SKEMA PENDANAAN: PENELITIAN MADYA

**LAPORAN KEMAJUAN PENELITIAN**



**ANALISA SEMIOTIKA MAKNA NASIONALISME MELALUI *TEXT MINING* PADA MEDIA SOSIAL TWITTER DI KEJUARAAN AFF TAHUN 2020**

**PENELITI:**

**Jeanie Annissa, S.IP., M.Si (100045)**

**FAKULTAS KOMUNIKASI DESAIN KREATIF**

**UNIVERSITAS BUDI LUHUR**

**JUNI 2022**

# HALAMAN PENGESAHAN

|  |  |  |
| --- | --- | --- |
| **Judul Penelitian** | : | Analisa Semiotika Makna Nasionalisme Melalui *Text Mining* Pada Media Sosial Twitter Di Kejuaraan AFF Tahun 2020 |
| **Bidang Penelitian** | : | Komunikasi Massa |
| **Ketua Peneliti** |  |  |
| 1. Nama Lengkap
 | : | Jeanie Annissa,S.IP.,M.Si |
| 1. NIP/NIDN/ID-SINTA
 | : | 100045/0328068503/6693721 |
| 1. Jabatan Fungsional
 | : | Asisten Ahli |
| 1. Program Studi
 | : | Ilmu Komunikasi |
| 1. Nomor HP
 | : | 081584785075 |
| 1. Alamat e-mail
 | : | jeanie.annissa@budiluhur.ac.id |
| **Anggota Peneliti (1)** |  |  |
| 1. Nama Lengkap
 | : | - |
| 1. NIP/NIDN/ID-SINTA
 | : | - |
| **Anggota Peneliti (2)** |  |  |
| 1. Nama Lengkap
 | : | - |
| 1. NIP/NIDN/ID-SINTA
 | : | - |
| **Mahasiswa yang terlibat** |  |  |
| 1. Nama Lengkap
 | : | - |
| 1. NIM
 | : | - |
| **Lama Penelitian** | : | 6 Bulan |
| **Biaya Penelitian** |  |  |
| 1. Sumber Universitas Budi Luhur
 | : | Rp. 6.500.000 |
| 1. Sumber Lain
 | : | Rp.  |

 Jakarta, 3 Juni 2022

Mengetahui,

Dekan Fakultas Komunikasi Desain dan Kreatif Ketua Peneliti

(Dr. Nawiroh Vera, M.Si) (Jeanie Annissa, S.IP., M.Si)

NIP: 030558 NIP: 100045

Menyetujui,

Direktur Riset dan Pengabdian Kepada Masyarakat

( Dr. Krisna Adiyarta M. )

NIP: 890001

# RINGKASAN

Penelitian ini membahas tentang analisa semiotika makna nasionalisme pada kejuaraan *Asean Football Federation* 2020 di akhir tahun 2021. Pada penelitian ini menggunakan paradigma konstruktivis yaitu pembentukan makna yang terjadi dari realita sosial yang ada. Teknik pengumpulan data primer melalui observasi *non-participant* melalui *text mining* pada media sosial twitter dengan menggunakan metodologi *Cross-Industry Standard Process For Data Mining* (CRISP-DM) dengan bantuan perangkat *rapid miner* untuk pengolahan datanya. Pada Metodologi CRISP-DM tahapan yang dilakukan meliputi pemrosesan text dan penggunaan teknik *clustering* K-Means. Hasil dari pemrosesan *Text Mining* adalah *clustering* text yaitu pengelompokan data yang memiliki kesamaan. Selanjutnya hasil *clustering* dianalisa menggunakan teori semiotika *triangle meaning* yang terdiri dari *sign*, *object*, dan *interpretant*. *Sign* dilakukan dengan melihat tanda dari *clustering* teks yang akan dihasilkan. *Object* adalah penjelasan atas tanda yang didapat dari *sign*. *Interpretant* merupakan tafsiran makna dari *sign* yang diperoleh. Penelitian ini diharapkan dapat memperoleh makna nasionalisme pada Kejuaran AFF 2020 melalui tanda-tanda berupa tekstual dari media sosial twitter.

**Kata Kunci:** semiotika, *text mining*, nasionalisme, piala aff, twitter

# PRAKATA

Puji syukur kehadirat Tuhan Yang Maha Esa, yang telah memberikan rahmat dan karunia-nya sehingga peneliti dapat menyelesaikan penyusunan penelitian yang berjudul “Analisa Semiotika Makna Nasionalisme Melalui *Text Mining* Pada Media Sosial Twitter Di Kejuaraan AFF Tahun 2020”

Dalam proses penyusunan penelitian ini, peneliti mengalami berbagai hambatan dan kesulitan, namun atas pertolongan-Nya serta dukungan, semangat dan bimbingan dari berbagai pihak sehingga peneliti ini dapat terselesaikan. Oleh karena itu, melalui kesempatan ini perkenakan peneliti untuk menyampaikan ucapan terima kasih sebesar-besarnya dan penghargaan yang setinggi-tingginya kepada yang terhormat:

1. Dr. Ir. Wendi Usino, M.Sc., MM., selaku Plt. Rektor Universits Budi Luhur
2. Dr. Nawiroh Vera, M.Si. selaku Dekan Fakultas Ilmu Komunikasi Universitas Budi Luhur
3. Dr. Umaimah Wahid, selaku Ketua Program Studi Magister Ilmu komunikasi
4. Bintarto Wicaksono, S.PT, M.Sn., selaku ketua Program Studi Komunikasi Universitas Budi Luhur
5. Rini Lestari, M.I.Kom, selaku Sekretaris Program Studi Ilmu Komunikasi Universitas Budi Luhur.
6. Amin Aminudin, M.I.Kom, selaku sekretaris Progam Studi Magister Ilmu Komunikasi.
7. Doddy Wihardi, M.I.Kom, selaku Kepala Konsentrasi Public Relations Program Studi Ilmu Komunikasi Universitas Budi Luhur.
8. Benny Muhdaliha, M.Sn. Ketua Program Desain Komunikasi Visual
9. Haronas Kutanto, S.PT., M.I.Kom., selaku Kepala Konsentrasi *Broadcast Journalism* Universitas Budi Luhur
10. Ricky Widyananda Putra, S.Ikom., M.Sn. selaku Kepala Konsentrasi Game dan Animasi
11. Artyasto Jati Sidi, M.I.Kom. Kepala Konsentrasi *Digital Advertising*

Mengingat keterbatasan pengetahuan dan pengalaman yang dimiliki, peneliti menyadari bahwa Proposal Penelitian ini masih belum sempurna maka kritik dan saran yang membangun sangat peneliti harapkan. Semoga tulisan ini bermanfaat bagi pembaca pada umumnya dan pihak-pihak yang membutuhkan.

Jakarta, 31 Januari 2022

Jeanie Annissa, S.IP., M.Si

# DAFTAR ISI

Halaman

[HALAMAN PENGESAHAN i](#_Toc105172928)

[RINGKASAN ii](#_Toc105172929)

[PRAKATA iii](#_Toc105172930)

[DAFTAR ISI iv](#_Toc105172931)

[DAFTAR TABLE v](#_Toc105172932)

[DAFTAR GAMBAR vi](#_Toc105172933)

[DAFTAR LAMPIRAN vii](#_Toc105172934)

[BAB 1. PENDAHULUAN 1](#_Toc105172935)

[1.1. Latar Belakang 1](#_Toc105172936)

[1.2. Masalah Penelitian 3](#_Toc105172937)

[1.3. Tujuan dan Manfaat Penelitian 3](#_Toc105172941)

[1.4. Target Luaran Penelitian 4](#_Toc105172944)

[BAB 2. TINJAUAN PUSTAKA 5](#_Toc105172945)

[2.1. Teori Semiotika 5](#_Toc105172946)

[2.2. Definisi *Data Mining* 7](#_Toc105172947)

[2.3. *Cross-Industry Standard Process For Data Mining (CRISP-DM*) 7](#_Toc105172948)

[2.4. Definisi *Text Mining* 8](#_Toc105172949)

[2.5. Tahapan *Pre-Processing* 9](#_Toc105172950)

[2.6. Algoritma K-Means 11](#_Toc105172951)

[2.7. RapidMiner 11](#_Toc105172952)

[BAB 3. METODE PENELITIAN 13](#_Toc105172953)

[3.1. Metode Penelitian 13](#_Toc105172954)

[3.2. Tahapan Penelitian 13](#_Toc105172955)

[3.3. Teknik Pengumpulan Data 14](#_Toc105172956)

[3.4. Teknik Analisa Data 17](#_Toc105172959)

[BAB 4. HASIL YANG DICAPAI 18](#_Toc105172960)

[4.1. Koleksi Data (*Data Collection*) 18](#_Toc105172961)

[4.2. Persiapan Data (Data Preparation) 19](#_Toc105172962)

[4.3. Pemodelan (Modelling) 22](#_Toc105172966)

[4.4. Evaluasi (Evaluation) 27](#_Toc105172970)

[DAFTAR PUSTAKA 28](#_Toc105172971)

[LAMPIRAN-LAMPIRAN 30](#_Toc105172972)

