# BAB I

# PENDAHULUAN

## 1.1. Latar Belakang

Perkembangan teknologi komunikasi dan informasi di dunia digital yang begitu pesat dirasakan saat ini, merupakan indikasi dari kebutuhan manusia sebagai makhluk sosial yang menginginkan cara efektif dan efisien dalam berkomunikasi dan mendapatkan informasi. Salah satu wadah komunikasi dan informasi dalam dunia digital antara lain adalah jejaring sosial. Berbagi penyedia layananan jejaring sosial bermunculan seiring berkembangnya cara manusia berkomunikasi. Oleh karenanya, setiap penyedia layanan jejaring sosial berlomba untuk menawarkan fitur-fitur yang berbeda antar satu dan lainnya, sehingga masing-masing penyedia layanan jejaring sosial memiliki fitur unik yang menjadi pilihan bagi penggunanya. Salah satu layanan jejaring sosial yang memiliki fitur unik dalam layanan komunikasi dan informasi adalah layanan jejaring sosial *twitter*. *Twitter* merupakan suatu wadah berkomunikasi dan berbagi informasi, dimana bentuk komunikasi dan informasi dapat disampaikan dalam sebuah *tweet*. Banyaknya karakter dalam sebuah *tweet* dibatasi sebanyak seratus empat puluh karakter, oleh karenanya pengguna layanan jejaring sosial *twitter* dituntut untuk menggunakan kata-kata yang singkat,

padat dan jelas dalam berkomunikasi dan berbagi informasi dengan sesama pengguna (Pratama, 2018).

Pengguna *twitter* dapat membuat sebuah pesan pendek yang disebut dengan *tweet*, dimana melalui *tweet* tersebut, pengguna *twitter* dapat saling berhubungan, berbagi pendapat, dan menemukan kabar dari berbagai penjuru dunia. Sebagaian besar pengguna *twitter* juga memanfaatkan media sosial ini untuk menemukan pelaku bisnis, dimana mereka akan menjadi pengikut (*followers*) dan berinteraksi dengan pelaku bisnis tersebut.

Shopee Indonesia merupakan salah satu pelaku bisnis di Indonesia yang mengusung model bisnis *marketplace* dengan menggunakan media sosial *twitter* sebagai sarana untuk melakukan promosi terhadap bisnisnya. Dengan menemukan jenis konten *tweet* yang banyak dilakukan *retweet* oleh *followers* sehinga Shopee Indonesia dapat menggunakan jenis konten *tweet* tersebut sebagai saran untuk melakukan promosi kepada pengguna *twitter* serta dapat dengan mudah mengetahui informasi dan membagikan saran maupun kritikan untuk membuat *shopee* menjadi lebih baik lagi dalam proses bisnisnya sehingga diharapkan semakin banyak pengguna *twitter* yang menjadi konsumen dari Shopee Indonesia (Indraloka, 2017).

Pengumpulan data *tweet* dari *twitter* dapat dilakukan dengan mengintegrasikan *Web Scraping* dan *tools Orange Data Mining*. Untuk mempermudah mengetahui jenis konten dari sejumlah data *tweet*, maka perlu dilakukan proses *text mining* terhadap data *tweet* tersebut dengan menerapkan teknik *clustering*. Pada *text mining*, teknik *clustering* digunakan

untuk mengelompokkan data tekstual berdasarkan kesamaan konten yang dimiliki ke dalam beberapa *cluster*, sehingga didalam setiap *cluster* akan berisi data tekstual dengan konten yang mirip.

*Text mining* adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang di ekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda (Rustiana, 2017). *Text mining* merupakan teknik yang digunakan untuk menangani masalah *classification, clustering, information extraction* dan *information retrival* (Kurniawan, 2012). *Clustering* merupakan salah satu metode *text mining* yang dikembangkan untuk mengefisiensikan pengelolaan teks serta peringkasan teks. Beberapa hal yang dapat meningkatkan kualitas *clustering* antara lain mengatasi dimensi tinggi yang diakibatkan besarnya jumlah kata dalam dokumen, meningkatkan skalabilitas agar mampu bekerja dengan jumlah dokumen dalam skala kecik ataupun besar (*scalable*), meningkatkan akurasi, memberikan label *cluster* yang bermakna, mampu mengatasi *overlapping*, serta memperhitungkan kesamaan konseptual istilah dari kata (Rozi, 2015).

