

**PEMILIHAN FITUR BAGI ANALISIS SENTIMEN
MENGUNAKAN TEKNIK NAIVE BAYES**

FATIMAH BINTI RAHMAT

UNIVERSITI KEBANGSAAN MALAYSIA

PEMILIHAN FITUR BAGI ANALISIS SENTIMEN MENGGUNAKAN TEKNIK
NAIVE BAYES

FATIMAH BINTI RAHMAT

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS KOMPUTER
(TEKNOLOGI PERISIAN)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

22 Januari 2018

FATIMAH BINTI RAHMAT
GP04124

PENGHARGAAN

Syukur Alhamdulillah saya panjatkan kesyukuran kepada Allah S.W.T kerana memberikan saya peluang dalam bentuk kesihatan, masa dan pemberian buah fikiran untuk menyiapkan kajian penyelidikan ini. Jutaan terima kasih yang tidak terhingga saya ucapkan kepada penyelia saya Dr. Lailatul Qadri Binti Zakaria atas bantuan yang begitu besar, bimbingan dan nasihat yang amat saya perlukan sepanjang kajian ini. Ucapan penghargaan ini juga saya tujukan kepada pihak pengurusan Fakulti Teknologi dan Sains Maklumat yang banyak membantu saya dari segi kemudahan fasiliti, pihak pensyarah dan staf pentadbiran sepanjang sesi pengajian saya di sini.

Selain itu, penghargaan ribuan terima kasih yang tidak terhingga kepada Jabatan Pendidikan Politeknik yang memberi peluang, kerjasama dan bantuan yang berterusan dalam membantu menjayakan penyelidikan saya. Terima kasih juga diucapkan kepada rakan-rakan di UKM yang sama-sama membantu dan memberikan sokongan sepanjang pengajian.

Teristimewa buat ibu, bapa, suami dan anak-anak yang sentiasa ada di sisi saya dalam memberikan sokongan dan dorongan secara berterusan sehingga siapnya kajian penyelidikan ini. Sesungguhnya hanya Allah sahaja yang mampu membalas budi baik yang telah diberikan kepada saya selama ini.

Akhir kata, terima kasih kepada semua pihak yang terlibat sama ada secara langsung atau tidak sepanjang saya menyiapkan kajian saya ini. Semoga jasa kalian diberkati dan dirahmati Allah.

ABSTRAK

Tweet digunakan untuk menyampaikan maklumat dan lazimnya maklumat tersebut boleh dikategorikan sama ada positif ataupun negatif. Secara umumnya, tweet yang dimuatnaik di kawasan Institusi Pengajian Tinggi Awam (IPTA) lebih menjurus kepada penyampaian maklumat atau perasaan yang negatif seperti kesedihan, kemarahan, dan perasaan tidak puas hati. Namun begitu, hasil pemerhatian awal terhadap tiga IPTA iaitu Universiti Sains Malaysia (USM), Universiti Teknologi Malaysia (UTM) dan Universiti Malaysia Sabah (UMS) menunjukkan tweet yang positif lebih tinggi berbanding tweet negatif. Oleh itu, kajian ini akan menganalisis sentimen tweet di sekitar Universiti Kebangsaan Malaysia (UKM) bagi mendapatkan maklumat sama ada tweet di UKM selari dengan corak sentimen di ketiga-tiga IPTA tersebut. Kaedah kajian ini dibahagikan kepada lima fasa utama iaitu pengumpulan data, pra-pemprosesan, pemilihan fitur, pengkelasan *Naive Bayes* (NB) dan pengujian. Pengumpulan data ialah fasa awal mengesktrak data daripada Twitter. Pra-pemprosesan pula adalah proses pembersihan data daripada teks yang tidak digunakan. Pemilihan fitur ialah fasa menyenarai dan memilih fitur-fitur yang bersesuaian dengan data. Pengkelasan NB adalah algoritma yang digunakan semasa membuat pengkelasan. Pengujian pula adalah untuk mengetahui keberkesanan hasil daripada proses pengkelasan. Hasil kajian mendapati fitur yang sesuai digunakan untuk pengkelasan tweet ialah fitur frekuensi perkataan (TF), kata henti, kata henti khusus dan pencantasan adalah fitur terbaik bagi analisis sentimen terhadap data tweet. Kajian akan datang boleh diteruskan dengan menggunakan data dari laman media sosial Facebook dan seumpamanya. Oleh itu, semoga kajian pengkelasan tweet ini dapat dikembangkan dan menjadi rujukan untuk penyelidikan akan datang.

FEATURE SELECTION IN SENTIMENT ANALYSIS USING NAIVE BAYES TECHNIQUE

ABSTRACT

Tweet is a medium to disseminate information either in a positive or negative way. In general, tweet uploaded in Higher Learning Institution (HLI) environment more prone to negative sentiment such as sadness, anger, and dissatisfaction. However, preliminary investigation revealed that positive tweets were higher compared to negative tweet in three HLI which are USM, UTM and UMS respectively. Hence, this study will analyze the tweet sentiment in UKM in order to determine whether it is parallel with the sentiment pattern of the three HLI. Research methodology consists of five main phases which are data gathering, pre-processing, feature selection, Naive Bayes (NB) classification and testing. Data gathering is the first phase which is involving data extraction from Twitter. Meanwhile, pre-processing is the data cleaning process by eliminating data of unused texts. Then, feature selection is the phase of identifying and choosing the suitable feature that is matched with the data. NB classification is the technique for classifying the feature. Lastly, testing phase will measure the effectiveness of classification process. The study resulted in producing the best feature of tweet classification analysis sentiment of tweet data. The feature is Term Frequency (TF), stopwords, specific stopwords and stemming. In future, this study can be expanded with another social media tool such as Facebook. Hopefully, this classification study can be constructed and become a reference for future research.

