Identifikasi Spam Tweet Komentar Pada Twitter Berbasis Ontologi (Studi kasus : Tweet / caption di twitter dengan tema "Pilpres 2019")

Jahtra Genio Muhammad[1], Anisa Herdiani, S.T., M.T.[2], Indra Lukmana Sardi, S.T., M.T.[3]

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

Abstrak

Dengan diadakannya pemilihan presiden 2019 banyak media sosial yang mengangkat berita tersebut, sejalan dengan hal itu tentu banyak postingan yang membahas berita pemilihan presiden 2019. Hal tersebut menyebabkan aktivitas sebuah media sosial menjadi cukup tinggi. Twitter merupakan salah satu media sosial yang sangat populer digunakan untuk menyampaikan pendapat untuk saat ini, adanya pemilihan presiden 2019 membuat twitter menjadi salah satu media sosial yang ramai. Tingginya aktivitas twitter dimanfatkan oleh *spammers* untuk menyebarkan *spam* khususnya *spam* pada kolom komentar yang tidak memiliki keterkaitan dengan *tweet* atau topik pembicaraan serta menimbulkan dampak yang tidak nyaman bagi pengguna lainnnya. Spam yang dimaksud pada penelitian ini adalah komentar yang tidak ada keterkaitan atau keterhubungan dengan caption atau postingan. Untuk mengatasi masalah tersebut dibutuhkan sebuah sistem yang bisa mendeteksi spam berbasis ontologi. Dengan menggunakan ontologi, proses identifikasi spam menjadi lebih efisien dan sederhana karena data dipisahkan berdasarkan domain tertentu yang didefinisikan. Berdasarkan hasil pengujian, proses identifikasi spam menggunakan metode ontologi pada penelitian ini menghasilkan nilai rata-rata *f1-score* sebesar 89.14% Hal ini menunjukan bahwa ontologi dapat diimplementasikan untuk mengidentifikasi spam komentar pada twitter.

Kata kunci : pemilihan presiden *2019, spam,* spammers*, ontologi, twitter*

Abstract

With the 2019 presidential election held a lot of social media that raised the news, many posts discuss the 2019 presidential election news. This causes social media activity to be quite high. Twitter is a one of very popular social media used to express opinions for the moment, The 2019 presidential election makes Twitter one of the popular social media. The high level of twitter activity is used by spammers to spread spam, especially spam, in the comments column that has no connection with tweets or topics of conversation and has an uncomfortable impact on other users. Spam generally in this research is illustrated by comments that have no relevance or connectedness with the caption or post. For solve this problem, a system that can detect ontology-based spam is needed. By using an ontology, the process of identifying spam becomes more efficient and simpler because data is separated based on certain domains defined. Based on the results of testing, the process of identifying spam using the ontology method in this study resulted in an average f1-score of 89.14%. This shows that ontology can be implemented to identify comment spam on twitter.

Keywords: ***presidential election, spam, spammers, ontologi, twitter***

1. Pendahuluan

Latar Belakang

Dilihat dari meningkatnya penggunaan twitter di Indonesia pada tahun 2018 yang sudah mencapai angka lebih dari 20 juta [1] pengguna menandakan bahwa twitter telah menjadi sebuah wadah bagi orang-orang untuk berkomunikasi, dari fakta tersebut menunjukan tingginya popularitas pengguna twitter di Indonesia. Salah satu topik yang menarik yaitu tentang pemilihan presiden, pemilihan presiden merupakan sebuah pemilihan umum kepala negara yang berjalan setiap lima tahun sekali [2].

Seiring dengan diadakannya pemilihan presiden pada tahun 2019, banyak media sosial yang ikut membicarakan berita tentang hal tersebut. Oleh karena itu mulai banyak pula muncul pendapat dan pembicaraan masyarakat Indonesia mengenai pemilihan presiden 2019 khususnya terhadap calon pasangan presiden yang ikut serta dalam pemilihan presiden 2019.

Twitter menjadi salah satu media sosial yang ikut ramai aktivitasnya ketika masa pemilihan presiden 2019 [3]. Hal tersebut karena banyak pengguna yang ingin mengikuti berita dan pembicaraan mengenai pemilihan presiden 2019 [4]. Adanya aktivitas tersebut juga bisa mengundang *spammers* yang menyebarkan spam untuk tujuan tertentu yang memiliki dampak yang mengganggu bagi pemilik akun atau pengguna lain. Spam secara

umum digambarkan sebagai pesan yang memiliki konten yang tidak sesuai dengan *caption* atau topik pembahasan dan dikirimkan ke sejumlah orang [5].

