

# BAB I

## PENDAHULUAN

Dalam bab ini akan dibahas latar belakang dilaksanakannya penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

### 1.1 Latar Belakang

Dalam beberapa tahun terakhir internet menjadi sumber informasi paling penting. Salah satu penyebabnya adalah banyaknya jejaring sosial. Jejaring sosial telah menjadi sumber informasi penting dan sumber terkini dalam pencarian informasi. Contoh dari jejaring sosial adalah Blog, Twitter, Facebook, Tumblr dan lain sebagainya. Keempat contoh jejaring sosial tersebut mempunyai peranan besar dalam menghasilkan informasi tentang peristiwa-peristiwa yang terjadi di masa lalu, sekarang dan masa yang akan datang.

Twitter merupakan salah satu situs *microblog* yang dioperasikan oleh Twitter, Inc. Disebut *microblog* karena situs ini memungkinkan penggunanya mengirim dan membaca pesan seperti blog pada umumnya, seperti pesan teks ataupun *micromedia* (foto dan *audio clip*). Pesan tersebut diberi nama *tweet*, yaitu teks tulisan dengan panjang maksimal 140 karakter yang ditampilkan pada halaman profil pengguna. Pengguna seringkali mengirim *tweet* yang berisi tentang apa yang pengguna lakukan dan apa yang pengguna rasakan.

Twitter berkembang sangat pesat, pada tahun 2007 terdapat 400.000 *tweet* yang diposting per kuartal, pada tahun 2008 jumlah *tweet* bertambah menjadi 100 juta kicauan per kuartal. Pada tahun 2014 twitter telah memiliki 645 juta pengguna aktif dan menghasilkan 9.100 *tweets* setiap detiknya (Statistic Brain, 2014). Indonesia sendiri merupakan negara ketiga di Asia Pasifik dengan jumlah pengguna aktif sebanyak 22,1 juta (Statista, 2015). Berdasarkan hal tersebut twitter berpotensi sebagai sumber informasi yang dapat dimanfaatkan.

Bandung memiliki 2,4 juta penduduk berada di posisi lima sebagai kota dengan pengguna twitter teraktif di dunia (SemioCast, 2012). Menurut data

pemerintahan kota Bandung, kota yang dikenal sebagai kota kembang dipenuhi dengan *mall*, *factory outlet*, tempat wisata dan kuliner. Berdasarkan hal tersebut informasi tentang Bandung sangat menarik untuk digali kembali lebih jauh. Informasi tersebut dapat diperoleh dari mana saja, salah satunya melalui twitter. Namun informasi yang didapatkan dari twitter sangatlah beragam, dibutuhkan pemrosesan khusus untuk dapat memperoleh topik-topik yang sedang diperbincangkan oleh pengguna tentang Bandung. *Social media monitoring* adalah kegiatan untuk mengetahui informasi dari percakapan-percakapan tentang suatu objek yang terjadi di jejaring sosial (Fensel, 2012). Terdapat beberapa fungsi dari *social media monitoring*, diantaranya untuk menemukan topik-topik dan memonitor percakapan pengguna jejaring sosial.

*Topic detection and tracking* (TDT) merupakan metode yang dapat digunakan untuk pengelompokan topik-topik pembicaraan secara otomatis (James Allan, 1998). Hal tersebut akan memudahkan pengguna dalam mencerna informasi. Penelitian tentang TDT dilakukan pertama kali oleh James Allan dkk pada tahun 1998. Data yang digunakan dalam penelitian ini adalah berita dan program *broadcast* berita. Tahapan yang digunakan dalam penelitian ini adalah sebagai berikut: segmentasi, *event detection*, *retrospective event detection*, *new event detection* dan *topic tracking*. Penelitian ini menggunakan teknik *clustering* dengan algoritma *group-average clustering* untuk proses *event detection* dengan hasil *recall* 61% dan *precision* 69%. Sementara pada tahap *topic tracking* menggunakan teknik klasifikasi dengan algoritma k-Nearest Neighbors (k-NN) dengan hasil *precision* 61%.

