ENHANCED ASPECT LEVEL OPINION MINING KNOWLEDGE EXTRACTION AND REPRESENTATION

MAQBOOL RAMDHAN IBRAHIM AL-MAIMANI

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARGATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT

Author's full name : <u>**MAQBOOL RAMDHAN IBRAHIM AL-MAIMANI**</u>

Date of Birth : <u>**1/1/1963**</u>

Title : <u>**ENHANCED ASPECT LEVEL OPINION MINING KNOWLEDGE EXTRACTION AND REPRESENTATION**</u>

Academic Session : **2014/2015-2**

I declare that this thesis is classified as:

☐	**CONFIDENTIAL**	(Contains confidential information under the Official Secret Act 1972)*
☐	**RESTRICTED**	(Contains restricted information as specified by the organization where research was done)*
✓	**OPEN ACCESS**	I agree that my thesis to be published as online open access (full text)

I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:

1. The thesis is the property of Universiti Teknologi Malaysia

2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.

3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

SIGNATURE	**SIGNATURE OF SUPERVISOR**
04035018	Prof. Dr. Naomie Salim
(NEW IC NO/PASSPORT)	**NAME OF SUPERVISOR**
Date: August 2015	Date: August 2015
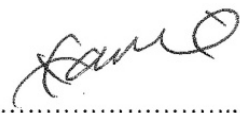
NOTES: *   If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Doctor of Philosophy (Computer Science)"

Signature   : ....

Name of Supervisor   : Prof. Dr. Naomie Salim

Date   : AUGUST 2015

BAHAGIAN A – Pengesahan Kerjasama*

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui kerjasama

antara _____ dengan _____

Disahkan oleh:

Tandatangan   :   ....   Tarikh :   ....

Nama   :   ....

Jawatan   :   ....
(Cop rasmi)

* *Jika penyediaan tesis/projek melibatkan kerjasama.*

BAHAGIAN B – Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar   :   **Prof. Dr. Narayanan Kulathuramaiyer**

Faculty of Computer Science and Information

Technology,

Universiti Malaysia Sarawak,

94300 Kota Samarahan,

Sarawak

Nama dan Alamat Pemeriksa Dalam   :   **Prof. Dr. Ali bin Selamat**

Pusat Teknologi Maklumat dan Komunikasi,

UTM Johor Bahru.

Nama Penyelia lain (jika ada)   :   -

Disahkan oleh Timbalan Pendaftar di Sekolah Pengajian Siswazah:

Tandatangan :   --   Tarikh :   ----------

Nama   :   **ASRAM BIN SULAIMAN @ SAIM**

ENHANCED ASPECT LEVEL OPINION MINING KNOWLEDGE EXTRACTION AND REPRESENTATION

MAQBOOL RAMDHAN IBRAHIM AL-MAIMANI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

AUGUST 2015

I declare that this thesis entitled *"Enhanced Aspect Level Opining Mining Knowledge Extraction and Representation"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature   :

Name   :   MAQBOOL RAMDHAN IBRAHIM AL-MAIMANI

Date   :   AUGUST 2015

To my father and mother

ACKNOWLEDGEMENT

ABSTRACT

There is a need to find more effective techniques to extract, classify, represent and summarize customers' online opinions on products and services for better sentiment analysis. The aim of this thesis is to enhance aspect level opinion extraction and representation. This study uses SentiWordNet lexical resource which is specifically built for opinion mining and widely used in sentiment analysis. This research introduces an approach using adjectives, verbs, adverbs and nouns (AVAN) which analyses all opinion word types for sentiment analysis and not only limited to adjectives and adverbs as have been conventionally done. SentiWordNet is used in this thesis to identify and analyze all word types for opinion extraction and representation. Opinion representation is enhanced by capturing key elements of opinions into predicates that consists of opinion word, strength, score and category in order to improve the opinion representation and classification. Then it further enhances the mining by introducing opinion accounting which summarizes opinion scores at various group levels. In addition, this thesis introduces a new concept called opinion strength which classifies opinions into degrees. An enhanced score is assigned to opinion based on the strength at which these opinions are expressed. Furthermore, as opinions are fuzzy in nature, this study shows that fuzzy logic is an effective technique to address opinion vagueness since human-like logic is fuzzy. This is important as opinions should not only be categorized in classical Boolean sentiments. This study identifies SentiWordNet, AVAN, Opinion Strength and fuzzy logic as classification features to classify customer reviews into a 5-class prediction model (Excellent, Good, Fair, Poor and Very Poor). The results show an accuracy of 92% using Sequential Minimal Optimization classifier for these features, outperforming previous works that implemented Support Vector Machine and Logistic Regression. Moreover, combination of AVAN, Opinion Strength and fuzzy logic outperformed SentiWordNet alone by a 30% accuracy.

ABSTRAK

Terdapat keperluan untuk mencari teknik-teknik yang lebih berkesan untuk mengekstrak, mengelaskan, mewakilkan dan merumuskan pendapat pelanggan dalam talian terhadap produk dan servis untuk analisis sentimen yang lebih baik. Tesis ini bertujuan untuk meningkatkan aspek pengekstrakan dan perwakilan pendapat. Kajian ini menggunakan sumber leksikal SentiWordNet yang khusus dibina bagi perlombongan pendapat dan digunakan secara meluas dalam analisis sentimen. Kajian ini memperkenalkan pendekatan yang menggunakan Kata Sifat, Kata Kerja, Kata Adverba dan Kata Nama (AVAN), bagi menganalisis kesemua jenis sentimen perkataan berkaitan pendapat dan ia juga tidak hanya terhad kepada Kata Sifat dan Kata Adverba seperti dalam pendekatan konvensional. Perwakilan pendapat dipertingkatkan dengan menawan elemen-elemen utama pendapat ke dalam predikat yang terdiri daripada perkataan pendapat, kekuatan, skor dan kategori bagi meningkatkan perwakilan pendapat dan klasifikasinya. Seterusnya, perakaunan pendapat telah diperkenalkan untuk meringkaskan skor pendapat pada tahap kumpulan yang pelbagai. Di samping itu, tesis ini memperkenalkan satu konsep baru yang dikenali sebagai kekuatan pendapat dengan mengklasifikasikannya kepada darjah pendapat tertentu. Skor diperuntukkan kepada pendapat berdasarkan kepada kekuatan pendapat itu dinyatakan. Selain itu, sebagaimana pendapat adalah kabur dalam alam semulajadi, kajian ini menunjukkan bahawa logik kabur adalah teknik yang efektif untuk digunakan kerana logik manusia terdapat kekaburan. Ini adalah penting kerana pendapat tidak seharusnya hanya boleh dikategorikan dalam sentimen Boolean klasik. Kajian ini mengenalpasti SentiWordNet, AVAN, Kekuatan Pendapat dan Logik Kabur sebagai ciri-ciri untuk mengelaskan ulasan pelanggan ke dalam model ramalan 5 kelas (Cemerlang, Baik, Sederhana, Kurang Baik dan Tidak Baik). Keputusan menunjukkan Pengelas Turutan Pengoptimuman Minimum menggunakan ciri-ciri ini memberikan 92% lebih ketepatan berbanding teknik sebelum ini iaitu Mesin Sokongan Vektor dan Regresi Logistik. Selain itu, pengabungan AVAN, Pendapat Kekuatan dan Logik Kabur mengatasi SentiWordNet secara bersendirian dengan ketepatan 30%.