ Coeffient + P(Concept) \quad (2.1)$$

Keterangan:

W = bobot dari kata kunci
Len = kedalaman konsep pada ontologi
Freq = berapa kali kata kunci muncul

Jika konsep terdapat di dalam ontologi maka korelasi koefisien = 1 dan jika tidak maka korelasi koefisien = 0 sedangkan untuk *probability* didasarkan pada kemungkinan konsep dalam dokumen. Berikut adalah fungsi untuk menghitung *probability* [10]:

$$P(concept) = \frac{jumlah\ kejadian\ dari\ konsep}{jumlah\ kejadian\ dari\ seluruh\ konsep} \quad (2.2)$$

Hasil dari langkah ini akan digunakan untuk mengelompokan dengan menggunakan algoritma Dbscan

2.5. *Evaluasi*

Tahap evaluasi bertujuan untuk mengukur performansi dari sistem yang telah dibangun. Evaluasi akan dilakukan menggunakan tabel yang berisi prediksi yang disebut *confusion matrix*. Pada evaluasi ini terdapat 4 kategori, yaitu [11]:

- *True Positive* (TP) yaitu hasil yang positif dan dilabeli positif.
- *False Negative* (FN) yaitu hasil yang negatif dan dilabeli negatif.
- *False Positive* (FP) yaitu hasil yang negatif tetapi dilabeli positif.
- *True Negative* (TN) yaitu hasil yang positif tetapi dilabeli negatif.

Dalam *Precision and Recall* terdapat 2 definisi *confusion matrix* yaitu *Precision* pada sumbu y dan *Recall* pada sumbu x, berikut adalah rinciannya :

- *Precision* digunakan untuk mengukur bagian-bagian kecil dari contoh-contoh yang diklasifikasikan sebagai positif dan mendapatkan label yang positif. *Precision* didapatkan dengan rumus sebagai berikut [11]:

$$Precision = \frac{TP}{TP + FP} \quad (2.3)$$

- *Recall* digunakan untuk mengukur bagian-bagian kecil dari contoh-contoh yang diklasifikasikan sebagai positif dan mendapatkan label yang benar. *Recall* didapatkan dengan rumus sebagai berikut [11]:

$$Recall = \frac{TP}{TP + FN} \quad (2.4)$$

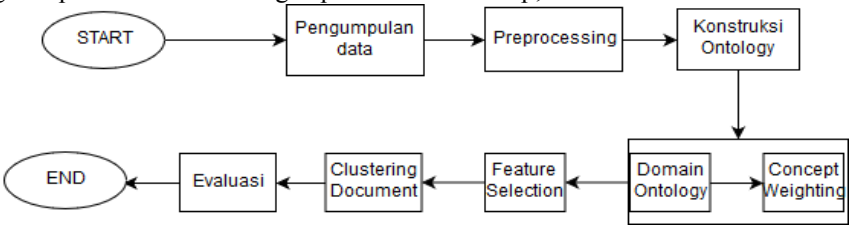
Setelah didapatkan nilai *precision* dan *recall*, maka dapat dihitung nilai kombinasi dari *precision* dan

recall menggunakan rumus F1 *Score* sebagai berikut [11]:

$$F1\ Score = \frac{2(Precision\ \times Recall)}{Precision + Recall} \quad (2.5)$$

3. Metodologi

Dalam tugas akhir ini dibangun sebuah sistem perangkat lunak yang digunakan untuk mengelompokkan *tweet* terkait tentang pemilihan presiden 2019 yang berhubungan dengan kedua calon pasangan presiden. Pada sistem ini hanya mengelompokkan tweet berbahasa Indonesia. Pengelompokan tweet ini tidak menggunakan label sehingga tergolong *unsupervised learning*. Sistem pengelompokan *tweet* ini mengunakan algoritma dbscan dan ontologi yang digabungkan dengan pembobotan konsep. Pada sistem ini terdapat 7 langkah utama yaitu pengumpulan data, *preprocessing*, kontruksi *ontology*, menghitung pembobotan konsep berdasarkan ontologi, *feature selection*, pengelompokan dokumen dengan pembobotan konsep, dan evaluasi.



