# BAB I

# PENDAHULUAN

## 1.1. Latar Belakang

Pada saat ini perkembangan hidup manusia sangat dipengaruhi oleh perkembangan media sosial. Dari media sosial manusia memperoleh banyak informasi karna ruang lingkup media sosial yang tak terbatas. Media sosial ada dalam berbagai bentuk yang berbeda, termasuk social network, forum internet, weblogs, social blogs, micro blogging, wikis, podcasts, gambar, video,rating dan bookmark social[Kaplan and Haenlein, 2010]. Dari informasi yang didapat membuat pekerjaan manusia menjadi mudah, sehingga membuat media sosial menjadi salah satu kebutuhan yang dibutuhkan manusia.

Dari banyaknya media sosial *Twitter* adalah salah satu yang paling populer.Dari banyaknya kegiatanyang biasanya dilakukan di *Twitter* adalah melakukanupdate status, melakukan *tweet*,comment ataupun *re-tweet* yang merupakankegiatan interaksi dalam kehidupan sehari-hari yang tuangkan melalui sosial media.

*Twitter* adalah layanan sosial media yang dimiliki oleh *Twitter* inc dan sudah sangat banyak melayani pengguna sehingga menjadikannya sebagai salah satu sosial media yang memiliki data yang sangat besar. Menurut Semiocast 2nd (2013), tercatat bahwa saat ini terdapat lebih dari 500 juta pengguna *Twitter* dari seluruh dunia[Antoni et al., 2015]. Dari data yang sangat besar tersebut dapat dimanfaatkan sebagai sumber data untuk web berita *online*. Akan tetapi dengan banyaknya topik yang di ektraksi pada data *Twitter* membuat data yang masuk memiliki beragam topik yang menyebabkan kesulitan dalam mengidentifikasi topik dari kumpulan data yang diambil dan akan membutuhkan waktu yang banyak jika harus dilakukan secara manual oleh manusia. Sedangkan, data tersebut berpotensi dibutuhkan untuk memberikan informasi secepat mungkin.

Pada umumnya situs-situs berita *online*belum melakukan pengklasifikasi otomatis sesuaidengan topik pembahasan berita tersebut sehingga terjadiketerlambatan dalam menyebarkan informasi. Hal ini disebabkan dokumen beritayang di ambil dari *Twitter* sangatlah banyak dalam rentangwaktu yang cepat, sedangkan data tersebut, berpotensi dibutuhkan secepat mungkin.

Klasifikasi merupakan salah satu metode dalam *Text mining* yang bertujuan untuk mendefinisikan kelas dari sebuah objek yang belum diketahui kelasnya. Penentuan obyek dapat menggunakan suatu model tertentu bebeapa model yang bisa digunakan antara lain: *classification* (*IF-THEN*) *rules,decision trees*, formula matematika atau *neural networks* (Han, J., Kamber, M., dan Pei, J., 2006)[Raharjo and Winarko, 2014].

Pembandingkan dua metode yaitu Naive Bayes dan *Support Vector Machine* (SVM) menggunakan*stemming Confix Stripping Stemmer* telah dilakukan oleh Ariadi dan Fithriasari (2015) membandingkan dua metode yaitu Naive Bayes dan *Support Vector Machine* (SVM) menggunakan stemming *Confix Stripping Stemmer*. Penelitian tersebut memberikan hasil akurasi, *precision*, *recall*, dan *f-measure* sebesar 82,2%, 83,9%, 82,2%, dan 82,4% untuk metode Naive Bayes, sedangkan untuk metode SVM memberikan hasil akurasi, *precision*, *recall*, dan *f-measure* sebesar 88,1%, 89,1%, 88,1%, dan 88,3%[Prasetyo and Kusumanigrum, 2018].

Dikarenakan beberapa metode klasifikasi teks yang telah disebutkan sebelumnya masih memiliki kelemahan-kelemahan. Klasifikasi menggunakan variabel bebas sudah banyak di implementasikan menggunkan metode Naive Bayes dan hasilnya bergantung pada fitur yang digunakan, dan tidak berlaku jika probabilitas kondisionalnya adalah nol. sedangan klasifikasi menggunakan metode SVM terdapat banyak teks yang tidak dapat di klasifikasikan dengan benar dikarenakan karakteristik dimensi yang tinggi, masih kaku dan kinerja tergantung pada pemilihan fungsi kernel menimbulkan masalah *sparseness* data (Kusumaningrum et al., 2016).

Dimana terdapat kumpulan data yang mengandung nilai mendekati nol lebih dominan disebut *Sparnesses*data[Prasetyo and Kusumanigrum, 2018].

