PEMILIHAN FITUR DALAM ANALISIS SENTIMEN MENGGUNAKAN TEORI SET KASAR DAN ALGORITMA PENGOPTIMUMAN BERASASKAN PENGAJARAN PEMBELAJARAN

Abdullah bin Muhammad, Salwani Abdullah

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia
43600, Bangi, Selangor.

Email: gp06690@siswa.ukm.edu.my, salwani@ukm.edu.my

ABSTRAK

Pemilihan fitur dan analisis sentimen merupakan dua kajian yang popular dilaksanakan pada masa sekarang. Ini adalah sejajar dengan kemajuan bidang pengkomputeran dan juga peningkatan penggunaan media sosial. Walau bagaimanapun, kurang kajian yang dilakukan terhadap pemilihan fitur dalam analisis sentimen. Masalah utama pemilihan fitur dalam analisis sentimen ialah ia mempunyai ruang fitur yang berdimensi besar. Ditambah pula dengan ulasan dari medan sosial ini kebanyakannya mempunyai banyak hingar dan maklumat yang tidak relevan untuk pengelasan sentimen. Oleh itu ulasan ini perlu melalui proses pembersihan awal dengan menggunakan kaedah pemprosesan teks. Kajian ini akan mengkaji kaedah pemprosesan teks daripada teknik pemprosesan bahasa tabii yang akan digabungkan dengan teknik pemprosesan linguistik untuk menghasilkan keputusan pengelasan yang tinggi. Set data hasil daripada pemprosesan teks ini seterusnya akan melalui proses pemilihan fitur. Pemilihan fitur amat penting kerana ia dapat membantu mengurangkan saiz dimensi fitur dan juga berupaya untuk memilih fitur yang benar-benar berupaya memberikan prestasi yang tinggi semasa proses pengelasan sentimen. Dalam kajian ini, gabungan teori set kasar (TSK) dan algoritma pengoptimuman berasaskan pengajaran pembelajaran (PBPP) atau dikenali sebagai TSKPBPP dicadangkan sebagai algoritma pemilihan fitur. Untuk membuktikan keupayaan algoritma ini, ujian ini dilakukan dengan menggunakan set data ulasan pengguna terhadap enam buah syarikat penerbangan Amerika Syarikat yang di ekstrak dari laman sosial Twiter. Keputusan kajian ini mendapati bahawa gabungan daripada pemprosesan teks yang terpilih dan juga algoritma pemilihan fitur TSKPBPP berupaya untuk menghasilkan prestasi analisis sentimen yang lebih baik atau setanding dengan algoritma pemilihan fitur lain dalam kajian literatur.

Kata kunci: Pemilihan Fitur, Analisis Sentimen, Teori Set Kasar, Algoritma Pengoptimuman Berasaskan Pengajaran Pembelajaran, Pemprosesan Teks

1. PENGENALAN

Analisis sentimen ialah sejenis pengelasan teks yang melibatkan penyataan subjektif. Ia berfungsi untuk menganalisis dan mengenal pasti jenis sentimen daripada ulasan pengguna yang berbentuk teks (Pang & Lee 2008). Sejajar dengan perkembangan dunia Internet dan juga media sosial, pengguna lebih kerap menggunakan laman sosial, laman web, blog dan forum sebagai medium utama untuk memberikan ulasan, pandangan dan pendapat mereka mengenai sesuatu perkhidmatan, isu, idea dan pelbagai perkara. Sehubungan dengan itu,

analisis sentimen memainkan peranan yang penting untuk menganalisis maklumat ulasan, pandangan dan pendapat yang disampaikan oleh pengguna melalui medium ini.

Analisis sentimen berasal daripada teknologi perlombongan teks, pemprosesan bahasa tabii dan pengelasan teks (Seerat & Azam 2012). Ia berfungsi untuk mengelaskan maklumat yang berbentuk teks seperti ulasan pengguna mengenai perkhidmatan, politik, produk dan sebagainya kepada kategori sentimen positif atau sentimen negatif (Arafat et al. 2014; Vinodhini & Chandrasekaran 2013). Analisis sentimen boleh dibahagikan kepada tiga teknik iaitu teknik analisis sentimen berasaskan leksikon, analisis sentimen berasaskan pembelajaran mesin dan teknik hibrid (Rustam et al. 2019). Teknik pembelajaran mesin merupakan salah satu kaedah yang paling popular untuk mengelaskan sentimen. Teknik ini digabungkan dengan kaedah bahasa bagi mengenal pasti kategori sentimen yang terkandung dalam teks (Abbasi et al. 2011). Terdapat tiga proses utama dalam analisis sentimen yang dilaksanakan dalam kajian ini iaitu pemprosesan data, pemilihan fitur, dan pengelasan sentimen.

Maklumat yang diperoleh daripada laman sosial dan Internet ini sentiasa berkembang dan berubah dari semasa ke semasa. Ini menyebabkan proses penganalisisan maklumat ulasan secara manual adalah tidak efektif, tidak relevan dan ia memerlukan usaha yang banyak dan sering kali menghasilkan keputusan yang tidak tepat. Oleh itu, satu mekanisme perlu dibangunkan untuk memudahkan proses penganalisisan ulasan pengguna ini. Justeru itu, kajian ini dijalankan untuk melihat kesan pemilihan fitur dengan menggunakan gabungan teknik teori set kasar (TSK) dan algoritma pengoptimuman berasaskan pengajaran pembelajaran (PBPP) atau dipanggil sebagai TSKPBPP untuk menganalisis sentimen dari ulasan pengguna dari media sosial Twiter.

1.1 Penyataan Masalah

Antara masalah utama dihadapi dalam analisis sentimen ialah kaedah untuk memproses teks ulasan pengguna yang diperoleh dari media sosial, web, forum atau blog yang mengandungi hingar dan maklumat yang tidak relevan. Data-data ini perlu diuruskan dan diproses secara teratur sebelum analisa sentimen dilaksanakan terhadapnya (Pradha et al. 2019). Dari kajian yang dilaksanakan oleh Pradha et al. (2019) teknik pemprosesan teks memainkan peranan penting dalam ketepatan ramalan sentimen dan juga mempengaruhi tempoh masa pengkomputeran semasa pengelasan sentimen dilaksanakan. Pemprosesan teks merupakan

proses yang dibangunkan untuk menganalisis dan menyediakan teks untuk klasifikasi sentimen (Haddi et al. 2013).