KANDUNGAN

| | | Halaman |
|--------------------------|---|----------------|
| PENGAKUAN | | ii |
| PENGHARGAAN | | iii |
| ABSTRAK | | iv |
| ABSTRACT | | v |
| KANDUNGAN | | vi |
| SENARAI JADUAL | | ix |
| SENARAI ILUSTRASI | | x |
| SENARAI SINGKATAN | | xi |
| | | |
| BAB I | PENDAHULUAN | |
| 1.1 | Pengenalan | 1 |
| 1.2 | Penyataan Masalah | 3 |
| 1.3 | Persoalan Penyelidikan | 5 |
| 1.4 | Objektif Penyelidikan | 5 |
| 1.5 | Skop Dan Batasan Kajian | 6 |
| 1.6 | Kepentingan Kajian | 6 |
| 1.7 | Organisasi Projek | 7 |
| 1.8 | Kesimpulan | 7 |
| | | |
| BAB II | KAJIAN SUSASTERA | |
| 2.1 | Pengenalan | 8 |
| 2.2 | Kandungan Maklumat Yang Dijana Oleh Pengguna Media Sosial | 9 |
| 2.3 | Analisis Sentimen | 10 |
| 2.4 | Analisis Sentimen Dalam Data Media Sosial | 12 |
| | 2.4.1 Pemprosesan Bahasa Tabii (NLP) | 14 |
| | 2.4.2 Pembelajaran Mesin (ML) | 14 |
| 2.5 | Proses Pengkelasan Teks | 15 |
| | 2.5.1 Pra-pemprosesan | 16 |
| | 2.5.2 Pengindeksan | 16 |
| | 2.5.3 Pemilihan fitur | 17 |
| | 2.5.4 Algoritma pengkelasan | 17 |
| | 2.5.5 Pengukuran prestasi | 17 |

| | | |
|----------------|--|----|
| 2.6 | Algoritma Pengkelasan Teks | 17 |
| | 2.6.1 Naive Bayes (NB) | 18 |
| | 2.6.2 Mesin Sokongan Vektor (SVM) | 19 |
| | 2.6.3 K-Jiran Terdekat (KNN) | 20 |
| 2.7 | Kajian Terdahulu | 21 |
| 2.8 | Kesimpulan | 29 |
| | | |
| BAB III | METODOLOGI KAJIAN | |
| 3.1 | Pengenalan | 30 |
| 3.2 | Rangka Kerja Kajian | 30 |
| 3.3 | Pengekstrakan Tweet | 32 |
| 3.4 | Penterjemahan Teks Tweet Ke Bahasa Inggeris Piawai | 33 |
| 3.5 | Penetapan Nilai Piawai (Gold Standard) | 34 |
| 3.6 | Pra-Proses Pembersihan Data | 37 |
| | 3.6.1 Bag Of Words (BOW) | 38 |
| | 3.6.2 Pentokenan | 39 |
| | 3.6.3 Mengeluarkan Perkataan Kata Henti Umum | 39 |
| | 3.6.4 Mengeluarkan Perkataan Kata Henti Khusus | 40 |
| | 3.6.5 Pencantasan | 41 |
| 3.7 | Proses Pengekstrakan Fitur | 42 |
| | 3.7.1 Document-Term Matrix (DTM) | 42 |
| | 3.7.2 Terma Pemberat | 44 |
| 3.8 | Pengujian Menggunakan Algoritma Naive Bayes | 46 |
| 3.9 | Pengukuran | 48 |
| 3.10 | Kesimpulan | 49 |
| | | |
| BAB IV | ANALISIS KAJIAN | |
| 4.1 | Pengenalan | 50 |
| 4.2 | Set Data | 50 |
| 4.3 | Proses Pembersihan Data | 53 |
| 4.4 | Perbincangan | 55 |
| | 4.4.1 Hasil Eksperimen Ujian Pertama: Ketepatan dan Dapatan Semula | 56 |

| | | |
|--------------|--|----|
| | 4.4.2 Hasil Eksperimen Ujian Kedua: Ketepatan dan Dapatan Semula | 59 |
| | 4.4.3 Kejituan | 62 |
| 4.5 | Kesimpulan | 64 |
| | | |
| BAB V | PERBINCANGAN DAN KESIMPULAN | |
| 5.1 | Pengenalan | 65 |
| 5.2 | Rumusan Dan Penemuan Kajian | 65 |
| 5.3 | Sumbangan Kajian | 67 |
| 5.4 | Kekangan Kajian | 68 |
| 5.5 | Cadangan Masa Hadapan | 69 |
| 5.6 | Penutup | 69 |
| | | |
| | RUJUKAN | 70 |
| | | |
| | LAMPIRAN | |
| Lampiran A | Senarai Kata Henti Khusus Berbentuk Kod ASCII | 73 |
| Lampiran B | Senarai Kata Henti Khusus | 77 |
| Lampiran C | Senarai Kata Henti Bahasa Inggeris | 79 |
| Lampiran D | Resume Penterjemah Bahasa | 80 |

SENARAI JADUAL

| No. Jadual | | Halaman |
|-------------------|--|----------------|
| Jadual 1.1 | Bilangan status tweet pada 20-26 Nov 2016 mengikut IPTA | 4 |
| Jadual 2.1 | Sampel data buah | 18 |
| Jadual 2.2 | Kelebihan dan kekurangan NB | 19 |
| Jadual 2.3 | Kelebihan dan kekurangan SVM | 20 |
| Jadual 2.4 | Kelebihan dan kekurangan KNN | 21 |
| Jadual 2.5 | Ringkasan kajian terdahulu | 26 |
| Jadual 3.1 | Contoh output selepas proses pra-pembersihan | 38 |
| Jadual 3.2 | Sebahagian Output DTM Menggunakan RStudio | 44 |
| Jadual 3.3 | Jadual Matriks Penilaian | 47 |
| Jadual 3.4 | Penerangan Istilah Pengelasan Dokumen | 48 |
| Jadual 4.1 | Jumlah asal keseluruhan data set | 51 |
| Jadual 4.2 | Pengagihan data | 51 |
| Jadual 4.3 | Contoh tweet asal dalam Bahasa Melayu | 52 |
| Jadual 4.4 | Contoh tweet selepas proses terjemah | 52 |
| Jadual 4.5 | Contoh tweet selepas proses pembersihan data | 54 |
| Jadual 4.6 | Hasil ketepatan eksperimen pertama menggunakan algoritma NB | 56 |
| Jadual 4.7 | Hasil dapatan semula eksperimen pertama menggunakan algoritma NB | 57 |
| Jadual 4.8 | Hasil ketepatan eksperimen kedua menggunakan algoritma NB | 59 |
| Jadual 4.9 | Hasil dapatan semula eksperimen kedua menggunakan algoritma NB | 60 |
| Jadual 4.10 | Hasil kejituan keseluruhan eksperimen menggunakan algoritma NB | 62 |

