Prediksi *Information Cascade* di dalam Twitter dengan Menggunakan Metode *Random Walk*

Proposal Tugas Akhir

Kelas MK Penulisan Proposal (CCH4A3)

1301198506

RAFI HAFIZHNI ANGGIA

SARJANA INFORMATIKA

Fakultas Informatika

Universitas Telkom

Bandung

2020

Daftar Isi

Lembar Persetujuan

Prediksi *Information Cascade* didalam Twitter dengan Menggunakan Metode *Random Walk*

Predicting Information Cascade on Twitter Using Random Walk Method

NIM :1301198506

RAFI HAFIZHNI ANGGIA

Proposal ini diajukan sebagai usulan pembuatan tugas akhir pada
Program Studi Sarjana Teknik Informatika
Fakultas Informatika Universitas Telkom

Bandung 2020
Menyetujui

Pembimbing 1

Jondri, S.Si.,M.Si.
NIP : 95700035-1

ABSTRAK

Perkembangan teknologi informasi terbilang meningkat sangat cepat, teknologi membantu kebutuhan manusia sebagai sarana informasi, berinteraksi, mengemukakan opini, sarana bisnis. Fitur *retweet* dianggap sebagai mekanisme penyebaran informasi yang mudah, dan data tersebut diambil secara *crawling* pada twitter untuk digali informasi menarik dengan *Information Cascade* menganalisis penyebaran informasi bermanfaat bagi bidang bisnis, social, ekonomi, Kesehatan, pemdidikan dll, dengan acuan *Content Based, User Based, Structural Features of the network* dan *Temporal features* menggunakan metode *Random Walk* sebagai metode probabilitas untuk menggambarkan jalur dari serangkaian langkah acak berturut-turut dalam suatu ruang matematis guna memprediksi sebuah difusi eksploitasi konten pengguna, kemudian dilakukan proses klasifikasi menggunakan *Support Vector Machine* yang dipercaya dapat menghasilkan performansi terbaik dalam hal akurasi.

Kata kunci : *Crawling, Information Cascade, Random Walk, Support Vector Machine,* Twitter

1. Pendahuluan

1.1 Latar Belakang

Perkembangan teknologi informasi saat ini terbilang meningkat dengan sangat cepat, selain itu teknologi sangat membantu kebutuhan manusia sebagai sarana informasi, berinteraksi, mengemukakan opini, sarana bisnis, hal ini memunculkan banyak akses media sosial salah satunya yaitu twitter dengan berbasis *microblog* [1], salah satu fitur dalam twitter adalah membuat sebuah tweet atau memposting ulang (*retweet*) [2] *retweet* bersifat subjektif [3]. Perilaku *retweet* dapat didefinisikan ketika satu individu mengamati postingan, tweet, dll. Namun orang lain membagikan atau memposting ulang. Proses ini terus terjadi sampai individu berhenti menyebarkan atau memposting ulang. Informasi menarik ini diitinjau dalam perilaku ekonomi dan teori jaringan dilakukan membuat keputusan dalam mode beruntut prediksi yang disebut *Information Cascade* [4].

Information cascade memiliki keunggulan dapat dijadikan tempat penyebaran berita yang bermanfaat dalam membentuk opini pengguna [5] Adanya data dari *information cascade* ini dapat menganalisis penyebaran informasi bisa bermanfaat bagi bisnis dalam meluncurkan produk [2]. Dapat digunakan untuk memahami pasar saham dan dapat menggunakan metode ini untuk mendapatkan dukungan pada saat pemilihan atau pencanangan kebijakan sebagai prediksi [2], namun memiliki masalah penting karena dapat memberikan pemahaman yang mendalam tentang pembentukan opini di pengguna sosial media, misal isu sensitif seperti SARA yang kerap terjadi.

Penelitian yang telah dibuat [4]untuk memecahkan masalah terkait dengan *information casecade* pada penyebaran informasi didalam media sosial dengan memiliki fitur seperti *Content Based, User Based, Structural Features of the network* dan *Temporal features*. Metode *Random Walk* digunakan sebagai metode probabilitas untuk menggambarkan sebuah jalur dari serangkaian langkah acak berturut-turut dalam suatu ruang matematis [6]. *Random Walk* juga digunakan untuk memprediksi sebuah difusi dengan mengeksploitasi konten pengguna [4], dari hal tersebut perlu proses klasifikasi guna mengetahui performansi sistem dengan tingkat akurasi menggunakan pendekatan *Macine Learning*.

Pendekatan ini dilakukan untuk proses klasifikasi dengan terdapat Terdapat beberapa metode klasifikasi yang salah satunya *Support Vector Machine (SVM). SVM* adalah metode *supervised learning* yang melakukan proses klasifikasi data secara *linier* [7]. Metode klasifikasi *Support Vector Machine* untuk mengetahui *performance* dan hasil dari *information cascade* menggunakan *Random Walk* pada twitter. Metode klasifikasi *Support Vector Machine* ini dipilih berdasarkan penelitian sebelumnya yang menghasilkan hasil akurasi yang baik dalam melakukan klasifikasi [8].