Selain itu, saiz fitur yang besar merupakan masalah biasa dihadapi dalam analisis sentimen. Saiz fitur yang besar mempengaruhi nilai prestasi pengelasan sentimen dan masa pemprosesan. Masalah ini dapat diatasi dengan menggunakan kaedah pembelajaran mesin yang menggunakan teknik pemilihan set fitur yang sesuai untuk menghapuskan fitur-fitur yang hingar dan tidak relevan (Abbasi et al. 2011; Agarwal & Mittal 2013; Arafat et al. 2014). Pemilihan fitur merupakan masalah polinomial yang tidak berketentuan dan memerlukan algoritma yang efisien seperti algoritma metaheuristik bagi membantu dalam pemilihan fitur (Kohavi & John 1997; Mafarja & Eleyan 2013; Unler & Murat 2010; Yusta 2009).

Menurut Rao et al. (2011) kebanyakan algoritma pemilihan fitur sedia ada sekarang memerlukan parameter spesifik. Algoritma yang menggunakan parameter spesifik perlu melaraskan nilai parameter ini kepada nilai yang tepat untuk mendapatkan keputusan yang baik. Sebagai contoh algoritma genetik memerlukan parameter spesifik seperti parameter kemungkinan silangan, parameter kadar mutasi dan parameter kaedah pemilihan. Sehubungan dengan itu Rao et al. (2011) memperkenalkan algoritma PBPP yang tidak memerlukan parameter spesifik untuk berfungsi dengan baik.

1.2 Objektif Kajian

Matlamat kajian ini adalah seperti berikut:

i. untuk mengkaji kesan kaedah pemprosesan teks daripada kategori pemprosesan bahasa tabii kepada ketepatan ramalan analisis sentimen.

ii. mencadangkan algoritma berasaskan TSKPBPP bagi pemilihan fitur dan pengurangan saiz fitur untuk pengelasan sentimen.

iii. menilai ketepatan fitur yang dipilih dengan menggunakan kaedah pemilihan fitur atau tanpa pemilihan fitur .

Seksyen seterusnya dalam kertas kajian ini disusun seperti berikut: Seksyen 2 akan menerangkan kajian yang telah dijalankan berkenaan topik yang dibincangkan. Seksyen 3 pula menerangkan mengenai metodologi kajian. Seksyen 4 akan membincangkan hasil keputusan kajian yang dijalankan. Manakala seksyen 5 pula merumuskan hasil kajian yang telah dijalankan.

2. KAJIAN BERKAITAN

Seksyen ini membincangkan tentang kajian yang telah dijalankan berkenaan analisis sentimen, pemprosesan teks, TSK dan algoritma PBPP.

2.1 Analisis Sentimen

Liu (2012) menerangkan sentimen sebagi perlombongan pendapat, iaitu bidang yang mengkaji ulasan pelanggan, pendapat, idea, penilaian, emosi manusia atau sikap terhadap sesuatu perkara seperti perkhidmatan, organisasi, produk, peristiwa, individu dan sebagainya. Liu (2008) pula memberi contoh satu set dokumen teks yang mengandungi ulasan pengguna mengenai sesuatu topik dan tujuan utama analisis sentimen adalah untuk mengenal pasti dan menilai fitur mengenai ulasan terhadap sesuatu objek yang terkandung dalam dokumen ini. Pang & Lee (2008) pula menerangkan analisis sentimen berfungsi dengan mengekstrak sentimen atau pendapat dengan menganalisis dokumen dalam bentuk teks.

Pandangan dan ulasan pelanggan daripada saluran media sosial dan internet amat penting kerana dapat membantu sesebuah organisasi menilai kualiti perkhidmatan yang diberikan dan juga mencari peluang untuk menambah baik mutu perkhidmatan sesuatu organisasi (Agarwal & Mittal 2013a). Sehubungan dengan itu teknologi analis sentimen amat diperlukan untuk menganalisis maklumat-maklumat ini (Ahmad et al. 2015).

2.2 Pemprosesan Teks

Pemprosesan teks adalah proses pembersihan dan penyediaan teks untuk pemilihan fitur dan pengelasan sentimen (Haddi et al. 2013). Tujuan pelaksanaan pemprosesan teks adalah untuk menghapuskan hingar dan data yang tidak relevan untuk mendapatkan keputusan pengelasan sentimen yang tepat (Pradha et al. 2019). Akan tetapi proses ini mengambil masa pengkomputeran yang agak panjang dan amat bergantung kepada saiz set data. Oleh itu,

adalah penting untuk memilih teknik pemprosesan yang cepat di samping boleh menghasilkan ketepatan yang tinggi.

Khader et al. (2018) menerangkan teknik pemprosesan teks ini terbahagi kepada dua kategori iaitu pemprosesan linguistik yang terdiri daripada proses penukaran ke huruf kecil, penghapusan URL, penghapusan @, penghapusan tanda pagar, penghapusan simbol, penghapusan ruang kosong, penghapusan format teks berkod dan penghapusan kata henti. Kategori pemproses bahasa tabii pula terdiri daripada *stemming*, *lemmatization* dan pembetulan ejaan.

2.3 Teori Set Kasar (TSK)

Teknik yang diperkenalkan oleh Pawlak (1982) berasal daripada model maklumat yang ringkas. TSK dianggap sebagai alatan matematik baru untuk memproses maklumat yang tidak pasti selepas teori ketidakpastian (Klir & Ramer 1990), teori set kabur (Zimmermann 2010) dalam bidang analisis data dan pemprosesan data. TSK dikenal pasti sebagai kaedah matematik yang efektif untuk meminimumkan data daripada sistem maklumat dan juga alatan matematik yang digunakan untuk menyelesaikan masalah yang tidak pasti (Han et al. 2016). TSK mempunyai prosedur yang tersusun rapi, algoritma dan alatan khusus untuk mengenal pasti pola dan berkebolehan untuk mencari pengurangan yang sah iaitu mencari set fitur yang paling minimum melalui fungsi *reduct*.

