Analysis of Aspect-Based Sentiment in Clinical Text from Electronic Medical Records

Nuttapong SANGLERDSINLAPACHAI

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Analysis of Aspect-Based Sentiment in Clinical Text from Electronic Medical Records

by

Nuttapong SANGLERDSINLAPACHAI

Supervisor: Associate Professor Hieu Chi DAM
Professor Tu Bao HO

School of Knowledge Science
Japan Advanced Institute of Science and Technology

September 2019

Abstract

Sentiment analysis is a process of understanding an opinion in a written or spoken language. It may be applied at different scales, ranging from phrases to a whole document. Instead of determining the sentiment of an entire text portion, aspect-based sentiment analysis addresses sentiments corresponding to parts, components, attributes, or aspects of an entity of interest, which are mentioned in the given text portion. This dissertation studies how a linguistic structure is used to improve aspect-based sentiment analysis and how to apply sentiment analysis to a document in medical domain, especially, a clinical narrative.

In our study, an aspect mentioned in a text portion is first detected, and elementary discourse units (EDUs) relevant to the aspect are then localized by using the linguistic structure, i.e., the rhetorical structure theory (RST). Using lexicon-based approaches, the polarity scores of terms occurring in an EDU are combined into the polarity score of the EDU. We propose a new score aggregation strategy that utilizes RST to aggregate scores from all EDUs relevant to the aspect. Experimental results on online product reviews demonstrate that our new score aggregation method improves sentiment classification at the level of local aspect segments.

To apply the proposed method to clinical text in electronic medical records (EMRs), some extensions are required. The medical-domain-knowledge corpus, i.e., the Unified Medical Language System (UMLS), is employed to detect aspects mentioned in a clinical narrative. Local aspect segments are then formed by using RST. However, occurrences of medicine-technical terms, e.g., disease names or treatment processes, make the sentiment on a clinical narrative hard to analyse. For example, the sentiment of the text portion "Appears to have premature atrial contraction with bundle showing" depends greatly on the meaning of the term "premature atrial contraction". Semantic types of technical terms, provided by UMLS, are incorporated into lexicon-based sentiment classification methods of two types, i.e., methods using a generic sentiment lexicon

and those using a trained sentiment lexicon. Preliminary results show that different classification methods are appropriate for text portions containing different semantic types. Classifier combination is then employed to select a classification method that is most suitable for an input text portion.

Keywords: Aspect-based sentiment analysis, Lexicon-based sentiment classification method, Rhetorical structure theory, Product review, Clinical narrative, Clinical text, Electronic medical record, Classifier combination

Acknowledgments

Firstly, I would like to express my sincere gratitude to my former supervisor, Professor Tu Bao Ho for the continuous support of my doctoral study and research. I am deeply grateful for his kindness to give me one of the greatest opportunities in my life to join his laboratory at Japan Advanced Institute of Science and Technology (JAIST). He has been giving me a lot of advice and knowledge for my study, research, and also daily life aboard.

Many thanks to my current supervisor, Associate Professor Hieu Chi Dam, for all his support and generosity. I am extremely grateful to other two supervisors in the Dual-Degree program, Associate Professor Ekawit Nantajeewarawat (Sirindhorn International Institute of Technology, SIIT), and Dr. Anon Plangprasopchok (National Electronics and Computer Technology Center, NECTEC), for their valuable advice, suggestions, and comments on my research.

Besides my supervisors, I would like to thank other members of my examination committee both at JAIST and SIIT for their valuable time and insightful comments on my work. The comments have been valuable to my research.

I would like to thank the SIIT-JAIST-NECTEC Dual Doctoral Degree program for the scholarship. Also, I would like to thank all friends in Ho Laboratory, at JAIST and SIIT for their help, cheers, and comments giving to me all times we are together.

Special thanks to my family, mother and father, for their love and support I have always gotten.

Contents

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Sirindhorn International Institute of Science and Technology, Thammasat University.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sentiment Analysis and Clinical Text

Sentiment analysis [1, 2] is a method for analysing given text and extracting attitudes or opinions from it. The analysis can be performed at various levels, e.g., words, clauses, sentences, paragraphs, or documents. The analysis may focus on parts or attributes of an object of interest instead of entire text portions. This type of sentiment analysis is called aspect-based or fine-grained sentiment analysis [3, 4, 5]. Output of sentiment analysis depends on applications. Some applications only require subjectivity detection, i.e., detecting whether a sentence contains opinions of its author. Some applications may need more details about opinions, e.g., polarity or emotion. While emotion can be divided into many classes, polarity is separated into only two or three classes, i.e., positive class, negative class, and sometimes neutral class.