Berdasarkan uraian diatas maka dibutuhkan analisis informasi pengelompokkan kata yang dominan (sering muncul) dalam akun *twitter* Shopee Indonesia (@*ShopeeID*) dengan judul "**Analisis *Text Clustering* Akun *Fanpage Shopee* Dengan Komentar *Followers* Menggunakan *Tools Orange Data Mining***".

## 1.2. Rumusan Masalah

Berdasarkan penjelasan dan uraian diatas, maka rumusan masalah pada penelitian ini yang muncul sebagai acuan untuk analisis adalah: Bagaimanakah menganalisis akun *twitter fanpage* Shopee Indonesia (@*ShopeeID*) dengan komentar *followers* menggunakan *tools Orange Data Mining* ?

## 1.3. Batasan Masalah

Agar pembahasan lebih terarah dan tidak menyimpang dari permasalahan yang ada, maka penulis hanya membatasi pembahasan permasalahan hanya pada:

1. Penelitian ini menggunakan sumber data dari media sosial yaitu *twitter* dengan menganalisis data akun *fanpage* Shopee Indonesia dengan komentar *followers*-nya sebanyak sampel data yang di dapat di akun media sosial *twitter* Shopee Indonesia (@*ShopeeID*).

2. Penggalian informasi analisis ini menggunakan *data mining* dengan teknik *text mining* dan metode *clustering*.

3. *Tools* yang digunakan dalam penelitian ini adalah *Orange Data Mining* sebagai analisis data dan *Web Scraper* untuk pengambilan data.

## 1.4. Tujuan dan Manfaat Penelitian

Dalam melakukan penelitian ini penulis memiliki tujuan penelitian yang mana dapat bermanfaat bagi orang lain. Berikut ini tujuan dan manfaat penelitian ini:

### 1.4.1. Tujuan

Berdasarkan permasalahan yang ada, maka tujuan dari penelitian ini diantaranya sebagai berikut:

1. Melakukan analisis *text* dari status dan komentar *followers* akun *fanpage twitter* Shopee Indonesia (@*ShopeeID*) untuk mengetahui kata yang dominan muncul (sering muncul).
2. Melakukan penggalian informasi analisis menggunakan *data mining* dengan teknik *text mining* dan metode *clustering*.
3. Melakukan analisis menggunakan *tools Orange Data Mining* dan *Web Scraper*.

### 1.4.2. Manfaat

Berikut ini manfaat yang diberikan setelah membaca penelitian ini diantaranya sebagai berikut:

**a. Bagi Mahasiswa**

1. Melatih mahasiswa untuk bisa memanfaatkan *text mining* dalam mengelola data berupa *text* dalam jumlah yang banyak dari media sosial.
2. Memberi pengetahuan kepada mahasiswa tentang kegunaan media sosial untuk analisis *data mining* dengan teknik *text mining* dan menggunakan metode *clustering*.

3. Memberi pengetahuan kepada mahasiswa tentang bagaimana menggunakan *tools Orange Data Mining* untuk analisis *data mining* dengan teknik *text mining* dan menggunakan metode *clustering*.

**b. Bagi Ilmu Pengetahuan**

1. Diharapkan hasil penelitian ini dapat menjadi kajian pada penelitian selanjutnya atau sejenisnya.

## 1.5. Metodologi Penelitian

### 1.5.1. Metode Penelitian

Pada dasarnya *text mining* didefinisikan sebagai proses penggalian informasi dimana pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu dengan menggunakan suatu alat analisis. *Text mining* mencari informasi dari sumber-sumber data melalui identifikasi dan eksplorasi pola tertentu, dalam kasus ini sumber data adalah kumpulan dokumen dengan pola yang ditemukan pada data teks yang tidak berstruktur. Praproses dari *text mining* sendiri berpusat pada identifikasi dan ekstraksi fitur representatif untuk dokumen *Natural Language* (Pratama, 2018).

Proses *text mining* membutuhkan penyusunan teks masukan berdasarkan tata bahasa, yang diikuti dengan menggali pola dari data yang sudah terstruktur, evaluasi dan interpretasi hasil. Proses ini biasanya digunakan untuk pengklasifikasian, pengelopokkan, analisis makna,

pengambil kesimpulan dari dokumen dan pemodelan hubungan objek yang berupa kata. Berikut merupakan tahapan dalam *text mining*:

1. *Information Retrieval*

   Yaitu tahapan untuk memperoleh dokumen yang sesuai dengan permintaan peneliti atau yang sesuai dengan permasalahan.