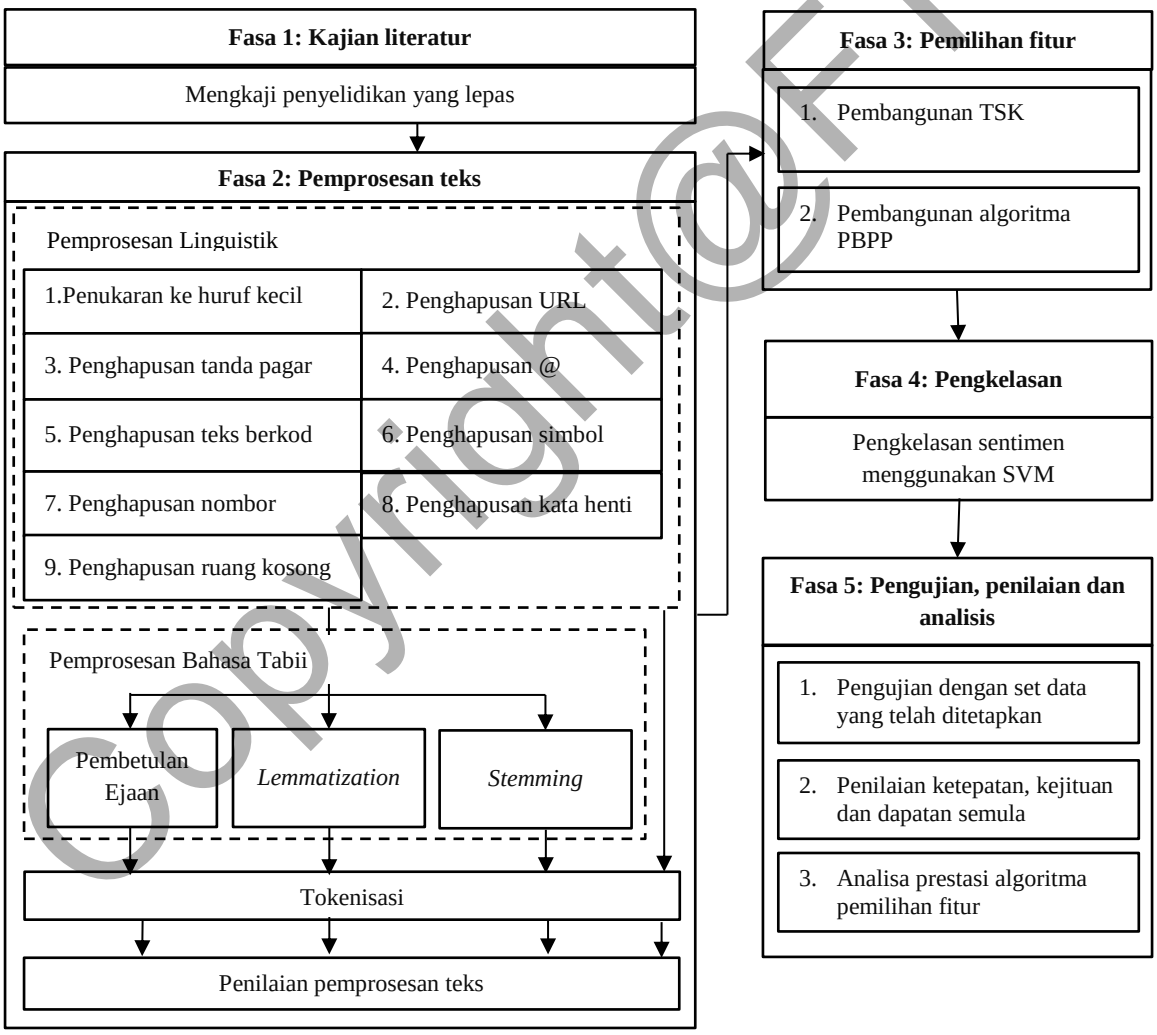
2.4 Algoritma Pengoptimuman Berasaskan Pengajaran Pembelajaran (PBPP)

Algoritma pemilihan fitur berasaskan populasi ini diperkenalkan oleh Rao et al. (2011). Ia diinspirasikan oleh proses pengajaran dan pembelajaran yang berlaku di dalam sesebuah kelas. Algoritma ini tidak memerlukan parameter kawalan spesifik, sebaliknya hanya memerlukan parameter kawalan umum seperti saiz populasi dan jumlah generasi.

Algoritma ini terbahagi kepada dua sesi iaitu sesi pengajaran dan sesi pembelajaran. Sesi pengajaran merupakan proses di mana pelajar akan belajar daripada pengajar. Pengajar akan cuba meningkatkan purata markah keputusan pelajar di dalam kelas dan berdasarkan kebolehannya melalui proses pengajaran. Sesi pembelajaran, pelajar akan meningkatkan pengetahuan melalui proses belajar bersama-sama. Pelajar akan berinteraksi secara rawak dengan pelajar lain untuk meningkatkan pengetahuan dan keputusan masing-masing.

3. METODOLOGI KAJIAN

Metodologi kajian ini terdiri daripada lima fasa iaitu fasa kajian literatur, fasa pemprosesan teks, fasa pembangunan algoritma pemilihan fitur, fasa pengelasan sentimen dan fasa pengujian, penilaian dan analisis seperti ditunjukkan dalam Rajah 1. Set data tanda aras yang digunakan dalam kajian ini diperoleh daripada laman web Kaggle yang boleh dicapai melalui https://www.kaggle.com/crowdflower/twitter-airline-sentiment. Ia berkenaan dengan ulasan pengguna terhadap enam buah syarikat penerbangan utama Amerika Syarikat yang diperoleh dari laman sosial Twiter pada Februari 2015. Set data ini mengandungi 14640 ulasan, 2363 daripada ulasan ini telah dilabelkan sebagai positif, 3099 sebagai neutral dan selebihnya 9178 sebagai negatif.



Rajah 1. Metodologi kajian

3.1 Fasa 1: Kajian Literatur

Fasa kajian literatur melibatkan kajian ke atas penyelidikan yang telah dilaksanakan oleh kajian lain sebelum ini. Kajian ini merangkumi pemahaman ke atas isu-isu semasa dalam analisis sentimen, teknik pemprosesan teks, teknik pemilihan fitur dan pengelasan sentimen.

