**BAB III**

**METODE PENELITIAN**

Pada BAB ini, akan dijelaskan tahapan metode penelitian yang terdiri dari pengolahan data dan perancangan sistem yang akan digunakan dalam penelitian.

* 1. **Pengolahan Data**

Pada tahap pengolahan data, akan diuraikan metode dalam mendapatkan dataset, penerjemahan dataset menjadi Bahasa Inggris, labelling dataset hingga preprocessing dataset. Tahapan pengolahan data ini bertujuan untuk mengolah data agar dapat digunakan secara optimal pada saat penelitian menggunakan sistem yang akan dirancang nantinya. Tahapan pengolahan data dapat dilihat dalam gambar 3.1.



Gambar 3.1 Tahapan Pengolahan Data

* + 1. **Pengumpulan Dataset**

Data yang digunakan dalam penelitian ini adalah hasil dari *web scrapping* di situs Twitter dengan mencari 20.000 total tweet yang mengandung kata “bjorka”. Peneliti melakukan *web scrapping* untuk mengambil tweet yang tidak mengandung link serta mention/reply dari pengguna lain sehingga tweet tersebut murni sebuah tweet yang berisi ide/gagasan/pendapat yang ingin dituliskan oleh pengguna. Setelah itu, data disatukan menjadi sebuah dataframe/table yang berisi username pengunggah tweet, waktu dibuatnya tweet tersebut serta isi tweet/teks tweet tersebut. Setelah itu, dataframe disimpan dalam format excel (xlsx) agar bisa dilihat dengan mudah dan dapat digunakan kembali ketika akan digunakan dalam penelitian. Berikut beberapa sampel dataset hasil web scrapping dapat dilihat pada table 3.1.

|  |  |  |
| --- | --- | --- |
| Tanggal dibuat | Username | Teks/tweets |
| 2022-11-02 | rudiartraa | yaah , udah selesai nih Bjorka Bjorkaan nya ? |
| 2022-11-02 | oi\_filix315 | Kadang gua mikir sekilas pp gua mirip bjorka, tapi kan gua udh pake pp ini lama :((( |
| 2022-11-01 | falahsanjaya\_ | Knp ga ada yg jadi Bjorka Halloween party kemaren? :(( |
| 2022-11-01 | namanyaajess | Beritanya taik, mau ferdi sambo, mau bjorka, mau apalah yang kelaperan ga mungkin jadi tranding topik, yang tertindas sulit untuk berdiri, yang melawan akan tetap dibungkam. |
| 2022-11-01 | namakubaim | Apa kabar Bjorka? |

Tabel 3.1 Sample dataset

Selain itu, peneliti mengumpulkan dataset yang berisi kata-kata alay yang nantinya digunakan untuk melakukan transformasi kata-kata alay yang mungkin saja di dapatkan di dalam tweet menjadi bentuk normalnya. Dataset alay ini didapatkan dari Kaggle.com dalam format csv (comma separated value). Berikut beberapa sample dari kata-kata alay yang dapat dilihat pada table 3.2.

|  |  |
| --- | --- |
| Kata Alay | Pembenaran |
| aje | aja |
| ajg | anjing |
| diem | diam |
| ente | kamu |
| gr | gede rasa |
| ijo | hijau |
| knp | kenapa |

Tabel 3.2 Sample kata-kata alay dan pembenarannya

Kata-kata alay ini nantinya akan berguna Ketika memasuki tahap preprocessing data untuk mentransformasi kata alay menjadi bentuk dasar/bentuk benar.

* + 1. **Penerjemahan Dataset**

Tahapan ini bertujuan untuk menerjemahkan data tweet dari Bahasa Indonesia menjadi Bahasa Inggris. Hal ini perlu dilakukan karena dalam pelabelan menggunakan Vader, dibutuhkan dataset dengan Bahasa inggris sehingga dapat diketahui kutub/kecenderungan sebuah kalimat, apakah positif, negative, atau bahkan netral. Dataset diterjemahkan melalui program python menggunakan libraries translate. Penggunaan libraries ini memungkinkan untuk melakukan translate dengan memanggil API libraries tersebut dan melakukan translate di dalam program python, kemudian menambahkan hasil translate menjadi kolom baru. Berikut beberapa sample hasil penerjemahan dataset Bahasa Indonesia menjadi Bahasa Inggris yang dapat dilihat pada table 3.3.

|  |  |
| --- | --- |
| Indonesia | Inggris |
| yaah , udah selesai nih Bjorka Bjorkaan nya ? | yeah, have you finished Bjorka Bjorkaan? |
| Apa kabar Bjorka? | How are Bjorka? |
| Bjorka kemana ya | Where is Bjorka |
| Andai bjorka bisa tolong,,, 😞 | If bjorka can help ,,, 😞 |
| Pagi semua, kecuali bjorka | Morning all, except Bjorka |

Tabel 3.3 Sample dataset setelah diterjemahkan menjadi Bahasa Inggris menggunakan libraries translator dari python.

Setelah diterjemahkan, maka akan melakukan pencarian kutub kecenderungan dataset, baik bermuatan positif, negative ataupun netral.