1.2 Perumusan Masalah

Penjelasan hasil pengematan dari latar belakang didapatkan rumusan masalah seperti berikut:

1. Bagaimana pengaruh *Random Walk* dalam menganalisa prediksi *retweet* dari data pada twitter dengan metode klasifikasi *Support Vector Machine*?
2. Bagaimana cara kerja metode *Random Walk* dalam penyeleseian masalah

1.3 Tujuan

Tujuan pencapaian yang saya inginkan dalam tugas akhir ini adalah sebagai berikut:

1. Melakukan implementasi *information cascade* dalam sebuah retweet dari data twitter dengan metode *Random Walk* dengan proses klasifikasi *Support Vector Machine*.
2. Mengetahui cara kerja metode *Random Walk* dalam menganalisa prediksi sebuah data pada twitter.

1.4 Batasan Masalah

Batasan masalah yang terkait tugas akhir ini sebagai berikut :

Data set yang diambil dari *keyword* dan akun dan yang sedang *trending* saja.

1.5 Rencana Kegiatan

Rencana kegiatan yang dilakukan dalam penyelesaian Tugas Akhir ini adalah:

1. **Studi Literatur**

   Pada tahap ini akan dilakukan proses pencarian dan pengumpulan referensi yang akan digunakan sebagai acuan dalam pembuatan Tugas Akhir ini.

2. **Pengumpulan Data**

   Pada tahap ini akan dilakukan pencarian data dan pengumpulan data terkait dengan masalah yang diangkat pada Tugas Akhir ini.

3. **Perancangan Sistem**

   Pada tahap ini akan dilakukan perancangan terhadap sistem yang akan dibangun serta pembahasan tiap gambaran umum yang dilakukan pada penelitian Tugas Akhir ini.

4. **Implementasi Sistem**

   Pada tahap ini akan dilakukan pengimplementasian sistem berdasarkan rancangan yang telah dibuat sebelumnya.

5. **Pengujian dan Analisis**

   Pada tahap ini hasil pengujian akan dianalisis untuk menyimpulkan hasil dari penelitian Tugas Akhir ini.

6. **Penyusunan Laporan**

   Pada tahap ini akan dilakukan pembuatan laporan akhir sebagai bentuk dokumentasi dari penelitian Tugas Akhir ini.

7. **Jadwal Kegiatan**

   Berikut jadwal kegiatan dalam penyelesaian Tugas Akhir ini:

Tabel 1 Jadwal Kegiatan

Kegiatan	**Bulan**					
	1	2	3	4	5	6
Studi Literatur						
Pengumpulan Data						
Perancangan Sistem						
Implementasi Sistem						
Pengujian dan Analisis						
Penyusunan Laporan						

2 Tinjauan Pustaka

2.1 Studi Terkait

Penelitian yang dilakukan oleh Syeda Nadia Firdaus, Chen Ding, Alireza Sadeghian menjelaskan cara memprediksi sebuah tweet dan bagaimana seseorang itu akan meretweet atau memposting ulang sebuah tweet, kemudian memahami sebuah informasi itu bisa tersebar dengan inti penelitian ditujukan untuk memberikan sebuah gambaran tentang prediksi sebuah *retweet* [2].

Penelitian yang dilakukan oleh Nidhi Singha, Anurag Singha, Rajesh Sharma yang dijadikan sebagai acuan dibuatnya tugas akhir ini menjelaskan bagaimana cara memprediksi sebuah *information cascade* bagaiman sebuah *retweet* akan berakhir dan pada metode yang digunakan untuk solusi permasalahan pada kasus ini dengan menggunakan metode *Random Walks* yaitu sebuah proses statistika yang dapat memprediksi sebuah probabilitas [4].

Penelitian yang dilakukan oleh José Luis Ortega yang menjelaskan analisis hubungan antara diseminasi penelitian sebuah makalah di twitter dan pengaruhnya terhadap dampak penelitian, digunakan untuk mendeteksi perbedaan secara signifikan antara masing-masing kelompok jurnal dan menggunakan metode regresi untuk mendeteksi sebuah *variable* yang nantinya akan mempengaruhi sebuah tweet [9].

Penelitian yang dilakukan oleh Feng Xia dan lainnya menjelaskan tinjauan komprehensif tentang *random walk* sebuah proses matematika yang digunakan untuk melakukan *review* sebuah algoritma dan aplikasi. Penelitian ini bertujuan untuk berkontribusi pada bidang penelitian yang berkembang dengan membedah sebuah metode *Random Walk* [10].

Tujuan tugas akhir ini digunakan untuk membuat sebuah prediksi *Information Cascade* dengan metode penyeleseian dengan menggunakan *Random Walk* sehingga dapat mengetahui prediksi tweet tersebut.