SENARAI ILUSTRASI

| No. Rajah | | Halaman |
|------------------|--|----------------|
| Rajah 2.1 | Teknik pengkelasan sentimen | 12 |
| Rajah 2.2 | Proses Pengelasan Dokumen | 15 |
| Rajah 2.3 | Kedudukan <i>hyperplane</i> dalam SVM | 20 |
| Rajah 3.1 | Rangka kerja kajian | 31 |
| Rajah 3.2 | Antaramuka perisian RStudio versi 0.99.903 | 32 |
| Rajah 3.3 | Aturcara koding untuk mengekstrak tweet daripada laman web Twitter | 33 |
| Rajah 3.4 | Cara Penetapan Nilai Piawai | 35 |
| Rajah 3.5 | Senarai Pengkelas Teks sedia ada di uclassify.com | 36 |
| Rajah 3.6 | Contoh output paparan pengelas sentimen | 36 |
| Rajah 3.7 | Aturcara koding pembersihan tweet daripada daripada data asal | 38 |
| Rajah 3.8 | Ilustrasi model BOW | 38 |
| Rajah 3.9 | Hasil output <i>wordcloud</i> | 43 |
| Rajah 3.10 | Aturcara koding menggunakan algoritma NB dalam RStudio | 47 |
| Rajah 4.1 | Ikon berbentuk emosi yang disediakan oleh Twitter | 54 |
| Rajah 4.2 | Carta bar ketepatan eksperimen pertama menggunakan algoritma NB | 57 |
| Rajah 4.3 | Carta bar dapatan semula eksperimen pertama menggunakan algoritma NB | 58 |
| Rajah 4.4 | Carta bar ketepatan eksperimen kedua menggunakan algoritma NB | 60 |
| Rajah 4.5 | Carta bar dapatan semula eksperimen kedua menggunakan algoritma NB | 61 |
| Rajah 4.6 | Carta bar kejituan keseluruhan eksperimen menggunakan algoritma NB | 63 |

SENARAI SINGKATAN

| | |
|------|------------------------------------|
| API | Application Programming Interface |
| BOW | Bags Of Words |
| DTM | Document Term Matrix |
| FN | Negatif Palsu |
| FP | Positif Palsu |
| IPTA | Institusi Pengajian Tinggi Awam |
| IR | Capaian Maklumat |
| IDE | Persekitaran Pembangunan Bersepadu |
| KNN | K-Jiran Terdekat |
| ML | Pembelajaran Mesin |
| NB | Naive Bayes |
| NLP | Pemprosesan Bahasa Tabii |
| POS | Part Of Speech |
| RF | Random Forest |
| RT | Retweet |
| SVM | Mesin Sokongan Vektor |
| TP | Positif Benar |
| TN | Negatif Benar |
| UGC | User Generated Content |
| UKM | Universiti Kebangsaan Malaysia |
| UMS | Universiti Malaysia Sabah |
| URL | Unified Resource Link |
| USM | Universiti Sains Malaysia |

BAB I

PENDAHULUAN

1.1 PENGENALAN

Perlombongan data ialah satu proses analisis data untuk mendapatkan maklumat tersembunyi yang berguna daripada koleksi set data yang besar (Sharma & Srivastava 2016). Koleksi set data tersebut boleh digunakan untuk tujuan pengelasan kepada beberapa kumpulan kecil yang mengikut tema tertentu. Contoh tema adalah seperti makanan, sukan, politik dan pendidikan.

Analisis sentimen menentukan pendapat mengenai objek atau subjek dalam sesuatu topik perbincangan yang dihasilkan sendiri oleh individu (Thakkar & Patel 2013). Perubahan dalam dunia aplikasi media sosial telah membolehkan setiap individu menulis status mereka dalam bentuk perbincangan, mengkritik, membuat ulasan dan juga memberi cadangan atau pendapat dalam pelbagai bidang pengetahuan. Contohnya, koleksi set data yang besar berkenaan ulasan sesuatu produk boleh dikelaskan pada ulasan positif atau ulasan negatif. Produk yang mendapat banyak reaksi negatif memerlukan penambahbaikan bagi meningkatkan prestasi produk tersebut. Manakala ulasan positif pula membolehkan pengeluar produk meningkatkan proses pengeluaran untuk memenuhi permintaan pengguna. Maklumat ulasan positif dan negatif bagi produk ini penting untuk pengeluar produk mengetahui maklumbalas pengguna secara telus dan terus daripada pengguna.

Perkembangan dan penggunaan media sosial di kalangan warga Universiti Kebangsaan Malaysia (UKM) mahupun seluruh Malaysia telah memberikan impak yang besar kepada kehidupan aktiviti sosial setiap warganya. Penggunaan aplikasi media sosial ini telah menjadi medium komunikasi utama untuk menyampaikan maklumat, berkongsi cerita dan sebagainya tanpa mengira taraf warga UKM sama ada pelajar dan juga pekerja. Menurut penyelidik Java et al. (2007) dan Thakkar & Patel (2013), istilah “*microblogging*” digunakan yang memberi maksud sebagai medium penyiaran komunikasi baharu yang membolehkan pengguna memuatnaik status berbentuk pendapat dan analisis sentimen dalam bentuk teks pendek dengan menggunakan laman web media sosial, emel atau khidmat pesanan ringkas telefon mudah alih. Antara laman web media sosial yang menggunakan alat web 2.0 seperti Facebook, Twitter, Instagram, Pininterest dan LinkedIn telah muncul sebagai “*microblogging*” yang paling dinamik digunakan oleh pengguna. Menurut laman web rasmi Twitter (<https://about.twitter.com/company>), seramai 313 juta individu mahupun organisasi di seluruh dunia telah dikenalpasti sebagai pengguna aktif Twitter pada setiap bulan. Ini membuktikan aplikasi Twitter merupakan salah satu antara alat media sosial yang telah menarik minat pengguna yang aktif sebagai medium untuk berkomunikasi pada masa kini (Java et al. 2007).

Secara lazimnya, berdasarkan kajian Java et al. (2007), Twitter digunakan untuk memuatnaik status individu atau organisasi kepada empat elemen yang berbeza. Elemen tersebut ialah perbualan harian, perkongsian maklumat, perkongsian alamat laman sesawang (URL) dan laporan berita. Kajian pengkelasan tweet telah banyak dibuat oleh pengkaji terdahulu seperti (1) pengkelasan lokasi tweet dimuatnaik (Alsudais, Leroy & Corso 2014); (2) pengkelasan tweet terhadap tingkah laku pengguna terhadap makanan (Vidal et al. 2015); dan (3) pengkelasan tweet secara umum (Selvaperumal & Suruliandi 2014). Hasil daripada dapatan pengkelasan tersebut telah membantu penyelidik berkongsi maklumat kepada umum mengikut skop kajian masing-masing berkenaan data berbentuk teks yang boleh diperolehi secara masa nyata mahupun secara pengumpulan secara berperingkat melalui laman media sosial Twitter.

Twitter menyediakan kemudahan yang membenarkan pengguna untuk menghantar teks yang dikenali sebagai tweet sehingga 140 karakter, imej, video dan pautan URL ke laman web yang lain. Tweet yang dihantar oleh pengguna kepada

pengguna lain telah banyak digunakan sebagai data mentah utama untuk dijadikan bahan kajian oleh para pengkaji analisis teks. Justeru, kajian ini fokus kepada data tweet yang dimuat naik di dalam lingkungan kawasan UKM seluas 2 kilometer radius. Ini akan meningkat kebarangkalian tweet yang dikumpul dan disebarkan oleh warga UKM.