Sentiment analysis is used in many domains. In the marketing domain [6, 4, 7, 8], from opinions about products or services expressed on web sites, e.g., Amazon[1] or TripAdvisor[2], feedbacks from customers can be automatically analysed using sentiment analysis. In the financial domain, sentiment analysis is applied to news articles or online posts, and the resulting sentiments may be used as features for predicting a stock price, e.g., in [9]. In the medical domain, sentiment analysis can be used for revealing the health status of a patient and the progress of a treatment [10].

[1] https://www.amazon.com

[2] https://www.tripadvisor.com

Text currently used in the medical research is collected from three main sources, i.e., literature, on-line sources, and electronic medical records (EMRs) [10]. Text collected from the literature, e.g., the publications on MEDLINE[3], is well-written, consisting of problems and findings from existing research works. Many works mined the abstracts of biomedical papers to obtain knowledge for a decision support system [11, 12, 13]. On-line sources include on-line health forums, e.g., PatientsLikeMe[4], or blogs and micro-blogs, e.g., Twitter[5]. Text collected from these sources is mainly provided by customers or patients. The topics of the text may vary from health status, medication effectiveness, to quality of medical providers [14, 15, 16, 17]. Under the Health Information Technology for Economic and Clinical Health (HITECH) Act[6], electronic medical records (EMRs) are increasing and become interesting for medical research community [10, 18]. A textual part in an EMR is called "clinical text", which is in the form of plain text describing a patient, written by medical staffs. Clinical text records activities and states of a patient from admission to discharge. It is a large resource for medical research from a real environment.

A sentence in clinical text is regularly an affirmative sentence. Opinion words, e.g., "good" or "bad", may not appear in the text and may be replaced by specific words in the medical domain. Consider the sentence "The neck was supple and without lymphadenopathy" taken from a discharge summary in the MIMIC II database [19], for example. This sentence gives information that the neck of a patient is normal or positive. To classify that the sentence is positive, we have to know that "supple" and "lymphadenopathy" have positive senses for the "neck". Consequently, aspect-based sentiment analysis seems to be useful for handling this issue.

[3]https://www.nlm.nih.gov/bsd/pmresources.html

[4]https://www.patientslikeme.com

[5]https://twitter.com

[6]https://www.fpc.gov/health-information-technology-for-economic-and-clinical-health-act-of-2009-hitech

1.2 Motivations and Research Problems

In this dissertation, aspect-based sentiment analysis on clinical text is studied, with two research problems being addressed as follows.

1. *Classifying the sentiment of a given aspect with respect to a linguistic structure*:

   Aspects of an object of interest may be obtained using various methods, e.g., manual defining [20], collecting with respect to statistical counts [3], and applying topic modelling algorithms [21]. However, an important issue is how to assign a sentiment to each aspect. Some works look for opinion terms related to aspects [3] or analyse the whole text portion (e.g., a clause and a sentence) that contains an aspect term [4]. In [22, 23, 24, 25], linguistic structures, e.g., the term dependency and the discourse structure, are first employed to identify the text portions which are relevant to an aspect. The sentiment of the aspect is then classified using those relevant portions.

2. *Classifying sentiments in clinical text*:

   Sentiment analysis on the medical domain was introduced in 2005 [11]. Most research works [11, 12, 26, 27, 13] conduct sentiment analysis on text from two data sources, i.e., literature and on-line sources. Few works are interested in clinical text, e.g., in [10, 18]. A major difficulty of sentiment analysis on the medical domain is a domain-specific meaning of a term. Considering the sentence "Your blood test on this disease is positive", for example, the term "positive", which generally gives a positive sentiment, makes the sentiment of the sentence negative.

1.3 Contributions

The main contributions of this dissertation are as follows:

1. *Aspect-based text segmentation using RST*:

   This dissertation proposes a method that uses the rhetorical structure theory (RST) for localising text portions (i.e., EDUs) relevant to a given aspect. Sen-

timent classification considering all relevant EDUs yield more accurate results than that considering only the EDU containing the aspect.

