**DETEKSI SARKASME DENGAN MENGGUNAKAN FITUR KLASIFIKASI**

**PROPOSAL TESIS**

****

**Oleh**

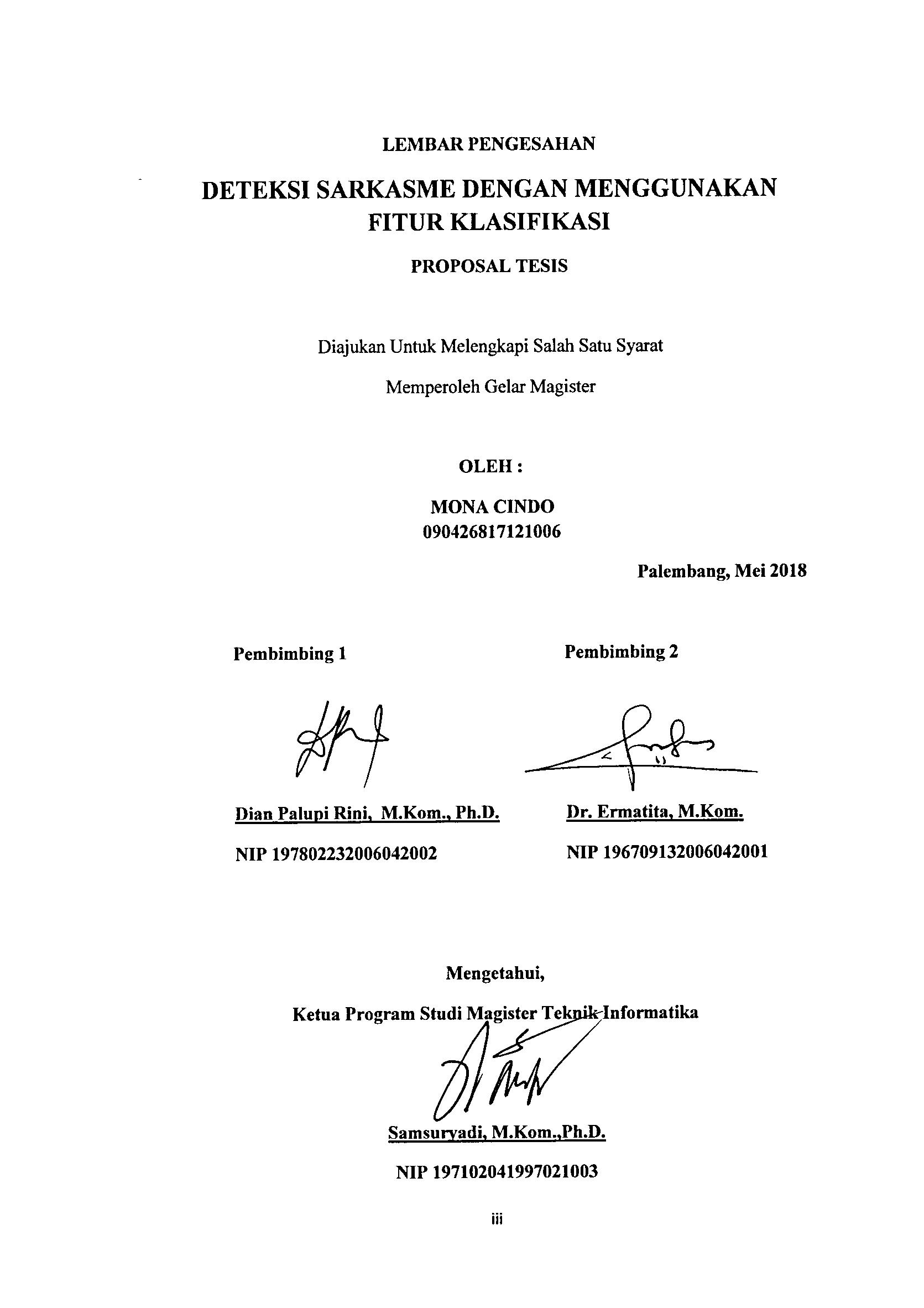
**090426817121006 MONA CINDO**

**MAGISTER TEKNIK INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS SRIWIJAYA**

**Mei 2018**

****

**DAFTAR ISI**

Halaman

**HALAMAN JUDUL**  i

**LEMBAR PENGESAHAN**  ii

**DAFTAR ISI**  iii

**BAB I PENDAHULUAN**  **1**

1.1Latar Belakang 1

1.2Perumusan Masalah 2

1.3 Batasan Masalah 3

1.4Tujuan Penelitian 3

1.5Manfaat Penelitian 3

1.6Sistematika Penulisan 3

**BAB II TINJAUAN PUSTAKA** **5**

2.1 Penelitian Terkait ............5

2.2 Landasan Teori ..............6

2.2.1Sarkasme 6

2.2.2Metode Klasifikasi 6

2.2.2.1Random Forest 6

**BAB III METODOLOGI PENELITIAN** **8**

3.1 Metodologi Penelitian 8

3.1.1 Metodologi Penelitian yang digunakan ............9

3.1.1.1 Pengumpulan Data ............9

3.1.1.2 Pra Proses Data ............9

3.1.1.3 Fitur Ekstraksi ............9

3.1.1.4 Klasifikasi ............10

3.1.1.5 Evaluasi ............11

3.1.1.6 Kesimpulan Saran ............11

3.2 Jadwal Penelitian 12

**DAFTAR PUSTAKA** 13

**BAB I**

**PENDAHULUAN**

Pada bab ini berisi latar belakang dilakukannya penelitian yang berjudul : “Deteksi Sarkasme Dengan Menggunakan Fitur Klasifikasi”. Kemudian dari latar belakang tersebut dapat dirumuskan permasalahan yang akan di angkat, agar permasalahan yang ada tidak meluas maka diberikan batasan masalah. Kemudian diberikan tujuan dan manfaat dari penelitian yang dibuat, dan metodelogi yang digunakan dalam penelitian tersebut.

* 1. **Latar Belakang**

Twitter menjadi salah satu tujuan web terbesar bagi orang-orang mengekspresikan pendapat mereka, berbagi pemikiran mereka dan melaporkan peristiwa-peristiwa real-time, dan lainnya (J. M. Soler, F Cuartero and M. Roblizo, 2012). Meningkatnya popularitas twitter dan media sosial lainnya banyak dimanfaatkan oleh perusahaan-perusahaan