3.2 Fasa 2: Pemprosesan Teks

Proses ini melibatkan pembersihan data dan penyediaan set data yang dipilih ke dalam bentuk yang membolehkan set data ini diproses oleh pemilihan fitur. Input kepada fasa ini merupakan set data tanda aras ulasan pengguna dari laman media sosial Twiter yang telah dilabelkan dengan maklumat sentimen.

3.2.1 Kaedah Pemprosesan Teks

Dalam proses ini, setiap ayat ulasan pengguna akan di ekstrak dan disimpan di dalam pangkalan data MySQL. Seterusnya set data ini kemudiannya melalui dua kategori pemprosesan teks iaitu pemprosesan linguistik dan pemprosesan bahasa tabii. Pemprosesan linguistik yang digunakan dalam kajian ini terdiri daripada sembilan teknik pemprosesan teks iaitu proses penukaran kepada huruf kecil, penghapusan URL, penghapusan tanda pagar, penghapusan @, penghapusan simbol, penghapusan nombor, penghapusan ruang kosong, penghapusan kata henti dan penghapusan teks berkod. Manakala untuk pemprosesan bahasa tabii pula, tiga teknik telah dipilih iaitu pembetulan ejaan, *stemming* dan *lemmatization*.

Terdapat empat model pemprosesan teks yang diuji dan dinilai iaitu Model A: gabungan pemprosesan linguistik dan pembetulan ejaan, Model B: gabungan pemprosesan linguistik dan *stemming*, Model C: gabungan pemprosesan linguistik dan *lemmatize* dan Model D: pemprosesan linguistik sahaja. Model yang menghasilkan keputusan pengelasan sentimen terbaik dipilih sebagai kaedah pemprosesan teks untuk menghasilkan senarai set fitur bersaiz kecil dalam kajian ini.

a. Pemprosesan Linguistik

Jadual 1 menunjukkan teknik dan contoh pemprosesan linguistik yang dilakukan dalam kajian ini.

Jadual 1. Teknik dan contoh pemprosesan linguistik

Bil	Teknik	Sebelum	Selepas
1.	Penukaran ke huruf kecil	*@**V**irgin**A**merica **S**eriously would pay $30 a flight for seaats that did not have this playing!!! **I**t is really the only bad thing about flying & **C**lick here http://t.co/**UT5G**r**R**w**A**a**A** #29**D**ays**T**o**G**o*	*@**v**irgin**a**merica **s**eriously would pay $30 a flight for seaats that did not have this playing!!! **i**t is really the only bad thing about flying & **c**lick here http://t.co/**ut5g**rw**a**a**a** #29**d**ays**t**o**g**o*
2.	Penghapusan URL	*@virginamerica seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying & click here **http://t.co/ut5grwaaa** #29daystogo*	*@virginamerica seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying & click here #29daystogo*
3.	Penghapusan tanda pagar	*@virginamerica seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying & click here **#29daystogo***	*@virginamerica seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying & click here*
4.	Penghapusan @	***@virginamerica** seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying & click here*	*seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying & click here*
5.	Penghapusan teks berkod	*seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying **&** click here*	*seriously would pay $30 a flight for seaats that did not have this playing!!! it is really the only bad thing about flying click here*
6.	Penghapusan simbol	*seriously would pay **$**30 a flight for seaats that did not have this playing**!!!** it is really the only bad thing about flying click here*	*seriously would pay 30 a flight for seaats that did not have this playing it is really the only bad thing about flying click here*
7.	Penghapusan nombor	*seriously would pay **30** a flight for seaats that did not have this playing it is really the only bad thing about flying click here*	*seriously would pay a flight for seaats that did not have this playing it is really the only bad thing about flying click here*
8.	Penghapusan kata henti	*seriously **would** pay **a** flight **for** seaats **that did not have this** playing **it is** really **the only** bad thing **about** flying click **here***	*seriously pay flight seaats playing really bad thing flying click*
9.	Penghapusan ruang kosong	*seriously pay flight seaats playing_really bad thing flying click*	*seriously pay flight seaats playing really bad thing flying click*

b. Pemprosesan Bahasa Tabii

Pemprosesan bahasa tabii yang dipilih untuk kajian ini terdiri daripada tiga teknik pemprosesan teks iaitu pembetulan ejaan, *lemmatization* dan *stemming* seperti ditunjukkan dalam Jadual 2.

Jadual 2. Teknik dan contoh pemprosesan bahasa tabii

Bil	Teknik	Sebelum	Selepas
1.	Pembetulan Ejaan	*seriously pay flight **seaats** playing really bad thing flying click*	*seriously pay flight **seats** playing really bad thing flying click*
2.	*Lemmatization*	*seriously pay flight seaats **playing** really bad thing **flying** click*	*seriously pay flight seaats **play** really bad thing **fly** click*
3.	*Stemming*	***seriously** pay flight **seaats playing really** bad thing **flying** click*	***serious** pay flight **seaat play realli** bad thing **fli** click*

3.2.2 Tokenisasi

Proses tokenisasi ialah satu proses di mana suatu ayat ulasan dari set data akan dipecahkan kepada beberapa siri fitur. Sebagai contoh, ayat "*seriously pay flight seats playing really bad thing flying click*" ditukarkan kepada siri perkataan "*seriously", "pay", "flight", "seats", "playing", " really", "bad", "thing", "flying", "click"*. Setiap perkataan yang terhasil daripada tokenisasi ini dianggap sebagai set fitur. Set fitur ini akan digunakan untuk langkah pengelasan teks bagi penilaian pemprosesan teks.

3.2.3 Penilaian Pemprosesan Teks

Proses penilaian dilakukan dengan menggunakan set fitur yang dihasilkan daripada proses tokenisasi dan dinilai melalui ketepatan pengelasan sentimen. Algoritma pengelasan yang digunakan dalam proses ini ialah *Support Vector Machine* (SVM). Penilaian yang dilakukan ini adalah untuk mengenal pasti gabungan kaedah pemprosesan teks yang dapat menghasilkan ketepatan pengelasan sentimen yang paling tinggi untuk dipilih dan digunakan dalam kajian ini.

3.3 Fasa 3: Pembangunan Algoritma Pemilihan Fitur

Fasa ketiga ini pula merupakan pembangunan algoritma pemilihan fitur TSKPBPP yang terdiri dari TSK dan algoritma PBPP. Ia dibangunkan dengan menggunakan pengaturcaraan PHP dan pangkalan data MySQL.