2. *Aspect-based polarity score aggregation using a rhetorical structure*:

   This dissertation proposes weighted-averaging methods and rule-based methods that exploit the RST structures of relevant EDUs and their relations to infer the sentiment of a given aspect. Incorporated with heuristics, i.e., removing conflicting rules and tuning the confidence threshold, the rule-based methods can improve the classification performance on the negative polarity class.

3. *Sentiment analysis on clinical text using UMLS semantic types*:

   This dissertation proposes strategies that exploit UMLS semantic types in the Disorders group to improve lexicon-based sentiment classification methods. Apart from increasing an accuracy, the proposed strategies reduce the size of a training set that is required for constructing a lexicon.

4. *Influence of UMLS semantic types in Disorders group on sentiment analysis*:

   The influence of each UMLS semantic type in the Disorders group on sentiment classification is studied. Not all semantic types in this group indicate negative sentiment. The semantic type "acquired abnormality", for example, does not express a negative sentiment even though the type is in the Disorder group. This obtained knowledge is important for employment of UMLS semantic types in applications involving sentiment classification.

1.4 Dissertation Organization

The remaining content of the dissertation is organized as follows:

Chapter 2 describes background knowledge and related works used in this research, i.e., aspect-based sentiment analysis, clinical text, and use of RST and UMLS for sentiment analysis.

Chapter 3 presents the proposed aspect-based sentiment classification methods

that exploit a linguistic structure, i.e., RST. All experiments in this chapter are conducted on product reviews collected from on-line sources.

Chapter 4 illustrates how to applied the developed methods in chapter 3 to clinical text by using a domain knowledge corpus, i.e., UMLS. The difference between clinical narratives and product reviews, in terms of a rhetorical structure, is also presented in this chapter.

Chapter 5 summarizes findings and knowledge received from this research, and describes further research that possibly employs the methods and the findings presented in this dissertation.

Chapter 2

Related Works

2.1 Aspect-based Sentiment Analysis

Sentiment analysis is a natural language processing component that extracts sentiments from given textual descriptions [1, 28]. It may be applied at different scales, ranging from phrases to a whole document [28, 29]. Recently, sentiment analysis at the aspect level [28, 8, 30] gains more attention since tons of textual documents are more available on the Internet, such as reviews of products or services, and each of them generally describes things in various aspects. For example, a review of a mobile phone may mention its several features (aspects), i.e., screen, camera, price, etc.

Essentially, aspect-based sentiment analysis performs two main tasks [28]: aspect extraction and sentiment classification. The first one is concerned with identifying the relevant aspects of a given text portion. In prior works, an aspect is identified using descriptive statistics, e.g., term frequencies [4], or using topic modelling techniques, e.g., Latent Dirichlet Allocation (LDA) [31]. Jo and Oh [8] applied LDA at the sentence level to detect aspects and their respective sentiment words. Moghaddam and Ester [30] proposed an LDA-based framework, in which an opinion phrase, a pair of a term and its modifier, was used as a substitute source for term occurrences in the LDA process. After processing LDA with bag-of-phrases, aspects and their associated sentiments, were then extracted from top opinion phrases.

Once textual parts relevant to an aspect are identified, the second task, sentiment

classification, is performed. Machine-learning-based and lexicon-based methods are the two mainstream approaches. In the former approach [32, 33, 34], text content is segmented into a bag of words, and a machine-learning technique, e.g., Support Vector Machine (SVM) [32, 33], is then applied to discover term occurrence patterns for each polarity class. In the latter approach, a sentiment score is calculated from the polarity scores of terms, which are obtained from a list of sentiment-carrying words or lexicon corpus, e.g., SentiWordNet [35]. A lexicon-based method seeks for a proper linear combination of term scores that represents an overall sentiment. Chamlertwat et al. [20] proposed a simple yet efficient method that sums the polarity scores of terms to represent the sentiment of a document. Negation terms (e.g., not, no) are used as triggers to flip the sentiment polarity. However, there is a common drawback to these two approaches. Since a document is disintegrated into a bag of words, ignoring linguistic structures, the two approaches typically perform poorly on textual documents with rich linguistic structures.

