# BAB III

## ***Data Collection***

Dalam prosedur pengumpulan data, penelitian ini memperoleh kumpulan tweet dengan memanfaatkan library yang tersedia di Python untuk mempermudah mengumpulkan tweet. Pengumpulan tweet dilakukan dengan pertama menentukan kata kunci yang akan dipakai untuk *crawling* tweet, setelah kata kunci ditentukan kemudian ditentukan **juga** batas tahun untuk pengambilan tweet. Untuk penelitian ini tweet diambil dengan kurun waktu 10 tahun terakhir. Dengan kata kunci sesuai tabel dibawah.

|  |  |
| --- | --- |
| Relevansi | Keyword |
| Evakuasi | evakuasi AND bencana |
| Makanan/Minuman | (butuh OR bantuan) AND (makanan OR minuman OR pangan) AND bencana |
| Pakaian | (butuh OR bantuan) AND (baju OR celana OR pakaian) AND bencana |
| Kesehatan | (butuh OR bantuan) AND (p3k OR medis OR dokter) AND bencana |
| Perlindungan | (butuh OR bantuan) AND tempat perlindungan |
| Umum | (tolong OR bantuan) AND bencana |
| Umum | (tolong OR bantuan) AND (banjir OR (gempa OR (gempa bumi)) OR tsunami OR longsor OR (gunung berapa OR gunung meletus) OR erupsi OR (angin puting beliung OR puting beliung) OR badai OR (angin kencang) OR kebakaran OR (gelombang pasang) OR abrasi) |

## ***Text Preprocessing***

Data yang sudah terkumpul melalui proses *crawling* kemudian akan di proses lebih lanjut untuk dibersihkan terlebih dahulu, tujuan utama dari *text preprocessing* adalah untuk membersihkan, mengubah, dan mengorganisasi data sehingga dapat digunakan untuk pelatihan model pembelajaran mesin yang lebih akurat. Beberapa tahapan *text preprocessing* pada penelitian ini diantara lainnya:

* + 1. Menghilangkan *escape sequences*, *escape sequenecs* adalah suatu karakter khusus dalam bahasa pemrograman yang digunakan untuk mewakili karakter khusus seperti karakter *newline* (\n), karakter tab (\t), dan lainnya.
    2. Menghilangkan tanda baca dan *lower casing*, berikutnya tanda baca harus dihilangkan dalam teks tersebut dan teks juga harus diubah dalam huruf kecil semua.
    3. *Tokenizing*, *tokenizing* adalah suatu proses memecah suatu kalimat dalam sebuah teks menjadi struktur terkecil seperti contohnya memecah kumpulan kalimat menjadi kumpulan kata.
    4. Menghapus *stopwords*, *stopwords* adalah kata-kata yang sering muncul didalam suatu dokumen dan dianggap tidak memiliki makna atau kontribusi penting dalam pemahaman suatu teks.
    5. *Lemmatizing* dengan POS-Tagging, pada bagian kumpulan kata pada suatu dokumen akan diubah ke dalam bentuk dasarnya dengan mempertimbangkan jenis kata tersebut. Dengan merubah semua kata menjadi kata dasarnya akan berguna untuk mengurangi kompleksitas teks dan membantu menganalisis kata dalam suatu dokumen.

## **Melatih Model *Word Embedding***

*Word embedding* adalah suatu teknik pengolahan bahasa yang digunakan untuk merepresentasikan kata-kata dalam suatu dokumen menjadi suatu bentuk vektor. Setiap kata dalam suatu dokumen direpresentasikan oleh vektor yang mengandung informasi tentang arti dan makna kata tersebut. Tujuan dari *word embedding* adalah untuk memudahkan menganalisis bahasa pada teks. Dengan merepresentasikan kata-kata dalam bentuk vektor, kata tersebut dapat dilakukan operasi matematika untuk menghitung kesamaan atau perbedaan antara dua kata, mencari kata-kata terdekat dalam suatu konteks, dan lainnya.

Algoritma *word embedding* yang akan digunakan pada penelitian ini adalah word2vec, word2vec adalah salah satu algoritma word embedding yang paling terkenal dan paling banyak digunakan dalam pengolahan bahasa alami. Word2vec bekerja dengan mempelajari hubungan antara kata-kata dalam sebuah korpus teks, dan menghasilkan representasi vektor untuk setiap kata. Algoritma ini memperhitungkan kemunculan kata-kata dalam konteks yang berbeda, sehingga kata-kata yang sering muncul akan memiliki representasi vektor yang mirip satu sama lain.

## **Melatih Model Relevansi**

Model relevansi dibuat dengan tujuan untuk menentukan dari kumpulan tweet yang sudah di *crawling*, tweet **manakah** yang sesuai atau relevan dengan studi kasus pada penelitian ini. Dengan memanfaatkan salah satu algoritma *semi-supervised* dan menggunakan model *word embedding*, pertama kumpulan teks akan dilabelkan secara *manual*, pelabelan manual ini hanya memakai Sebagian kecil dari total teks yang tersedia. Setelah 1000 teks pertama dilabelkan, kemudian seluruh teks akan di representasikan dalam suatu vektor melalui model *word embedding*. Kemudian kumpulan vektor tersebut akan dilatih untuk menentukan apakah vektor yang merepresentasikan teks tersebut termasuk relevan atau tidak.

Tujuan dari proses melatih model relevansi adalah untuk memilah teks **manakah** yang sesuai dengan tema bantuan untuk korban bencana alam. Dari kumpulan teks yang sudah diperolah melalui proses *crawling* tidak semua teks itu berisi tentang bantuan untuk korban bencana alam, oleh karena itu dengan melatih model relevansi akan membantu untuk memilah teks **manakah** yang relevan atau tidak.

## **Melatih Model Klasifikasi Tipe Bantuan**

Model klasifikasi tipe bantuan bencana dibuat dengan tujuan untuk menentukan dari kumpulan tweet yang sudah di *crawling*, tweet tersebut masuk ke kategori bantuan apa yang dibutuhkan. Model klasifikasi tipe bantuan akan dibangun menggunakan algoritma *mean teacher* dan transformer. Pertama, dari kumpulan tweet, sebagian kecil tweet akan dilabelkan sesuai dengan kategori tipe bantuan bencana, kemudian menggunakan algoritma *mean teacher* tweet yang belum memiliki label dapat digunakan menggunakan pendekatan *semi-supervised learning.*

Algoritma *mean teacher* memiliki 2 arsitektur model didalamnya yang disebut *student* dan *teacher*, dimana untuk penelitian ini akan digunakan arsitektur transformer sebagai arsitektur dasar *mean teacher*, kemudian data yang memiliki label akan diinput ke arsitektur *student* dimana *student* akan coba mempelajari secara *supervised* dari data dan label tersebut kemudian akan dibandingkan dengan data aktual menggunakan *classification loss*. Kemudian *weight* dari *student* akan ditransformasi menggunakan *Exponential Moving Average* dan *weight* tersebut akan digunakan sebagai *weight* dari *teacher*. Kemudian model *teacher* akan melakukan prediksi konsistensi dari data yang tidak memiliki label dengan pendekatan *unsupervised*, dan akan dibandingkan dengan data aktual menggunakan *consistency loss*. Terakhir *loss* dari *classification* dan *consistency* akan ditambahkan sebagai *loss* total.