untuk meningkatkan layanan produknya, dan memberikan layanan *after sales* yang memungkinkan pelanggannya dapat me-*review* kualitas produknya sehingga banyak perusahaan yang mengandalkan *tools* HootSuite untuk menganalisa data dan menyediakan layanan *costumer services* seperti konten manjement, sentiment analisis. Akan tetapi *tools* ini memiliki kekurangan dalam mendeteksi bentuk kalimat sarkasme maupun humor sehingga hasil dari *tools* ini tidak selalu tepat dan jelas (Rajadesingan, 2014).Sarkasme merupakan ungkapan dari seseorang yang memiliki makna yang tidak jelas.Salah satu dari ciri sarkasme adalah sangat sulit untuk dipahami.Kesusahan dalam memahami sarkasme ini menjadikan permasalahan dalam komunikasi sehari-hari dan juga dalam sistem *Natural Language Processing* (Davidov, 2010). Beberapa paper seperti (Davidov, 2010) melakukan dan membuat algoritma yang disebut dengan *semi-supervised sarcasm identification algorithm* (SASI). Algoritma ini mempelajari tentang susunan pola per-kalimat sarkasme dan membuatnya sebagai suatu fitur untuk melakukan klasifikasi.

Berbeda dengan paper sebelumnya (M. Bouazizi and T. Otsuki, 2016) mengusulkan pendekatan berbasis pola untuk mendeteksi sarkasme di Twitter untuk mengetahui kritikan atau ejekan pada sebuah jejaring sosial twitter.Berdasarkan dari paper tersebut dapat ditarik kesimpulan, yaitu untuk dapatmempelajari dan mendeteksi sarkasme pada twitter memerlukan pehaman baik dari pola teks sarkasme maupun dari kebiasaan suatu user.Maka dari itu penulis akan melakukan suatu percobaan dengan fitur tersebut untuk mendeteksi sarkasme yang nantinya akan di uji dengan melakukan metode rendom forest. Tujuan dari penelitian ini adalah membuat suatu penelitian yang mampu mendeteksi suatu sarkasme dengan adanya penelitian ini diharapkan akan mampu meningkatkan hasil akurasi tools analisis sentiment agar mendapatkan data yang lebih akurat dan jelas.