# DAFTAR TABLE

Halaman

[Tabel 1. Jumlah Data Tweet Hasil *Crawling* 18](#_Toc105172973)

[Table 2. Sample Isi Data *Crawling* 18](#_Toc105172974)

[Table 3. Sample Perubahan Sebelum dan Sesudah Preprocessing 20](#_Toc105172975)

[Tabel 4. Sampel Data Tweet 21](#_Toc105172976)

[Tabel 5. Penentuan Bobot *Term* 21](#_Toc105172977)

[Table 6. Sample Data Wordcloud 23](#_Toc105172978)

[Table 7. Sampe Data Tweet Cluster 1 25](#_Toc105172979)

[Table 8. Samle Data Tweet Cluster 2 25](#_Toc105172980)

[Table 9. Sampe Data Tweet Cluster 3 26](#_Toc105172981)

[Table 10. Sampe Data Tweet Cluster 4 26](#_Toc105172982)

[Table 11. Sample Data Tweet Cluster 5 26](#_Toc105172983)

# DAFTAR GAMBAR

Halaman

[Gambar 1. Triangle Meaning Charles S. Pierce 7](#_Toc105172984)

[Gambar 2. Contoh *Remove Duplicate* 9](#_Toc105172985)

[Gambar 3. Contoh *Replace* 9](#_Toc105172986)

[Gambar 4. Contoh *Case Folding* 10](#_Toc105172987)

[Gambar 5. Contoh *Tokenizing* 10](#_Toc105172988)

[Gambar 6. Contoh Stop Removing 11](#_Toc105172989)

[Gambar 7. Tahapan Penelitian 13](#_Toc105172990)

[Gambar 8. Wordcloud 24](#_Toc105172991)

[Gambar 9. Elbow Method 25](#_Toc105172992)

# DAFTAR LAMPIRAN

Halaman

[Lampiran 1. Biodata Peneliti 30](#_Toc105172993)

[Lampiran 2. Catatan Harian 32](#_Toc105172994)

# PENDAHULUAN

## Latar Belakang

Komunikasi di era digital telah mengalami perkembangan dalam penyampaian pesan dan efek pesan yang dapat ditimbulkan di masyarakat. Meskipun komunikator dan komunikan berinteraksi dalam ruang maya, *platform* media baru sebagai sarana berkomunikasi dan berinteraksi dapat memberikan dampak besar terhadap situasi dan kondisi yang terjadi di masyarakat. Melalui keberagaman jenis *platform* media yang terkategorisasi berdasarkan karakteristik pengguna dapat mempengaruhi pesan-pesan tertentu yang berdampak pada karakter bangsa.

Twitter misalnya, menurut data di laman resmi Kominfo dikategorisasikan sebagai sebuah *platform* media yang digunakan paling sering di Indonesia, dengan peringkat negara kelima terbesar sebagai pengguna Twitter di dunia, dengan total sebanyak 19,5juta pengguna. Menurut data antara *news*, Managing Director Twitter wilayah India, Timur Tengah, Afrika dan Asia Tenggara yakni Parminder Singh menyatakan bahwa terdapat tiga karakteristik pengguna twitter di Indonesia. *Pertama*, pengguna *entrepreneurial* yang memanfaat twitter sebagai media berwirausaha secara kreatif. *Kedua*, Pengguna Ekspresif yang mampu memanfaatkan video dengan *all out* karena 40 persen pengguna Indonesia mengunggah video. *Ketiga*, pengguna twitter sangat *engaged* atau aktif terhadap topik-topik yang berkembang, walaupun ketersediaan penggunaan kata hanya 140 karakter dalam sekali unggah tetapi pengguna Indonesia selalu kreatif dan efektif dalam memanfaatkannya untuk menyampaikan pesan. Berdasarkan data tersebut dapat kita ketahui bahwa Twitter merupakan salah satu *platform* media *digital* yang sangat diminati di Indonesia dan menjadi wadah komunikasi yang ekspresif dalam menyampaikan pesan-pesan tertentu terhadap sebuah topik yang sedang berkembang di masyarakat (Prabancono, 2015).

Memasuki akhir tahun 2021, Masyarakat Indonesia antusias terhadap penyelenggaraan Kejuaraan AFF (*Asean Football Federation*), yakni sebuah kejuaraan sepak bola internasional antar negara di kawasan Asia Tenggara. Topik kegiatan tersebut menyita perhatian masyarakat Indonesia khususnya setelah kemenangan Tim Indonesia sebagai pemenang *runner-up* melawan Thailand dengan skor 6-2. *Euphoria* ini menjadi topik bahasan yang menarik di media sosial termasuk twitter. Dalam kondisi ini, olah raga sepak bola dapat memicu rasa nasionalisme masyarakat yang bangga atas pencapaian tim nasionalnya. Untuk menguji potensi tersebut, maka diperlukan sebuah pola *text mining* untuk menganalisa makna nasionalisme melalui analisis semiotika.

*Data mining* adalah proses menggali kumpulan data besar untuk mendapatkan pola (*pattern*) dan aturan (*rule*) yang berarti dengan menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning*. *Text mining* merupakan salah satu rumpun ilmu komputer dari *data mining* yang digunakan untuk mengekstrak informasi berupa teks dari berbagai sumber berbeda. Dalam *text mining* dapat dilakukan kegiatan untuk menganalisa teks yang disebut dengan *text analytic*. *Text analytic* adalah kegiatan untuk mengkonversi data yang tidak terstruktur menjadi data yang memiliki arti sehingga dapat digunakan untuk mengindentifikasi pola, tema, dan topik. Penerapan yang dapat dilakukan dengan *Text Mining* antara lain ekstraksi informasi (*information* *extraction*), pelacakan topik (*topic tracking*), perangkuman (*summarization*), kategorisasi (*categorization*), penggugusan (*clustering*), penautan konsep (*concept* *linking*), dan penjawaban pertanyaan (*question answering*).

Penggugusan (*clustering*) merupakan salah satu penerapan yang bisa dilaksanakan terhadap *Text Mining*. *Clustering* adalah pengelompokan data tekstual menjadi beberapa cluster. Sebuah cluster merupakan sekumpulan objek yang digabung karena memiliki kedekatan atau kesamaan. Ketika melakukan pengelompokan data yang tidak terstruktur dan tidak diketahui bagaimana cara mengelompokan data tersebut maka teknik *clustering* merupakan cara yang tepat untuk melakukan pengelompokan. Data yang diolah pada teknik cluster menggunakan teknik pembobotan kata (*term weighting*) untuk menghitung frekuensi kemunculan setiap kata atau *term*. Terdapat berbagai macam teknik *clustering* yang dapat digunakan, teknik tersebut terkait dengan Algoritma yang akan digunakan untuk melakukan *clustering* data. *Clustering* dilakukan dengan cara memberikan bobot pada setiap kata yang muncul, selanjutnya dengan menggunakan sebuah algoritma akan dikelompokan berdasarkan kemungkinan kluster yang terbentuk. Algoritma yang banyak digunakan adalah algoritma K-Means, algoritma k-means bekerja dengan cara melakukan *clustering* berdasarkan kemungkinan K yang dapat diberikan.

Dalam melakukan teknik *clustering* maka diperlukan perangkat lunak yang menyediakan fitur untuk memproses text dan *clustering*. Pemrosesan text seperti mengkonversi data tidak terstruktur menjadi data yang memiliki arti untuk dilakukan teknik *clustering*. RapidMiner merupakan perangkat lunak open-source yang menyediakan fitur prosesing bahasa, statistic, dan teknik machine learning, dan lainnya terkait dengan data mining.

*Text Mining* untuk digunakan dalam proses *clustering,* memungkinkan untuk dilakukan kategorisasi makna pada setiap teks yang telah diolah untuk dimasukan ke dalam penggunaan semiotika Pierce. Hasil dari proses *clustering* dapat digunakan untuk mengetahui secara jelas *sign* (tanda) dari teks yang digunakan, objek yang menjadi pembahasan, dan interpretasi yang saling berkolerasi untuk memaknai nilai-nilai yang mengandung unsur nasionalisme dalam perhelatan kejuaraan AFF 2020.

## Masalah Penelitian

### Identifikasi Masalah

Permasalahan penelitian mengacu kepada penggunaan semiotika untuk melihat makna nasionalisme yang dilakukan melalui pengolahan data pengguna twitter dengan menggunakan teknik text mining dengan pola clustering untuk melihat tanda, objek, dan interpretasi melalui kegiatan kejuaraan piala AFF di tahun 2020.

### Batasan Masalah

Permasalahan penelitian ini akan dibatasi pada pencarian makna nasionalisme menggunakan teori semiotika C.S.Pierce yang diolah dari koleksi data text twitter dalam bentuk klasterisasi.