2.2 Twitter

Pada dasarnya media sosial merupakan perkembangan dari teknologi web berbasis internet digunakan untuk komunikasi, pertisipasi, berbagi dan membentuk sebuah jaringan secara *online* untuk menyebarluaskan konten mereka sendiri. Memposting pada twitter dapat dilihat oleh jutaan orang secara gratis [11]. Twitter

menjadi situs *microblogging* yang merupakan kategori sosial media. Konten dari twitter berisi teks yang menampung 280 karakter setiap *updatenya* karena sifat twitter yang singkat dan langsung sehingga mudah untuk penyampaian informasi. Pengguna menulis pesan dan dapat dikelompokkan berdasar dengan *topic*, jenis pesan atau dengan menggunakan # (*hashtag*) [11].

Twitter telah banyak dimanfaatkan dengan keperluan pribadi, media promosi produk dan jasa bahkan pesan resmi dari suatu otoritas. Twitter dimanfaatkan menjadi jejaring sosial yang efisien dan efektif [12]. Fitur dalam twitter adalah memposting ulang atau yang biasa disebut *retweet* pesan dianggap sebagai mekanisme penyebaran informasi yang tersedia, penyebaran secara luas secara cepat dan semakin besarnya informasi tersebut merupakan cerminan dari ramainya informasi yang sedang dibicarakan akan menjadi *trending* [13]. Pengguna twitter di Indonesia terus meningkat, seperti contoh pada tahun 2014 jumlahnya ±12 juta, angka tersebut diprediksi akan terus meningkat hingga mencapai ±22,8 juta pengguna di tahun 2019 [14].

2.3 *Crawling* Data

Crawling adalah tahap pengumpulan data atau pengunduhan data dari suatu *database*. Data tersebut dikumpulkan dari pengunduhan *server* pada twitter berupa *user* dan atribut tweetnya [15]. Proses *crawling* mengambil data berukuran besar maupun kecil pada halaman *web* yang dapat disimpan dan dicari menggunakan kata kunci [16].

Proses pengunduhan atau *crawling* menggunakan suatu program aplikasi telah tersedia dari twitter yang biasanya disebut dengan API *(Application Programming Interface)* untuk mempermudah *developer* untuk pengaksesan informasi yang tersedia di *website* twitter [8]. *Developer* mendaftar pada aplikasi twitter pada halaman https://dev.twitter.com untuk penggunaan API dan akan mendapatkan *token* dan *key* sebagai syarat otentifikasi dalam pengunduhan data yang tersedia pada twitter [15].

2.4 Preprocessing

Pada tahap *preprocessing* dilakukan olah data mentah sebelum ke proses selanjutnya, dengan melakukan eliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses [17]. *Preprocessing* sangat penting

terutama digunakan untuk olah data pada hasil *crawling* berita *hoax* di twitter yang sebagian besar terdapat kata-kata atau kalimat ambigu atau tidak terstruktur serta memiliki *noise* yang sangat besar [18]. Teknik yang dapat dilakukan pada proses *preprocessing* yaitu [18]:

1. *Case folding*, membuat data input yang berupa teks menjadi huruf kecil.
2. *Remove punctuation*, menghapus semua karakter *non alphabet* misalnya simbol, spasi, dan lain-lain.
3. *Remove username*, menghapus nama *user* yang biasanya diawali dengan simbol "@".
4. *Remove hashtag*, memiliki "#" yang diikuti dengan topik yang dimaksud.
5. *Clean number*, berfungsi menghapus angka yang ada di awal atau akhir kalimat. Penggunaan angka yang salah dalam bahasa Indonesia biasanya ada pada kata perulangan. Contoh: yang seharusnya "ramai-ramai" justru ditulis "ramai2".
6. *Clean one character*, jika hanya ada satu karakter saja, dan tidak memiliki arti yang penting, bisa dihapus karena akan sulit untuk dideklarasikan.
7. *Removal URL*, seringnya terdapat *url* di sela-sela tweet, membuat data yang ada menjadi kurang efektif atau bahkan tidak memiliki arti. Seperti orang yang menyisipkan *url* promosi produk yang tidak bersangkutan dengan tweet yan dibuat.
8. *Convert number*, pemakaian bahasa gaul pada twitter mengubah penulisan beberapa huruf menjadi angka, seperti "S14p4" untuk kata "Siapa", namun dalam kasus layanan sinyal atau produk operator, maka bisa saja proses ini merubah arti, seperti "sinyal 3G".
9. *Remove negation word*, proses ini akan menilai apakah kalimat yang diproses mengandung kalimat negative.
10. *Convert word*, berfungsi untuk mengubah kata yang tidak baku, seperti bahasa gaul dan sejenisnya.
11. *Convert emoticon*, orang mengekspresikan perasaannya terkadang tidak hanya lewat kata, namun juga bisa melalui *emoticon*. *Emoticon* ini dapat dikonversi untuk dapat mengetahui maknanya dalam bentuk teks.

2.5 TF-IDF

TF-IDF merupakan metode untuk melakukan pembobotan pada setiap kemunculan kata dalam dokumen [19]. Nilai bobot menunjukan seberapa penting sebuah istilah *(term)* dalam sebuah dokumen [8].