* + 1. **Pelabelan Dataset**

Pelabelan data berfungsi untuk memberikan label yang terdiri dari positif, negative maupun netral terhadap dataset. Hal ini perlu dilakukan karena model deep learning yang akan dibangun membutuhkan sebuah label untuk dapat berlatih dan mengklasifikasikan ujaran kebencian twitter. Untuk melakukan pelabelan terhadap dataset, menggunakan Vader Lexicon. Vader (Valence Aware Dictionary and Sentiment Reasoner) Lexicon merupakan sebuah tools/libraries yang berfungsi untuk melakukan sentiment analysis berbasis aturan yang secara khusus disesuaikan dengan sentiment yang diungkapkan di media sosial (MIT, 2014). Tools ini bersifat open-source dan dibawah lisensi MIT. Vader dapat menangani beberapa kasus khusus dalam sentiment analysis, seperti memahami negasi, memahami arti emoticon/emoji, memahami arti sebuah akronim, penggunaan tanda baca konvensional menandakan peningkatan intensitas sentiment, dan masih banyak lagi.

Dengan menggunakan Vader, data teks akan diproses dan dicari nilai sentimennya. Output yang dihasilkan oleh Vader merupakan sebuah probabilitas dari 3 kemungkinan yaitu positif, negative dan netral (2 yaitu positif dan negative jika menggunakan 2 kutub saja). Dari probabilitas tersebut, untuk melakukan labelling terhadap suatu teks di dalam dataset, maka diambil nilai probabilitas paling tinggi diantara ketiga kutub tersebut. Setelah menentukan nilai probabilitas paling tinggi untuk setiap teks dalam dataset dan melakukan pelabelan terhadap dataset, maka dataset hasil pelabelan siap digunakan untuk diproses ditahap selanjutnya yaitu tahap preprocessing. Berikut contoh output dari Vader sentiment analysis.

|  |  |  |
| --- | --- | --- |
| **Teks** | **Vader Output** | **Label** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Tabel 3.4 Labeling Data

* + 1. **Preprocessing Dataset**

Tahapan ini bertujuan untuk membuat kualitas data yang akan digunakan dalam training/analisis menjadi baik. Dalam kasus sentiment analysis ujaran kebencian ini, untuk membuat kualitas dataset teks menjadi baik, maka perlu dilakukan beberapa tahapan preprocessing diantaranya adalah lowercasing, pembersihan karakter selain alphanumeric, pembersihan karakter khusus/tidak biasa, normalisasi kata-kata alay, stemming, membersihkan stopword, tokenizing serta tranformasi data.

1. Lowercasing

Lowercasing merupakan tahapan transformasi huruf menjadi huruf kecil. Input dari tahapan ini adalah sebuah kalimat yang terdiri dari huruf besar dan kecil yang akan diproses menjadi sebuah output kalimat yang terdiri dari huruf kecil saja. Berikut input dan output setelah dilakukan lowercasing dapat dilihat pada table 3.5.

|  |  |
| --- | --- |
| **Sebelum** | **Sesudah** |
| yaah , udah selesai nih Bjorka Bjorkaan nya ? | yaah , udah selesai nih bjorka bjorkaan nya ? |
| Apa kabar Bjorka? | apa kabar bjorka? |
| Kadang gua mikir sekilas pp gua mirip bjorka, tapi kan gua udh pake pp ini lama :((( | kadang gua mikir sekilas pp gua mirip bjorka, tapi kan gua udh pake pp ini lama :((( |

Tabel 3.5 Lowercasing

1. Pembersihan Non-Alphanumeric

Alphanumeric merupakan karakter yang teridiri dari huruf dan angka. Pembersihan Non-Alphanumeric bertujuan untuk membersihkan karakter selain huruf dan angka, seperti tanda baca dan karakter khusus lainnya. Tahapan ini dilakukan agar data yang akan digunakan dalam penelitian dapat bekerja secara maksimal karena dalam membuat model deep learning, nantinya setiap kata akan dijadikan sebuah token yang di transformasi menjadi sebuah angka. Berikut input dan output untuk proses pembersihan Non-Alphanumeric karakter yang dapat dilihat pada table 3.6.

|  |  |
| --- | --- |
| **Sebelum** | **Sesudah** |
| yaah , udah selesai nih bjorka bjorkaan nya ? | yaah udah selesai nih bjorka bjorkaan nya |
| apa kabar bjorka? | apa kabar bjorka |
| kadang gua mikir sekilas pp gua mirip bjorka, tapi kan gua udh pake pp ini lama :((( | kadang gua mikir sekilas pp gua mirip bjorka tapi kan gua udh pake pp ini lama |

Tabel 3.6 Pembersihan Non-Alphanumeric

1. Pembersihan karakter khusus

Pembersihan karakter khusus ini bertujuan untuk menghapus beberapa karakter khusus yang ada di dalam kalimat/teks, seperti kalimat yang mengandung kata “\n”, url, serta emoji. Hal ini akan membuat dataset yang nantinya akan digunakan kualitasnya menjadi baik karena kata-kata khusus tersebut yang berasal dari media sosial twitter cenderung tercantum di dalam kalimat/dataset tersebut sehingga akan membuat model machine learning kurang baik dalam melakukan klasifikasi. Berikut contoh input dan output untuk pembersihan karakter khusus yang dapat dilihat pada table 3.7.