Penelitian lain dilakukan oleh Gabor Cselle dkk pada tahun 2007 tentang *Topic detection and tracking* dengan menggunakan email. Metode yang digunakan pada penelitian ini adalah *topic detection* dan *topic tracking*. Pada tahap *topic detection* penelitian ini menggunakan teknik *clustering* algoritma *clustering single linkage* dengan hasil *recall* 93% dan *precision* 61%. Sementara pada tahap *topic tracking* menggunakan nilai *threshold* untuk menentukan apakah email yang baru masuk ke dalam topik yang sudah ada atau email tersebut merupakan topik yang baru. Tahapan ini mendapatkan hasil *recall* 71% dan

precision 13%. Walaupun berbagai penelitian TDT telah dilakukan, tapi penelitian TDT tentang Bandung belum pernah dilakukan sebelumnya. Oleh karena itu dengan data dari twitter akan dibuat sistem TDT dengan menggunakan metode *clustering* dan klasifikasi.

*Clustering* merupakan salah satu teknik dari *text mining* yang dapat mengelompokkan dokumen berdasarkan kemiripannya. Klasifikasi merupakan salah satu teknik dari *text mining* yang memiliki fungsi untuk mengelompokkan suatu teks ke dalam suatu kategori atau kelas tertentu. Secara teknis, tugas klasifikasi teks adalah memetakan fungsi tujuan (Arifin, D. dkk, 2012).

Metode yang digunakan pada penelitian ini yaitu menggunakan teknik *clustering* dengan menggunakan algoritma K-Means dan teknik klasifikasi dengan menggunakan algoritma k-Nearest Neighbors (k-NN). K-Means dipilih karena memiliki performa yang baik dan kemampuannya untuk melakukan *clustering* dengan data besar dengan sangat cepat (Narwati, 2012). Sementara itu, dengan mengacu kepada penelitian yang dilakukan James Allan dkk. K-NN dipilih karena memiliki hasil yang cukup baik dalam melakukan proses klasifikasi, selain itu menurut Riyad Al-Shalabi pada tahun 2005 k-NN memiliki ketangguhan dalam mengolah dokumen yang besar dan memiliki banyak *noise*.

Berdasarkan penjelasan di atas penulis bermaksud melakukan penelitian tentang *topic detection and tracking* dengan data twitter menggunakan algoritma K-Means dan k-NN.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, ada beberapa permasalahan yang timbul dalam *topic detection and tracking* ini :

1. Bagaimana melakukan TDT pada data twitter berkata kunci “bandung” dengan menggunakan algoritma *clustering* K-Means dan algoritma klasifikasi k-NN?
2. Apakah aplikasi TDT dapat dimanfaatkan sebagai alat untuk menemukan topik pembicaraan tentang Bandung di twitter?

### 1.3 Batasan Masalah

Untuk memfokuskan penelitian, ditetapkan beberapa batasan masalah, yaitu sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah *tweet* berbahasa Indonesia.
2. *Tweet* yang digunakan dalam penelitian ini adalah *tweet* dengan kata kunci “bandung”.

### 1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Dapat melakukan proses *topic detection* atau mengelompokkan *tweet* dengan topik yang sama menggunakan algoritma K-Means dan melakukan proses *topic tracking* atau proses klasifikasi *tweet* baru dengan menggunakan algoritma k-NN.
2. Aplikasi TDT dapat bermanfaat sebagai alat untuk mengetahui topik pembicaraan tentang Bandung di twitter.

### 1.5 Manfaat Penelitian

Dengan dilakukannya penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan alat yang memudahkan dalam *topic detection and tracking tweet* berbahasa Indonesia.
2. Dapat menjadi rujukan bagi penelitian selanjutnya yang memiliki keterkaitan dengan penelitian ini baik secara langsung maupun tidak langsung.

### 1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut.

## BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian yang akan dilakukan, dan sistematikan penulisan.

## BAB II TINJAUAN PUSTAKA

Bab ini berisi penjelasan tentang teori-teori dan konsep algoritma yang digunakan dalam penelitian.

## BAB III METODOLOGI PENELITIAN

Bab ini berisi penjelasan langkah-langkah yang akan dilakukan dalam penelitian.

## BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi uraian tentang hasil penelitian dan pembahasan terhadap hasil penelitian yang dilakukan

## BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari keseluruhan penelitian yang telah dilakukan, serta saran dari penulis untuk kegiatan penelitian selanjutnya terkait dengan topik yang sedang dibahas.