Secara umumnya, setiap tweet yang dimuatnaik dibahagi pada tiga kategori utama iaitu tweet positif, negatif dan neutral (Bouchlaghem, Elkhelifi & Faiz 2016). Tweet positif mewakili tweet yang menunjukkan emosi positif seperti perkataan *suka* atau *gembira*. Tweet negatif mewakili tweet yang menunjukkan emosi negatif seperti perkataan *sedih* atau *kecewa*. Tweet neutral pula tidak mewakili mana-mana emosi positif atau negatif kerana tiada nilai tersembunyi disebalik tweet tersebut. Contoh tweet neutral seperti kenyataan "*saya di sini*". Oleh itu, kajian ini fokus kepada ketiga-tiga kategori tersebut.

1.2 PENYATAAN MASALAH

UKM sebagai Institusi Pengajian Tinggi Awam (IPTA) yang mempunyai bilangan warga kerja dan pelajar dari pelbagai bidang pengajian tidak ketinggalan untuk mempunyai akaun media sosial tersendiri. Antara kegunaan penggunaan media sosial ini bertujuan untuk berkongsi maklumat dengan seluruh warga Malaysia dan seluruh dunia tentang aktiviti, berita, kejayaan dan peristiwa penting yang berlaku di dalam kampus UKM. Namun demikian, tiada kajian dibuat yang mengkhususkan pengkelasan analisis sentimen tweet dalam kategori IPTA di Malaysia.

Pemerhatian awal telah dibuat oleh pengkaji sendiri terhadap tweet dengan membuat analisis awal ringkas. Analisis awal telah dibuat oleh pengkaji melalui kaedah pemerhatian secara umum terhadap 100 tweet yang dimuatnaik oleh pengguna Twitter terhadap tiga buah universiti tempatan yang mewakili zon utara Universiti Sains Malaysia (USM), zon selatan Universiti Teknologi Malaysia (UTM) dan zon borneo Universiti Malaysia Sabah (UMS) di Malaysia. Contoh tweet seperti berikut:

- Tweet positif : "*Marketers need to understand that not all millennials are created equal*".

- Tweet negatif : “*UTM has gotten a lot better since I left so I have high expectations*”.

Hasil pemerhatian dan analisis awal dalam kajian ini mendapati bahawa status tweet yang mempunyai kenyataan positif adalah berada dalam bilangan yang tinggi (Jadual 1.1). Menurut Bouchlaghem, Elkhelifi & Faiz (2016), kenyataan positif ialah kenyataan yang menggunakan perkataan positif seperti *terima kasih, tahniah, gembira*; kenyataan negatif pula menggunakan perkataan negatif seperti *benci, letih, putus*; dan kenyataan neutral pula berbentuk ayat penyata yang tidak menggambarkan emosi dan perbuatan individu (contoh ayat: *saya di sini, saya di dewan kuliah, saya ada kelas*).

Jadual 1.1 Bilangan Status Tweet Pada 20-26 Nov 2016 mengikut IPTA

| Universiti | | | Kenyataan Positif | Kenyataan Negatif | Kenyataan Neutral | Jumlah tweet |
|---------------------|-----------|----------|----------------------|----------------------|----------------------|-----------------|
| Universiti (USM) | Sains | Malaysia | 33 | 6 | 61 | 100 |
| Universiti (UTM) | Teknologi | Malaysia | 40 | 11 | 49 | 100 |
| Universiti (UMS) | Malaysia | Sabah | 39 | 7 | 54 | 100 |
| Jumlah Peratus | | | 37% | 8% | 55% | 100% |

Berdasarkan Jadual 1.1 sebagai taburan corak status tweet mengikut kategori secara umum, timbul persoalan, adakah UKM juga tergolong dalam institusi yang mempunyai status tweet berbentuk kenyataan negatif yang paling rendah atau paling tinggi bilangannya? Justeru, tindakan yang sewajarnya ialah dengan melakukan kajian terperinci untuk mengetahui jawapan bagi persoalan tersebut. Oleh yang demikian, tweet yang dimuatnaik oleh warga UKM akan dianalisa dengan menggunakan teknik analisis teks dan pembelajaran mesin (ML).

Penggunaan bahasa atau ayat dalam tweet lazimnya tidak formal kerana kekangan jumlah perkataan ialah 140 karakter. Ini menyebabkan pengguna menggunakan singkatan perkataan tertentu seperti perkataan “*otw*” yang bermaksud “*on the way*” atau ikon berbentuk emosi seperti “😊” yang bermaksud senyum untuk menyampaikan maklumat. Selain itu, masalah penggunaan campuran lebih daripada satu bahasa atau dikenali sebagai bahasa rojak dalam tweet juga merupakan masalah yang perlu diselesaikan melalui proses penterjemahan bahasa kepada bahasa bentuk piawai.

Pemilihan fitur memainkan peranan yang penting dalam analisis pengelasan. Antara fitur yang digunakan adalah TF, TFIDF, pencantasan dan kata henti khusus. Kajian ini cuba mendapatkan fitur yang terbaik bagi pengelasan tweet tersebut.

1.3 PERSOALAN PENYELIDIKAN

Terdapat tiga persoalan kajian dalam penyelidikan ini iaitu:

- a) Apakah corak tweet di UKM sama seperti tiga IPTA yang lain?
- b) Bagaimanakah pemilihan fitur terbaik dibuat?

1.4 OBJEKTIF PENYELIDIKAN

Matlamat utama kajian ini adalah untuk mengetahui adakah warga UKM memuat naik status di Twitter yang berunsurkan kenyataan positif atau negatif terhadap institusi tempat mereka bekerja mahupun tempat menimba ilmu pengetahuan? Bagaimana struktur pengelasan ini dibuat dapat membantu memudahkan pengelasan teks pada masa akan datang? Oleh itu, objektif kajian penyelidikan ini perlu menguasai perkara berikut bagi memenuhi keperluan objektif penyelidikan:

- a. Mengenalpasti gabungan fitur yang terbaik bagi pengelasan tweet dari fitur Frekuensi Perkataan (TF), *Term Frequency-Inverse Document Frequency* (TFIDF), pencantasan dan kata henti khusus.

- b. Membangunkan sistem pengkelas tweet menggunakan kaedah *Naive Bayes* (NB).
- c. Menguji keberkesanan sistem yang dibangunkan.