Some works[36, 37] investigated relationships between text clauses to better calculate a sentiment score. Specifically, they hypothesized that the relationships could boost sentiment classification accuracy by appropriately aggregating sentiments across interrelated clauses. Rhetorical Structure Theory (RST) [38], which defines several types of clause relations, has been adopted in many studies including ours [39]. In the sentence-level sentiment analysis method presented by Chenlo et al. [40], RST was applied to fragment a sentence into a nucleus part and a satellite part. After calculating polarity scores from the two parts separately, the resulting scores were weighted according to the relationship between the nucleus and satellite, and are then aggregated.

Recently, some researchers proposed hybrid methods, combining the machine-learning and the lexicon-based approaches with linguistic structural information. Chenlo and Losada [41] extended their lexicon-based approach [40] by treating extracted RST relations as features for SVM and Logistic Regression for sentiment classification.

In this dissertation, we incorporates the hierarchical structure of an RST tree. The structure is used for aggregating polarity scores of EDUs that are relevant to a given aspect. More issues about the use of RST structure are investigated, i.e., (1) different contributions of each part in a RST tree to the polarity score of a relevant segment,

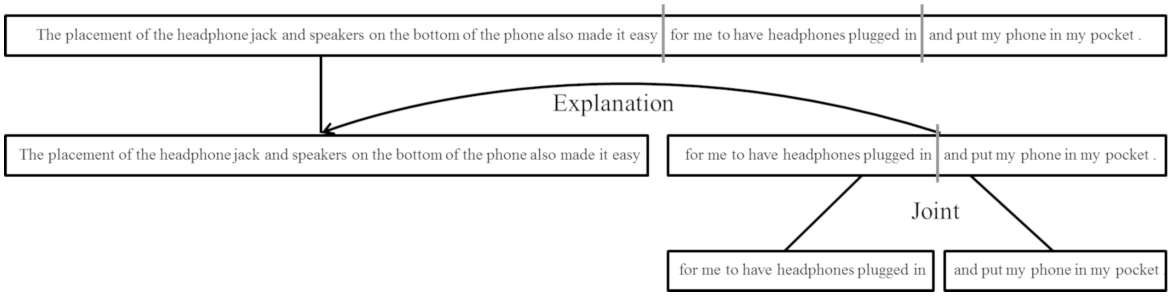


Figure 2.1: An example of an RST tree

and (2) aggregation methods for determining the overall polarity of an aspect.

2.2 Rhetorical Structure Theory in Sentiment Analysis

The rhetorical structure theory (RST) [38] is a linguistic theory that defines relations between phrases or clauses in a document. The relation usually connects two parts, each of which is either a nucleus or a satellite. A nucleus is the main part or the core of an entire given text portion, while a satellite is a complement of the nucleus. Most relation types connect a satellite to a nucleus, while some of them connect two nuclei together, e.g., joint and contrast relations. An elementary discourse unit (EDU) is a leaf node of an RST tree; it is the smallest unit in RST, representing a phrase or a clause. An example of an RST is given in Figure 2.1, which is generated from the sentence "The placement of the headphone jack and speakers on the bottom of the phone also made it easy for me to have headphones plugged in and put my phone in my pocket." Due to its ability to identify important parts of text, RST is applied to analyse the sentiment of a document in several research works, which are summarized in Table 2.1.

The rhetorical structures were used to expand parts of content that are related to a key EDU, i.e., an EDU that contains a keyword of an aspect [39]. The expanded group of EDUs is called a *local aspect segment*. The average polarity score of all EDUs in a local aspect segment often fails to represent the overall polarity of the segment. For example, consider a local aspect segment consisting of the two clauses "On the

hardware front I was not sure about 5-inch screen at first," and "but I wanted that special antenna and the stereo speakers," connected by a "contrast" relation, which is relevant to the "sound" aspect. Its nucleus is the EDU containing the second clause, with a positive polarity score of 0.029 (calculated based on SentiWordNet [35]). Its satellite part is the first clause, with a negative polarity score of -0.030. The average value of these polarity scores is negative, whereas the actual polarity concerning the "sound" aspect of this local aspect segment should be positive.

As studied in recent related works, e.g., [42, 40, 41, 43, 44] in Table 2.1, RST is influential in identifying parts of content that are relevant to the overall sentiment of a document. Different RST components may differently affect how the polarity scores of the EDUs in a local aspect segment should be aggregated. With RST components, we expect that a classification rule such as $((relation = contrast) \wedge (nucleus = negative) \wedge (satellite = positive)) \Rightarrow (segment = negative)$ could be useful for more accurate sentiment prediction.