3.3.1 Pembangunan TSK

TSK dibangunkan untuk mengurangkan dimensi set fitur yang diperoleh daripada fasa pemprosesan teks dengan menggunakan fungsi *reduct*. Fungsi ini akan menilai dan menghapuskan fitur-fitur bertindan dan tidak relevan untuk mengurangkan saiz dimensi fitur.

Untuk proses ini, set fitur yang dihasilkan daripada pemprosesan teks diwakilkan dengan jadual sistem maklumat seperti yang ditunjukkan dalam Jadual 3. Nilai 1 menunjukkan atribut fitur yang dipilih, dan 0 menunjukkan atribut fitur tersebut tidak dipilih dalam sesuatu ulasan.

Jadual 3. Perwakilan fitur dalam jadual sistem maklumat

Ulasan	**a:** *seriously*	**b:** *flying*	**c:** *really*	**D (Kelas)**
S1	0	1	1	Negatif
S2	1	0	0	Negatif
S3	1	1	1	Positif
S4	0	1	1	Negatif
S5	1	0	1	Positif
S6	0	0	0	Negatif

Dalam jadual contoh ini, mempunyai tiga atribut fitur iaitu *seriously* (a), *flying* (b) dan *really* (c) yang terkandung di dalam enam ulasan pengguna (S1 sehingga S6) yang telah diberikan kelas (d) positif atau negatif. Dengan anggapan $U = \{S1, S2, S3, S4, S5, S6\}$, $W = \{a, b, c, d\}$, $C = \{a, b, c\}$ dan $D = \{d\}$. Langkah yang akan dilakukan oleh TSK adalah menghapuskan atribut fitur melalui pengiraan darjah kebergantungan dengan menggunakan fungsi *reduct*. Sebagai contoh, darjah kebergantungan atribut {d} daripada atribut {a,b} dikira seperti berikut:

$$k = \gamma_{\{a,b\}}(\{d\}) = \frac{|POS\ \{a,b\}(\{d\})|}{|U|} = \frac{|\{S1,S3,S4,S6\}|}{|\{S1\},\{S2\},\{S3\},\{S4\},\{S5\},\{S6\}|} = \frac{|4|}{|6|} = \frac{2}{3}$$

Darjah kebergantungan D = *{d}* ke atas semua subset adalah seperti berikut:

$\gamma_{\{a\}}(\{d\}) = \frac{|3|}{|6|}$ $\gamma_{\{b\}}(\{d\}) = 0$ $\gamma_{\{c\}}(\{d\}) = \frac{|2|}{|6|}$

$\gamma_{\{a,b\}}(\{d\}) = \frac{|4|}{|6|}$ $\gamma_{\{a,c\}}(\{d\}) = 1$ $\gamma_{\{b,c\}}(\{d\}) = \frac{|3|}{|6|}$

$\gamma_{\{a,b,c\}}(\{d\}) = 1$

Daripada pengiraan di atas, didapati bahawa atribut {b} boleh dihapuskan kerana ia tidak mempunyai nilai kebergantungan atau kebergantungan paling minimum iaitu kosong. Hasil terakhir jadual maklumat ditunjukkan pada Jadual 4.

Jadual 4. Perwakilan fitur dalam jadual sistem maklumat selepas *reduct*

Ulasan	**a:** *seriously*	**c:** *really*	**D (Kelas)**
S1	0	1	Negatif
S2	1	0	Negatif
S3	1	1	Positif
S4	0	1	Negatif
S5	1	1	Positif
S6	0	0	Negatif

3.3.2 Pembangunan Algoritma PBPP

Pembangunan algoritma PBPP adalah berdasarkan kod pseudo yang ditunjukkan dalam Rajah 2. Secara umumnya, pembangunan algoritma PBPP terdiri daripada tiga bahagian utama iaitu pengisytiharan parameter umum, penjanaan subset fitur penyelesaian awal dan pembaikan.

a. Bahagian Pengisytiharan Parameter Umum

Dalam bahagian pengisytiharan parameter umum, terdapat dua jenis parameter yang perlu ditetapkan iaitu saiz populasi dan juga bilangan generasi. Dalam kajian ini penetapan saiz parameter populasi diset kepada 30 (Allam & Nandhini 2018) dan bilangan generasi diset kepada 100 (Amiri 2012).

b. Bahagian Penjanaan Subset Penyelesaian Awal

Subset penyelesaian awal dijana secara rawak dan disimpan dalam bentuk tatasusunan satu dimensi seperti dalam Rajah 3. Sebagai contoh satu subset penyelesaian terdiri dari enam atribut fitur yang dilabelkan daripada F1 sehingga F6. Sel yang bernilai 1 mewakili atribut fitur yang dipilih dan sel bernilai kosong menggambarkan atribut fitur tersebut tidak dipilih.

Satu populasi boleh dianggap sebagai sebuah kelas dan setiap satu subset penyelesaian dianggap sebagai seorang pelajar. Manakala atribut fitur pula dianggap sebagai mata pelajaran yang diambil oleh pelajar tersebut. Nilai skor ialah markah purata bagi setiap mata pelajaran yang diambil oleh pelajar tersebut. Kemudian kedudukan pelajar di dalam kelas akan disusun mengikut purata markah mata pelajaran yang diambil.

```
    Bahagian pengisytiharan parameter umum
1   Pengisytiharan saiz populasi
2   Pengisytiharan kriteria pemberhentian (jumlah generasi/iterasi)
    Bahagian penjanaan subset penyelesaian awal
3   Jana populasi penyelesaian awal secara rawak
4   Penilaian set penyelesaian awal
5   Susun penyelesaian mengikut skor penilaian
    Bahagian pembaikan
6   Semak jika kriteria pemberhentian belum dipenuhi
7   {
        Sesi Pengajaran
7       Pilih penyelesaian yang terbaik sebagai pengajar
8       Pilih secara rawak satu penyelesaian sebagai pelajar
9       Proses pengajaran berlaku (penyilangan) di antara pengajar dan pelajar
10      Penilaian penyelesaian baru yang dihasilkan
11      Jika penyelesaian baru menghasilkan keputusan lebih baik:
12          Masukkan penyelesaian ke dalam populasi penyelesaian
13          Susun penyelesaian mengikut skor penilaian
14      Sebaliknya:
15          Penyelesaian baru tidak diterima
        Sesi Pembelajaran
16      Pilih dua penyelesaian secara rawak sebagai pelajar 1 dan pelajar 2
17      Proses pembelajaran berlaku (penyilangan) di antara pelajar yang dipilih
18      Penilaian penyelesaian baru yang dihasilkan
19      Jika penyelesaian baru menghasilkan skor lebih baik:
20          Masukkan penyelesaian ke dalam populasi penyelesaian
21          Susun penyelesaian mengikut skor penilaian
22      Sebaliknya:
23          Penyelesaian baru tidak diterima
24      Ulang semula langkah 6
25  }
26  Pulangkan subset fitur yang mempunyai skor terbaik sebagai set fitur berkualiti.
```