1.5 SKOP DAN BATASAN KAJIAN

Kajian ini memfokuskan kepada tweet yang diperoleh mengikut radius sekitar kawasan kampus UKM sahaja di mana data tersebut dihasilkan oleh warga yang mendiami di kawasan kampus dan isi kandungannya juga memperihalkan aktiviti di dalam kampus.

Analisis sentimen dalam domain Pemrosesan Bahasa Tabii (NLP) mempunyai tiga tahap kebutiran (*granularity*) iaitu pengkelasan tahap dokumen, pengkelasan tahap ayat dan pengkelasan tahap frasa perkataan (Troussas et al. 2013). Skop kajian ini fokus kepada pengkelasan tweet tahap dokumen. Pengkelasan tahap dokumen ini ialah setiap satu tweet mewakili satu dokumen pengkelasan. Justifikasinya ialah untuk memudahkan proses pengkelasan terhadap keseluruhan data tweet secara pukal.

Sebanyak 80% data mentah yang diperolehi dalam kajian ini adalah dalam Bahasa Melayu, 10% menggunakan campuran Bahasa Melayu dan Inggeris, sementara 10% lagi menggunakan Bahasa Inggeris sepenuhnya. Dengan kekangan tersebut, kajian ini mengambil inisiatif dengan penambahan proses terjemahan tweet kepada Bahasa Inggeris piawai.

1.6 KEPENTINGAN KAJIAN

Kepentingan kajian ini boleh dilihat dari dua sudut berbeza iaitu kepada pengkaji dan kepada institusi UKM. Proses mendapatkan data dalam kajian ini amat sesuai bagi pengkaji yang memerlukan data yang dimuatnaik oleh pengguna secara sukarela tanpa paksaan dengan menggunakan platform media sosial yang sedia ada. Kajian ini boleh dijadikan sumber rujukan kepada pengkaji untuk generasi akan datang.

Kajian ini membolehkan pihak pengurusan UKM untuk membuat pemerhatian, memantau, menyelidik, menyiasat dan mengenalpasti status individu yang boleh mencemarkan nama baik UKM sekiranya berlaku kejadian yang tidak diingini. Bukan itu sahaja, dengan cara ini juga pihak pengurusan UKM mampu mengetahui maklumbalas berbentuk tidak formal yang dimuatnaik oleh para pengguna Twitter.

1.7 ORGANISASI PROJEK

Projek ini menjurus kepada lima bab perbincangan. Setiap bab diterangkan secara terperinci bagi menjelaskan keseluruhan perjalanan kajian.

Bab I mengandungi pengenalan, pernyataan masalah, objektif kajian, persoalan kajian, skop dan batasan kajian dan kepentingan kajian.

Bab II menerangkan kajian susastera yang dibuat oleh kajian terdahulu dalam bidang klasifikasi teks.

Bab III pula mengkhususkan kepada metodologi yang digunakan untuk mencapai objektif sepanjang kajian dijalankan.

Bab IV membincangkan analisis kajian yang diperolehi hasil daripada metodologi yang digunakan.

Bab V pula merumuskan keseluruhan proses kajian yang telah dijalankan.

1.8 KESIMPULAN

Bab ini telah membincangkan berkenaan pengenalan, pernyataan masalah, objektif kajian, persoalan kajian, skop kajian, batasan kajian dan kepentingan kajian. Beberapa persoalan yang timbul dalam persoalan kajian ini akan dibincangkan dalam bab seterusnya. Semoga kajian ini memberi manfaat kepada semua warga UKM secara umum.

BAB II

KAJIAN SUSASTERA

2.1 PENGENALAN

Kajian penyelidikan ini menggunakan media sosial Twitter sebagai capaian data mentah yang utama untuk diproses bagi menghasilkan maklumat yang baru kepada pihak organisasi UKM. Pemilihan media sosial Twitter dilakukan kerana (1) saiz tweet yang kecil (2) tidak berat sebelah (3) capaian yang mudah dengan *Application Programming Interface* (API) dan (4) tweet yang dimuatnaik terdiri daripada pelbagai domain sosio-budaya (Bakliwal et al. 2012). Kajian ini fokus kepada dua pengkelasan tweet yang utama iaitu positif dan negatif. Bab ini bertujuan untuk membincangkan bagaimana proses dan kaedah pengkelasan tweet dilakukan berdasarkan kepada kajian penyelidikan yang lepas. Memandangkan proses pengkelasan tweet ini adalah sebahagian daripada dua bidang utama dalam sains komputer iaitu bidang Pemrosesan Bahasa Tabii (NLP) dan Pembelajaran Mesin (ML), penerangan terperinci perlu dibincangkan dan diperluaskan.

Bab ini akan menerangkan dengan terperinci berkenaan bagaimana media sosial menjadi pilihan pengguna sebagai medium komunikasi sehingga terhasilnya maklumat yang berlebihan. Dengan maklumat yang berlebihan ini, membolehkan kajian analisis teks dilakukan bagi mengetahui jenis-jenis maklumat berlebihan yang ada di media sosial. Dengan menggunakan metodologi tertentu yang sedia ada dalam ML dan NLP, maklumat yang berlebihan ini akan diproses dengan berdasarkan kajian penyelidikan yang lepas sebagai penanda aras untuk menghasilkan pengkelasan tweet

seperti yang dikehendaki. Hasil daripada pengelasan tweet ini, pihak organisasi UKM boleh menggunakan maklumat ini untuk membuat sesuatu keputusan mahu pun sebagai makluman untuk penambahbaik mutu perkhidmatan.

2.2 KANDUNGAN MAKLUMAT YANG DIJANA OLEH PENGGUNA MEDIA SOSIAL

Pada asalnya, media sosial diguna sebagai medium komunikasi yang membolehkan keluarga dan rakan-rakan terdekat untuk sentiasa berhubung antara satu sama lain dengan berkongsi aktiviti yang dilakukan sepanjang hari (Hiltz & Plotnick 2013). Sebagai contoh, sebuah keluarga yang menetap jauh di luar negara memuatnaik gambar perkembangan anak menggunakan media sosial untuk tatapan keluarga di Malaysia. Tetapi kini, perkongsian itu boleh dikongsi oleh sesiapa sahaja yang berdaftar dengan media sosial. Tambahan pula dengan kemudahan telefon pintar yang menyediakan aplikasi media sosial yang membolehkan capaian Internet dicapai pada bila-bila masa dan di mana sahaja.