Oleh sebab itu, metode reduksi diperlukan pada dimensi data yang besar dan implementasi klasifikasi pada dokumen berita Bahasa Indonesia dengan konsep topik *modelling*, antara lain seperti *Probabilistic Latent* Semantic *Analysis* (*PLSA*), *Latent Semantic Analysis* (*LSA*), dan *Latent Dirichlet Allocation* (*LDA*). *LSA* memiliki kekurangan dalam pengengolah data dalam jumlah yang besar.*PLSA* merupakan pembaharuan dari *LSA*, *PLSA* dapat mencari topik yang tersembunyi pada korpus(Hofmann, 1999)[Prasetyo and Kusumanigrum, 2018]. *Latent Dirichlet Allocation* (*LDA*) menggunakan model hirearki sehingga lebih stabil dan dapat mengolah data dalam jumlah besar (Liu, 2013)[Prasetyo and Kusumanigrum, 2018].

Samuel Adi Prasetyo (2018) telah melakukan penelitian dengan jumlah data pelatihan sebanyak 1000 berita (200 berita per kategori) mendapatkan hasil menggunakan metode gabungan *LDA* dan *Word2Vec* sudah cukup baik dalam melakukan klasifikasi dengan nilai akurasi tertinggi sebesar 73,4%. Menyimpulkan akurasi lebih baik didapatkan oleh metode *LDA* murni tanpa *Word2Vec* dengan nilai akurasi sebesar 87,5% sehingga memiliki selisih akurasi sebesar 14,1%. Kedua perbandingan metode tersebut sama-sama diperoleh pada kombinasi parameter alpha 0,1; beta 0,01; dan jumlah topik sebanyak 300 topik[Prasetyo and Kusumanigrum, 2018].

I Made Kusnanta Bramantya Putra (2017) melakukan penelitian pengujian secara mesin dengan nilai *Perplexity* terbaik sebesar 213.41 dan diuji kemudahannya untuk diinterpretasi oleh manusia melalui uji koherensi topik yang terdiri dari *Twitter* Word *Intrusion task* dan Topik *Intrusion Task*. Kesimpulan dari uji koherensi topik menyatakan bahwa model yang dihasilkan dengan metode *LDA* pada studi kasus ini dapat diinterpretasi manusia dengan baik[Putra, 2017].

Dokumen berita yang akan digunakan dalam penelitian ini antara lain adalah hasil dari *crawling* data *Twitter* dengan menggunakan *Application*

Programming Interface (API) yang telah disediakan oleh *Twitter* menghasilkan kumpulandata text berdasarkan update yang telah di unggah oleh pengguna *Twitter*. Proses *crawling* berhasil terhadap data *Twitter* dengan menggunakan Application*Programming Interface* dan telah menghasilkan data yang informatif melalui proses *Crawling*[Negara et al., 2016].

Dari hasil dari beberapa penelitian yang disebutkan sebelumnya, peneliti memutuskan untuk menggunakan metode *Latent Dirichlet Allocation* (*LDA*) dengan judul **Klasifikasi Topik pada Web Berita *Online* dengan Pendekatan Teks Mining.**

## 1.2. IdentifikasiMasalah

Berdasarkan pada latar belakang yang telah dijelaskan, disimpulkan bahwa identifikasi masalah terdiri dari :

1.2.1. Web berita yang terkendala dengan banyaknya waktu digunakanuntuk klasifikasi topik berita.

1.2.2. Belum adanya klasifikasi topik otomatis pada web berita *online*.

## 1.3. RumusanMasalah

Berdasarkan pada uraikan dalam identifikasi masalah diatas, maka dari itu permasalahan yang terjadi yaitu :

1.3.1. Bagaimana mengembangkan Aplikasi pengumpulan pada data *Twitter.*

1.3.2. Bagaimana melakukan klasifikasi topik berita pada web berita *online* menggunakan metode *Latent Dirichlet Allocation*(*LDA*).

## 1.4. Batasan Masalah

Batasan masalah yang akan dibahas dalam penelitian yang akan dilakukan adalah :

1. Data yang digunakan adalah tweet dari media sosal online twitter dengan hastag Indonesia.

2. Pengelolahan data ini menggunakan text preprosesing, stemming, stopword, text mining dan Latent Dirichlet Allocation (LDA).
3. Bahasa pemrograman yang digunakan adalah python.

## 1.5. Manfaat Penelitian

Dengan dilakukannya penelitian ini diharapkan hasilnya dapat memberikan kontribusi pada pengembangan aplikasi pengumpulan data *Twitter* untuk web berita *online* mengenai klasifikasi dokumen berita bahasa Indonesia dengan menggunakan metode Latent Dirichlet Allocation (LDA).

## 1.6. TujuanPenelitian

Tujuan yang ingin dicapai pada penelitian ini yaitu sebagai berikut :

1.6.1. Untuk mengembangkan aplikasi pengumpulan dokumen berita data Twitter untuk berita pada web berita online.

1.6.2. Mengklasifikasi topik pada dokumen berita untuk web berita online menggunakan metode Latent Dirichlet Allocation (LDA).