Topik Masalah

Sejalan dengan adanya kegiatan tersebut, tidak sedikit pengguna twitter ikut berpendapat tentang pemilihan presiden 2019, hal itu menyebabkan banyak pengguna yang ingin mengikuti berita mengenai pemilihan presiden 2019 yang menjadikan aktivitas pada twitter menjadi cukup tinggi [6]. Adanya aktivitas tersebut dimanfaatkan oleh para *spammers* yang menyebarkan spam untuk tujuan tertentu yang memiliki dampak yang mengganggu bagi pemilik akun atau pengguna lain yang melihat keadaan tersebut [7].

Masalah tersebut menjadi dasar penelitian ini, pada penelitian ini akan dirancang sebuah alat pendeteksi *spam* yang dapat membedakan komentar *tweet reply* yang tidak memiliki keterhubungan dengan *tweet* atau topik pembahasan dengan memanfaatkan ontologi. Metode ontologi dipilih karena dapat menangani klasifikasi ke dalam beberapa kelas yang didefinisikan secara dinamis [8].

Tujuan dan Batasannya

Tujuan pada penelitian ini berdasarkan latarbelakang yang telah dijelaskan diantaranya:

1. Mengaplikasikan ontologi untuk mengidentifikasi *tweet* yang dikategorikan sebagai spam pada kolom komentar twitter
2. Mengukur hasil performansi dari identifikasi spam menggunakan ontologi

Adapun batasan masalah pada penelitian ini diantaranya :

1. Spam yang diidentifikasi hanya pada bagian *reply*
2. Tidak ada keterkaitan dengan *reply* yang berbentuk gambar
3. *Tweet* yang diidentifikasi hanya *tweet* yang berkaitan dengan visi misi paslon

Sistematika Penulisan

Sistematika penulisan jurnal ini sebagai berikut: Pada bagian 2 (Studi terkait) menunjukan hasil studi terkait penelitian ini. Pada penelitian ini sistem yang akan diajukan yaitu Identifikasi Spam Tweet Komentar Pada Twitter Berbasis Ontologi (Studi kasus : Tweet / caption di twitter dengan tema “Pilplres 2019”) yang akan dijelaskan pada bagian 3. Pada bagian 4 akan dijelaskan mengenai pengujian, hasil pengujian dan evaluasi hasil sistem. Kemudian pada bagian 5 akan dijelaskan kesimpulan dan saran.

2. Studi Terkait

2.1 Spam

Spam merupakan sebuah pesan yang sama sekali tidak diinginkan oleh user atau penerima, spam juga sebuah kiriman yang tidak memiliki keterhubungan dengan topik diskusi [5]. Pelaku spam disebut *spammer* dan tindakan penyebaran tersebut disebut *spamming*. Spam memiliki banyak dampak buruk untuk pengguna walaupun sebagian tidak bertujuan untuk melakukan tindakan kejahatan namun terbilang mengganggu dan meresahkan. Kenyamanan pengguna layanan akan sedikit terganggu oleh aktivitas spammer yang tidak bertanggung jawab [9]. Media sosial seperti twitter tak luput dari adanya spam yang berada di dalamnya. Umumnya kegiatan penyebar spam seperti membuat sebuah akun palsu untuk mendapatkan sebuah informasi dari yang di targetkan. Spam pada twitter adalah segala bentuk postingan, tweet atau retweet yang didalamnya terdapat komentar tidak penting seperti konten yang tidak sesuai dengan topik pembahasan atau menawarkan sebuah iklan komersial. Definisi spam pada twitter tersebut berdasarkan hasil dari kuisioner dapat dilihat pada lampiran 4.

2.2 Ontologi

Ontologi merupakan sebuah ilmu tentang makna dari sebuah objek, properti dari sebuah objek dan realsi dari objek tersebut yang mungkin terjadi pada suatu domain [10]. Ontologi menyediakan sebuah hubungan konseptual yang luas dalam sebuah skema [8] sehingga dapat menampung kata-kata kunci. Melalui skema tersebut ontologi akan menjadi kamus data dari definisi spam yang ditentukan. Ontologi menyatakan representasi formal dari suatu *knowledge* dan dapat mendeskripsikan sebuah domain dengan membaginya ke dalam beberapa konsep dan mendeskripsikannya ke relasi yang ada di dalamnya [10].

2.3 *Ontology Web Language*

OWL (*Ontologi web language*) merupakan bahasa semantik untuk merepresentasikan pengetahuan, OWL merupakan bahasa standar yang dapat digunakan untuk merepresentasikan ontologi. Pengetahuan yang kompleks seperti kumpulan term benda hingga relasi antar term tersebut dapat direpresentasikan oleh OWL [11]. OWL juga menjadi salah satu rekomendasi dari W3C yang didesain untuk mendukung web semantik. OWL sendiri terdiri dari RDF (*Resource Description Framework*) yang berfungsi sebagai model data dan relasinya kemudian ditambahkan kontruksi ontologi berbasis pengetahuan dan aksioma oleh OWL [12].