2. *Natural Language Processing*

   Yaitu tahapan untuk mentransformasi kata-kata yang terdapat dalam dokumen yang telah diperoleh sebelumnya. Dimana dari dokumen awalnya yang tidak terstruktur menjadi lebih terstruktur, sehingga dapat diperoleh informasi yang lebih akurat dan berguna.

3. *Information Extraction*

   Yaitu tahapan dimana informasi yang sudah diperoleh sebelumnya akan diesktrak sehingga peneliti akan lebih mudah memahami permasalahan yang diteliti melalui visualisasi yang ditampilkan.

4. *Knowledge Discovery*

   Pada tahapan ini pola dari suatu dokumen mulai teridentifikasi dan pengetahuan untuk mengatasi permasalahan telah didapat.

### 1.5.2. Metode Pengumpulan Data

Dalam penelitian ini bagian-bagian penelitian akan dilakukan secara komputerisasi. Mulai dari pengambilan data dengan teknik *Web Scraping* hingga analisis *text* menggunakan *tools Orange Data Mining*.

1. ***Web Scraping***

Aplikasi *web scraping* (juga disebut *intelligent, automated,* atau *autonomous agents*) hanya fokus pada cara memperoleh data melalui pengambilan dan ekstraksi data dengan ukuran data yang bervariasi. Pada kasus penelitian ini mencoba mencari informasi data yang hanya berupa konten status dan komentar (sebagai fokus) dari sebuah akun *twitter*. Metode *web scraping* dalam penelitian ini menggunakan aplikasi *scraping* yaitu *Web Scrap* yang hanya bisa digunakan di *Google Chrome* untuk *windows* yang memerlukan akun pengguna *twitter* untuk dapat mengambil informasi (konten status dan komentar) dari sebuah *link* postingan kemudian diekstrak dalam bentuk *file* format *.csv.

2. ***Orange Data Mining***

Aplikasi *Orange Data Mining* hanya fokus pada cara menganalisis *text clustering* dengan hasil data yang didapatkan melalui *scraping* dari *Web Scraper*. Pada kasus penelitian ini *Orange Data Mining* menampilkan beberapa *widget* untuk mencari informasi data kata yang dominan muncul (sering muncul) dari konten status dan komentar akun *twitter* yang akan menghasilkan tampilan *word cloud* dari *widget Orange Data Mining*.

### 1.5.3. Metode Pengolahan Data

Sebelum analisis *text* data masih berupa data mentah hasil *crawling* dari *web* dan masih mengandung beberapa simbol, aksen, dan lain-lain yang akan diproses menggunakan metode pengolahan data *Preprocess Text*. Berikut ini metode pengolahan data dalam penelitian ini:

1. ***Preprocess Text***

*Preprocess Text* membagi teks sehingga menjadi unit-unit yang lebih kecil seperti *transformation, tokenization, filtering* dan melakukan normalisasi (*stemming dan lemmatization*). Langkah-langkah dalam analisis adalah diterapkan secara berurutan dan dapat diaktifkan atau dinonaktifkan. *Preprocess text* akan melaporkan beberapa hal seperti jumlah dokumen (tentang jumlah dokumen yang di masukkan), total token (menghitung semua token dalam *corpus* (kumpulan dokumen)), dan token yang dilaporkan hanya pada token unik di *corpus* bukan token duplikat. Dalam *preprocess text* data penelitian ini akan melalui tahap-tahap berikut:

a. ***Transformation***

Mengubah data input. Meliputi: ***Lowercase***, akan mengubah semua teks menjadi huruf kecil. ***Remove Accents***, akan menghapus semua dikritik/aksen dalam teks; contoh: *naïve* → *naive*. ***Parse html*** akan mendeteksi *tag html* dan menguraikan teks saja .: <*a href..*> beberapa teks </a> → beberapa teks. ***Remove url*** akan menghapus *url* dari teks. Ini sebuah *url http://orange.biolab.si/.* → ini sebuah *url*.