Gambar 3 1 Gambaran Umum Alur Kerja Sistem

Berikut adalah penjelasan dari Gambar 3.1:

3.1. Pengumpulan Data

Pengumpulan data pada twitter ini dilakukan dengan menggunakan fungsi *crawling search* dalam format API Twitter untuk mendukung pengambilan data. API ini memiliki beberapa parameter, salah satunya adalah *query* yang digunakan untuk mencari *tweet* berdasarkan *query*. Berikut adalah beberapa *query* yang digunakan untuk mengumpulkan data dari twitter:

Tabel 3 1 *Query* Pengumpulan Data

Keyword			
Ekonomi	Infrastruktur	Politik	Ham
Korupsi	Pangan	Pendidikan	Hubungan internasional
Terorisme	Sumber daya alam	Sosial	ketenagakerjaan
Ideology	Hukum	Energy	Pemilihan presiden
Jokowi	Maruf amin	Sandiaga uno	Prabowo

Query yang disebutkan di atas dapat mengumpulkan data yang diharapkan. Data yang dikumpulkan adalah sekitar 600 *tweet* yang berhubungan dengan *query* tersebut dan menyaring *tweet* dari hasil *retweet* yang menyebabkan duplikasi *tweet*. Pengumpulan data ini menggunakan bahasa pemrograman python.

3.2. *Preprocessing*

Pada tahapan ini dilakukan preprocessing yang bertujuan mengubah data lebih terstruktur. Pada *preprocessing* ini dilakukan 3 proses yaitu *case folding, tokenizing,* dan *stemming*.

a. Tahap *case folding* ini dilakukan pengubahan semua karakter menjadi huruf kecil dan menghilangkan angka, tanda baca, dan URL yang dianggap tidak valid. Contoh cara kerja *case folding*:
   Sebelumnya: @jokowi pembangunan jalan tol ditargetkan mencapai 1.851 km di tahun 2019. Semangat Pak Jokowi!!
   Sesudahnya: jokowi pembangunan jalan tol ditargetkan mencapai km di tahun semangat pak Jokowi

b. Tahap *tokenizing* ini dilakukan pemotong kalimat yang diubah menjadi kata yang menyusunnya. Contoh cara kerja *tokenizing*:
   Sebelumnya: @jokowi pembangunan jalan tol ditargetkan mencapai 1.851 km di tahun 2019. Semangat Pak Jokowi!!

Sesudahnya:			
	jokowi	ditargetkan	tahun
	pembangunan	mencapai	semangat
	jalan	km	pak
	tol	di	Jokowi

c. *Stemming* adalah proses yang digunakan untuk mengembalikan kata yang berimbuhan ke kata dasarnya. Contoh cara kerja *stemming*:

Sebelumnya:	pembangunan	km	Sesudahnya:	bangun	km
	jalan	tahun		jalan	tahun
	tol	semangat		tol	semangat
	ditargetkan	mencapai		target	capai

3.3. Pembangunan *Ontology*

Pada tahapan ini dilakukan untuk membangunan *ontology*. Pembangunan *ontology* ini mengacu pada penelitian "*Ontology Development 101: Guide to Creating Your First Ontology*" oleh Natalya F. Noy dan Deborah L McGuinness [12]. Berikut adalah tahapan yang dilakukan:

a. Penentuan domain dan *scope* dalam *ontology*
   Dalam penelitian ini, *ontology* yang dibuat memiliki domain yaitu pemilihan presiden dengan memiliki 15 *class* yaitu hubungan internasional, energy, pendidikan, terorisme, sumber daya alam, sosial, politik, pangan, korupsi, ketenagakerjaan, infrastruktur, hukum, ideologi, ham, dan ekonomi [1].
b. Pertimbangan *ontology* yang sudah ada
   Setelah menentukan domain, proses selanjutnya adalah mempertimbangkan *ontology* yang sudah ada sebelumnya. Dari pencarian yang sudah dilakukan, tidak ditemukan *ontology* yang bisa digunakan kembali karena tidak ditemukannya penelitian sejenis.
c. Penulisan istilah-istilah penting yang berhubungan dengan domain pada *ontology*
   Term tweet yang telah melewati *preprocessing* dilakukan pendefinisian term apa saja pada suatu *tweet* yang terkait dengan pemilihan presiden. Contoh *tweet* hasil penyaringan:

   Keuangan meroket sangat tinggi

   Dari kalimat di atas, kemudian dilakukan analisis pada setiap kata yang terdapat pada *tweet*. Pada penelitian ini mengambil setiap nama dari setiap calon dan mengambil kata dari contoh *tweet* di atas yang dapat dilihat pada tabel 3.2