Terdapat dua parameter untuk melakukam pembobotan kata atau W_i yaitu nilai *TF* dan nilai *IDF*. *TF* adalah frekuensi kemunculan istilah *t* dalam dokumen $d_{i,}$ dimana nilai $f_t(x)$ bernilai 1 jika x = t dan 0 jika x ≠ t. Untuk mencari nilai TF maka dapat dihitung dengan persamaan rumus **(2.1)**:

$$TF_{t,d} = \sum_{x \in d} f_t(x) \quad \textbf{(2.1)}$$

IDF adalah banyaknya dokumen yang mengandung istilah *t.* Asumsikan total dokumen dalam sebuah koleksi adalah *D* dan dokumen yang masuk dalam koleksi *D* diasumsikan d_i dokumen. Setiap d_i dokumen terdapat teks yang berisi istilah (*term*) t_i. Jadi, di dalam dari kumpulan *D* dokumen terdiri dari d_i dokumen dan setiap d_i dokumen berisikan istilah (*term*) t_i. *IDF* dapat dihitung dari banyaknya dokumen $|D|$ dibagi banyaknya dokumen d_i yang mengandung istilah (*term*) t_i atau disebut dengan df (t) Untuk mencari nilai *IDF* dapat dirumuskan dengan persamaan **(2.2)** [8]**.**

$$IDF_t = log\left(\frac{|D|}{df_t}\right) \quad \textbf{(2.2)}$$

Nilai bobot W_i merupakan hasil kali nilai *TF* (t, d) dengan nilai *IDF*(t). Jadi nilai bobot W_i dapat dihitung dengan Persamaan **(2.3**) berikut ini [19]:

$$W_i = TF_{t,d} \times IDF_t \quad \textbf{(2.3)}$$

2.6 Random Walk

Random Walk merupakan suatu proses secara acak yang dapat mendeskripsikan suatu jalur termasuk serangkaian yang dilakukan secara acak di ruang matematika [20]. *Random walk* ini semakin populer di berbagai ilmu seperti matematika dan ilmu komputer. Lebih jauh lagi, dalam mekanika kuantum jalan bisa terjadi dianggap sebagai analog kuantum dari *Random Walk* klasik. *Random walk* klasik dapat digunakan hitung kedekatan antara node dan ekstrak dalam suatu topologi

jaringan [10]. Untuk memprediksi sebuah *information cascade* dengan mertimbangkan dimulai dari simpul ni dan berakhir di simpul n j,maka n dapat direpresentasikan dalam persamaan **(2.4)** [4]:

$$pi = (pi1, pi2, ... pin) \quad \textbf{(2.4)}$$

Probabilitas untuk berada dalam status sendiri adalah pii = 1 - α di mana α mewakili prior bahwa probabilitas pasti akan berjalan meninggalkan statusnya saat ini. Probabilitas sebelumnya sebanding dengan derajat keluar dari *node* [4]. Kesamaan matrik dan probabilitas dihitung sebagai S, CS dan P, *random walker* akan tetap berada di simpul j karena kesamaan antara dua *node* dan diberikan sebagai : P = DS , Maka kondisi terakhir dari *random walk* dapat dihitung dilihat dari persamaan **(2.5):**

$$R (t + 1) = \alpha PR (t) + (1 - \alpha) I \quad \textbf{(2.5)}$$

R (t), R (t + 1) masing-masing didefinisikan sebagai matriks probabilitas dari keadaan pada t dan t + 1.

2.7 Informasi Cascade

Information cascade merupakan suatu perilaku seorang individu yang mengamati suatu postingan dan melakukan sebuah *retweet* atau membagikan ulang informasi tersebut. Proses ini terus terjadi sampai individu berhenti menyebarkan atau memposting ulang. *Platform* media sosial aktivitas dapat sesuai dengan berbagi pandangan, posting, multimedia (seperti foto, video) dan dapat dilakukan penyebaran berita dapat berguna dalam membentuk opini pengguna [21]. Tahapan -tahapan yang digunakan [4].

1. *Users' Similarity*

   Pada twitter akan dapat ditemukan kesaman antara pengguna jika saling mengikuti satu sama lain dan mereka saling membalas atau meretweet [4].

$$S_{i,j} = \begin{cases} 1, if\ user\ ui, similiar\ to\ the\ user\ uj \\ 0, otherwise \end{cases} \quad \textbf{(2.6)}$$

2. *Content Similarity*

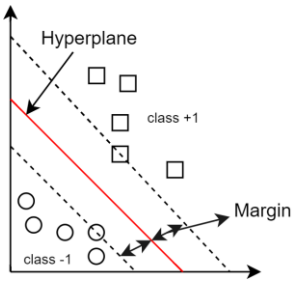
   Kesamaan tweet dihitung menggunakan *cosine similarity*. Berdasarkan kata Koleksi dan vektor dibuat. Kemudian, vektor biner untuk setiap tweet

ditemukan yang digunakan dalam penghitungan *Cosine similarity*. Setiap tweet dapat dianggap sebagai dokumen individual. Itu diberikan sebagai [4]:

$$Cs_{f1f2=\frac{\langle \widehat{t1}-\widehat{t2}\rangle}{\|f1\|\|f2\|}} \quad \textbf{(2.7)}$$

2.8 Support Vector Machine (SVM)

Proses selanjutnya adalah melakukan klasifikasi berupa data input yang sudah berpola dan diolah menjadi vektor/*binary* dengan menggunakan metode *SVM* yang dapat mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space*. *Pattern* yang merupakan anggota dari dua buah kelas +1 dan -1 sebagai alternatif garis pemisah. *Margin* adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing kelas. *Pattern* yang paling dekat disebut sebagai *SVM* [22].