b. ***Tokenization***

Tokenisasi adalah metode memecah teks menjadi komponen yang lebih kecil (kata, kalimat, *bigrams*). Meliputi: ***Word dan Punctuation*** akan membagi teks berdasarkan kata dan tetap membiarkan *symbol* tanda baca (tidak menghilangkannya); contoh: *This Sample*. (*This*),

(*sample*),(.). ***Tweet,*** yang akan membagi teks dengan model *Twitter pra-trained*, yang memuat *hashtag*, *emoticons* dan simbol khusus lainnya. Contoh: *This words*. :-) *#simple* → (*This*), (*Words*), (.), (:-)), (*#simple*). Pada dasarnya *Word* dan *Punctuation* serta *tweet* memiliki kesamaan sifat proses, akan tetapi *word* dan *punctuation* menjadi proses utama dalam *tokenization*. *Word* dan *punctuation* juga digunakan untuk analisis tren.

**c. *Normalization***

Berlaku untuk *stemming* (menguraikan) dan *lemmatization* (memilih) kata-kata. (contoh *I've always loved cats → I have always loved cats*). Proses ini akan cenderung menggunakan ***wordnet lemmatizer*** proses ini mencoba mencocokan teks (sinonim/prediksi *text*) berdasarkan *database lexicon* yang besar dari NLTK. Proses ini juga mencoba menganalisis *typo* dalam *text* (karena penggunaan MT (*Machine Translate*).

**d. *Filtering***

Pemfilteran menghapus atau menyimpan pilihan kata. Meliputi: ***Stopwords*** menghapus kata-kata penutup dari teks (misalnya menghapus *'and', 'or', 'in'* ...). Dapat juga memuat daftar kata-kata sandi sendiri yang disediakan dalam format **.txt file* dengan satu *stopword* per baris. Dengan opsi pilihan bahasa *filtering* dapat digunakan untuk banyak bahasa, bahasa Indonesia ditetapkan sebagai

*default*. ***Regexp*** menghilangkan kata-kata yang cocok dengan ekspresi reguler ini \.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\’|…|\-|–|—\$|&|*|>|<|\/| Dan secara *default* diatur untuk menghapus tanda baca. *Most Frekuensi Token* (kata kemunculan terbanyak) dengan *wordcloud* pada *Orange Data Mining* pada dasarnya untuk melihat kalimat apa yang sering muncul dalam sebuah dokumen dan menentukan berapa banyak yang akan ditampilkan hal ini menggunakan *Most Frekuensi Token*.

### 1.5.4. Metode Analisis Data

Metode yang digunakan untuk penerapan *data mining* menggunakan metode *clustering* (pengelompokkan). Dalam penerapan *data mining* ini menggunakan tahapan KDD (*Knowledge Discovery in Database*) dari beberapa tahapan, yaitu: (Yunita, 2018)

1. ***Data Selection***

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *KDD* dimulai. Data hasil seleksi yang digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. ***Pre-processing/Cleaning***

   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus *KDD*. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

3. ***Transformation***

   Proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses dalam *KDD* merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. ***Data Mining***

   *Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *KDD* secara keseluruhan.

5. ***Interpretation/Evalution***

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses *KDD* yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

## 1.6. Sistematika Penulisan

Pada penelitian ini sistematika penulisan dijelaskan pada bagian ini. Berikut sistematika penulisan beserta penjelasannya.

**BAB I PENDAHULUAN**

Berisi tentang latar belakang penelitian ini, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian, metodologi penelitian, dan penjelasan sistematika penulisan.

**BAB II TINJAUAN PUSTAKA**

Memuat hasil-hasil penelitian sejenis terdahulu yang menginspirasi atau melandasi pelaksanaan penelitian ini, dan juga mengulas landasan teoritik yang berhubungan dengan penelitian yang akan dilakukan.

**BAB III ANALISIS DAN RANCANGAN**

Memuat analisis penelitian, meninjau data, mendesain penelitian, serta instrumen penelitian dan juga berisi analisis keperluan alat dan bahan penelian ini.

**BAB IV HASIL DAN PEMBAHASAN**

Bagian yang memuat berlangsungnya penelitian, pengolahan data, dan menyajikan hasil-hasil yang diperoleh dan cara pencapaiannya serta membahas hasil analisis penelitian ini.

**BAB V PENUTUP**

Berisi rangkuman hasil penelitian sebagai jawaban rumusan masalah, serta saran yang perlu diperhatikan berdasarkan asumsi.