2.4 *Performance Evaluation*

Sistem pemrosesan teks dapat dihitung nilai performansinya menggunakan *precision* dan *recall* [13]. Dokumen relevan dan tidak relevan harus ditentukan sebelum menentukan *precision* dan *recall*, dokumen relevan ditentukan sebagai event target dan dokumen tidak relevan bukan sebagai *event* target.
Berikut merupakan unsur-unsur yang mempengaruhi *precision* dan *recall* diantaranya [14]:

1. *True Positive* (TP) merupakan dokumen relevan yang teridentifikasi dengan benar sebagai dokumen relevan.
2. *True Negative* (TN) merupakan dokumen tidak relevan yang teridentifikasi sebagai dokumen yang tidak relevan.
3. *False Positive* (FP) merupakan dokumen tidak relevan tapi teridentifikasi sebagai dokumen relevan
4. *False Negative* (FN) merupakan dokumen relevan yang teridentifikasi sebagai dokumen yang tidak relevan.

Performansi yang diuji diantaranya:

1. *Precision*
   Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Rumus pernyataan *precision* sebagai berikut.

$$\textbf{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} * 100\% \quad (1)$$

2. *Recall*
   Recall merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Rumus pernyataan *recall* sebagai berikut.

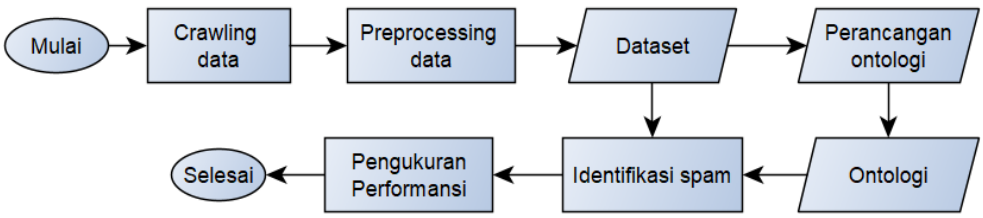
$$\textbf{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} * 100\% \quad (2)$$

3. *F1-Score*
   F1-score merupakan hasil akhir dari perhitungan nilai dari precision dan recall untuk mencari nilai rata-rata dari *precision* dan *recall. F1-score* juga diartikan sebagai penyetaraan nilai *precision* dan *recall*. Rumus pernyatannya sebagai berikut.

$$\textbf{F1-Score} = 2 * \frac{Precision * Recall}{Precision + Recall} * 100\% \quad (3)$$

3. Sistem yang Dibangun

Gambar di bawah ini merupakan alur metodologi penyelesaian yang digunakan pada penelitian ini.



Gambar 1. Alur metodologi penelitian

Berikut merupakan penjelasan setiap tahap dari gambar 1 yang menjelaskan tentang alur metodologi pada penelitian.

3.1. *Crawling* data

Proses *crawling* data untuk mendapatkan dataset yang dilakukan pada penelitian ini menggunakan *python* dengan mengambil data *tweet* beserta *reply* nya yang terkait dengan pemilihan presiden 2019. Berdasarkan hasil kuisioner pada lampiran 7 yang telah dilakukan, dataset yang digunakan yaitu dataset yang berhubungan dengan visi misi para paslon karena dataset yang berhubungan visi misi dari para paslon cenderung memiliki antusias yang tinggi dari masyarakat.

Proses *crawling* dataset untuk pembangunan ontologi menggunakan beberapa *keyword* utama yang berbentuk *hashtag* diantaranya

*IndustriIndonesia, PembangunanIndustri, IndustriIndoneisa(*Mobil listrik*), HukumIndonesia, HukumJokowi, HukumEraJokowi, HukumPrabowo, PembangunanInfrastruktur, InfrastrukturEraJokowi, InfrastrukturUntukIndonesiaMaju, InfrastrukturUntukNegri, InfrastrukturJokowi, SumberDayaManusia, JokowiMembangunSDM, PembangunanSumberdayaManusia, EkonomiIndonesia, EkonomiEraJokowi EkonomiJokowi, EkonomiPrabowo*.

Dataset yang dihasilkan untuk pembangunan ontologi sebanyak 1825 *tweet*. Kemudian untuk dataset uji diambil dari setiap topik yang diwakili oleh 1 data *tweet* beserta data *reply* nya, untuk *tweet* yang mewakili topik dapat dilihat pada lampiran 3. Pada tabel 1 berikut merupakan daftar topik pada dataset yang digunakan sistem pada penelitian ini.

Tabel 1. Daftar dataset yang digunakan sistem

No	Topik	Jumlah *reply*
1	Industri	161
2	Infrastruktur	160
3	Ekonomi pariwisata	161
4	Ekonomi umkm	161
5	Sumber daya manusia	160
6	Hukum	164
Jumlah		967

Pada tabel 2 berikut merupakan contoh dari dataset yang digunakan oleh sistem untuk melakukan identifikasi spam pada penelitian ini.