Gambar 1 *Hyperplane* terbaik dari SVM

Gambar 1 merupakan pencarian *hyperplane* terbaik. Garis merah yang terletak di tengah adalah garis *hyperplane,* lalu garis putus-putus yang berada di sebelah garis *hyperplane* adalah *margin* dimana garis tersebut adalah jarak dari *hyperplane* pada data yang terdekat [23]. Prinsip dasar yang digunakan adalah klasifikasi linier dan klasifikasi *non-linier* dengan memasukkan kernel trik pada ruang dimensi tinggi [24].

Data yang tersedia dinotasikan sebagai $\overrightarrow{x_i} \in R^d$ sedangkan label masing-masing dinotasikan $y_i \epsilon \{-1, +1\}$ untuk i = 1, 2, … ,*l* , yang mana *l* adalah banyaknya data. Diasumsikan kedua kelas -1 dan +1 dapat terpisah secara sempurna oleh *hyperplane* berdimensi *d*, yang didefinisikan [25]. Untuk mencari nilai data tersedia dapat dilihat pada rumus **(2.8)** dibawah ini:

$$\vec{w}.\vec{x} + b = 0 \quad \textbf{(2.8)}$$

Pattern $\overrightarrow{x_i}$ yang termasuk kelas -1 (sampel negatif) dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan. Untuk mencari nilai data sampel negatif dapat dilihat pada rumus **(2.9)** dibawah ini [25] :

$$\vec{w}.\vec{x} + b \leq -1 \quad \textbf{(2.9)}$$

Pattern $\overrightarrow{x_i}$ yang termasuk kelas +1 (sampel positif) dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan. Untuk mencari nilai data sampel positif dapat dilihat pada rumus **(2.10)** dibawah ini [25]:

$$\vec{w}.\vec{x} + b \leq +1 \quad \textbf{(2.10)}$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara *hyperplane* dengan titik terdekatnya, Untuk mencari titik terdekat margin dapat dilihat rumus **(2.11)** dibawah ini :

$$\frac{1}{\|\vec{w}\|} \quad \textbf{()(}\ \textbf{(2.11)}$$

2.9 Confusion Matrix

Confusion Matrix diartikan sebagai alat yang berfungsi untuk melakukan analisis terhadap *classifier* dalam mengenali *tuple* dari kelas berbeda yang digunakan dalam *Superviced Learning* yang digunakan untuk melihat hasil tes dari model yang telah diprediksi [26]. Tahap ini menghitung nilai dari *accuracy, precision, dan recall.* Tabel 1 merupakan *Confusion Matrix* [19].

Tabel 2 *Confusion Matrix*

	Prediksi		
Aktual		Kelas 1	Kelas 0
	Kelas 1	TP	FN
	Kelas 0	FP	TN

Dimana :

a. *True Positive* (TP) berarti kelas yang di prediksi sesuai dengan target yaitu positif.
b. *False Positive* (FP) berarti kelas yang di prediksi (positif) tidak sesuai dengan target (negatif).
c. *False Negative* (FN) berarti kelas yang di prediksi (negatif) tidak sesuai dengan target (positif).
d. *True Negative* (TN) berarti kelas yang di prediksi sesuai dengan target yaitu negatif.

Dari tabel diatas dapat dihitung nilai *precision, recall* dan *accuracy. Precision* digunakan untuk mengukur pola positif yang diprediksi dengan benar dari total pola prediksi dalam kelas positif [27]. Berikut pada persamaan rumus **(2.12)** yang digunakan untuk mencari nilai *Precision* :

$$Precision = \frac{TP}{TP+FP} \times 100 \quad \textbf{(2.12)}$$

Recall digunakan untuk mengukur pola positif yang diklasifikasikan dengan benar [27]. Berikut pada persamaan rumus **(2.13)** yang digunakan untuk mencari nilai *Recall* :