### Perumusan Masalah

Berdasarkan latar belakang yang telah peneliti buat, maka peneliti merumuskan masalah dengan pertanyaan penelitian sebagai berikut: **Bagaimana analisis semiotika pada makna nasionalisme pada kejuaraan AFF 2020 melalui media sosial twitter dengan menggunakan *text* *mining*?**

## Tujuan dan Manfaat Penelitian

### Tujuan Penelitian

Tujuan penelitian yang hendak dicapai dalam penelitian ini adalah sebagai berikut:

1. Untuk mengetahui pengembangan ilmu komunikasi dan teknologi (*new* *media*).
2. Untuk mengetahui representasi makna nasionalisme pada media sosial twitter

### Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah:

1. Secara Teoritis

Secara teoritis penelitian ini diharapkan dapat ikut serta menambah kajian mengenai penelitian yang terkait semiotika melalui media sosial

1. Secara Praktis

Hasil penelitian ini diharapkan menjadi materi pengembangan keilmuan sebagai sebuah pemikiran baru dalam menambah informasi kepada pembaca guna mengetahui bahasan representasi makna nasionalisme melalui media sosial pada kejuaraan AFF 2020. Data diperolah diperoleh dari penambangan tweet di social media twitter (*crawling*), kemudian data diproses untuk dianalisa dengan menggunakan teori semiotika.

## Target Luaran Penelitian

Target luaran yang direncanakan adalah memasukan hasil penelitian ke dalam bentuk jurnal ilmiah yang berstandarisasi sinta.

# TINJAUAN PUSTAKA

## Teori Semiotika

Semiotika adalah semiosis sebagai hubungan diantara tanda,benda, dan arti. Dalam buku Teori Komunikasi dijelaskan model *triadic* dari Peirce sering juga disebut sebagai *“triangle meaning semiotics”* atau dikenal dengan teori segitiga makna, yang dijelaskan secara sederhana: “tanda adalah sesuatu yang dikaitkan seseorang untuk sesuatu dalam beberapa hal atau kapasitas (Littlejohn & Foss, 2011). Tanda menunjuk pada seseorang, yakni, menciptakan di benak orang tersebut suatu tanda yang setara, atau suatu tanda yang lebih berkembang, tanda yang menciptakannya dinamakan *interpretant* dari tanda pertama. Tanda itu menunjukan sesuatu, yakni objeknya” (Nawiroh, 2014).

Tanda- tanda (*signs*) adalah basis dari seluruh komunikasi. Manusia dengan perantaraan tanda-tanda dapat melakukan komunikasi dengan sesamanya. Semiotika adalah suatu ilmu atau metode analisis untuk mengkaji tanda. Tanda- tanda adalah perangkat yang kita pakai dalam upaya berusaha mencari jalan di dunia ini, di tengah- tengah manusia dan bersama- sama manusia (Sobur, 2013).

Semiotika pada dasarnya hendak mempelajari bagaimana kemanusiaan (*humanity*) memaknai hal-hal (*things*). Memaknai (*to sinify*) dalam hal ini tidak dapat dicampur-adukkan dengan mengkomunikasikan (*to communicate*) (Sobur, 2013). Analisis semiotika berupaya menemukan makna tanda termasuk hal-hal yang tersembunyi di balik sebuah tanda (teks, gambar,dll). Karena sistem tanda sifatnya amat kontekstual dan bergantung pada pengguna tanda tersebut. Pemikiran pengguna tanda merupakan hasil pengaruh dari berbagai konstruksi sosial dimana pengguna tanda tersebut berbeda. Misalnya, Apa makna sosial lirik lagu? Mengapa berita menggunakan frase atau kalimat tertentu ketika menggambarkan kelompok tertentu? dan sebagainya (Kriyantono, 2014).

Semiotika komunikasi menekankan pada produksi teori tanda. Semiotika mempunyai tiga bidang utama, yaitu:

Tanda itu sendiri, terdiri atas aturan tentang berbagai tanda yang berbeda, cara-cara tanda yang berbeda dalam menyampaikan makna dan cara tanda-tanda itu terkait dengan manusia yang menggunakannya.

Kode atau sistem yang mengorganisasikan tanda. Studi ini mencakup cara berbagai kode dikembangkan untuk memenuhi kebutuhan suatu masyarakat atau budaya atau untuk mengeksploitasikan selama komunikasi yang tersedia untuk mentransmisikannya.

Kebudayaan tempat kode dan tanda bekerja. Pada gilirannya bergantung pada penggunaan kode-kode dan tanda-tanda tersebut untuk keberadaan dan bentuknya sendiri.

Simbol merupakan tanda yang memiliki hubungan dengan objeknya berdasarkan konvensi, kesepakatan atau aturan. Simbol adalah sebuah label arbiter atau representasi dari sebuah fenomena. Kata adalah simbol untuk konsep dan benda. Label dapat bersifat ambigu, dapat berupa verbal dan nonverbal dan dapat terjadi dalam komunikasi tatap muka dan komunikasi dengan menggunakan media. Simbol merupakan gerakan, gambar atau objek yang memiliki budaya yang sama. Simbol pada dasarnya telah disepakati bersama dalam sebuah kelompok, tetapi tidak jarang sebuah simbol tidak dimengerti di luar lingkup kelompok tertentu. Oleh karena simbol disebut arbiter. Simbol merupakan tanda yang memiliki hubungan dengan obyeknya berdasarkan konvensi, kesepakatan, aturan kata-kata umumnya adalah simbol. Simbol juga berarti sesuatu yang diberi makna oleh manusia, yang mereka gunakan untuk berkomunikasi. Simbol itu berdiri dari gerak isyarat, bahasa, norma, nilai, sanksi, adat istiadat dan peraturan rakyat.

Menurut teori semiotika Charles Sanders Pierce, semiotika didasarkan pada logika, karena logika mempelajari bagaimana orang bernalar, sedangkan penalaran menurut Pierce dilakukan melalui tanda-tanda. Tanda-tanda ini menurut Pierce memungkinkan kita berpikir, berhubungan dengan orang lain dan memberi makna pada apa yang ditampilkan oleh alam semesta . Dalam hal ini manusia mempunyai keanekaragaman akan tanda-tanda dalam berbagai aspek di kehidupanya. Dimana tanda linguistik menjadi salah satu yang terpenting. Dalam teori semiotika ini fungsi dan kegunaan dari suatu tanda itulah yang menjadi pusat perhatian. Tanda sebagai suatu alat komunikasi merupakan hal yang teramat penting dalam berbagai kondisi serta dapat dimanfaatkan dalam berbagai aspek komunikasi. Klasifkasi tanda menurut Peirce menghendaki agar teori semiotikanya ini menjadi rujukan umum atas kajian berbagai tanda-tanda. Oleh karenanya itu memerlukan kajian lebih mendalam mengenai hal tersebut. Terutama mengenai seberapa luas jangkauan teorinya ini.

*Grand theory* Pierce meliputi *sign, object dan interpretation* yang dikenal dengan *triangle meaning* (Kriyantono, 2014) yang bersifat menyeluruh, deskripsi struktural dari semua sistem penandaan. Sebuah tanda atau *representamen* adalah sesuatu yang bagi seseorang mewakili sesuatu yang lain dalam beberapa hal atau kapasitas. Sesuatu yang lain tersebut disebut sebagai interpretan yang pada gilirannya memacu objek tertentu. Secara sederhana dapat dikatakan bahwa sebuah tanda (representamen) memiliki relasi langsung dengan interpretan dan objeknya, dan inilah yang disebut sebagai proses semiosis atau signifikasi (Wahjuwibowo, 2013).

Pierce menyatakan bahwa penalaran manusia senantiasa dilakukan melalui tanda yang terdiri dari tiga elemen yang digambarkan sebagai berikut :



Gambar 1. Triangle Meaning Charles S. Pierce

***Sign*** atau tanda merupakan sesuatu yang berbentuk fisik yang dapat ditangkap oleh panca indera manusia dan merupaka sesuatu yang merepresentasikan hal lain di luar tanda itu sendiri. Tanda terdiri dari symbol (tanda yang muncul dari kesepakatan), Ikon (tanda yang muncul dari perwakilan fisik) dan Indeks (tanda yang muncul dari sebab akibat). Sedangkan acuan dari tanda tersebut disebut sebagai ***Object*** yang merupakan konteks sosial yang menjadi referensi dari tanda atau sesuatu yang dirujuk tanda. ***Interpretant*** sendiri adalah pengguna tanda yang merupakan konsep pemikiran dari orang yang menggunakan tanda dan menurunkannya ke suatu makna tertentu atau makna yang ada dalam benak seseorang tentang objek yang dirujuk pada sebuah tanda. Selanjutnya perlu diketahui bahwa hal yang terpenting dalam sebuah semiosis adalah bagaimana makna muncul dari sebuah tanda, ketika tanda itu digunakan orang saat berkomunikasi.

Dengan demikian, dapat dipahami pola kerja dari ketiga unsur tersebut untuk memahami semiotika dalam pandangan Charles S.Pierce. Tetapi, perlu diketahui bahwa syarat represantemen dapat menjadi sebuah tanda yakni dengan adanya Ground, yaitu pengetahuan yang ada pada pengirim dan penerima tanda sehingga representament dapat dipahami (Zaimar, 2008).