Rajah 2. Kod pseudo algoritma PBPP

F1	F2	F3	F4	F5	F6
1	0	1	0	1	1

Rajah 3. Contoh tatasusunan subset penyelesaian

c. Bahagian Pembaikan

Dalam bahagian pembaikan, algoritma ini terbahagi kepada sesi pengajaran dan sesi pembelajaran. Sebelum meneruskan ke sesi pengajaran dan pembelajaran, kriteria pemberhentian akan disemak terlebih dahulu.

i. Sesi Pengajaran

Dalam sesi pengajaran, satu subset penyelesaian yang mempunyai skor pengelasan sentimen yang terbaik akan dipilih sebagai pengajar. Kemudian, satu subset penyelesaian dipilih secara rawak untuk dijadikan sebagai seorang pelajar. Dalam contoh ini, set fitur diwakilkan oleh F1, F2, F3, F4, F5 dan F6. Sesi pengajaran di antara pengajar dan pelajar berlaku dengan kaedah penyilangan seperti ditunjukkan dalam Rajah 4.

INPUT:

Pengajar

Titik penyilangan

F1	F2	F3	F4	F5	F6
1	0	1	0	1	1

Pelajar

F1	F2	F3	F4	F5	F6
1	1	0	0	1	0

OUTPUT:

Penyelesaian baru 1

F1	F2	F3	F4	F5	F6
1	0	1	0	1	0

Penyelesaian baru 2

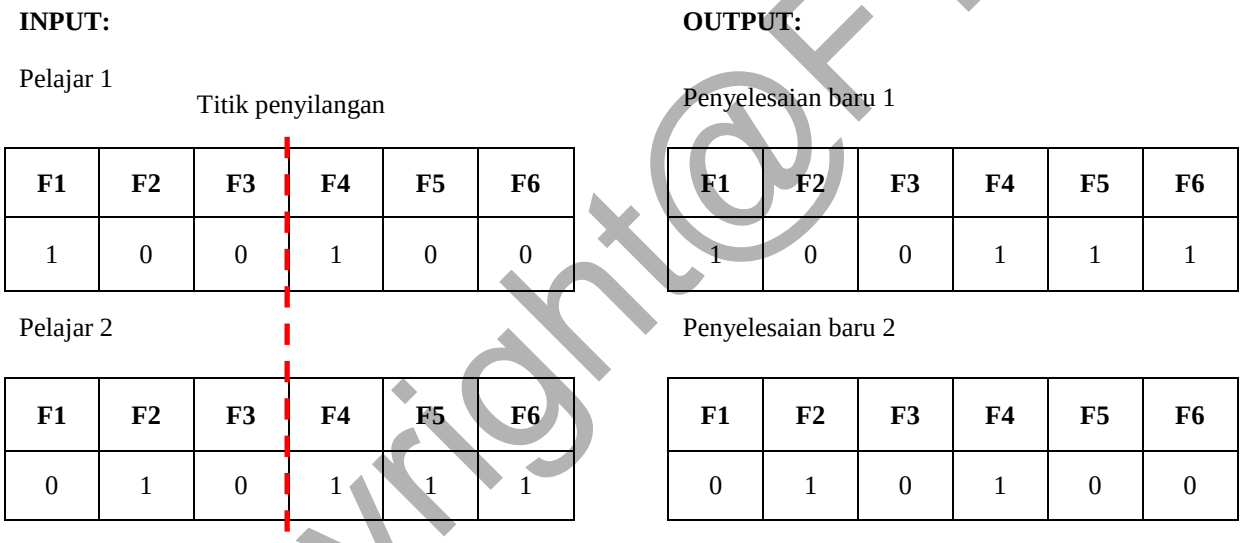
F1	F2	F3	F4	F5	F6
1	1	0	0	1	1

Rajah 4. Sesi pengajaran

Sesi pengajaran ini akan menghasilkan dua penyelesaian baru dan kedua-duanya akan melalui proses penilaian pengelasan sentimen. Berdasarkan keputusan penilaian ini, hanya penyelesaian yang menghasilkan skor yang lebih baik daripada senarai skor pelajar sedia ada sahaja yang diterima ke dalam populasi dan kedudukannya dalam populasi akan disusun berdasarkan skor yang diperoleh.

ii. Sesi Pembelajaran

Dalam sesi pembelajaran, pelajar akan belajar bersama-sama dengan pelajar lain untuk meningkatkan skor masing-masing. Oleh itu dalam sesi ini, dua orang pelajar akan dipilih secara rawak daripada populasi penyelesaian sedia ada. Kemudian kedua-dua pelajar ini akan melalui proses pembelajaran bersama-sama. Dalam algoritma ini, proses pembelajaran di antara kedua-dua pelajar berlaku melalui kaedah penyilangan seperti ditunjukkan dalam Rajah 5. Hasil daripada pembelajaran ini akan menghasilkan dua penyelesaian baru dan kemudian ia dinilai berdasarkan prestasi pengelasan sentimen. Berdasarkan skor penilaian sentimen yang diperoleh, sekiranya penyelesaian baru ini lebih baik daripada penyelesaian sedia ada, ia akan diterima dan kedudukannya dalam populasi akan disusun berdasarkan skor.



Rajah 5. Sesi pembelajaran

Kedua-dua sesi pengajaran dan pembelajaran ini akan diulang sehingga kriteria berhenti dipenuhi. Sekiranya kriteria berhenti tidak dipenuhi, proses pembaikan ini akan diulang semula. Output kepada algoritma ini adalah subset fitur berkualiti dan bersaiz kecil yang akan digunakan dalam fasa pengelasan sentimen.