Tabel 2. Contoh dataset yang digunakan sistem

Tweet	*Reply*
67 persen daripada ekonomi rumah tangga ditopang oleh emak emak oleh karena itu umkm harus kita dorong ke depan kita harus permudah perizinan dalam melakukan usaha kita tidak ingin ekonomi kita dikuasai oleh usaha besar kita harus memberi kesempatan pada pengusaha pemula kecil	tolong bedakan ekonomi makro dengan mikro pak
	umkm butuh dukungan permodalan tapi saya sebagai pelaku umkm ingin bantuan permodalan non riba pak semoga pak dan pak bisa menerapkan syirkah dalam permodalan
	import ikan juga jadi salah satu solusi buat usaha mikro bahkan bisa makro pak
	pak tolong permudah perizinan untuk bahan ekspor hasil umkm
	kalo infrastruktur untuk komoditas tidak jalan mending turun aja pak

3.2. *Preprocessing* data

Dataset yang telah diambil kemudian dilakukan *preprocessing*, tahap *preprocessing* pada penelitian ini diantaranya *case folding, tokenizing, character removal, phrase lookup.*

a. *Case folding*
   Pada tahap ini merupakan proses mengkonversi semua huruf kapital yang ada pada dataset menjadi huruf kecil [15].
b. *Tokenizing*
   Pada tahap ini dilakukan proses pemotongan sebuah kalimat menjadi menjadi kata perkata pada setiap kata yang menyusunnya.
c. *Character removal*
   Pada tahap ini dilakukan proses penghilangan karakter di luar abjad alfabhet untuk mempermudah pemrosesan data.
d. *Phrase lookup*
   Pada tahap ini berfungsi untuk menghubungkan kumpulan kata/frasa yang masih bisa memiliki makna apabila digabungkan. Pada penelitian ini dibuat kamus kata untuk menggabungkan frasa [16].

Pada tabel 3 berikut ini merupakan contoh hasil dari *preprocessing* data yang dilakukan pada penelitian ini.

Tabel 3. Contoh hasil *preprocessing* data

Tahap	Data sebelum diproses	Setelah diproses
Case Fold	**P**ak ini **UMKM** mulai berjalan	pak ini umkm mulai berjalan
Tokenizing	Semoga tahun ini makin baik ekonomi indonesia	'Semoga ',' tahun ',' ini','makin','baik','ekonomi','indonesia'
Characrer removal	Bagaimana dengan Bandung **???** sudah 2 tahun **!**	bagaimana dengan bandung sudah 2 tahun
Phrase lookup	harapan, membuka, **lapangan, kerja**, baru	harapan, membuka, **lapangan_kerja**, baru

3.3. Pembangunan Ontologi

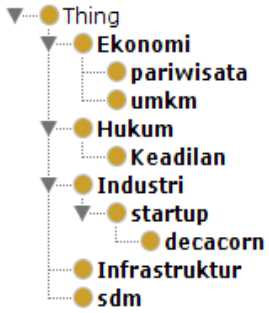
Pada pembangunan ontologi ini mengacu kepada referensi sebuah paper dengan judul "*Ontology Development 101 : A Guide to Create Your First Ontology" oleh Natalya F. Noy dan Deborah L. McGuinness"* [17]. Berikut merupakan tahapannya.

a. Penentuan domain dan ruang lingkup.
   Ontologi yang dibangun memiliki domain tentang pemilihan presiden 2019.
b. Menentukan penggunaan ontologi yang sudah ada.
   Pada sistem ini ontologi yang dibutuhkan yaitu ontologi yang memuat tentang pemilihan presiden 2019 khususnya pada topik umum visi misi paslon. Namun ontologi tersebut tidak ditemukan sehingga digunakan sumber lain sebagai acuan, sumber yang dapat digunakan salah satunya adalah dataset pada twitter yang berkaitan dengan visi misi pada pemilihan presiden 2019.
c. Mengidentifikasi istilah penting
   Istilah penting yang berhubungan dengan domain yang telah ditentukan mengacu kepada topik umum visi misi utama yang sebagai nilai jual utama dari para paslon yaitu tentang ekonomi, hukum, industri, infrastruktur, dan sumber daya manusia. Pada tabel 4 berikut merupakan contoh penentuan istilah penting pada ontologi yang digunakan pada penelitian ini.