## Definisi *Data Mining*

*Data Mining* merupakan proses menemukan korelasi baru yang bermanfaat, pola dan trend dengan menambang sejumlah repository data dalam jumlah besar, menggunakan teknologi pengenalan pola seperti statistik dan teknik matematika (Fatmawati & Windarto, 2018). *Data mining* merupakan proses analisis data menggunakan perangkat lunak untuk menemukan pola dan aturan (*rules*) dalam himpunan data (Arifin, 2020). Data Mining merupakan suatu metode untuk menemukan pengetahuan dalam suatu tumpukan data yang cukup besar (Rodiyansyah, 2015).

## *Cross-Industry Standard Process For Data Mining (CRISP-DM*)

CRISP-DM menyediakan standar proses baku untuk data mining yang dapat diterapkan ke dalam strategi pemecahan masalah umum pada bisnis atau pada unit penelitian. Dalam CRISP-DM sebuah proyek data mining memiliki siklus hidup yang terbagi dalam enam fase, berikut merupakan gambaran secara umum mengenai siklus hidup dalam CRISP-DM (Feblian & Daihani, 2016):

1. **Pemahaman Bisnis (*Business Understanding*)**

Pemahaman bisnis adalah fase penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan. Selanjutnya menerjemahkan tujuan dan batasan menjadi formula dan definisi dari permasalahan data mining.

1. **Pemahaman Data (*Data Understanding*)**

Pemahaman data adalah fase mengumpulkan data menggunakan analisis penyelidikan data untuk mengenali lebih lanjut mengenai data dan pencarian awal pengetahuan yang ada didalamnya selanjutnya mengevaluasi kualitas data.

1. **Pengolahan data (*Data Preparation*)**

Pengolahan data adalah fase menyiapkan data yang tersedia dari awal, karena merupakan kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif, menyiapkan data awal sehingga siap untuk perangkat pemodelan.

1. **Pemodelan (*Modeling*)**

Pada fase ini memilih dan mengaplikasikan teknik pemodelan yang sesuai, selanjutnya mengkalibrasi aturan model untuk mengoptimalkan hasil.

1. **Evaluasi (*Evaluation*)**

Pada fase ini mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan, menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak ditertangani dengan baik dan mengambil keputusan berkaitan dengan penggunan hasil dari data mining.

## Definisi *Text Mining*

*Text Mining* adalah menambang data berupa teks dimana sumber data biasanya didapatkan dari dokumen (Yulian, 2018). *Text Mining* adalah ilmu yang mempelajari bagaimana menarik informasi yang menarik, sesuatu yang baru, pola yang belum diketahui sebelumnya atau menemukan kembali informasi tersirat yang berasal dari kumpulan sumber-sumber data text yang berbeda-beda (Jumeilah, 2017).

Tujuannya adalah untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen (Yulian, 2018).

## Tahapan *Pre-Processing*

Tahap pre-processing adalah suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur kedalam bentuk data yang terstruktur. Proses-processing dilakukan agar data yang digunakan bersih dari noise, memiliki dimensi yang lebih kecil, serta lebih terstruktur, sehingga dapat diolah lebih lanjut. Tahap pre-processing memiliki beberapa proses, yaitu *case folding*, *tokenizing*, *stopwords removing*, dan *stemming* (Jumeilah, 2017).

1. ***Remove Duplicates***

*Remove duplicates* adalah proses untuk menghapus data yang muncul secara berulang. Seperti yang terlihat pada Gambar 2 dimana dari tweet yang terletak disebelah kanan yaitu terdapat dua tweet yang sama, kemudian setelah dilakukan *Remove Duplicate* menjadi satu tweet.



Gambar 2. Contoh *Remove Duplicate*

1. ***Replace***

*Replace* adalah proses untuk menghilangkan *link* dan *mention* seperti *retweet, https, dan @username* yang terdapat pada *tweet*. Pada Gambar 3 adalah contoh dari tweet yang telah dilakukan *replace*. Tweet sebelah kiri terdapat tanda # (*hastag*) dan juga *link*, setelah dilakukan replace maka tweet akan berubah seperti tweet sebelah kanan yaitu hilang tanda # (*hashtag*) dan juga *link*.



Gambar 3. Contoh *Replace*

1. ***Case Folding***

*Case Folding* adalah proses mengubah setiap huruf didalam post menjadi lowercase atau semuanya dijadikan huruf kecil. Seperti terlihat pada Gambar 4 tweet sebelah kiri merupaka tweet yang belum di *case folding*, di sebelah kanan adalah tweet yang sudah dilakukan *case folding* sehingga semua tweet yang berada di sebelah kanan berubah menjadi huruf kecil.



Gambar 4. Contoh *Case Folding*

1. ***Tokenizing***

*Tokenizing* adalah proses tahapan pemotong *string* masukkan berdasarkan kata-kata yang menyusunnya atau dengan kata lain pemecahan kalimat menjadi kata. Seperti yang terlihat pada Gambar 5 merupakan contoh dari *tokenizing* yaitu memecah tweet menjadi perkata.



Gambar 5. Contoh *Tokenizing*

1. ***Stopword Removing***

*Stopwords Removing* adalah proses menghapus kata-kata yang sangat umum contohnya seperti : dan, di, itu, dengan. Pada Gambar 6 merupakan hasil setelah dilakukan *stopword removing*.



Gambar 6. Contoh Stop Removing

1. ***Term Weighting***

*TF-IDF* merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*Term*) terhadap langkah-langkah sebagai berikut :

1. Mencari nilai *Term Frequency(TF)*, dengan cara menghitung jumlah yang muncul pada dokumen
2. Mencari nilai *Document Frequency*(*DF*), dengan cara menghitung jumlah dokumen yang mengandung sebuah kata
3. Mencari nilai *Invers Document Frequency*(*IDF*), dengan cara jumlah seluruh dokumen dibagi jumlah dokumen yang mengandung sebuah kata (*DF*).
4. Menghitung bobot = *Term Frequency* (*TF*) x *Invers Document Frequency* (*IDF*)

## Algoritma K-Means

Algoritma K-means diperkenalkan oleh MacQueen tahun 1967 sebagai metode klasterisasi data nonhierarki yang mempartisi objek data ke dalam kelompok-kelompok. Kmeans adalah proses klasterisasi tanpa supervisi di mana objek data ditempatkan secara ‘alami’ dalam sebuah kelompok dengan tidak mengetahui pola atau pengetahuan yang dimiliki untuk memandu proses klasterisasi (Miner et al., 2012). Dalam metode ini, data yang memiliki karakteristik yang sama dimasukan ke dalam satu kelompok yang sama dan data yang memiliki karakteristik berbeda dimasukan ke dalam kelompok yang berbeda. Proses klasterisasi k-means dilakukan dengan cara meminimalkan variasi objek data dalam kelompok yang sama dan memaksimalkan variasi data antarkelompok yang berbeda (Hardi et al., 2019).

## RapidMiner

Rapidminer adalah platform pernagkat lunak data ilmu pengetahuan yang dikembangkan oleh perusahaan dengan nama yang sama, yang menyediakan lingkungan terpadu untuk pembelajaran mesin (*machine learning*), pembelajaran mendalam (*deep learning*), penambangan teks(*text mining*), dan analisis prediktif (*predictive analytics*). Aplikasi ini digunakan untuk aplikasi bisnis dan komersial serta untuk penelitian, pendidikan, pelatihan, pembuatan *prototype* dengan cepat, dan pengembangan aplikasi serta mendukung semua langkah proses pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan. Rapidminer dikembangkan dengan model open core (Nofitri & Irawati, 2019).

# METODE PENELITIAN

## Metode Penelitian

Penelitian ini menggunakan paradigma konstruktivis. Paradigma ini digunakan untuk melihat bahwa sesuatu realitas kehidupan sosial bukanlah terjadi secara alamiah tetapi bagian dari hasil konstruksi sosial. Dengan menggunakan paradigma konstruktivis, maka analisis dilakukan menggunakan teori semiotika untuk fokus memahami permasalahan tentang bagaimana sebuah realitas sosial dapat terbentuk. Dalam arti, semiotika sebagai sebuah teori digunakan untuk menganalisa makna nasionalisme dengan menggunakan media twitter di Indonesia pada objek pembahasan pertandingan Kejuaraan AFF tahun 2020. Pendekatan penelitan menggunakan pendekatan kualitatif dengan tujuan untuk memahami fenomena tentang apa yang dialami oleh subyek penelitian misalnya, perilaku, persepsi, motivasi, tindakan dan lain – lain, secara holistic, dan dengan cara deskripsi dalam bentuk kata – kata dan Bahasa pada suatu konteks yang alamiah dan dengan memanfaatkan metode ilmiah (Moleong, 2016).

Penelitian ini juga menggunakan teknik pengumpulan data berupa data primer yang bersumber dari data observasi non-partisipan yang dilakukan melalui pengolahan data (*text mining)*, selain itu juga menggunakan data sekunder yang berasal dari data penunjang seperti buku, jurnal dan data-data *online*.