3.4 Fasa 4: Pengelasan Sentimen

Fasa 4 ialah fasa pengelasan sentimen. Dalam fasa ini, pengelasan sentimen akan dilaksanakan melalui algoritma SVM menggunakan set fitur yang dipilih oleh fasa pemilihan fitur. Pengelasan sentimen ini dilakukan dengan menggunakan perisian WEKA (*Waikato*

Environment for Knowledge Analysis) versi 3.8. Dalam WEKA, algoritma SVM dikenali sebagai LibSVM. Dalam kajian ini, parameter untuk algoritma SVM adalah seperti berikut; parameter c diset kepada 1, fungsi *gamma* di set kepada 0 , parameter *kernel* diset kepada *linear* dan menggunakan kaedah 10 lipatan pengesahan silang (Khalid et al. 2020).

3.5 Fasa 5: Pengujian, penilaian dan analisis

Fasa 5 ialah fasa pengujian, penilaian dan analisis yang dilakukan berasaskan keputusan pengelasan sentimen. Keputusan prestasi pengelasan diuji berdasarkan matriks kekeliruan yang diperoleh daripada keputusan pengelasan. Matriks kekeliruan menunjukkan maklumat mengenai jumlah sebenar sesuatu kelas dan jumlah ramalan yang dijana oleh algoritma pengelasan seperti ditunjukkan dalam Jadual 5.

Jadual 5. Matriks kekeliruan

		Keputusan Pengelasan	
		Ya	**Tidak**
Kelas sebenar	**Ya**	TP	FN
	Tidak	FP	TN

Positif benar (TP) ialah keadaan di mana kes positif berjaya kelaskan sebagai positif. Negatif benar (TN) ialah keadaan di mana kes negatif berjaya dikelaskan sebagai negatif. Positif palsu (FP) ialah kes negatif tetapi disalah kelaskan sebagai positif. Negatif palsu (FN) ialah kes positif tetapi disalah kelaskan sebagai kes negatif.

Terdapat tiga kriteria pengujian prestasi yang digunakan iaitu ketepatan, kejituan (*p*) dan dapatan semula (*r*) berdasarkan persamaan berikut:

$$ketepatan = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$Kejituan\ (p) = \frac{TP}{TP + FP}$$

$$Dapatan\ semula\ (r) = \frac{TP}{TP + FN}$$

Bagi tujuan perbandingan dan penilaian ke atas algoritma yang digunakan, pelaksanaan pengujian dilakukan dengan beberapa kajian yang telah dipilih. Akhir sekali,

analisis dilakukan ke atas keputusan eksperimen bagi melihat prestasi algoritma TSKPBPP untuk pemilihan fitur dalam pengelasan sentimen.

4. KEPUTUSAN DAN PERBINCANGAN

Bahagian ini menerangkan keputusan kajian pemilihan kaedah pemprosesan teks dan pemilihan fitur dengan menggunakan algoritma TSKPBPP.

4.1 Pemilihan Kaedah Pemprosesan Teks

Daripada kajian yang dijalankan, didapati bahawa model A iaitu gabungan pemprosesan bahasa tabii dan pembetulan ejaan daripada pemprosesan linguistik memperoleh bacaan kadar ketepatan, kejituan dan dapatan semula yang terbaik iaitu 76.9%, 76.3 dan 76.9% seperti ditunjukkan dalam Jadual 6.

Jadual 6. Perbandingan prestasi model pemprosesan teks

Model	Ketepatan	Kejituan	Dapatan Semula
Model A	**76.9 %**	**76.3%**	**76.9%**
Model B	75.1 %	75.6%	75.1%
Model C	74.2 %	73.2%	74.2%
Model D	73.4%	67.2%	73.4%

Selepas perbandingan prestasi dilakukan, pemilihan model terbaik ditentukan berdasarkan jumlah skor kedudukan bagi setiap model berdasarkan metrik penilaian ketepatan, kejituan dan dapatan semula. Jadual 7 menunjukkan model A berada di kedudukan yang terbaik dengan skor 3 berbanding model lain. Daripada proses pemprosesan teks model ini juga, sebanyak 8424 fitur telah dikenal pasti.

Jadual 7. Skor kedudukan prestasi model pemprosesan teks

Model	Ketepatan	Kejituan	Dapatan Semula	Jumlah Skor	Skor Akhir
Model A	**1**	**1**	**1**	**3**	**1**
Model B	2	2	2	6	2
Model C	3	3	3	9	3
Model D	4	4	4	12	4

4.2 Pemilihan Fitur Dengan Menggunakan TSKPBPP

Matriks kekeliruan hasil pengelasan sentimen dengan menggunakan algoritma TSKPBPP adalah seperti ditunjukkan dalam Jadual 8. Daripada jadual ini, didapati sentimen negatif memperoleh ketepatan pengelasan paling tinggi iaitu sebanyak 95%, diikuti sentimen positif sebanyak 86% dan diikuti sentimen neutral sebanyak 28%. Jurang perbezaan ketepatan yang besar ini menunjukkan set data yang digunakan tidak seimbang.

Jadual 8. Matriks kekeliruan keputusan pengelasan

		Keputusan pengelasan			**Jumlah sebenar**	**Peratus ketepatan**
		Negatif	**Neutral**	**Positif**		
Sebenar	**Negatif**	*8802	20	356	9178	95%
	Neutral	1137	***889**	1073	3099	28%
	Positif	278	30	***2055**	2363	86%

Nota: * ramalan yang tepat

Jadual 9 pula memaparkan perbandingan keputusan kajian ini dengan kajian-kajian lain. Daripada jadual ini, didapati pengelasan sentimen menggunakan teknik pemilihan fitur TSKPBPP mendapat kadar ketepatan dan dapatan semula yang terbaik iaitu 80.2% berbanding teknik pemilihan fitur lain. Kadar kejituan pula menunjukkan algoritma ini memperoleh 83.6%. Bilangan fitur berjaya dikurangkan sebanyak 98.9% iaitu 95 fitur dipilih berbanding 8424 fitur asal. Daripada keputusan ini didapati algoritma TSKPBPP mampu berfungsi dengan baik sebagai algoritma pemilihan fitur dalam pengelasan sentimen. Kemampuan ini didorong oleh algoritma TSKPBPP yang tidak memerlukan pelarasan parameter spesifik untuk beroperasi secara optimum.