Terdapat sumber yang memberi kemudahan paparan statistik penggunaan media sosial melalui laman sesawang <https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts-for-2016/> bertarikh 7 Mac 2016. Berdasarkan sumber tersebut, 500 juta tweet dihantar setiap hari, 300 juta foto dimuat naik di Facebook, lebih 80 juta foto dimuat naik di Instagram setiap hari dan 300 jam video dimuat naik di YouTube setiap minit. Media sosial telah mengubah tingkah laku seseorang individu mahupun organisasi dalam konteks perkongsian maklumat. Berdasarkan kajian Jalonen (2014), media sosial dilihat sebagai perubahan komunikasi secara global yang mana pengguna dengan secara sukarela berkongsi pelbagai pengalaman yang penuh dengan pelbagai emosi dan situasi sama ada dalam bentuk teks, gambar mahupun video. Dengan bilangan pengguna media sosial yang sentiasa bertambah seiring dengan perkongsian kandungan (*content-sharing*) yang dikongsi ini menyebabkan maklumat yang berlebihan telah berlaku dalam dunia digital media sosial ini. Dengan perkembangan "*User-Generated Content*" (UGC) ini, maklumat berlebihan ini mempunyai potensi yang besar untuk digunakan oleh para penyelidik untuk mendalami tentang bagaimana untuk mengekstrak maklumat tersebut dan digunakan dalam

pelbagai aplikasi kajian yang bersesuaian (Huang et al. 2013). Walau bagaimanapun, bentuk maklumat yang berlebihan ini adalah dalam bentuk tidak berstruktur kerana ia dihasilkan untuk kegunaan manusia dan bukan untuk diproses oleh komputer.

Maklumat yang berlebihan ditakrifkan sebagai maklumat yang diberikan pada kadar yang terlalu cepat bagi seseorang untuk memproses (Hiltz & Plotnick 2013). Menurut kajiannya lagi, kadar maklumat yang dimuatnaik tidak mengikut topik atau mempunyai isi kandungan yang tidak penting dalam keadaan semasa. Bukan itu sahaja, Hiltz & Plotnick (2013) juga mendapati maklumat yang berlebihan ini juga kerap berlaku dan sentiasa memenuhi “*news feed*” pengguna apabila terdapat situasi berisiko tinggi seperti peristiwa kecemasan kebakaran, gempa bumi dan banjir. Dengan kebanjiran maklumat yang berlebihan ini, membolehkan kajian analisis teks diketengahkan memandangkan maklumat ini boleh dicapai dengan mudah tanpa perlu menggunakan kaedah penyelidikan secara tradisional seperti mengguna borang soal selidik.

2.3 ANALISIS SENTIMEN

Dengan kemajuan perkembangan penggunaan media sosial, segala bentuk perkongsian status pengalaman dan pendapat individu boleh dicapai secara atas talian. Perkongsian status secara atas talian ini menjadi medium utama yang popular masa kini (Thakkar & Patel 2013). Analisis sentimen melibatkan kajian berkenaan perlombongan pendapat yang biasanya digunakan dalam bidang pengiklanan, pencipta filem dan pengeluar produk yang memerlukan reaksi secara telus berbentuk maklumbalas pelanggan sama ada berbentuk positif, negatif atau neutral (Troussas et al. 2013). Analisis sentimen didefinisikan sebagai luahan perasaan individu sama ada dalam bentuk perasaan positif atau perasaan negatif (Go, Bhayani, & Huang, 2013). Menurut kajian Troussas et al. (2013), terdapat cabaran utama yang harus diutamakan sebelum proses analisis sentimen dibuat iaitu:

- Tidak semua perkataan yang dikongsikan oleh pengguna media sosial adalah signifikan dengan perkara yang dibincangkan

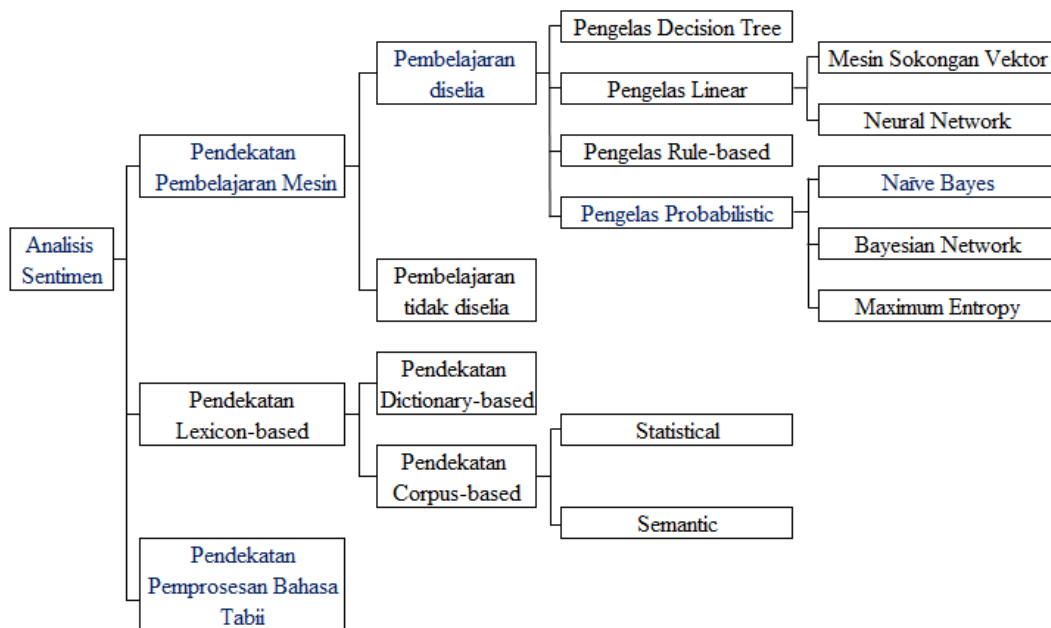
- Ada perkataan yang telah dikenalpasti sebagai gangguan (*noise*) kerana perkataan ini tidak memberi apa-apa maksud
- Perkataan seperti “*not*” apabila digunakan bersama perkataan positif secara automatik akan menghasilkan kenyataan negatif
- Penggunaan simbol sedih “☹” atau simbol gembira “😊” perlu diproses dengan cara khusus yang signifikan dengan kaedah pemprosesan NLP

Oleh itu, analisis sentimen ini perlu menumpukan kepada pemilihan teks yang tepat dan ia perlu dibuat secara terperinci bagi memastikan data yang digunakan adalah data yang berguna. Teks merupakan medium utama yang digunakan dalam penggunaan media sosial dalam menyebarkan maklumat. Secara umumnya, semasa memuatnaik status dalam media sosial Twitter, pengguna Twitter akan memuatnaik status bersama empat elemen berikut :

- Lokasi di mana mereka berada dengan menggunakan simbol @ seperti @ UKM, Bangi.
- Berkongsi status dengan akaun media sosial yang lain dengan menyertakan pautan URL ke media sosial lain oleh pengguna yang sama seperti <https://t.co/YYQohh3OIc>
- Menunjukkan emosi dalam bentuk ikon berbentuk emosi
- Memuatnaik imej gambar

Empat elemen ini tidak akan diambilkira dalam kajian ini. Justeru, kajian ini hanya fokus kepada teks status pengguna sahaja. Untuk itu, analisis teks merupakan elemen penting dalam kajian ini.