## Tahapan Penelitian



Gambar 7. Tahapan Penelitian

Pada Gambar 7 merupakan tahapan penelitian yang dimulai dari peneliti melakukan identifikasi masalah, kemudian dilanjutkan ke tahap berikutnya adalah pengumpulan data. Pengumuplan data menggunakan metodologi CRIPS-DM yang terdiri dari Business Understanding, Data Understanding, Data Collection, Data Preparation, Modeling, dan Evaluasi. Setelah pengumpulan data, data yang didapat masuk ke tahap analisa data yaitu analisa semiotika menggunakan teori semiotika pierce yaitu melihat text yang sudah dioleh berdasarkan sign, object, dan intepretant

## Teknik Pengumpulan Data

### Data Primer

#### Observasi

Umi Narimawati menyebutkan bahwa data primer ialah data yang berasal dari sumber asli atau pertama. Data ini tidak tersedia dalam bentuk terkompilasi ataupun dalam bentuk *file-file*. Data ini harus dicari melalui narasumber atau dalam istilah teknisnya responden, yaitu orang yang kita jadikan objek penelitian atau orang yang kita jadikan sebagai sarana mendapatkan informasi ataupun data (Narimawati, 2008). Berdasarkan pendapat di atas dapat diketahui bahwa data primer dalam penelitian kualitatif dapat diperoleh observasi yang bertujuan untuk mengamati dan menganalisa objek agar dapat memahami data yang diperoleh.

Menurut Kriyantono dalam buku Teknis Praktis Riset Komunikasi, Observasi diartikan sebagai kegiatan mengartikan secara langsung (tanpa mediator) suatu objek untuk melihat dengan dekat kegiatan yang dilakukan tersebut (Kriyantono, 2014). Observasi sendiri memiliki berbagai jenis yang diklasifikasikan sebagai berikut:

1. **Pengamatan Partisipasi**

Penelitian melakukan observasi dengan cara melibatkan diri atau menjadi bagian lingkungan sosial (organisasi) tengah diamati melalui teknik partisipasi dapat memperoleh data relatif lebih akurat dan lebih banyak, karena peneliti secara langsung mengamati perilaku dan kejadian atau peristiwa dalam lingkungan sosial tertentu.

1. **Pengamatan Non-partisipasi**

Peneliti melakukan pengamatan nonpartisipan ini, melakukan observasi pengumpulan data dan informasi tanpa melibatkan diri, atau tidak menjadi bagian dari lingkungan sosial/organisasi yang diamati.

1. **Pengamatan Pasif**

Peneliti hadir dan tidak berinteraksi dengan pihak lainnya pengamat hanya bersikap pasif dalam mengamati dan hanya merekam data atau informasi apa yang sedang terjadi pada peristiwa tersebut tanpa diketahui pihak lain.

1. **Pengamatan aktif**

Pengamatan yang aktif memiliki peran dalam situasi lingkungan sosial tertentu. Selanjutnya, peneliti secara aktif melakukan wawancara dengan pihak orang-orang yang terlibat untuk memperoleh informasi dan data lebih lengkap.

1. **Pengamat Moderat**

Peneliti mempertahankan adanya keseimbangan antara sebagai orang dalam (*insider*) dan pihak luar (*outsider*) atau berinteraksi antara pengamat dan partisipan.

Dalam penelitian ini penggunaan data primer yang digunakan adalah melalui observasi non-partisipan dengan melakukan *text mining* pada media sosial yang memiliki relevansi dengan *keywords* yang sedang *trending topic* seperti #AFFSuzukiCup2020, Final Piala AFF, #kitagaruda, timgaruda, timnasgaruda, PSSI, timnas Indonesia. Pada proses penelitian ini peneliti mengumpulkan data dari *twitter* dengan menggunakan *tools rapidminer*.

#### Metodologi *CRISP-DM*

Pada tahapan ini terdapat sub-tahapan dengan menerapkan metodologi CRISP-DM. Model proses CRISP-DM memiliki beberapa fase diantaranya adalah sebagai berikut :

1. **Pemahaman Bisnis (*Business Understanding)***

Di dalam tahapan business understanding peneliti mencoba untuk memahami permasalahan yang ingin diangkat yaitu analisis semiotika nasionalisme pada kejuaran AFF 2020, pertama peneliti akan mencoba mempelajari artikel atau berita yang ada di internet tentang informasi performa timnas, kemudian mencoba melihat trending topik yang berkaitan dengan timnas Indonesia pada media sosial twitter.

1. **Pemahaman Data (*Data Understanding)***

Di dalam tahapan data understanding yang dilakukan peneliti adalah mencoba untuk menentukan tweets yang relevan dengan topik yang ingin di angkat dengan keywords #AFFSuzukiCup2020, Final Piala AFF, kitagaruda, timgaruda, timnasgaruda, PSSI, timnas Indonesia. Kemudian keywords tersebut akan dilakukan pencarian pada media twitter untuk menampilkan tweet sesuai dengan topik permasalahan jika penentuan keywords sudah dilakukan maka selanjutnya peneliti akan menentukan lama periode data tweet yang akan digunakan.

1. **Koleksi Data (*Data Collection)***

Di dalam tahapan data collection peneliti akan melakukan pengambilan data pada media sosial twitter. Sebelum proses pengambilan tweet dilakukan, peneliti diwajibkan untuk membuat koneksi yang menghubungkan antara rapidminer dengan twitter. Data yang diambil hanya data yang sesuai dengan keywords yang sebelumnya telah ditentukan.

1. **Persiapan Data (*Data Preparation)***

Pada tahapan ini data yang sudah didapatkan melalui tahapan *data collection* akan dilakukan data preprocessing, dengan tahapan sebagai berikut:

1. *Remove Duplicate* : Pada tahap ini dilakukan proses menghilangkan duplikasi data yang ada pada data hasil crawling dengan menggunakan *RapidMiner*.
2. *Case Folding* : Pada tahap ini dilakukan proses mengubah semua huruf pada data *tweets* yang sudah didapat menjadi bentuk *lower case* atau huruf kecil.
3. *Cleansing / Filtering* : Pada tahap ini dilakukan proses menghapus karakter atau tanda baca yang tidak diperlukan dalam proses analisis sentimen. Contoh seperti URL, *hastag* (#), *username*, dan *mention*.
4. *Tokenizing* : Pada tahap ini dilakukan proses memecah data tweets yang masih dalam bentuk kalimat, menjadi kata individual
5. *Stopwords* : Pada tahap ini dilakukan proses *stopwords* untuk menghilangkan kata sambungan / *stopwords* yang ada pada data.

Data *tweets* yang sudah melalui proses data *preprocessing* nantinya akan berubah bentuknya dari kalimat akan menjadi token / kata. Lalu data akan diberikan pembobotan (*term weighting*) yang kemudian digunakan untuk proses *clustering.*

1. **Pemodelan (*Modeling)***

Di dalam tahapan Modeling peneliti akan melakukan pemodelan terhadap dataset yang sudah dilakukan preprocessing. Kemudian dataset tersebut akan dilakukan clustering dengan menggunakan Algoritma K-Means dimana data yang memiliki karakteristik yang sama dimasukan kedalam kelompok yang sama. Algoritma K-Means clustering dilakukan dengan memilih bilangakn K secara acak, kemudian menentukan centroid, dilanjutkan dengan menghitung jarak semua titik terhadap centroid, diakhiri dengan melakukan clustering. Algoritma K-Means dapat dilakukan secara berulang hingga mendapat data kluster dengan performat terbaik.

1. **Evaluasi (*Evaluation)***

Pada tahapan ini peneliti akan melakukan evaluasi terhadap hasil *clustering* dan diolah menggunakan teori semiotika untuk menentukan tanda-tanda melalui teks yang telah diklasifikasi, lalu memastikan objek dari tanda tersebut, kemudian ditemukan makna teks sesuai dengan interpretasi makna yang dihasilkan.

### Data Sekunder

Sugiyono (2017:137) menjelaskan bahwa data sekunder adalah sumber data yang tidak langsung. Data sekunder ini merupakan data yang sifatnya mendukung keperluan data primer seperti buku-buku, literatur, dan bacaan yang berkaitan dan menunjang penelitian ini. Selain itu, data sekunder dapat diperoleh secara tidak langsung melalui media perantara (dihasilkan pihak lain) atau digunakan oleh lembaga lainnya yang bukan merupakan pengolahnya,tetapi dapat dimanfaatkan dalam suatu penelitian tertentu.

## Teknik Analisa Data

Hasil dari pengumpulan data pemrosesan *Text Mining* berupa *clustering* kata kemudian dilanjutkan dalam pengolahan semiotika menurut Pierce yaitu *sign*, *object*, dan *interpretant*. *Sign* adalah tanda dalam penelitian ini dilihat dalam bentuk *text* hasil dari proses *clustering*. *Object* merupakan penjelas dari *text* yang diperoleh. Kemudian *interpretant* adalah penafsiran makna terhadap kata yang dihasilkan.

# HASIL YANG DICAPAI

## Koleksi Data (*Data Collection*)

Pada tahapan *data* collection peneliti mengumpulkan data yang akan digunakan. Data yang didapatkan sebanyak 62.236 data. Peneliti melakukan pengambilan data *tweet* yang berkaitan dengan timnas Indonesia pada tanggal 30 November 2021 hingga 06 Januari 2022 dengan *keywords* yang sedang *trending topic* seperti AFFSuzukiCup2020, Final Piala AFF, Kita Garuda, Tim Garuda, dan Timnas Garuda. Peneliti menggunakan *hastag* tersebut karena kaitannya dengan pelaksanaan Final Piala AFF dan sepak bola Indonesia.