$$Recall = \frac{TP}{TP + FN} \times 100 \quad \textbf{(2.13)}$$

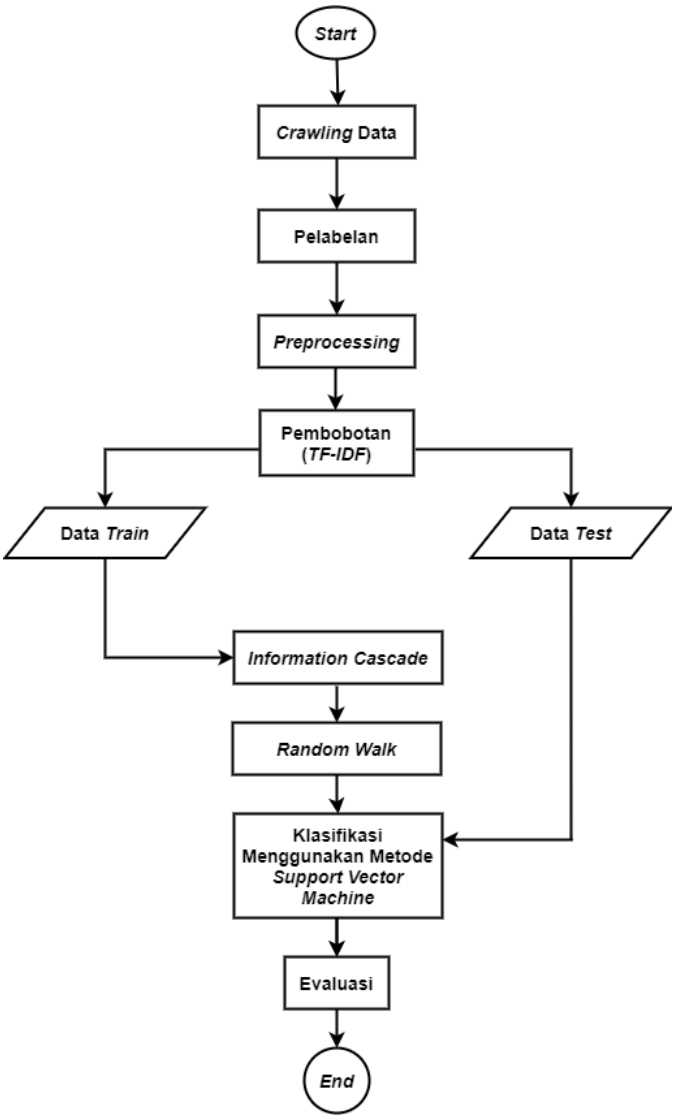
Untuk mencari nilai Akurasi maka akan digunakan persamaan rumus **(2.14)**.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad \textbf{(2.14)}$$

3. Perancangan Sistem

3.1 Gambaran Umum Sistem

Perancangan sistem menggambarkan proses penelitian tugas akhir berbentuk *flowchart*, untuk mengetahui bagaimana tahapan yang akan dibangun. Gambar 2 merupakan gambaran umum sistem.



Gambar 2 Alur Pemodelan Sistem

Proses yang akan dilakukan dalam pembangunan sistem pada tugas akhir yaitu *crawling* data yang akan dilakukan pada media sosial twitter, setelah data dikumpulkan dilakukan proses pelabelan dengan beberapa kategori, tahap berikutmya adalah proses *preprocessing*, dimana proses ini memiliki beberapa

tahap yaitu Data *Cleaning, case folding, Tokenization, Stemming, Stopword Removal, Normalization, Convert Negation*, proses selanjutnya yaitu pemisahan data yaitu data *training* dan data *testing*. Data yang telah selesai dari *preprocessing* akan dilakukan pembobotan menggunakan *TF-IDF* dari setiap kata yang terdapat didalam data tweet, selanjutnya untuk menghasilkan nilai probabilitas dari hasi *information cascade* menggunakan *random walk*. setelah semua data dilakukan pembobotan dan menghasilkan nilai berbentuk vektor, dan akan dilakukan proses klasifikasi menggunakan metode *Support Vector Machine* berdasarkan kategori pertweet.

3.2 Pengumpulan Data

Pada tahap pengumpulan data akan dilakukan dengan cara *crawling* otomatis pada twitter, proses dari *crawling* menggunakan *keyword* dapat mengambil maksimal 100 tweet dalam sekali *crawling*, berbeda dengan menggunakan *crawling account* yang bisa mengambil 200 dalam sekali *crawling* [19]. Batasan ini diterapkan pada tingkatan pengguna dan tingkatan aplikasi [21]. Kemudian data yang diambil menggunakan *keyword* yang sudah ditentukan. Berikut adalah data yang saya akan gunakan sebagai *keyword* dan data ini akan di lakukan pelabelan dijadikan sebagai *dataset* dengan beberapa katergori seperti ekonomi, sosial, politik, budaya, kesehatan, olahraga, seni, dll. Contoh pengumpulan data pada tabel 2 dan dataset pada tabel 3.

Tabel 3 Jumlah Pengumpulan Data

Keyword	Jumlah
Covid-19	1000
Work From Home	1000
New Normal	1000

Tabel 4 Contoh Dataset yang sudah dilabeli

No	*Tweet*	*Class*
1	Indonesia sebetulnya mampu menyelesaikan permasalahan melonjaknya pasien covid-19 dalam waku beberapa hari terakhir	Kesehatan
2	Pemain timnas Indonesia U-17 mempersiapkan diri untuk pertandingan Piala Dunia 2022 nanti	Olahraga
3	Ada Pesan dari Pemimpin Tentara Pembebasan Papua Sebelum Kerusuhan di Wamena, Warga Pendatang Akan Dieksekusi #GoRiau	Politik
4	Tragedi kecelakaan pesawat Sriwijaya air SJ-182 mengalami kerugian secara fisik dan materil	Ekonomi
5	Rewind youtube Indonesia 2020 menjadi trending 1 dunia karena dikatakan sangat kreatif dan menarik	Seni