2.3 Clinical Text from MIMIC II Database

Clinical text in this research is from MIMIC II database [19]. The MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) database is a open resource to do research on realistic patient data. It contains data of more than ten-thousand intensive care unit (ICU) patients from hospital medical information systems. Personal data, e.g., patient name, and hospital data, e.g., hospital name, location, or date of admission, are de-identified to preserve a privacy of the patient.

Clinical text from MIMIC II database is corresponding to a visit of a patient. One visit record may contain more than one clinical narrative of four types as follows.

1. **Radiology report** is a report written by a radiologist as shown in Figure 2.2a. It contains a description regarding results from radiology examination of a patient. Suggestions or comments of the radiologist are also written in this report.

2. **Discharge summary** records all activities about a patient from admission to discharge. Some medical information of the patient, e.g., allergies, are included

INDICATION: Cough, shortness of breath, rule out infiltrate or effusions.

FINDINGS: Chest PA and lateral radiographs are compared with the PA and lateral study dated [**2562-3-16**]. There is bibasilar collapse/consolidation which is new from the prior examination. There is prominence of the central pulmonary vasculature with upper zone redistribution. There are small bilateral pleural effusions. Areas of discoid atelectasis are also appreciated in both lung bases. The patient is somewhat rotated. No evidence for pneumothorax. The osseous and soft-tissue structures are unremarkable.

(a) Radiology report

PHYSICAL EXAMINATION: Examination on admission revealed the following: Temperature 96.3, pulse 88, blood pressure115/78, respiratory rate 18, oxygen saturation 98% on room air. GENERAL: The patient was an elderly woman lying in bed in no acute distress. She was very irritable. HEENT: Revealed dry mucous membranes, oropharynx clear. There was slight icterus of the sclera. There was no lymphadenopathy. Extraocular muscles are intact. Pupils were equal, round, and reactive. There was jugulovenous distention to the level of the jaw. Lungs were clear to auscultation with decreased breath sounds and egophony at the right base. Heart was regular rate and rhythm, no murmur or rub was appreciated. Abdomen was soft, nontender, and nondistended with good bowel sounds. Extremities showed 2+ edema on the right and 3+edema on the left. Neurological examination revealed the Cranial nerves II to XII grossly intact. Strength was [**6-10**] in all muscle groups tested. Coordination was intact.

LABORATORY DATA: On admission the laboratory data revealed the white count of 9.0 with a differential of 25 neutrophils,19 lymphocytes, 4 monocytes, 1.3 eosinophils. The hematocrit was 41 and the platelet count was 184. Chem 7 revealed the sodium of 140, potassium 4.0, chloride of 100, bicarbonate22, BUN 32, creatinine 1.5, glucose 134. The creatinine of1.5 is double her baseline of 0.7 in [**2562-3-9**]. The CK was 434, with MB index of 3. Troponin was less than 0.3.The ionized calcium was 0.44. Albumin was 40. Total calcium5.3, ALT 288, AST 275, total bilirubin 0.7.

(b) Discharge summary

I have reviewed history and examined infant. 35 week male with respiratory distress admitted for further evaluation.

2210 gram 35 week male born to a 24 yo multip with negative PNS except GBS unknown. Pregnancy c/b PL. Intrapartum abx given > 4 hrs ptd. ROM 7 hrs ptd. Vaginal delivery. Apgars [**7-9**].

Mild to moderate respiratory distress prompted initiation of prong bubble CPAP. Currently on CPAP 67 28%.

Exam Premature male with mod retractions and intermittent gruntingAF soft, flat, nondysmorphic, poor aeration, moderate retractions, no murmur, normal pulses, soft abd, 3 vessel cord, no hsm, normal male, testes in upper scrotum, patent anus, no hip click, no sacral dimple, + Mongolian spot, normal tone

(c) Medical doctor note

Please see carevue for all objective data. Please see FHPA for list of multiple allergies and complicated PMH. Essentially, this is a very ill appearing 57 year old woman admitted from the OR. She had an elective parathyroidectomy done and the procedure had been completed. She developed VEA (bigeminy), which progressed to VT. CPR was initiated and NSR resumed without any medications, then started on amiodarone.(?QTC prolongation at baseline, ?VEA with r on t phenomenon causing polymorphic VT). Pt transferred to [**Wardname 1626**] for monitoring and electrolyte replacement.