Tabel 1. Jumlah Data Tweet Hasil *Crawling*

|  |  |  |
| --- | --- | --- |
| ***Keywords*** | **Tanggal** | **Jumlah** |
| **AFF Suzuki Cup** | 25 Desember 2021 – 3 Januari 2022 | 12.000 |
| **Final Piala AFF** | 26 Desember 2021 – 4 Januari 2022 | 12.730 |
| **Kita Garuda** | 30 November 2021 – 6 Januari 2022 | 11.525 |
| **PSSI** | 30 November 2021 – 6 Januari 2022 | 12.997 |
| **Timnas Indonesia** | 30 November 2021 – 6 Januari 2022 | 12.984 |
| **Jumlah** | 62.236 |

Pada Tabel 1 merupakan hasil *crawling* yang berhasil dikumpulkan oleh peneliti dengan menggunakan rapidminer berdasarkan kata kunci (*keywords*) yang dipilih beserta dengan tanggal dari tweet dan jumlahnya.

Dalam penelitian ini, peneliti menggunakan *attribute Created-At, From-User* dan *Text* untuk isi dari data *tweet* yang diambil dalam proses *crawling* data yang dilakukan pada tanggal 30 November 2021 hingga 6 Januari 2022 dengan jumlah data 62.236 data. Isi dari kolom *Created-At* yaitu tanggal dimana *tweet* tersebut dibuat atau di posting oleh pengguna *twitter.* Isi dari kolom *From-User* adalah nama pengguna *twitter*. Dan isi dari kolom *Text* yaitu *tweet* yang diungkapkan oleh si pengguna *twitter*. Tabel 2 merupakan sample hasil *crawling* menggunakan rapidminer.

Table 2. Sample Isi Data *Crawling*

| ***Cread-At*** | ***From-User*** | ***Text*** |
| --- | --- | --- |
| 04/01/2022 05:23:55 PM | DJ Kasino | Atas permintaan Shin Tae-yong, PSSI berupaya agar Timnas Indonesia gelar pertandingan persahabatan pada akhir Januari untuk memperbaiki peringkat FIFA.#storybola #bolaindonesia #timnas #liga1 #beritabola #quotesanakbola #quotessepakbola #katasepakbola #ligaindonesia #anakbola https://t.co/Z8ifzYVCch |
| 04/01/2022 11:50:14 AM | Delhen | @PSSI pelatih terbaik sepanjang sejara sepak bola indonesia! |
| 30/12/2021 05:57:06 AM | Sjahriel Purwan | Apapun hasilnya, kalian sudah berjuang dan menampilkan yang terbaik.Love you all...#KitaGaruda #TimnasDay |
| 01/01/2022 09:33:11 PM | Galang | memang jelas timnas indonesia kekurangan striker, dulu masih jamannya Bepe bagus secara dia kapten tapi minim goal. Setelahnya kayanya ga ada lagi deh, mau ngomongin gonzales ga juga, beto gonzalves, osvaldo hay apa lagi |
| 01/01/2022 09:29:17 PM | Rizky Adnan | Wasit dari timur tengah ada masalah apa sih sama timnas indonesia? Keputusannya kontroversial mulu |

## Persiapan Data (Data Preparation)

Pada tahapan *data preparation* ini, peneliti akan membagi menjadi 3 tahapan sebagai berikut :

* + - 1. Tahapan *Remove Duplicate*, tahapan ini merupakan proses yang dilakukan peneliti untuk menghilangkan *link, hashtag*, dan *mention* yang terdapat pada *tweet*.
			2. Tahapan *preprocessing,* tahapan ini merupakan proses yang akan dilakukan oleh peneliti setelah data *tweet* sudah di *crawling*.
			3. Tahapan pembobotan, tahapan ini merupakan proses untuk menentukan pembobotan dari frasa dengan menggunakan TF-IDF*.* TF (*Term Frequency)* adalah frekuensi dari kemunculan sebuah *term*  dalam dokumen yang bersangkutan. DF (*Dokumen Frequency)* adalah jumlah dokumen yang ada pada TF. IDF merupakan sebuah perhitungan dari bagimana *term* didistirbusikan secara luas pada koleksi dokumen yang bersangkutan.

### Tahapan Remove Duplicate

Pada tahap ini dari 62.236 data berkurang menjadi 29.293 data setelah dilakukan *Replace* yaitu menghilangkan *hashtag*, *link*, *mention*, serta *remove duplicate*.

### Tahapan Preprocessing Data

Membersikan data dilakukan dengan menggunakan tahapan *preprocessing* yang terdiri dari *transform case, tokenize, filter tokens, filter stopword,* dan *steam*. Dimana dalam penggunakan *filter stopword* peneliti menggunakan kamus *stopwords* milik(SokKana, 2021) dan penggunaan *steam* peneliti menggunakan kamus *steam* milik(Andylibrian, 2015). Simpan kata-kata yang ada dalam *dataset* dalam bentuk file *notepad* atau .txt yang kemudian dimasukan kedalam *rapid miner* yang selanjutnya membuat kalimat menjadi kata dasar.

Table 3. Sample Perubahan Sebelum dan Sesudah Preprocessing

|  |  |  |
| --- | --- | --- |
| **Preprocessing** | **Sebelum** | **Sesudah** |
| *Transform Case* | Terima kasih sudah berjuang sampai detik terakhir, kalian luar biasa Timnas Indonesia.Kami Bangga | terima kasih sudah berjuang sampai detik terakhir, kalian luar biasa Timnas Indonesia. kami bangga |
| *Tokenize* | terima kasih sudah berjuang sampai detik terakhir, kalian luar biasa Timnas Indonesia. kami bangga | “terima”,”kasih”,”sudah”,”berjuang”,”sampai”,”detik”,”terakhir”,”kalian”,”luar”,”biasa”,”timnas”,”Indonesia”,”kami”,”bangga” |
| *Filter Tokens* | “terima”,”kasih”,”sudah”,”berjuang”,”sampai”,”detik”,”terakhir”,”kalian”,”luar”,”biasa”,”timnas”,”Indonesia”,”kami”,”bangga” | “terima”,”kasih”,”sudah”,”berjuang”,”sampai”,”detik”,”terakhir”,”kalian”,”luar”,”biasa”,”timnas”,”Indonesia”,”kami”,”bangga” |
| *Filter Stopwords* | “terima”,”kasih”,~~”sudah”~~,”berjuang”,~~”sampai”~~,”detik”,~~”terakhir”~~,~~”kalian”~~,~~”luar”~~,~~”biasa”~~,”timnas”,”Indonesia”,~~”kami”~~,”bangga” | “terima”,”kasih”,”berjuang”,”detik”,”timnas”,”Indonesia”,”bangga” |
| *Steam* | “terima”,”kasih”,”berjuang”,”detik”,”timnas”,”Indonesia”,”bangga” | “terima”,”kasih”,”juang”,”detik”,”timnas”,”Indonesia”,”bangga” |

Pada Tabel 3 terjadi perubahan data pada saat *preprocessing* dengan rincian*:*

1. *Transform Cases* adalah proses mengubah semua huruf pada data *tweet* menjadi bentuk *lower cases.* Pada proses ini semua huruf dirubah kedalam huruf kecil karena mayoritas text sebagian besar merupakan huruf kecil semua.
2. *Tokenize* adalah proses memecah data *tweet* yang masih dalam bentuk kalimat agar sistem dapat melakukan pengecekan 1 per 1 terhadap tiap-tiap text yang ada pada kalimat. Pada proses ini terdapat 5.993 kata yang telah dipecah.
3. Filter Tokenize adalah proses untuk memfilter *token* berdasarkan panjang karakter. Pada proses ini terdapat 492 kata yang terfilter maka kata yang dipecah menjadi 5.501 kata.
4. *Filter Stopwords* adalah proses untuk menghilangkan kata sambung, setelah proses ini dijalankan terdapat 658 kata yang dihilangkan sehingga menjadi 4.843 kata.
5. *Steamming* adalah proses untuk mengembalikan kata menjadi bentuk kata dasar dari sebuah kata.

### Tahapan Pembobotan TF-IDF

Tahapan selanjutnya adalah bobot *term* dengan menggunakan *term weighting(TF-IDF)*. Diambil secara acak sebanyak 5 data *tweet* yang dapat dilihat pada Tabel 4 yang dimana sample *tweet* D1 Sampai D4.

Tabel 4. Sampel Data Tweet

|  |  |
| --- | --- |
| **Sample** | ***Tweet*** |
| D1 | Terima kasih perjuangannya timnas indonesia! Kalian luar biasa! |
| D2 | Terima kasih timnas Indonesia |
| D3 | Cinta sama timnas indonesia itu rumit. Kecewa ada tapi senengnya lebih banyak. |
| D4 | Wasit dari timur tengah ada masalah apa sih sama timnas indonesia?keputusannya kontoversial mulu |

Setelah menentukan *sample tweet* yang ingin digunakan, maka selanjutnya adalah menentukan bobot *term* dari 4 *tweet* menggunakan *term weighting(TF-IDF)* seperti yang ditunjukan pada Tabel 5.