2.4 ANALISIS SENTIMEN DALAM DATA MEDIA SOSIAL



Rajah 2.1 Teknik pengelasan sentimen

Sumber: Medhat, Hassan & Korashy (2014)

Rajah 2.1 menunjukkan pecahan teknik pengelasan sentimen yang akan dibincangkan secara menyeluruh di dalam kajian ini. Secara keseluruhannya, kajian ini membincangkan khusus kepada jenis pendekatan yang digunakan, jenis pembelajaran, jenis pengelasan dan kaedah pengelasan tweet.

Menurut kajian Korde & Mahendar (2012), lebih 80% maklumat disimpan dalam bentuk teks. Menurut kajiannya lagi, perlombongan teks ini memberi nilai komersial yang tinggi kerana sifat teks yang diterokai ini boleh diperolehi melalui pelbagai sumber maklumat yang sedia ada.

Teks ini terdiri daripada dua jenis teks iaitu teks berstruktur dan teks tidak berstruktur. Menurut Sharma & Srivastava (2016), teks berstruktur ialah teks yang dipaparkan dalam bentuk baris dan lajur yang memudahkan proses capaian menggunakan alat perlombongan data. Contohnya seperti katalog perpustakaan (mengandungi baris seperti tarikh, nama penulis, tajuk dan penerbit) dan rekod banci (mengandungi baris seperti tarikh lahir, jantina, alamat dan pekerjaan). Manakala teks tidak berstruktur pula ialah teks yang tergolong dalam kumpulan UGC seperti emel, khidmat pesanan ringkas dan status yang dimuatnaik dalam media sosial. Teks tidak

berstruktur ini merupakan sumber maklumat terbesar yang mudah dicapai melalui media sosial. Oleh itu, kajian ini menggunakan metodologi analisis teks tidak berstruktur dalam data media sosial Twitter sebagai data utama kajian.

Dalam bidang sains komputer, terdapat kajian lepas seperti Bouchlaghem, Elkhelifi & Faiz (2016) dan Go, Bhayani & Huang (2013) yang membolehkan pengkelasan automatik terhadap tweet dihasilkan oleh manusia diklasifikasikan sebagai ayat positif, negatif dan neutral. Ayat positif didefinisikan sebagai ayat yang memberi maksud positif dan menggunakan perkataan positif seperti '*Saya berjaya mengurangkan berat badan*', di mana perkataan '*berjaya*' ialah katakunci untuk perkataan positif. Manakala ayat negatif pula bermaksud ayat yang memberi maksud negatif dan menggunakan perkataan negatif seperti '*Saya sudah serik dengan sambungan wifi yang sentiasa terputus sepanjang malam ini*', di mana perkataan '*serik*' ialah katakunci untuk perkataan negatif. Sementara ayat neutral tidak memberi maksud yang menjurus kepada positif atau negatif. Sebagai contoh, ayat '*Saya ada di sini malam ini*'. Pengkelasan tweet secara automatik ini amat berguna kepada pengguna mahu pun organisasi yang mahukan maklumbalas pelanggan terhadap sesuatu perkhidmatan atau produk (Go, Bhayani & Huang 2013).

Menurut kajian Jo (2008), terdapat dua jenis pendekatan klasifikasi teks iaitu "*rule based*" dan pembelajaran mesin. Pendekatan "*rule based*" bermaksud proses pengkelasan teks ditakrifkan secara manual menggunakan format *if-then-else*. Oleh itu, dokumen dikelaskan berdasarkan undang-undang tersebut. Manakala pendekatan ML pula bermaksud peraturan klasifikasi ditakrifkan secara automatik menggunakan sampel dokumen yang telah dilatih terlebih dahulu. Pendekatan ML ini telah memberi nilai dapatan semula yang tinggi dan nilai ketepatan yang sedikit rendah berbanding dengan pendekatan "*rule based*". Oleh itu, pendekatan ML telah menggantikan pendekatan "*rule based*".

Dua bidang utama yang diperlukan semasa proses analisis teks ini ialah dengan menggunakan Pemprosesan Bahasa Tabii (NLP) dan ML. Menurut Medhat, Hassan & Korashy (2014), gabungan teknik NLP dan ML amat berkesan semasa proses pengkelasan dokumen. Justeru, kajian ini menggunakan dua pendekatan NLP dan ML sebagai domain semasa proses kajian dibuat.

2.4.1 Pemrosesan Bahasa Tabii (NLP)

NLP diperlukan untuk mendapatkan ciri penting yang boleh digunakan dalam proses pengkelasan dokumen. NLP didefinisikan sebagai bagaimana komputer berinteraksi dengan manusia sama ada melalui penulisan mahu pun percakapan yang diperkaya dengan maklumat linguistik. Dalam kajian Alsudais, Leroy & Corso (2014), pendekatan NLP digunakan untuk mengenalpasti kandungan teks dalam setiap tweet. Contohnya seperti pendekatan *Part Of Speech (POS)*. POS ialah elemen kecil penting yang mewakili setiap perkataan yang ada di dalam sesebuah dokumen. Dalam tatabahasa Bahasa Inggeris, POS dibahagikan kepada lapan bahagian iaitu kata nama (*noun*), kata ganti nama (*pronoun*), kata kerja (*verb*), kata keterangan (*adverb*), kata sifat (*adjective*), kata penghubung (*conjunction*), kata depan (*preposition*) dan kata seru beremosi (*interjection*). Setiap perkataan akan dipisahkan mengikut sifat POS perkataan masing-masing. Contoh ayat:

Ayat asal: *She likes big shirts but I hate them*

Hasil POS: *She/PRONOUN likes/VERB big/ADJECTIVE shirts/NOUN
but/CONJUNCTION I/PRONOUN hate/VERB them/PRONOUN*

2.4.2 Pembelajaran Mesin (ML)

ML menyediakan beberapa algoritma pengelasan dokumen yang telah dibangunkan oleh pengkaji terdahulu. Algoritma seperti *Random Forest (RF)* dan K-Jiran Terdekat (KNN) adalah antara algoritma pengelasan dokumen yang telah digunakan oleh pengkaji terdahulu. ML didefinisikan sebagai bagaimana manusia mengajar komputer untuk belajar daripada data yang diberikan sebagai data latihan dan kemudian dengan menggunakan input data baru yang dikenali sebagai data ujian, komputer tersebut akan mempelajari daripada data tersebut dan menghasilkan output sebagaimana yang telah dipelajarinya. Terdapat pelbagai algoritma ML sedia ada yang telah dibangunkan oleh beberapa penyelidik berdasarkan masalah analisis teks yang telah dikenal pasti dalam kajian terdahulu. Menurut kajian Domingos (2012), ML telah digunakan dalam aplikasi

seperti carian laman sesawang, “*spam filters*”, “*fraud detection*”, “*stock trading*” dan “*drug design*”.