Jadual 9. Perbandingan keputusan algoritma pemilihan fitur

Pemilihan Fitur	**Ketepatan**	**Kejituan**	**Dapatan Semula**
TSKPBPP	**80.2 %**	83.6%	**80.2%**
IG	63.1 %	**97.9%**	63.1%
TF	77.3%	77.0%	70.0%
TF-IDF	72.0%	86.7%	72.0%

5 RUMUSAN

Kajian ini menunjukkan bahawa kaedah pemprosesan teks pembetulan ejaan daripada kategori pemprosesan bahasa tabii yang digabungkan dengan teknik pemprosesan linguistik tabii memberi kesan yang signifikan terhadap ketepatan pengelasan sentimen berbanding dengan kaedah pemprosesan bahasa tabii yang lain.

Kajian ini juga menunjukkan bahawa algoritma TSKPBPP mampu melaksanakan pemilihan fitur untuk pengelasan sentimen lebih baik atau setanding dengan algoritma pemilihan fitur sedia ada sekarang dengan kadar ketepatan pengelasan sebanyak 80.2%, kejituan 83.6% dan dapatan semula 80.2%. Dalam masa yang sama algoritma ini berjaya mengurangkan saiz fitur sebanyak 98.9% iaitu 95 fitur daripada 8424 fitur asal.

Seterusnya, hasil kajian ini menunjukkan bahawa kadar ketepatan pengelasan dengan menggunakan gabungan kaedah pemprosesan teks dan algoritma pemilihan fitur TSKPBPP adalah lebih tinggi iaitu 80.2% berbanding pengelasan sentimen yang hanya menggunakan pemprosesan teks dengan 76.9% sahaja. Ini adalah disebabkan oleh algoritma pemilihan fitur ini mampu mengurangkan dimensi set fitur dan mampu memilih set fitur yang berkualiti untuk digunakan dalam pengelasan sentimen.

Kajian ini juga menunjukkan bahawa metodologi pemprosesan teks memberikan kesan yang amat signifikan dalam analisis sentimen. Ini disebabkan pemprosesan teks banyak membantu dalam menghapuskan hingar dan maklumat yang tidak relevan daripada data dan mampu mendapatkan keputusan pengelasan yang agak baik. Sekiranya hingar dan maklumat yang tidak relevan ini tidak dihapuskan, maka proses pemilihan fitur akan menjadi lebih sukar, memerlukan sumber pemprosesan yang tinggi dan secara tidak langsung ketepatan pengelasan sentimen juga akan berkurangan. Algoritma pemilihan fitur pula berfungsi membantu meningkatkan lagi ketepatan pengelasan sentimen berdasarkan hasil daripada pemprosesan teks ke satu tahap yang lebih tinggi.

PENGHARGAAN

Pengarang ingin mengucapkan ribuan terima kasih kepada pensyarah Fakulti Teknologi dan Sains Maklumat (Prof Dr Salwani Abdullah), Universiti Kebangsaan Malaysia (UKM), keluarga dan juga rakan-rakan atas segala sokongan dalam menjalankan kajian ini.

RUJUKAN

Abbasi, A., France, S., Zhang, Z. & Chen, H. 2011. Selecting Attributes for Sentiment Classification using Feature Relation Networks. *IEEE Transactions on Knowledge and Data Engineering* 23(3): 447–462.

Allam, M. & Nandhini, M. 2018. Optimal Feature Selection using Binary Teaching Learning Based Optimization Algorithm. *Journal of King Saud University - Computer and Information Sciences* 1–13.

Amiri, B. 2012. Application of Teaching-Learning-Based Optimization Algorithm on Cluster Analysis. *Journal of Basic and Applied Scientific Research* 2(11): 11795–11802.

Arafat, H., Elawady, R. M., Barakat, S. & Elrashidy, N. M. 2014. Different Feature Selection for Sentiment Classification. *International Journal of Information Science and Intelligent System* 3(1): 137–150.

Haddi, E., Liu, X. & Shi, Y. 2013. The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science* 17(December 2014): 26–32.

Han, Z., Zhang, Q. & Wen, F. 2016. A Survey on Rough Set Theory and its Application. *Kongzhi Lilun Yu Yingyong/Control Theory and Applications* 16(2).

Khader, M., Awajan, A. & Al-naymat, G. 2018. The Effects of Natural Language Processing on Big Data Analysis : Sentiment Analysis Case Study. *2018 International Arab Conference on Information Technology (ACIT)* 1–7.

Khalid, M., Ashraf, I., Mehmood, A., Ullah, S., Ahmad, M. & Choi, G. S. 2020. GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier. *Applied Sciences (Switzerland)* 10(8): 1–20.

Klir, G. J. & Ramer, A. 1990. Uncertainty in the Dempster-Shafer theory: A critical re-examination. *International Journal of General Systems* 18(2): 155–166.

Pang, B. & Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(2): 1–135.

Pawlak, Z. 1982. Rough Sets. *International Journal of Computer & Information Sciences* 11(5): 341–356.

Pradha, S., Halgamuge, M. N. & Tran Quoc Vinh, N. 2019. Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. *Proceedings of 2019 11th International Conference on Knowledge and Systems Engineering, KSE 2019* 1–8.

Rao, R. V., Savsani, V. J. & Vakharia, D. P. 2011. Computer-Aided Design Teaching - Learning-Based Optimization : A Novel Method for Constrained Mechanical Design Optimization Problems. *Computer-Aided Design* 43(3): 303–315.

Rustam, F., Ashraf, I., Mehmood, A., Ullah, S. & Choi, G. S. 2019. Tweets Classification on the Base of Sentiments for US Airline Companies. *Entropy* 21(11).

Seerat, B. & Azam, F. 2012. Opinion Mining: Issues and Challenges (A survey). *International Journal of Computer Applications* 49(9): 42–51.

Vinodhini, G. & Chandrasekaran, R. M. 2013. Effect of Feature Reduction in Sentiment analysis of

online reviews. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 2(6): 2165–2172.

Zimmermann, H. J. 2010. Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(3): 317–332.