* 1. **Perumusan Masalah**

Terdapat beberapa isu yang akan dibahas dalam penelitian ini: (i) bagaimana cara mengimplementasikan pattern-based approach, (ii) bagaimana cara melakukan klasifikasi data menggunakan metode machine learning rendom forest, (iii) bagaimana cara meningkatkan akurasi data dengan menggunakan fitur enrichment yang terdapat dalam pattern-based approach.

**1.3 Batasan Masalah**

Berdasarkan latar belakang penelitian dan rumusan masalah, maka batasan masalah pada penelitian ini, antara lain:

1. Dataset yang digunakan merupakan dataset dari twitter yang berhastag #sarkasme dan dataset yang diuji tidak diambil secara real-time.
2. Dataset yang digunakan hanya sebatas dataset sarkasme peneliti tidak menggunakan dataset mengandung ironi ataupun hiperbola.
3. Penelitian ini hanya menggunakan dataset bahasa indonesia.

**1.4 Tujuan Penelitian**

Tujuan penelitian ini adalah sebagai berikut:

1. Menghasilkan dataset sarkasme yang baik agar dapat dengan mudah dipelajari oleh sistem.
2. Melakukan proses enskripsi dataset dengan menggunakan berbasis pola setiap kalimat.
3. Menguji hasil metode yang diusulkan kemudian melakukan enrichment data agar dapat menghasilkan tingkat akurasi yang lebih maksimal.

**1.5 Manfaat Penelitian**

Manfaat yang ingin dicapai dari penelitian ini adalah sebagai berikut:

1. Dapat mencari dataset yang mempunyai makna sarkasme di dalamnya.
2. Dapat merancang metode pendekatan yang terbaik untuk mendeteksi sarkasme dalam suatu kalimat.
3. Dapat menambah dataset (enrichment data) dari data yang sudah tersedia, sehingga sistem dapat membaca setiap kemungkinan pola dalam suatu kalimat.

**1.6 Sistematika Penulisan**

Untuk mempermudah dalam penyusunan tesis dan memuat uraian secara garis besar isi dari setiap bab, maka dibuatlah sistematika penulisan:

BAB I - PENDAHULUAN

Bab ini berisi tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II - TINJAUAN PUSTAKA

Bab ini berisi tentang seluruh penjelasan mengenai tinjauan pustaka yag berhubungan dengan permsalahan yang dibahas pada penulisan tesis ini.

BAB III – METODOLOGI PENELITIAN

Bab ini berisi penjelasan secara bertahap dan terperinci tentang langkah-langkah (metodologi) yang digunakan untuk membuat kerangka berfikir dan kerangka kerja dalam meyelesaikan tesis.

BAB IV – ANALISA DAN PEMBAHASAN

Bab ini berisi tentang analisa dan pembahasan dari tiap-tiap blok diagram perencanaan rangkaian data-data hasil pengukuran.

BAB V – KESIMPULAN (SEMENTARA)

Bab ini berisi kesimpulan tentang hasil yang telah diperoleh serta merupakan jawaban dari tujuan yang ingin dicapai pada BAB I (Pendahuluan), akan tetapi masih bersifat sementara.

**BAB II**

**TINJAUAN PUSTAKA**

**2.1 Penelitian Terkait**

Davidov dkk. (Davidov et al., 2010) merancang teknik Semi Supervised untuk mendeteksi sarkasme dalam tinjauan prodak amazon dan tweets. Mereka menggunakan pola yang menarik (menggunakan kata-kata frekuensi tinggi dan kata-kata konten) dan fitur berbasis tanda baca untuk membangun model klasifikasi model menggunakan sebuah classifier K-nearest untuk melakukan deteksi sarkasme.

(M. Bouazizi dan T. Ohtsuki, 2015) mendeteksi sarkasme di Twitter dengan memanfaatkan berbagai komponen tweet. menghasilkan metode baru untuk mendeteksi sarkasme di twitter. Sarkasme adalah bentuk canggih dari ironi banyak digunakan dalam jaringan sosial dan situs microblogging. Hal ini biasanya digunakan untuk menyampaikan informasi tersirat dalam mentransmisikan pesan seseorang. Sarkasme dapat digunakan untuk tujuan yang berbeda, seperti kritik atau ejekan. Mereka mengusulkan sebuah pendekatan berbasis pola untuk mendeteksi sarkasme di Twitter. mereka mengusulkan empat set fitur yang mencakup berbagai jenis mendefinisikan sarkasme. (M. Bouazizi dan T. Ohtsuki, 2016).