Tabel 4. Contoh penentuan istilah penting

Dataset	Istilah yang diambil
Salah satu point penting dari visi misi kami adalah membuat peningkatan kualitas **ekonomi** pada 4 tahun ke depan	Ekonomi
Dalam 4 tahun depan kita akan bekerja keras mati matian demi terciptanya **inftastruktur** yang memadai agar Indonesia bisa bersaing dengan negara lain	Infrastrukur
Salah satu visi misi andalan paslon 1 yg paling penting buat negri ini adalah pembangunan **sdm**	Sdm
Perusahaan gojek menjadi salah satu decacorn di Indonesia dan ini menjadi salah satu keberhasilan **industri** bidang teknologi di Indonesia dari kerja kerasnya jokowi	Industri
Berharap **hukum** 4 tahun kedepan tidak setumpul yang sekarang dan Bersama Prabowo Sandi akan menangani masalah hukum dengan seadil-adilnya	Hukum

d. Mendefinisikan kelas dan hirarki kelas
   Pada tahap ini dilakukan pemetaan terhadap istilah yang sudah ditentukan sebelumnya. Istilah-istilah penting yang sudah ditentukan kemudian dimasukan ke dalam kelas dan sub kelas yang saling berhubungan. Pada pembangunan ontologi ini menggunakan pendekatan gabungan antara *top-down* dan *bottom-up*. Pendekatan *top-down* dipilih karena ontologi memiliki acuan yang pasti untuk menentukan kelas utama yaitu mengenai pemilihan presiden 2019 yang berisi topik umum visi misi paslon. Pendekatan *bottom-up* digunakan untuk menentukan sub kelas dan *instance*, pendekatan ini mengambil istilah yang spesifik ke yang paling umum. Ontologi yang dibangun pada penelitian ini memiliki 5 *class* utama dan *subclass* nya. Pada gambar 2 berikut merupakan contoh kelas pada ontologi yang digunakan oleh sistem untuk melakukan proses identifikasi spam pada penelitian ini. Anotasi *class* dari ontologi dilampirkan pada lampiran ke 1



Gambar 2. Kelas dan hirarki ontologi

e. Mendefinisikan properti
Setelah kelas didefinisikan, dilanjutkan dengan menentukan properti kelas. Properti ini merupakan sebuah objek yang memberikan informasi tambahan mengenai kelas yang ada. Properti digunakan apabila terdapat struktur internal yang perlu dideklarasikan dari kelas tersebut. Contoh properti pada ontologi ini adalah salah satu program kerja dari paslon yaitu digital melayani.

f. Mendefinisikan *facet* dari properti
Setelah menentukan properti selanjutnya menentukan facet atau *value* yakni merupakan isi dari properti yang didefinisikan mulai dari tipe data, isi dari properti, domain dan fitur lainnya. Sebagai contoh properti program kerja memiliki tipe *string* dan berisi "*e-goverment, e-budgeting, e-procurement*" yang memiliki domain ekonomi dan industri.

g. Membuat *instance*
Berdasarkan kelas dan hirarki kelas yang sudah ditentukan kemudian kata-kata penting yang sudah dipilih dapat dijadikan sebuah *instance*. Ontologi yang dibangun pada penelitian ini memiliki 220 *instance* berdasarkan kelasnya masing masing. Pada gambar 3 berikut merupakan contoh *instance* pada ontologi yang digunakan oleh sistem untuk melakukan proses identifikasi spam pada penelitian ini

Gambar 3. *Instance* pada ontologi

Secara keseluruhan, kesimpulan dari pembangunan ontologi yang dibuat untuk penelitian ini memiliki 5 kelas utama dan memiliki total *instance* sebanyak 220. Pada tabel 5 berikut merupakan gambaran statistik dari ontologi yang digunakan sistem pada penelitian ini.

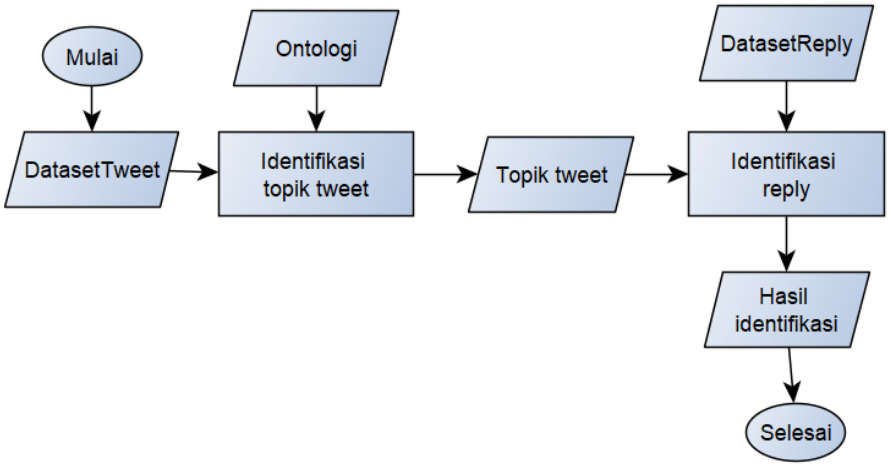
Tabel 5. Statistik ontologi

Kelas	**Jumlah *instance***
Ekonomi	72 *instance*
Hukum	47 *instance*
Industri	35 *instance*
Infrastruktur	36 *instance*
Sdm	30 *instance*

3.4. Identifikasi Spam

Pada tahap identifikasi spam ini menggunakan sistem deteksi spam yang dibuat menggunakan Bahasa pemrograman *java* dan menggunakan *tools netbeans*. Berikut merupakan tahap proses sistem dalam mengidentifikasi spam.