   Tabel 3 2 Contoh cara pengambilan kata penting

Kata	Diambil
Keuangan	Ya
Roket	Tidak
tinggi	Tidak

   Setiap kata penting di atas yang terdapat pada *tweet* akan dianalisis apakah setiap term memiliki kaitan dengan informasi pemilu yang menyangkut ke kelas yang sudah didefinisikan.
d. Pendefinisian kelas dan hierarki yang akan digunakan
   Pembuatan kelas hierarki dalam penelitian ini menggunakan pendekatan *top-down*. Proses pengembangan *top down* ini dimulai dengan mendefinisikan konsep umum dalam domain kemudian dilanjutkan dengan konsep yang lebih spesifik. Pada tahapan sebelumnya, dilakukan pemilihan kata penting kemudian kata tersebut dikelompokan ke dalam beberapa kelas.
e. Pendefinisian relasi dan hubungan antar domain
   Pendefinisian relasi ini dilakukan untuk mendefinisikan relasi, mempresentasikan relasi atau hubungan antar domain atau konsep. Pada penelitian ini tidak dilakukan pendefinisian relasi karena system yang dibangun hanya mendeteksi kemunculan dari setiap kata pada konsep.
f. Pembuatan *instance*.
   Instance adalah sebuah objek yang dibuat oleh sebuat *class*. Dalam penelitian ini, tidak mendefinisikan *instance* hanya mendefinisikan konsep.

3.4. *Ontology-based Concept Weighting*

Pada tahap ini dilakukan penghitungan bobot konsep berdasarkan *ontology*. Berikut adalah contoh data dokumen yang akan digunakan:

Dokumen 1 = Macan mati meninggalkan belang, Jokowi turun tahta meninggalkan janji dan hutang

Dokumen di atas dihitung berdasarkan rumus yang sudah dijelaskan pada bab 2 dengan *equaption* 2.1 dan *equaption* 2.2 maka akan dihasilkan bobot seperti yang tertera di tabel 3.3.

[1] https://www.jawapos.com/nasional/pemilihan/19/12/2018/ini-lima-tema-debat-capres-cawapres-pilpres-2019/

Tabel 3 3 Menghitung pembobotan konsep

nama *feature*	*depth*	*freq*	*correlation coeffisien*	*probability of concept*	bobot
macan	0	1	0	0.1	0.1
mati	0	1	0	0.1	0.1
tinggal	0	2	0	0.2	0.2
belang	0	1	0	0.1	0.1
jokowi	4	1	1	0.1	4.1
turun	0	1	0	0.1	0.1
tahta	0	1	0	0.1	0.1
janji	5	1	1	0.1	5.1
hutang	5	1	1	0.1	5.1

Setelah melakukan penghitungan bobot tersebut, didapatkan vector konsep yang akan digunakan pada tahap *feature selection*.

3.5. *Feature selection*

Pada tahap ini dilakukan tahap pemilihan *feature* yang akan digunakan. Tahapan ini digunakan untuk mengurangi dimensi dan memilih *feature* berdasarkan bobot dari perhitungan *ontology-based concept weighting*. Proses yang dilakukan pada tahapan ini menggunakan *variance threshold* untuk memilih *feature* berdasarkan bobot. Berikut adalah perhitungan *feature selection* menggunakan *variance threshold* dengan *threshold* 0.3 yang dijelaskan pada tabel 3.4:

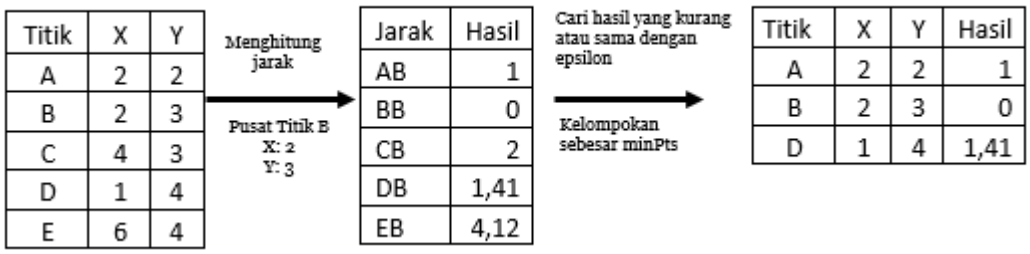
Tabel 3 4 Perhitungan *Feature Selection*

dokumen*feature*	A	B	C	D
1	0	2	0	3
2	0	1	4	3
3	0	1	1	3
variance	0	0,333333	4,333333	0

Dari tabel di atas, maka *feature* yang dipilih adalah *feature* B yang memiliki *variance* 0.3 dan C yang memiliki *variance* 0.4. Nilai *variance* tersebut lebih dari *threshold* yang telah ditentukan yaitu 0.3.