Bertujuan untuk mengklasifikasikan tweet, khususnya ketika klasifikasi melampaui klasifikasi biner atau terner. Dengan mengusulkan pendekatan berbasis pola untuk kuantifikasi sentimen di Twitter. Dengan kuantifikasi, Penulis merujuk pada pendeteksian sentimen-sentimen yang ada di dalam tweet dan pendeteksian berat dari sentimen-sentimen tersebut. Dengan mengklasifikasikan tweet menjadi positif, negatif, atau netral (M. Bouazizi dan T. Ohtsuki, 2016). Memperkenalkanmetodependekatansarkasmebaruyaitu SCUBA (Sarcasm Classification Using a Behavorialmodelin Approach) merupakanpendekatan yang menganalisahinggakehistoritwitterseorang user agar dapatmemahamikebiasaan user tersebutdalammenuliskalimat(A.Rajadesingan, R. Zafarani and H. Liu, 2015). Dalam penelitian (S. K. Bharti, R. Pradhan, K.S Babu, and S.K Jena) bertujuan untuk mengidentifikasi sarkasme dalam tweet dan menganalisa kinerja dari berbagai pengklasifikasi, yaitu Naive Bayes, SVM, Maximum entropy dan Decision tree untuk menganalisis kinerja berbagai pendekatan machine learning.

**2.2 Landasan Teori**

**2.2.1 Sarkasme**

Sarkasme merupakan suatu acuan yang lebih kasar dari ironi dan sinisme. Ia adalah suatu acuan yang mengandungkepahitan dan celaaan yang getir (Keraf, 2000:143). Contoh : Mulut kau harimau kau. Lihat sang raksasa itu (maksudnya si cebol).

**2.2.2Metode Klasifikasi**

**2.2.2.1 Random forest**

Random forest (atau random forests) adalah sekumpulan classifier yang terdiri dari banyak pohon keputusan dan melakukan klasifikasi berdasarkan keluaran dari hasil klasifikasi setiap pohon keputusan anggota. Metode random forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection (Breiman 2001). Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas n amatan dan p peubah penjelas, random forest dilakukan dengan cara (Breiman 2001; Breiman & Cutler 2003):

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.

2. Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut.Tahapan ini adalah tahapan random feature selection.

3. Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon. Respons suatu amatan diprediksi dengan menggabungkan (aggregating) hasil prediksi k pohon. Pada masalah klasifikasi dilakukan berdasarkan majority vote (suara terbanyak). Error klasifikasi random forest diduga melalui error OOB yang diperoleh dengan cara (Breiman 2001; Breiman & Cutler 2003; Liaw & Wiener 2002):

1. Lakukan prediksi terhadap setiap data OOB pada pohon yang bersesuaian. Data OOB (out of bag) adalah data yang tidak termuat dalam contoh bootstrap.

2. Secara rata-rata, setiap amatan gugus data asli akan menjadi data OOB sebanyak sekitar 36% dari banyak pohon. Oleh karena itu, pada langkah 1, masing-masing amatan gugus data asli mengalami prediksi sebanyak sekitar sepertiga kali dari banyaknya pohon. Jika a adalah sebuah amatan dari gugus data asli, maka hasil prediksi random forest terhadap a adalah gabungan dari hasil prediksi setiap kali a menjadi data OOB.

3. Error OOB dihitung dari proporsi misklasifikasi hasil prediksi random forest dari seluruh amatan gugus data asli.

**BAB III**

**METODOLOGI PENELITIAN**

* 1. **Metodologi Penelitian**
     1. **Metodologi Penelitian yang digunakan**

Dalam penelitian ini penulis berencana menggunakan 5 tahap metode, seperti paada Gambar 3.1.Tahap pertama merupakan tahap pengumpulan data.Data yang digunakan merupakan data jejaring social Twitter yang menggunakan hastag #sarkasme.Setelah itu langkah selanjutnya adalah praproses yang merupakan normalisasi data sekaligus pembuangan stopwords.Normalisasi teks bertujuan untuk mengubah tweet menjadi lebih simple dan mudah di mengerti. Setelah pra proses maka tahapan selanjutnya merupakan ekstraksi fitur, tahapan ini terbagi empat tahapan yaitu: sentiment-related feature, punctuation-related feature, syntactic/semantic feature, dan pattern related feature. Pada bagian selanjutnya merupakan bagian klasifikasi sarkasme. Metode yang akan digunakan dalam penelitian ini merupakan metode random forest. Selanjutnya tahapan terakhir merupakan evaluasi dari hasil klasifikasi.