Tabel 5. Penentuan Bobot *Term*

| **Term** | **Tf** | **Idf** | **Wdt=tf.idf** |
| --- | --- | --- | --- |
| **D1** | **D2** | **D3** | **D4** | **DF** | **Log(d/df)** | **D1** | **D2** | **D3** | **D4** |
| terima | 1 | 1 | 0 | 0 | 2 | 0.301 | 0.301 | 0.301 | 0 | 0 |
| kasih | 1 | 1 | 0 | 0 | 2 | 0.301 | 0.301 | 0.301 | 0 | 0 |
| juang | 1 | 0 | 0 | 0 | 1 | 0.602 | 0.602 | 0 | 0 | 0 |
| timans | 1 | 1 | 1 | 1 | 4 | 0 | 0 | 0 | 0 | 0 |
| indonesia | 1 | 1 | 1 | 1 | 4 | 0 | 0 | 0 | 0 | 0 |
| kalian | 1 | 0 | 0 | 0 | 1 | 0.602 | 0.602 | 0 | 0 | 0 |
| luar | 1 | 0 | 0 | 0 | 1 | 0.602 | 0.602 | 0 | 0 | 0 |
| biasa | 1 | 0 | 0 | 0 | 1 | 0.602 | 0.602 | 0 | 0 | 0 |
| cinta | 0 | 0 | 1 | 0 | 1 | 0.602 | 0 | 0 | 0.602 | 0 |
| rumit | 0 | 0 | 1 | 0 | 1 | 0.602 | 0 | 0 | 0.602 | 0 |
| kecewa | 0 | 0 | 1 | 0 | 1 | 0.602 | 0 | 0 | 0.602 | 0 |
| senang | 0 | 0 | 1 | 0 | 1 | 0.602 | 0 | 0 | 0.602 | 0 |
| wasit | 0 | 0 | 0 | 1 | 1 | 0.602 | 0 | 0 | 0 | 0.602 |
| timur | 0 | 0 | 0 | 1 | 1 | 0.602 | 0 | 0 | 0 | 0.602 |
| tengah | 0 | 0 | 0 | 1 | 1 | 0.602 | 0 | 0 | 0 | 0.602 |
| masalah | 0 | 0 | 0 | 1 | 1 | 0.602 | 0 | 0 | 0 | 0.602 |
| keputusan | 0 | 0 | 0 | 1 | 1 | 0.602 | 0 | 0 | 0 | 0.602 |
| kontroversial | 0 | 0 | 0 | 1 | 1 | 0.602 | 0 | 0 | 0 | 0.602 |

Berdasarkan Tabel 5 diperlihatkan contoh data *tweet*. *Term* adalah kata atau frasa yang menjadi subjek atau predikat dari sebuah proprosisi, Tf(*Term Frequency)* adalah frekuensi dari kemunculan sebuah *term*  dalam dokumen yang bersangkutan, Df(*Dokumen Frequency)* adalah jumlah dokumen yang ada pada Tf D1,D2,D3, dan D4, Idf merupakan sebuah perhitungan dari bagimana *term* didistirbusikan secara luas pada koleksi dokumen yang bersangkutan. Nilai Idf didapatkan dari log(D/df) dimana D adalah total dokumen dan df adalah banyak dokumen yang mengandung kata yang dicari, Wdt merupakan hasil dari perkalian Tf\*Idf.

## Pemodelan (Modelling)

Pada tahapan *Modelling* ini, peneliti akan membagi menjadi 2 tahapan sebagai berikut :

* + - 1. Tahapan wordcloud, yaitu untuk mendapatkan total jumlah dari setiap term
			2. Tahapan penentuan nilai K yang digunakan untuk menentukan jumlah cluster yang optimal dengan menggunakan metode elbow
			3. Tahapan *clustering,* setelah mendapatkan nilai K maka akan dilakukan *clustering* sesuai dengan jumlah k yang diperoleh..

### Wordcloud

Word Cloud adalah salah satu metode untuk memvisualisasikan data teks secara visual. Visualisasi pada word cloud berdasarkan frekuensi kata-kata dapat ditampilkan dalam bentuk yang menarik namun tetap informatif. Ukuran gambar teks dalam Word Cloud menyesuaikan dengan frekuensi data, semakin banyak frekuensi kata digunakan, maka semakin besar pula ukuran kata tersebut ditampilkan dalam Word Cloud (Pradana, 2020)

Table 6. Sample Data Wordcloud

|  |  |  |  |
| --- | --- | --- | --- |
| **word** | **total** | **word** | **total** |
| timnas | 22653 | kalah | 950 |
| dukung | 11747 | kasih | 915 |
| piala | 5134 | bola | 907 |
| pssi | 5097 | tonton | 826 |
| final | 4351 | semoga | 787 |
| thailand | 3679 | tanding | 744 |
| pemain | 2734 | hasil | 736 |
| juara | 2200 | ketua | 713 |
| semangat | 1893 | terima | 711 |
| menang | 1395 | pelatih | 707 |
| main | 1370 | ganti | 704 |
| singapura | 1317 | laga | 680 |
| garuda | 1170 | malam | 635 |
| ketum | 1106 | bangga | 612 |
| lawan | 1039 | babak | 544 |
| juang | 982 | baik | 515 |
| liga | 952 | harap | 506 |

Tabel 6 merupakan data word/kata beserta dengan total jumlahnya yang akan digunakan untuk membuat visualisasi wordcloud. Data yang ditampilkan adalah kata yang totalnya diatas 500 sebanyak 34 Kata. Total kata yang terbanyak adalah timnas dengan jumlah 22.653, diikuti dengan dukung sebanyak 11.747, dan selanjutnya adalah piala sebanyak 5.134



Gambar 8. Wordcloud

Gambar 8 adalah visualisasi dari Tabel 6 berupa *worldcloud*. Kata “timnas” yang memiliki total terbanyak maka terlihat lebih besar daripada kata lainnya. Selanjutnya kata “dukung” yang memiliki total kedua terbanyak juga terlihat dengan ukuran lebih kecil dari kata timnas. Kata “piala”, “pssi”, “final”, “thailand”, “pemain”, “juara”, “semangat” menampati posisi ketiga hingga kesembilan juga dapat terlihat pada wordcloud.

### Metode Elbow

Sebelum *clustering* dilakukan maka perlu diketahui nilai k yang optimal untuk digunakan dalam tahap *clustering*. Salah satu cara untuk mengetahui nilai k optimal adalah menggunakan Metode Elbow. Elbow yang dimaksud adalah posisi antara *average within centroid* dengan nilai k yang membentuk sudut siku (*elbow*).



Elbow

Gambar 9. Elbow Method

Gambar 9 merupakan hasil perbadingan antara nilai K dengan *average within centroid distance*. Pada nilai k 28 dan 29 membentuk siku dari *average within centroid distance*, sehingga nilai K yang optimal adalah 28 atau 29

### K-Means Clustering

Setelah mendapatkan nilai k maka proses *clustering* menghasilkan sebanyak jumlah 29, dengan sample tweet pada beberapa *cluster* sebagai berikut:

Cluster 1:

Table 7. Sampe Data Tweet Cluster 1

|  |  |
| --- | --- |
| **No** | **Tweet Asli** |
| 1 | Pemain Thailand dan Indonesia yang intens dulu, tidak bisa disamakan dengan Thailand dan Vietnam saat ini |
| 2 | Kartu merah dan juga gol untuk Singapura dipenghujung babak pertama |

Cluster 2:

Table 8. Samle Data Tweet Cluster 2

| **No** | **Tweet Asli** |
| --- | --- |
| 1 | Skor identik tahun baru |
| 2 | Era 70an timnas indonesia identik dengan persija, hampir semua bintang2 timnas berasal dari persija, tapi sekarang tak ada satupun pemain timnas yang berlaga di AFF tahun ini dari persija |

Cluster 3:

Table 9. Sampe Data Tweet Cluster 3

|  |  |
| --- | --- |
| **No** | **Tweet Asli** |
| 1 | Timnas Indonesia melangkah ke final Piala AFF 2020 setelah menorehkan kemenangan 4-2 atas Singapura dalam sebuah laga dramatis dan menguras emosi. |
| 2 | Timnas Indonesia menorehkan catatan kemenangan yang fantastis saat menghadapi Kamboja. Itu modal bagus untuk memenangkan pertandingan. |

Cluster 4:

Table 10. Sampe Data Tweet Cluster 4

|  |  |
| --- | --- |
| **No** | **Tweet Asli** |
| 1 | Tang mentang sama2 taeyong hahahahahah kudu banged lagunya favorit ya hahahaahahaha saranghae uri taeyong |
| 2 | Cuma di Indonesia pemain timnas masuk final tapi tang dikulik berita keluarganya. Waktu Euro kemarin gue kaga lihat berita serupa yang bahas keluarga pemain Inggris dan Italia. |

Cluster 5:

Table 11. Sample Data Tweet Cluster 5

|  |  |
| --- | --- |
| **No** | **Tweet Asli** |
| 1 | Pemain sayap Timnas Indonesia Rizky Ridho mendapat pesan khusus dari orang tua jelang laga Timnas Indonesia vs Thailand di Final Piala AFF 2020.  |
| 2 | Gelandang serang timnas Indonesia Egy Maulana Vikri angkat bicara terkait mental skuad Garuda jelang laga final kontra Thailand leg pertama Piala AFF 2020 di National Stadium, Singapura, Sabtu (29/12/2021). |

## Evaluasi (Evaluation)

Evaluasi dilakukan menggunakan teori semiotika *Triangle Meaning* Charles Sanders Pierce terhadap data *wordcloud* (Gambar 8) dan hasil *K-Means Clustering* (Tabel 7 sampai Tabel 11).