3.3 *Preprossecing* Data

Preprocessing merupakan tahapan awal untuk mempersiapkan data agar dapat dipergunakan ke tahap berikutnya, pada preprocessing ini mengurangi atribut yang tidak relevan. Pada proses ini melakukan eliminasi data yang tidak sesuai atau merubah bentuk data yang tidak terstruktur menjadi sebuah data yang terstruktur dan dapat digunakan ke proses selanjutnya.

a. **Data *Cleaning***

Proses yang dilakukan yaitu menghilangkan karakter, angka-angka, *url*, *#hashtag*, tanda baca yang terdapat pada data (…,@,$,*,&,<,>,-,#,%).

b. Stop Word

Teknik menghilangkan kata-kata yang tidak berarti dan tidak berguna untuk proses klasifikasi.

c. Case Folding

Terdapat perbedaan huruf, pada tahap ini melakukan proses merubah bentuk huruf kecil (*lowercase*).

d. Tokenization

Proses membagi dokumen menjadi kata-kata atau istilah, membangun vektor kata, yang dikenal sebagai *bag-of-word*

Tabel 5 Contoh Proses *Preprocessing*

Preprocessing	***Input***	***Output***
Cleaning	@Lambeturah Covid-19 semakin meningkat pada setiap harinya dan merenggut banyak sekali nyawa pada seluruh dunia	Lambeturah Covid19 semakin meningkat pada setiap harinya dan merenggut banyak sekali nyawa pada seluruh dunia
Case Folding	Lambeturah Covid19 semakin meningkat pada setiap harinya dan merenggut banyak sekali nyawa pada seluruh dunia	lambeturah covid19 semakin meningkat pada setiap harinya dan merenggut banyak sekali nyawa pada seluruh dunia
Stopword Removal	lambeturah covid19 semakin meningkat pada setiap harinya dan merenggut banyak sekali nyawa pada seluruh dunia	lambeturah covid semakin meningkat setiap hari dan merenggut banyak nyawa seluruh dunia
Tokenization	lambeturah covid semakin meningkat setiap hari dan merenggut banyak nyawa seluruh dunia	‘lambeturah’ ‘covid’ ‘semakin’ ‘meningkat’ ‘setiap’ ‘hari’ ‘dan’ ‘merenggut’ ‘banyak’ ‘nyawa’ ‘seluruh’ ‘dunia’

3.4 Pembobotan *TF-IDF*

Tahap selanjutnya yaitu pembobotan menggunakan *TF-IDF*, pembobotan ini digunakan untuk mengkur seberapa penting kata dalam suatu dokumen. *TF* merupakan frekuensi kemunculan suatu kata dalam satu dokumen, sedangkan *IDF* merupaka ukuran dari kemampuan kata untuk melakukan pembedaan kategori. Untuk rumus perhitumgan TF, IDF dan perhitungan *TF-IDF* dapat dilihat dari rumus **(2,1)**, **(2.2)**, **(2.3)** [19]**.**

3.5 Prediksi *Information Cascade*

Pada twitter akan dapat ditemukan kesaman antara pengguna jika saling mengikuti satu sama lain dan mereka saling membalas atau meretweet, maka menggunakan *user similarity*, *Content Similarity* yaitu kesamaan tweet dihitung menggunakan *cosine similarity dapat dilihat pada persamaan* **(2.6), (2.7).**

3.6 Random Walk

Random Walks digunakan untuk memprediksi sebuah *information cascade* dengan mertimbangkan dimulai dari simpul ni dan berakhir di simpul n j maka n dapat direpresentasikan dalam persamaan **(2.4), (2.5).**

3.7 Support Vector Machine

Pada proses klasifikasi yaitu dilakukannya proses klasifikasi berupa data input yang sudah berpola/terstruktur yang berasal dari perhitungan *Random Walk* yang akan diolah menjadi vektor atau berbentuk *biner* dengan menggunakan metode *Support Vector Machine* (SVM) dengan menggunakan persaman **(2.8), (2.9), (2.10), (2.11).**

3.8 Pengujian dan Analis

Dari hasil data yang sudah diklasifikasi, kemudian dilakukannya pengujian ke model hasil klasifikasi yang diperoleh dari hasil prediksi yang dilakukan oleh *confusion matrix.*

Daftar Pustaka

[1] E. S. Pandu Adi Cakranegara, “Analisis Strategi Implementasi Media Sosia Studi Kasus UKM "XYZ",” *Journal President,* pp. 1-16, 2019.

[2] C. D. A. S. Syeda Nadia Firdaus, Retweet: A popular information diffusion mechanism – A survey paper, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada: Department of Computer Science, 2018.

[3] T. B. N. H. J. Mothe, “Predicting Information Diffusion on Twitter - Analysis of predictive features,” *Computational Science,* pp. 1-11, 2017.