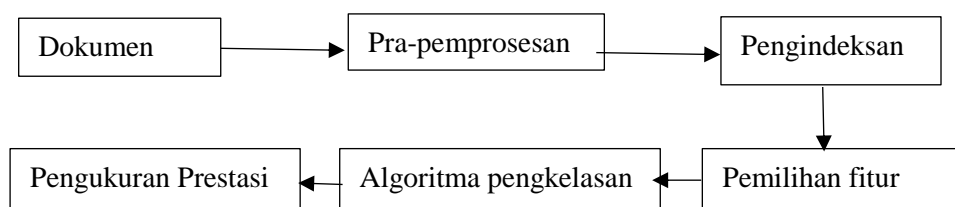
Terdapat dua kaedah utama dalam ML iaitu pembelajaran diselia dan pembelajaran yang tidak diselia :

- **Pembelajaran diselia** : kaedah yang digunakan dengan menggunakan pemerhatian dan pembelajaran terhadap data sedia ada yang telah diklasifikasikan secara manual dan ia dikenali sebagai set latihan.
- **Pembelajaran yang tidak diselia** : kaedah yang tidak merujuk kepada set latihan sebaliknya pengklasifikasian berlaku secara automatik menggunakan algoritma tertentu.

Secara ringkasnya, kaedah ML digunakan untuk memproses maklumat berbentuk linguistik. Aplikasi yang menggunakan kaedah ML ini memberikan output berdasarkan data yang telah dipelajari sebelum ianya memproses data yang baru.

2.5 PROSES PENGKELASAN TEKS

Sebelum sesuatu dokumen dipecahkan mengikut kelas kategori yang dikehendaki, terdapat enam proses asas pengkelasan teks yang harus diberi tumpuan semasa melakukan kajian pengkelasan dokumen. Menurut kajian Korde & Mahendar (2012), enam proses utama yang harus dilakukan oleh setiap pengkaji seperti Rajah 2.1 adalah perlu bagi memenuhi mana-mana tujuan dalam kajian pengkelasan dokumen berbentuk teks.



Rajah 2.2 Proses pengkelasan dokumen

Sumber : Korde & Mahendar (2012)

2.5.1 Pra-pemrosesan

Pra-pemrosesan ialah proses pembersihan data daripada data yang tidak diperlukan dalam kajian. Proses ini bertujuan untuk menghasilkan dokumen teks dalam format teks yang lebih jelas. Proses ini membolehkan hanya perkataan yang mempunyai nilai sahaja yang akan digunakan dalam kajian ini. Menurut Asghar et .al (2014), proses pra-pemrosesan ini menggunakan beberapa langkah pra-pemrosesan seperti :

- Penghapusan perkataan kata henti umum : Perkataan kata henti umum seperti perkataan “*the*”, “*a*” dan “*and*” adalah perkataan yang tergolong mempunyai kekerapan penggunaan yang paling banyak digunakan dalam sesebuah dokumen
- Penghapusan perkataan kata henti khusus : perkataan kata henti khusus seperti perkataan “UKM”, “Bangi” dan “Selangor” adalah perkataan yang perkataan yang tergolong mempunyai kekerapan penggunaan paling banyak khusus dalam persekitaran dokumen kajian ini sahaja.
- Pencantasan : menggunakan algoritma pencantasan yang membolehkan kata dasar sesuatu perkataan dalam sesebuah dokumen dikenalpasti. Melalui kata dasar perkataan tersebut, apa-apa perkataan yang berasal daripada perkataan tersebut secara automatik akan dihapuskan. Contohnya kata dasar “argu” yang diperoleh daripada perkataan “*argue*”, “*argues*” dan “*arguing*”.

2.5.2 Pengindeksan

Proses pengindeksan merupakan perwakilan dokumen yang digunakan untuk mengurangkan kekompleksan dokumen (Korde & Mahendar (2012). Proses ini dibuat untuk memudahkan pengurusan teks. Dokumen perlu diubah daripada versi asal kepada bentuk vektor dokumen. Vektor dokumen ini adalah dalam bentuk “*document matrix*”. Proses ini akan menyenaraikan senarai teks yang ada dalam setiap dokumen. Setelah terhasilnya senarai teks setiap dokumen ini, proses ini akan memudahkan proses pemilihan fitur untuk proses seterusnya.

2.5.3 Pemilihan fitur

Proses pemilihan fitur ialah untuk memilih sebahagian fitur yang ada dalam keseluruhan data asal yang sedia ada. Fitur ini akan mengambil teks yang mempunyai nilai skor yang paling tinggi dalam jadual “*document matrix*” semasa proses pengindeksan. Fitur ini juga mengambil kira nilai kekerapan teks paling tinggi yang terdiri daripada perkataan kata nama, kata kerja dan adjektif.

2.5.4 Algoritma pengkelasan

Setelah melepasi proses pra-pemprosesan dan pemilihan fitur, proses pengelasan dokumen boleh dibuat. Proses ini membolehkan teks dikelaskan menggunakan kaedah ML seperti *Naive Bayes* (NB), Mesin Sokongan Vektor (SVM) dan *Maximum Entropy* (MAXENT) yang berasaskan data latihan sebagai rujukan kepada data ujian. Setiap kaedah ML ini memberikan output yang berbeza kerana setiap satunya menggunakan algoritma yang berbeza walaupun dokumen yang digunakan adalah sama.

2.5.5 Pengukuran prestasi

Proses pengukuran prestasi dibuat untuk melihat hasil keputusan ramalan pengelasan dokumen. Pengukuran diukur melalui hasil eksperimen pengelasan semasa proses algoritma pengelasan. Output hasil pengukuran yang diperolehi berbentuk peratus ketepatan, peratus dapatan semula dan peratus kejituan.

2.6 ALGORITMA PENKELASAN TEKS

Tujuan utama pengkelasan ialah untuk membuat ramalan kelas target dengan menganalisis data daripada data latihan (Sassano 2003). Secara umumnya, set data latihan diperlukan untuk mengetahui garisan sempadan kelas yang wujud. Apabila