Gambar 3.1 Tahapan metode yang diusulkan

**3.1.1.1Pengumpulan Data**

Pada bagian ini penulis melakukan pengumpulan data dengan menggunakan twitter API, adapun data yang akan di ambil terbagi 2 set yaitu:

1. **Training set**

Dataset ini merupakan dataset yang berisi tentang koleksi datasetsarkasme dan dataset yang bukan sarkasme. Pada data ini akandilakukan pemberian skor secara manual. Skor yang diberikanadalah 1 sampai dengan 6 dimana 1 merupakan twitter yang pasti bukan sarkasme dan 6 merupakan skor untuk twitter yang pastisarkasme. Dataset ini akan digunakan sebagai model training pada saat klasifikasi nanti.

1. **Test set**

Data ini merupakan data yang berbentuk raw dan acak yang akan digunakan untuk melakukan evaluasi pada tahapan proses klasifikasi nantinya.

**3.1.1.2 Pra Proses Data**

Pada proses ini dilakukan normalisasi data yang bertujuan agarmempermudah proses analisis terhadap entitas-entitas mengingatbanyak kata yang tidak baku seperti singkatan, tanggal, jumlah mata uang. Selain itu penulis juga berencana membuang user pada data set twitter seperti @monacindo, dan juga URL (http-address) dan juga membuang twitter yang diluar dari Bahasa Indonesia dan memiliki jumlah suku kata kurang dari 3.

* + - 1. **Fitur Ekstraksi**

Pada fitur ekstraksi ini penulis memiliki 4 fitur yaitu:

1. **Sentiment-related Feature**

Bentuk sarkasme yang banyak digunakan baik dalam percakapan biasa maupun pesan singkat.Sarkasme merupakan ekspresi emosi negative yang digunakan dalam konteks negatif.Sebagai contoh (“terima kasih selalu ada disisiku #akubencikamu”) merupakan bentuk kata-kata yang bukan bertujuan untuk berterima-kasih tetapi menunjukan suatu kekecewaan.Atas dasar ini penulis menggunakan fitur sentiment ini untuk mendeteksi kalimat positif maupun negatif.

1. **Punctuation-related Feature**

Fitur ini merupakan fitur yang menghitung jumlah panjang kata-kata dalam suatu kalimat dan juga jumlah tanda baca pada kalimat seperti tanda kutip, seru, tanda tanya, huruf kapital pada kalimat.

1. **Syntatic/semantic-related Feature**
2. Penggunaan kata-kata yang tidak umum
3. Nomor dari kata-kata yang tidak umum
4. Eksistensi dari ekspresi sarkasme yang umum
5. Jumlah dari penggunaan kata seru
6. Jumlah dari ekspresi ketawa
7. **Patter-related Feature**

Bagian ini merupakan fitur utama yang akan diajukan oleh peneliti yaitu fitur relasi pola kalimat. Pada bagian berfungsi untuk membedakan kata-kata menjadi dua kategori yaitu kata berfrekuensi tinggi dan kata-kata konten berdasarkan frekuensi kemunculan mereka dalam pernyataan dan mendefinisikan pola sebagai urutan urutan kata-kata frekuensi tinggi.Kelas pertama merupakan “CI” berisi kata-kata yang dinilai penting dan kelas kedua adalah “GFI” berisi kata-kata yang penting secara gramatikal.

**3.1.1.4 Klasifikasi**

Klasifikasi menggunakan random forest yaitu proses klasifikasi dilakukan terhadap data tweet hasil praproses untuk mengklasifikasikan entitas berdasarkan respon terhadap entitas tersebut atau analisis sarkasme. Pada penelitian ini terdiri atas 2 kelas, yaitu sarkasme dan tidak sarkasme. Tweet yang termasuk di dalam sarkasme contohnya: @BeresNews: Mulut kau harimau kau. Lihat sang raksasa itu (maksudnya si cebol). Yang bukan sarkasme: @erikadrenata: Pagi semua jangan lupa senyum ya.. Sebelum dilakukan klasifikasi, data dibagi 2 menjadi data latih dan data uji. Data latih digunakan untuk memberi pengetahuan kepada sistem tentang ciri-ciri suatu sarkasme. Untuk mengambil ciri-ciri suatu sarkasme, dilakukan proses ekstraksi fitur yang di dalamnya terdapat pembobotan kata berdasarkan kemunculan kata. Data uji diperoleh dari data yang tersisa pada proses sebelumnya. Fungsi klasifikasi yang digunakan adalah fungsi klasifikasi dengan Random Forest. Penggunaan klasifikasi Random Forest dikarenakan prosesnya yang sederhana dan mudah diaplikasikan pada berbagai keadaan serta tidak akan mengalami kegagalan secara keseluruhan pada hasilnya.