1. Tahap pertama yaitu memasukan file, pada tahap ini ada 2 file yang harus di inputkan ke dalam sistem. File ontologi yang bernama Pilpres.owl beserta dataset uji yang telah didapatkan pada proses *crawling* sebelumnya.
2. Tahap kedua yaitu sistem membaca file ontologi secara bertahap menggunakan *library* pada bahasa pemrograman java yaitu *Apache Jena*, pembacaan dimulai dari kelas yang paling umum menuju kelas yang paling spesifik hingga ke dalam *instance*. Kemudian hasil dari pembacaan ontologi tersebut akan di simpan ke dalam variabel yang bertipe *array*.
3. Tahap ketiga yaitu sistem akan melakukan proses identifikasi, gambar 4 menunjukan alur proses identifikasi spam yang digunakan pada penelitian ini.



Gambar 4. Alur proses identifikasi spam

Berikut merupakan tahap-tahap dalam proses identifikasi spam.

1. Dataset diinputkan dan dibaca oleh sistem.
2. Sistem melakukan proses identifikasi topik tweet terhadap ontologi dengan cara melakukan pengecekan. Apabila ditemukan term pada *tweet* yang sama pada sebuah kelas atau *instance* di ontologi kemudian dihitung seberapa banyak term yang ditemukan. Pada tabel 6 berikut merupakan contoh pengecekan term pada dataset *tweet* yang digunakan pada penelitian ini untuk mendapatkan topik *tweet*.

Tabel 6. Contoh identifikasi topik tweet

Tweet	**Term terkait**	**Topik**
67 persen daripada **ekonomi** rumah tangga ditopang oleh **emak** **emak** oleh karena itu **umkm** harus kita dorong ke depan kita harus permudah **perizinan** dalam melakukan **usaha** kita tidak ingin **ekonomi** kita dikuasai oleh **usaha_besar** kita harus memberi kesempatan pada **pengusaha** pemula	Ekonomi, emak emak, umkm, perizinan, usaha, ekonomi, usaha besar, pengusaha	Ekonomi

3. Dataset *tweet* tersebut masuk ke dalam kelas pada ontologi yang jumlah kelas dan instancenya paling banyak ditemukan. Pada tahap ini *tweet* mendapatkan topik berdasarkan kelas yang didapatkan.
4. Setelah topik dari *tweet* didapatkan kemudian dataset *reply* dari *tweet* tersebut dilakukan proses identifikasi berdasarkan kelas yang didapatkan sebelumnya.
5. Sistem melakukan proses identifikasi spam terhadap *reply* dengan melakukan pengecekan term pada ontologi. Apabila ditemukan term yang sama pada *reply* di ontologi maka dikategorikan sebagai bukan spam dan jika tidak ditemukan term yang sama pada *reply* di ontologi maka dikategorikan sebagai spam. Pada tabel 7 berikut merupakan contoh pengecekan term pada dataset *reply* yang digunakan pada penelitian ini untuk mendeteksi spam atau bukan spam.

Tabel 7. Contoh identifikasi spam pada komentar atau *reply*

Reply	**Term terkait**	**Hasil**
tolong bedakan **ekonomi** **makro** dengan **mikro** pak	Ekonomi, makro, mikro	Bukan spam
umkm butuh dukungan **permodalan** tapi saya sebagai pelaku **umkm** ingin bantuan **permodalan** non **riba** pak semoga pak dan pak bisa menerapkan syirkah dalam **permodalan**	Umkm, permodalan, riba	Bukan spam
import ikan juga jadi salah satu solusi buat usaha **mikro** bahkan bisa **makro** pak	Mikro, makro	Bukan spam
pak tolong permudah **perizinan** untuk bahan ekspor hasil **umkm**	Perizinan, umkm	Bukan spam
kalo infrastruktur untuk komoditas tidak jalan mending turun aja pak	-	Spam

4. Pengujian dan Analisis

4.1. Pengujian Sistem

Pada tahap pengujian ini memiliki tujuan untuk mengukur hasil performansi dari penerapan metode ontologi untuk mengidentifikasi spam yang telah ditentukan. Berikut merupakan tahapan skenario pada pengujian.

a. Sistem diuji dengan data uji yang berupa 6 tweet beserta *reply* nya.
b. Dataset diberi label oleh pakar bahasa untuk mendapatkan keputusan spam atau bukan spam.
c. Kemudian sistem melakukan identifikasi terhadap data uji tersebut untuk membandingkan dengan hasil sistem. Sistem akan mengidentifikasi apakah *reply* dari data uji tersebut spam atau bukan spam.
d. Menghitung nilai performasi hasil dari data uji pada proses pengujian dengan parameter nilai *precision, recall* dan *f1-score.*

4.2. Analisis Pengujian

Proses pengujian sistem telah menghasilkan nilai *precision, recall* dan *f1-score,* berikut merupakan tabel yang memuat informasi hasil dari pengujian sistem yang dilakukan pada penelitian ini.