3.6. Algoritma Dbscan

Pada tahap ini dilakukan pengelompokan *tweet* menggunakan dbscan berdasarkan data yang sudah diproses melalui tahap *ontology-based concept weighting*. Berikut adalah contoh penghitungan dbscan dengan parameter epsilon 4 dan minPts 3 pada iterasi I:



Gambar 3 2 Perhitungan membentuk *cluster* pada Dbscan

4. Evaluasi

Setelah sistem selesai dibuat, maka langkah selanjutnya adalah pengujian terhadap sistem yang sudah selesai dibuat. Pengujian ini dilakukan untuk menguji seberapa baik algoritma ini mengelompokan *tweet* dan mengetahui hasil performansi dari sistem. Skenario pengujian adalah mengelompokan dataset ke beberapa *cluster* dan membandingkan hasil pengelompokan dari sistem dengan hasil pengelompokan yang telah dilakukan oleh pakar. Pengelompokan ini menggunakan data pasangan calon nomor urut 1 dan pasangan calon nomor urut 2.

4.1. Hasil Pengujian

Skenario pengujian adalah mengelompokan dataset ke beberapa *cluster* dan membandingkan hasil pengelompokan dari sistem dan hasil pengelompokan dari pakar. Pengelompokan ini menggunakan data pasangan calon nomor urut 1 dan pasangan calon nomor urut 2. Sedangkan untuk parameter yang digunakan pada dbscan untuk pasangan calon nomor urut 1 menggunakan epsilon 2.1 dan minPts 4 sedangkan untuk pasangan calon nomor urut 2 menggunakan epsilon menggunakan epsilon 2.0 dan minPts 3. Penentuan epsilon dan minPts di atas karena menyesuaikan dengan jumlah kelompok yang telah ditentukan oleh pakar. Pengujian ini menggunakan *F1-score.* Berikut adalah tabel 4.1 yang berisi hasil pengujian:

Tabel 4 1 Hasil Pengujian Capaslon 1

Data Pasangan Calon Nomor Urut	Precision	Recall	F-1 Measure	Akurasi
1	52.98%	26.50%	29.90%	26.50%
2	64.72%	44.16%	50.94%	44.16%

Hasil dari pengujian ini menghasilkan akurasi untuk data calon pasangan nomor urut 1 sebesar 26.50% dan nilai akurasi untuk data calon pasangan nomor urut 2 sebesar 44.16%.

4.2. Analisis Hasil Pengujian

Berdasarkan hasil pengujian di atas, diperoleh performansi antar kedua data tidak jauh berbeda yaitu 26.5% dan 44.16%. Hal ini disebabkan karena pembentukan pola pada dbscan, jika memiliki pola yang sama maka dijadikan satu *cluster*. Jika tidak maka akan membuat satu *cluster* yang berbeda. Berikut adalah contoh *tweet* yang memiliki pola berbeda:

Tabel 4 2 Contoh pembentukan pola

No	Kata	*Cluster*
1	bilang kalau gak dukung jokowi jangan mudik via jalan tol buat jokowi mbok sis ingetin presiden jokowi jangan terbang pake pesawat ga presiden itu era sby wong sama pake duit pajak rakyat kog biaya kog sama buzzer sederhana penting pilih presiden	0
2	gara atur jokowi menhub harga tiket pesawat mahal n hapus atur bagasi harus saing gara jokowi jeblog pilih presiden	1
3	masayarakat kecewa harga tiket pesawat mahal apa kata menhub	1
4	wooi pak jokowi tarif pswt naik bagasi bayar diam saja menteri budikaryas goblok banget sih gak punya otak menteri kok sumpah suara gerus kalah pilih presiden gara harga tiket pswt mahal ini bisa saya tuju ganti presiden kalau gin sih	2

Setelah dilakukan *feature selectio*n, *feature* yang terpilih adalah pada kata ‘jalan tol’ dan kata ‘pesawat’. Pada contoh nomor 1 terdapat kata ‘jalan tol’ dan kata ‘pesawat’ yang merupakan kata yang terdapat dalam *ontology*, akan tetapi memiliki *cluster* yang berbeda. Setelah dilakukan *feature selection*, *tweet* nomor 1 memiliki 2 nilai yang besar yaitu pada kata ‘jalan tol’ dan kata ‘pesawat’. Sedangkan *tweet* 2 dan *tweet* 3 hanya memiliki nilai yang besar pada kata ‘pesawat’, maka dibentuk cluster yang berbeda dengan *tweet* 1. Selain itu pada *tweet* nomor 4 memiliki *cluster* yang berbeda karena setelah dilakukan *feature selection*, *tweet* nomor 4 tidak memiliki nilai pada kedua kata tersebut. Sehingga akan dibentuk sebuah *cluster* baru.