**3.1.1.5 Evaluasi**

Tahap terakhir merupakan evaluasi. Tahapan ini bertujuan untuk mengetahui tingkat akurasi dari hasil klasifikasi data pada data test. Perhitungan akurasi data dilakukan dengan membandingkan antara hasil klasifikasi dengan menggunakan random forest dan klasifikasi secara manual.

**3.1.1.6 Kesimpulan dan Saran**

Kesimpulan dan Saran.Tahap ini merupakan tahap terakhir, pada tahap ini dirumuskan suatu kesimpulan berdasarkan permasalahan, studi pustaka, metodologi penelitian dan hasil pengujian. Kemudian akan diberikan beberapa saran yang dapat dijadikan landasan untuk penelitian-penelitan selanjutnya.

**3.2 Jadwal Penelitian**

Penelitian akan dilaksanakan pada rentang waktu bulan agustus 2018 sampai dengan bulan februari 2019.adapun jadwal kegiatan adalah sebagai berikut :

**Tabel 3.1 Jadwal Penelitian**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Waktu Penelitian (Tahun 2018/2019) | | | | | | | |
| Agustus | September | Oktober | November | Desember | Januari | Februari | Maret |
| 1. | Tahap Persiapan Penelitian |  |  |  |  |  |  |  |  |
| 1. Penyusunan dan pengajuan judul |  |  |  |  |  |  |  |  |
| 1. Pengajuan proposal |  |  |  |  |  |  |  |  |
| 1. Perijinan penelitian |  |  |  |  |  |  |  |  |
| 2. | Tahap Pelaksanaan |  |  |  |  |  |  |  |  |
| 1. Pengumpulan Data |  |  |  |  |  |  |  |  |
| 1. Pra Proses Data |  |  |  |  |  |  |  |  |
| 1. Fitur Ekstraksi |  |  |  |  |  |  |  |  |
| 1. Klasifikasi |  |  |  |  |  |  |  |  |
| e. Evaluasi |  |  |  |  |  |  |  |  |
| 3. | Tahap penyusunan laporan |  |  |  |  |  |  |  |  |

**DAFTAR PUSTAKA**

A. Rajadesingan, R. Zafarani and H. Liu, “Detecting Sarcasm on Twitter: A Behavior Modeling Approach” (2015).

Breiman L, Cutler A. 2001. Random Forest. [terhubung berkala]. http://www.stat.berk eley.edu/~breiman/RandomForests/cc\_home.ht m#intro [20 April 2018].

Breiman L, Cutler A. 2003. Manual on Setting Up, Using, and Understanding Random Forest V4.0. [terhubung berkala]. http://oz.berkeley.edu/users/bre iman/Using\_random\_forests\_v4.0.pdf [20 April 2018].

Keraf, Gorys. Diksi dan Gaya Bahasa, Jakarta: Rinneka Cipta, (2000).

Liaw A, Wiener M. Des, “Classification and Regression by randomForest. RNews “Vol. 2/3:18-22, (2002).

M. Bouazizi and T. O. Ohtsuki, ‘‘A pattern-based approach for sarcasm detection on Twitter,’’ IEEE Access, vol. 4, pp. 5477–5488, (2016).

M. Bouazizi and T. Ohtsuki, ‘‘Sentiment analysis in Twitter: From classification to quantification of sentiments within tweets,’’ in Proc. IEEE GLOBECOM, pp. 1–6 (2016).

S. K. Bharti, R. Pradhan, K.S Babu, and S.K Jena, Sarcasm Analysis on Twitter Data Using Machine Learning Approaches”, (2017).

Tsur, O., D. Davidov and A. Rappoport, “Icwsm-a great catchy name: Semisupervised recognition of sarcastic sentences in online product reviews.”, in “ICWSM”, (2010).