Tabel 8. Hasil pengujian sistem

Hasil	**Data uji 1**	**Data uji 2**	**Data uji 3**	**Data uji 4**	**Data uji 5**	**Data uji 6**
Precision	90.16 %	87.25 %	84.72 %	84.93 %	83.65 %	66.67 %
Recall	98.21 %	98.88 %	92.42 %	98.41 %	97.77 %	95.23 %
F1-Score	94.01 %	92.70 %	88.40 %	91.17 %	90.15 %	78.43 %

Pada tabel 8 menunjukan hasil performansi setelah dilakukan pengujian oleh 6 data uji kepada sistem. Beberapa hal yang sangat mempengaruhi hasil performansi diantaranya.

1. Hasil performansi pada tabel 4 memperlihatkan bahwa hasil identifikasi spam menggunakan ontologi untuk keenam dataset uji memiliki rata-rata *f1-score* sebesar 89.14%. Namun dapat dilihat nilai *precision* pada data uji 6 yang merupakan nilai paling rendah dibandingkan data uji lainnya. Hal tersebut dikarenakan hasil prediksi sistem kurang tepat dalam mengidentifikasi spam. Hal-hal yang mempengaruhi diantaranya :
   a. Terdapat *tweet reply* yang di dalamnya tidak ada term yang sesuai pada ontologi, namun *tweet reply* tersebut masih berhubungan dengan topik pembicaraannya yaitu mengenai hukum. Pada tabel 9 merupakan contoh *reply* yang tidak memuat istilah pada ontologi.

Tabel 9. contoh hasil data uji

Hasil sistem	**Hasil aktual**	**Konten**
Spam	Bukan	tragedy 98 juga seharusnya diperhatikan pak

Hal tersebut mempengaruhi nilai pada *false positive* yang menjadikan nilai *precision* menjadi lebih kecil. Untuk kekurangan ini dapat diatasi dengan menambahkan istilah-istilah yang baru yang ditemukan pada dataset sebelumnya ke dalam ontologi.

b. Terdapat *tweet reply* yang *typo* sehingga tidak terdeteksi oleh sistem. Pada tabel 10 merupakan contoh salah satu *reply* yang ditemukan dari data uji 6 yang terdapat *typo* atau kesalahan penulisan.

Tabel 10. Contoh hasil data uji

Hasil Sistem	Hasil aktual	Konten
Spam	Bukan	memang harusnya ini sudah dibawa ke **mahkamsh**

Pada komentar tersebut "memang harusnya ini sudah dibawa ke mahkamsh ", term "mahkamsh" yang seharusnya "mahkamah" membuat sistem menilai bahwa *reply* tersebut bukan spam dikarenakan term"mahkamah" tersebut merupakan salah satu istilah pada ontologi. Hal tersebut juga menyebabkan nilai *false positive* bertambah sehingga berpengaruh terhadap perhitungan nilai *precision*.

2. Pengidentifikasian spam untuk dataset uji 1-5 memiliki rata-rata nilai f1-score 91.26%. Hal tersebut dikarenakan banyak istilah pada dataset yang terdapat juga pada ontologi, sehingga dapat mempermudah sistem dalam mengidentifikasi dengan baik apakah *reply* tersebut spam atau bukan spam.

3. Pada sistem juga ditemukan sebuah *reply* yang cenderung ambigu seperti contoh pada tabel 11 berikut.

Tabel 11. Contoh hasil data uji

Hasil sistem	Hasil aktual	Konten
Bukan	Spam	apakah **pembangunan** cinta juga salah satu program bpk hehe kiding yah pak vis

Hal ini dapat diatasi oleh proses *preprocessing* dan juga dapat berperan dalam meningkatkan performansi, dengan mengurangi kemabiguan menggunakan proses *phrase lookup* yang berfungsi untuk menghubungkan kata atau frasa yang masih bisa memiliki makna apabila digabungkan. Sebagai contoh "pembangunan_indonesia" adalah sebuah istilah pada ontologi yang dimana ketika tidak dilakukan proses *phrase lookup* maka akan menjadi "pembangunan" saja tidak digabung dengan "Indonesia". Kemudian berdasarkan cara kerja sistem, dataset yang mengandung term "pembangunan" akan dikategorikan sebagai bukan spam dikarenakan term tersebut adalah istilah yang terdapat pada ontologi. Namun hal tersebut bisa saja tidak benar karena term pembangunan pada dataset tersebut belum tentu memiliki makna tentang pembahasannya.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berikut merupakan hasil kesimpulan yang diambil dari penelitian dan analisis hasil pengujian:

1. Berdasarkan hasil pengujian dan analisis yang telah dilakukan, menunjukan bahwa *reply* pada twitter memungkinkan memiliki beberapa topik pembicaraan dimana terdapat sebagian dari *reply* yang tidak memiliki keterkaitan dengan pembicaraan atau topik. Ontologi dapat membantu untuk melakukan klasifikasi terhadap *reply* yang memiliki keterhubungan dengan topik dan *reply* yang tidak memiliki hubungan denagn topik. Dalam proses pengidentifikasian spam pada penelitian ini, ontologi menjadi sebuah kumpulan term yang didalamnya terdapat istilah yang berhubungan dengan topik umum visi misi paslon pilpres 2019. Apabila *reply* tersebut mengandung term yang ada pada ontologi maka komentar tersebut teridentifikasi sebagai bukan spam.
2. Ontologi yang dibuat pada penelitian ini mampu mengidentifikasi spam yang ditentukan dengan nilai rata-rata *f1-score* sebesar 89.14%. Selain itu pada pengujian ini ditemukan beberapa hal yang menjadi kekurangan pada sistem yaitu sistem belum mampu mendeteksi *tweet* yang di dalamnya terdapat kata-kata yang *typo* atau salah penulisan.

5.2. Saran

Berikut merupakan hasil saran yang diambil dari penelitian dan analisis hasil pengujian:

1. Penggunaan algoritma *Levenstein Distance* untuk membantu mengidentifikasi jika terjadi *typo* atau salah penulisan untuk meningkatkan nilai performansi yang dihasilkan oleh sistem.
2. Pembangunan ontologi yang lebih luas dari kelas, sub kelas hingga *instance* sehingga dapat meningkatkan nilai performansi yang didapat.
3. Penambahan kamus pada *phrase lookup* untuk mengurangi keambiguan makna dan juga dapat meningkatkan performansi yang didapat.

Daftar pustaka

[1] Statista, "www.statista.com," Statista, 2019. [Online]. Available: www.statista.com/statistics/490548/twitter-users-indonesia/. [Accessed 02 January 2019].

[2] B. Madiong, Z. Mustapa and A. G. Ratu Chakti, Pendidikan Kewarganegaan (Civic Education), Celebes Media Perkasa, 2017.

[3] Khoirunnisa, "Selular.id," Selular, 15 April 2019. [Online]. Available: https://selular.id/2019/04/lebih-dari-13-juta-tweet-ramaikan-debat-pilpres-2019-terakhir/. [Accessed Juli 2019].

[4] A. M. Damar, "Liputan6," Liputan6, 30 Maret 2019. [Online]. Available: www.liputan6.com/tekno/read/3930030/antusiasme-warganet-sambut-debat-pilpres-keempat-2019. [Accessed 06 Juli 2019].

[5] M. T. Banday and J. Qadri, "SPAM - Tecnological and legal aspects," 2006.

[6] J. Iqbal, "Liputan6," 15 April 2019. [Online]. Available: www.liputan6.com/tekno/read/3941954/13-juta-twit-ramaikan-debat-terakhir-pilpres-2019. [Accessed 06 Juli 2019].

[7] Y. Zhu, X. Wang, E. Zhong, N. N. Liu, H. Li and Q. Yang, "Discovering Spammers in Social Networks," Hongkong, 2012.

[8] I. Horrocks, "Ontologies and the Semantic Web," Oxford, 2008.

[9] W. Thorkildssen, SPAM-Different Approach to Fighting, 2004.

[10] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator and S. William, Enabling Technology for Knowledge Sharing, vol. III, 1991.

[11] "W3C," OWL Working Group, 11 12 2012. [Online]. Available: https://www.w3.org/2001/sw/wiki/OWL. [Accessed 20 October 2018].

[12] L. Ding, P. Kolari, Z. Ding, S. Avanca and A. Joshi, Using Ontologies in the Semantic Web: A Survey, 2005.

[13] jesse, J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," Department of Computer Sciences and Department of Biostatistics and Medical Informatics, Madison, 2006.

[14] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," in *Proceedings of the 27th European conference on Advances in Information Retrieval Research*, Santiago de Compostela, 2005.

[15] C. D. Manning, P. Raghavan and . H. Schütze, Introduction to Information Retrieval, Cambrigde: Cambridge University Press, 20018.

[16] D. A. Nughroho, "Implementasi Rule-Based Classifier pada Analisis Sentimen dengan Metode Ontology Supported Polarity Mining," p. 5, 2018.

[17] N. F. noy and D. L. Mcguinness, A Guide to Create Your First Onthology, 2001.