5. Kesimpulan

5.1. Kesimpulan

Berdasarkan pengujian dan analisis yang dilakukan pada tugas akhir ini, maka dapat diambil beberapa kesimpulan:

a. Algoritma dbscan dan *ontology-based concept weigthing* tidak dapat mengelompokan *tweet* yang terkait dengan pemilihan presiden 2019**.**
b. Berdasarkan penelitian ini dihasilkan nilai akurasi untuk data calon pasangan nomor urut 1 sebesar 26.5% dan data calon pasangan nomor urut 2 sebesar 44.16%.

5.2. Saran

Saran untuk penelitian selanjutnya:

a. Data yang digunakan lebih baik seperti data berita yang menggunakan Bahasa yang lebih baku dibandingkan data di twitter yang banyak menggunakan Bahasa kurang baku agar dapat menghindari singkatan dan kesalahan dalam pengetikan kata.
b. Menanggulangi kata yang disingkat dapat menggunakan normalisasi dengan *levenshtein distance*.

Daftar Pustaka

[1] Herman, "Indonesia Masuk Lima Besar Pengguna Twitter," Berita Satu, 3 Mei 2017. [Online]. Available: https://www.beritasatu.com/iptek/428591/indonesia-masuk-lima-besar-pengguna-twitter. [Accessed 20 Juni 2019].

[2] Ontotext, "What are Ontologies?," Ontotext, [Online]. Available: https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies . [Accessed 20 Mei 2019].

[3] K. Rahmat, B. Aryo Putro and Shaufiah, "Implementasi Density based Clustering Application with Noise (DBSCAN) dalam perkiraan Terjadinya Banjir di Bandung," 2011.

[4] Y. J, "The Top 3 Twitter Shareholders (TWTR)," Investopedia, 30 Juli 2018. [Online]. Available: https://www.investopedia.com/articles/insights/060916/top-3-twitter-shareholders-twtr.asp. [Accessed 24 Oktober 2018].

[5] Suryadi, "5 Fitur Utama Twitter yang Bikin Kamu Betah!," Rocket Manajemen, 21 Agustus 2018. [Online]. Available: https://rocketmanajemen.com/fitur-dasar-twitter/#a. [Accessed 23 Oktober 2018].

[6] T. N. Ravishankar and S.R, "Ontology based Clustering Algorithm for Information Retrieval," 2013.

[7] A. T. and A. R, "ONTOLOGY AND ONTOLOGY CONSTRUCTION: BACKGROUND AND PRACTICES," 2012.

[8] I. M. Suwija Putra, "ALGORITMA DBSCAN (DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE) DAN CONTOH PERHITUNGANNYA," 2018.

[9] N. M. A. S. Devi, I. K. G. D. Putra and I. M. Sukarsa, " N. M. A. Santika Devi, I. K. G. Darma Putra dan I. M. Sukarsa, "Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan," *Lontar Komputer,* vol. 03, pp. 655-661, 2015 .

[10] H. T. Hmway and T. T. Soe Nyaunt, "Ontology-based Concept Weighting for Text Documents," *International Journal of Computer and Information Engineering,* vol. 5, p. 9, 2011.

[11] J. D. a. M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," *Appearing in Proceedings of the 23 rd International Conference on Machine Learning,* pp. 233-240 , 2006. .

[12] F. N. N. a. L. M. Deborah, Ontology Development 101: A Guide to Creating Your First Ontology,, Stanford University, CA, 2001.

[13] M. T. F. d. L. Muflikhah, "Clustering The Potential Risk of Tsunami Using Density-Based Spatial Clustering of Applications with Noise (DBSCAN)," *Environmental Engineering & Sustainable Technology,* vol. 03, no. 01, pp. 1-8, 2016.

[14] Twitter, "Twitter," 21 Maret 2006. [Online]. Available: http://twitter.com. [Accessed 24 Oktober 2018].

[15] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," *Conference on Computational Intelligence and Multimedia Applications,* vol. 2, 2007.

[16] J. Han and M. Kamber, Data Mining Concept and Techniques Second Edition, Burlington: Morgan Kaufman Publishers, 2006.

[17] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," *Appearing in Proceedings of the 23 rd International Conference on Machine Learning,* pp. 233-240, 2006.

[18] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," *Appearing in Proceedings of the 23 rd International Conference on Machine Learning,* pp. 233-240, 2006.