|  |  |
| --- | --- |
| **Sebelum** | **Sesudah** |
| rame rame pada hapus twit yg menuduh msf adalah sosok bjorka 🤣🤣 | rame rame pada hapus twit yg menuduh msf adalah sosok bjorka |
| bjorka kembali tebar ancaman, kecam semua akun palsuhttp://inet.detik.com/security/d-6303238/bjorka-kembali-tebar-ancaman-kecam-semua-akun-palsu?bytedance=true | bjorka kembali tebar ancaman, kecam semua akun palsu |
| bjorka, aku padamu🫶🏻🖤 | bjorka, aku padamu |

Tabel 3.7 Pembersihan karakter khusus

1. Normalisasi Kata-Kata Alay

Normalisasi kata-kata alay bertujuan untuk mengubah kata-kata yang dikategorikan menjadi kata-kata alay menjadi bentuk dasarnya. Hal ini bertujuan untuk mengurangi variansi kata yang ada di dalam dataset yang akan membuat kualitas dataset menjadi lebih baik dan optimal dalam klasifikasi nantinya. Tahapan ini akan menghilangkan kata alay yang sudah terdaftar pada dataset kata-kata alay yang sudah didapatkan melalui Kaggle. Berikut beberapa contoh normalisasi kata-kata alay menjadi bentuk dasarnya yang dapat dilihat pada table 3.8.

|  |  |
| --- | --- |
| **Sebelum** | **Sesudah** |
| anjirrr kok bisa adek gw percaya kalo bjorka itu gus syamsudin | anjirrr kok bisa adik gue percaya kalau bjorka itu gus syamsudin |
| knp ga ada yg jadi bjorka halloween party kemaren | kenapa tidak ada yang jadi bjorka halloween party kemarin |
| pengen gitu bikin sih bjorka biar gak bikin orang gak gagal fokus rakyat karena saat banyak kasus yg gagal fokus karena hacker bjorka bayaran pemerintah | ingin begitu bikin sih bjorka biar tidak bikin orang tidak gagal fokus rakyat karena saat banyak kasus yang gagal fokus karena hacker bjorka bayaran pemerintah |

Tabel 3.8 Normalisasi kata alay

1. Stemming word

Stemming merupakan Teknik menguraikan bentuk suatu kata menjadi bentuk asalnya/dasarnya. Hal ini dilakukan untuk mengubah kata berimbuhan menjadi bentuk dasarnya agar mempermudah model dalam melakukan klasifikasi, mengurangi variansi serta meningkatkan kualitas dataset menjadi lebih optimal. Sebagai contoh, untuk kata membenci akan ditransformasi menjadi benci, menyukai menjadi suka dan lebih banyak lagi. Berikut beberapa contoh stemming yang dapat dilihat pada table 3.9

|  |  |
| --- | --- |
| **Sebelum** | **Sesudah** |
| anjirrr kok bisa adek gw percaya kalo bjorka itu gus syamsudin | anjirrr kok bisa adik gue percaya kalo bjorka itu gus syamsudin |
| knp ga ada yg jadi bjorka halloween party kemaren | kenapa tidak ada yang jadi bjorka halloween party kemarin |
| pengen gitu bikin sih bjorka biar gak bikin orang gak gagal fokus rakyat karena saat banyak kasus yg gagal fokus karena hacker bjorka bayaran pemerintah | ingin gitu bikin sih bjorka biar tidak bikin orang tidak gagal fokus rakyat karena saat banyak kasus yang gagal fokus karena hacker bjorka bayaran pemerintah |

Tabel 3.9 Stemming word

1. Pembersihan stopword

Tahapan pembersihan stopword berfungsi untuk menghilangkan kata umum (common words) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Tujuan dari menghilangkan stop words adalah mengurangi jumlah kata dalam sebuah dokumen yang akan berpengaruh terhadap kecepatan dan performa model. Stop word yang akan dijadikan sebagai acuan didapatkan dari libraries python Sastrawi (Sastrawi, 2017). Berikut beberapa contoh data sebelum dan sesudah dilakukannya tahapan pembersihan stopword yang dapat dilihat pada tabel 3.10

|  |  |
| --- | --- |
| **Sebelum** | **Sesudah** |
| anjirrr kok bisa adek gw percaya kalo bjorka itu gus syamsudin | anjirrr kok bisa adik gue percaya bjorka gus syamsudin |
| knp ga ada yg jadi bjorka halloween party kemaren | bjorka halloween party kemarin |
| pengen gitu bikin sih bjorka biar gak bikin orang gak gagal fokus rakyat karena saat banyak kasus yg gagal fokus karena hacker bjorka bayaran pemerintah | gitu bikin sih bjorka biar bikin gagal fokus rakyat gagal fokus hacker bjorka bayaran pemerintah |

Tabel 3.10 Stopword Removal