# DAFTAR PUSTAKA

Arifin, M. (2020). Implementasi Data Mining Pada Prediksi Pemesanan Menggunakan Algoritma Apriori (Studi Kasus : Kimia Farma). *Pelita Informatika: Informasi Dan Informatika*, *8*(3), 353–356.

Fatmawati, K., & Windarto, A. P. (2018). Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (DBD) Berdasarkan Provinsi. *CESS (Journal of Computer Engineering System and Science)*, *3*(2), 173–178. https://doi.org/10.24114/cess.v3i2.9661

Feblian, D., & Daihani, D. U. (2016). Implementasi Model Crisp-Dm Untuk Menentukan Sales Pipeline Pada PP X. *Jurnal Teknik Industri*, *6*(1), 1–12. https://doi.org/10.25105/jti.v6i1.1526

Hardi, W., Kusuma, W. A., & Basuki, S. (2019). Pengelompokan Topik Dokumen Berbasis Text Mining Dengan Algoritme K-Means : Studi Kasus Pada Dokumen Kedutaan Besar Australia Jakarta. *Visi Pustaka*, *21*(1), 67–76.

Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *1*(1), 19–25. https://doi.org/10.29207/resti.v1i1.11

Kriyantono, R. (2014). *Teknik Praktis Riset Komunikasi*. Kencana.

Littlejohn, S. W., & Foss, K. A. (2011). *Teori Komunikasi*. Salemba Humanika.

Miner, G., Elder, J., Hill, T., Delen, D., Fast, A., & Nisbet, R. A. (2012). *Practical Text Mining And Statistical Analysis for Non-Structured Text Data Applications*. Elsevier Science. https://doi.org/https://doi.org/10.1016/C2010-0-66188-8

Moleong, L. J. (2016). Metodologi Penelitian Kualitatif Edisi Revisi. In *Bandung: PT Remaja Rosdakarya*. PT. Remaja Rosdakarya.

Narimawati, U. (2008). *Metodologi Penelitian Kualitatif dan Kuantitatif, Teori dan Aplikasi*. Agung Media.

Nawiroh, V. (2014). *Semiotika Dalam Riset Komunikasi*. Ghalia Indonesia.

Nofitri, R., & Irawati, N. (2019). Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *5*(2), 199–204. https://doi.org/10.33330/jurteksi.v5i2.365

Prabancono, H. (2015). *Perilaku Netizen: Inilah 3 Karakter Pengguna Twitter Indonesia*. Solopos.Com.

Pradana, M. G. (2020). Penggunaan fitur wordcloud dan document term matrix dalam text mining. *Jurnal Ilmiah Informatika (JIF)*, *8*(1), 38–43.

Rodiyansyah, S. (2015). Algoritma Apriori untuk Analisis Keranjang Belanja pada Data Transaksi Penjualan. *Infotech Journal*, *1*(2), 36–39.

Sobur, A. (2013). *Semiotika Komunikasi*. PT. Remaja Rosdakarya.

Wahjuwibowo, I. S. (2013). *Semiotika Komunikasi*. Mitra Wacana Media.

Yulian, E. (2018). Text Mining Dengan K-Means Clustering Pada Tema LGBT Dalam Arsip Tweet Masyarakat Kota Bandung. *Jurnal Matematika MANTIK*, *4*(1), 53–58. https://doi.org/10.15642/mantik.2018.4.1.53-58

Zaimar, O. K. (2008). *Semiotik dan Penerapannya Dalam Karya Sastra*. Pusat Bahasa, Departemen Pendidikan Indonesia.

# LAMPIRAN-LAMPIRAN

1. Biodata Peneliti
2. Identitas Diri
3. Nama Lengkap (dengan gelar) : Jeanie Annissa, S.IP.,M.Si
4. Jenis Kelamin : Perempuan
5. Jabatan Fungsional : Asisten Ahli
6. NIP/NIDN/ID-SINTA : 100045
7. Tempat, Tanggal Lahir : Jakarta, 28 Juni 1985
8. E-Mail : jeanie.annissa@budiluhur.ac.id
9. Nomor Handphone : 081584785075
10. Alamat : Jl. Raya Lenteng Agung no. 56 RT.006/RW.008
11. Riwayat Pendidikan

|  |  |  |  |
| --- | --- | --- | --- |
|  | S1 | S2 | S3 |
| Nama Perguruan Tinggi | Universitas Nasional | Universitas Indonesia |  |
| Bidang Ilmu | Ilmu Politik | Ilmu Politik |  |
| Tahun Masuk-Lulus | 2002-2006 | 2007-2010 |  |

1. Pengalaman Penelitian (5 Tahun Terakhir)

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Tahun** | **Judul Penelitian** | **Pendanaan** |
| **Sumber\*** | **Jumlah (Rp)** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

\* *Tuliskan sumber pendanaan baik dari Universitas Budi Luhur maupun dari sumber lainnya*

1. Publikasi Artikel Ilmiah dalam Jurnal (5 Tahun Terakhir)

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Judul Artikel Ilmiah** | **Nama Jurnal** | **Volume/Nomor/****Tahun** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. Pemakalah Seminar Ilmiah (5 Tahun Terakhir)

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Nama Temu Ilmiah/Seminar** | **Judul Artikel Ilmiah** | **Waktu dan Tempat** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. Karya Buku (5 Tahun Terakhir)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Judul Buku** | **Tahun** | **Jumlah Halaman** | **Penerbit** |
| 1. | Pengantar Kewarganegaraan | 2021 |  | Dot Publisher |
| 2. | Pengantar Ilmu Komunikasi | 2021 |  | Dot Publisher |
|  |  |  |  |  |

1. Perolehan HKI (10 Tahun Terakhir)

| **No.** | **Judul Tema/HKI** | **Tahun** | **Jenis** | **Nomor P/ID** |
| --- | --- | --- | --- | --- |
| 1. | Pendidikan Kewarganegaraan | 2021 | Buku | 000299343 |
| 2. | Pengantar Ilmu Komunikasi | 2021 | Buku | 000287994 |
|  |  |  |  |  |

 Jakarta, 3 Juni 2022

 Pengusul,



 (Jeanie Annissa,S.IP.M.Si)

1. Catatan Harian

|  |  |  |
| --- | --- | --- |
| **No** | **Jenis Kegiatan** | **2022** |
| **Jan** | **Feb** | **Mar** | **Apr** | **Mei** | **Juni** |
| 1 | Identifikasi Masalah, Pemahaman Bisnis | X |  |  |  |  |  |
| 2 | Pemahaman Data | X |  |  |  |  |  |
| 3 | Penyusunan Proposal Penelitian | X |  |  |  |  |  |
| 4 | Data Collection/Pencarian Data Sekunder | X | X |  |  |  |  |
| 5 | Persiapan Data/Text Processing |  | X | X |  |  |  |
| 6 | Modeling/*Clustering*  |  |  | X | X |  |  |
| 7 | Pembuatan Laporan Kemajuan Penelitian |  |  |  | X | X |  |
| 8 | Analisa Data |  |  |  |  | X |  |
| 9 | Evaluasi |  |  |  |  | X |  |
| 10 | Penyusunan Laporan Akhir Penelitian |  |  |  |  |  | X |

|  |  |  |
| --- | --- | --- |
| **No** | **Tanggal** | **Kegiatan** |
| 1 | 3 Februari 2022 | *Data collection* dan pencarian data sekunder mengolah crawling data tweet yang sudah didapat sebelumnya. Berdasarkan crawling dari bulan November hingga Januari |
| 2 | 15 Februari 2022 | Text processing data tweet untuk persiapan pengolahan *clustering.* Melakukan Replace dan Remove duplicate dari data tweet yang diperoleh |
| 3 | 9 Maret 2022 | Text processing data tweet untuk persiapan pengolahan *clustering*. Melakukan tokenizing, filtering, stopword, stem dari data tweet. |
| 4 | 24 Maret 2022 | Modeling untuk *clustering* menentukan nilai k untuk *clustering* |
| 5 | 5 April 2022 | Modeling melakukan *clustering* data sesuai dengan nilai k yang didapat |
| 6 | 3 Mei 2022 | Pembuatan draft Laporan Kemajuan  |
| 7 | 18 Mei 2022 | Finalisasi Laporan Kemajuan |