[4] A. S. ,. R. S. Nidhi Singha, “Predicting Information Cascade on Twitter Using Random Walk,” *Procedia Computer Science,* vol. 173, pp. 201-209, 2020.

[5] H. a. M. G. Allcott, “Social Media and Fake News in the 2016 Election,” *Journal of Economic Perspectives,* vol. 31, no. 2, pp. 211-236, 2017.

[6] G. R. Setyawa, Model Pergerakan Harga Saham Menggunakan Random Walk dan Gerak Brown, Yogyakarta: Repository USD, 2014, pp. 6-112.

[7] M. S. M. A. Muhammad Hilman Aprilian Nurjaman, “Analisis sentimen pada ulasan buku berbahasa inggris menggunakan Information Gain dan Suport Vector Machine,” *e-Proceeding of Engineering* , vol. 4, no. 3, p. 4900, 2017.

[8] E. B. S. Isep Mumu Mubaroq, “The Effect of Information Gain Feature Selection for Hoax Identification in Twitter Using Classification Method Support Vector Machine,” *INDOJC,* vol. 5, no. 2, pp. 107-118, 2020.

[9]]. J. L. Ortega, The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations), Madrid Spain: Cybermetrics Lab, 2017.

[10] J. L. H. N. Y. F. L. W. Feng Xia, “Random Walks: A Review of Algorithms and Applications,” *IEEE.*

[11] O. Z. Tane, “Analisis Sentimen pada Twitter Tentang Calon Presiden 2019 Menggunakan Metode SVM,” *Telkom University,* 2019.

[12] E. W. Arief Wibowo, “Paper Review: Data Mining Twitter,” *ResearchGate,* 2018.

[13] R. Noviana, Fenomena Celebritism di Twitter, Makassar: CORE AC, 2018, pp. 1-134.

[14] Satista, “Number of Twitter users in Indonesia from 2014 to 2019,” https://www.statista.com/statistics/490548/twitter-users-indonesia/., 2019.

[15] E. B. S. Z. A. B. Jaka Eka Sembodo, “Data Crawling Otomatis pada Twitter,” *Researchgate,* p. 12, 2016.

[16] E. B. S. Eias Raihandtsa Mamuri, “Mendeteksi Pesan Berita Palsu (Hoax) pada Twitter dengan Algoritma AdaBoost dan ANP,” *Universitas Telkom,* p. 1, 2019.

[17] E. B. S. Y. S. Lailis Sa'adah, “Analisis Sentimen Review E-Commerce pada Twitter Menggunakan dan Metode Klasifikasi Support Vector Machine,” *Open Library,* pp. 1-14, 2020.

[18] S. Mujilahwati, “Pre-Processing Text Mining pada Data Twitter,” *Sentika,* p. 50, 2016.

[19] E. B. S. Z. K. A. B. Achmad Fauzi, “Deteksi Berita Hoax di Twitter dengan Metode Term Frequency Inverse Document Frequency dan Support Vector Machine,” *Universitas Telkom,* p. 2, 2019.

[20] S. W. Yusuf S, Random Walk Langkah Pergerakan Acak, vol. 2, Bogor: docplayer IPB, 2017, pp. 1-112.

[21] J. W. Jenks, “The guidance of public opinion,” *American Journal of Sociology,* 2015.

[22] R. S. Dina Maulina, “Klasifikasi Artikel Hoax Menggunakan Suupport Vector Machine Linier dengan Pembobotan Term Frequency – Inverse Document Frequency,” *Jurnal Mantik Penusa,* vol. 2, no. 1, p. 35, 2018.

[23] S. A. F. A. Irene Mathilda Yulietha, “KIasifikasi Sentimen Review Film Menggunakan Algoritma,” *Telkom University,* 2017.

[24] L. M. M. A. F. Dimas Joko Haryanto, “Analisis Sentimen Rivew Barang Berbahasa Indonesia dengan Metode Support Vector Machine dan Query Expansion,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 2, no. 9, pp. 2909-2916, 2018.

[25] S. F. A. I.M Yulietha, “Klasifikasi Sentimen Riview Film Menggunakan Algoritma Support Vector Machine,” *e-Proceeding,* vol. 4, no. 3, pp. 4740-4750, 2017.

[26] C. O. a. P. Ti, “Performance Evaluation of the Data Mining Classification Method,” pp. 249-253, 2015.

[27] I. T. a. I.Technology, “Review on Evaluation Metrics For Data Classification Evaluations,” *Int J. Data Min. Knowl. Manag,* vol. 5, no. 2, pp. 1-11, 2015.

[28] S. Chakrabarti, “Crawling the Web,” *Min. Web,* pp. 17-43, 2018.

[29] M. A. F. I. L. S. D. Ni Made Gita Dwi Purnamasari, “Identifikasi Tweet Cyberbullying pada Aplikasi Twitter menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur,” *e-ISSN,* vol. 2, no. 11, p. 5328, 2018.

[30] A. N. Amalia, Implementasi Support Vector Machine pada Klasifikasi Laporan Skripsi (Studi Kasus : Teknik Informatika Unikom), Bandung: elib.unikom, 2016.