CVS: Heart rate has been SB without VEA since admission with B/P 90-120/systolic. Amiodarone dc'ed. Pt. is being ruled out for an MI...ECHO done with results pending. K+=4.2, Mg=1.9 (prior to replacement with 4grams), ionized Ca of 1.08 (prior to replacement of 2 Grams). Repeat bloods will be sent at 1830.

(d) Nursing/others

Figure 2.2: Example part of clinical text from MIMIC II database

Table 2.1: Related works with application of RST to sentiment analysis

Author(s)	Granularity	Dataset	RST parser	RST usage	Relation type usage?
Hogenboom et al. [42]	Document / paragraph / sentence	Movie reviews	SPADE (sentence-level parsing of discourse), HILDA (High-level discourse analyser)	Use RST to select text segments relevant to the overall sentiment of each granularity	No
Chenlo et al. [40]	Document / sentence	BLOGS06 (blog posts), MOAT (news articles), FSD (product reviews)	SPADE	Use the topmost relation type of a document to adjust how the polarity of a satellite part affects the overall sentiment.	Yes
Chenlo and Losada [41]	Sentence	MOAT (news articles), FSD (product reviews), MPQA (news articles)	SPADE	Use relation types as features for machine-learning methods	Yes
Wachsmuth et al. [43]	Document	Movie reviews, ArguAna TripAdivor (hotel reviews)	Lightweight lexicon-based discourse relation extractor (self-developed)	Extract sentiment flow patterns and use them as features for a machine-learning method (support vector machine)	Yes
Wang and Wu [44]	Document	Product reviews written in Chinese with 10 product domains	A self-developed parser from their previous work	Use RST tree structures to aggregate the polarity score of a document	Yes

in the summary. Discharge summary may be departed into many sections, e.g., history of present illness, physical examination, brief hospital course, and etc. Some text part of the summary is shown in Figure 2.2b.

3. **Medical doctor note** is a text given by the physician who takes care of a patient. All observed issues when the doctor visits are recorded in the note as shown in Figure 2.2c. The doctor notes do not appear in every medical records from MIMIC II. More than half of them do not contain the note.

4. **Nursing/others** clinical text is the most occurring type in one medical record, according to a number of visits of nurses. Nursing clinical text or nurse narrative records observations and opinions about patient health status by nurses or other medical staffs. All activities of treatment and also vital signs of a patient are records in this text. The text is useful for the medical staffs to know what

happened in the previous shifts. The example of text is shown in Figure 2.2d.

As seen in Figure 2.2, clinical text is different from general text, e.g., product reviews. For example, sentiment in clinical text is not expressed directly. Considering the sentence "On admission the laboratory data revealed the white count of 9.0 with a differential of 25 neutrophils, 19 lymphocytes, 4 monocytes, 1.3 eosinophils." from discharge summary (Figure 2.2b), this sentence gives details about patient blood status on admission. However sentiment of the sentence, i.e., positive(normal) or negative(abnormal), cannot be known if the standard values of blood components, e.g., "lymphocyte" and "eosinophil", are not provided. Some issues made sentiment analysis difficult are listed as follows.

- **Requirement of standard value range**: Many parts of clinical text, especially records of vital signs or laboratory examination results, contain numeric values. To assess sentiment of text containing numeric values, the standard ranges of value is required. For example, consider the sentence "The hematocrit was *41* and the platelet count was *184*." in discharge summary (Figure 2.2b). The standard amount ranges of "hematocrit" and "platelet" are required to analyse status of patient.

- **Usage of many abbreviations which are ambiguous**: Abbreviation is a key characteristic of an informal writing. In clinical text, abbreviations are usually used in note-type parts, i.e., medical doctor notes and nurse narratives, to save the time when the medical staffs record them. Some abbreviated terms are not used commonly both in general and in the same domain. To get correct meaning for text analysis, abbreviation ambiguity problems have to be solved. Considers example sentences from nurse narrative (Figure 2.2d) "Heart rate has been *SB* without *VEA* since admission with *B/P* 90-120/systolic." and "*Pt.* is being ruled out for an *MI...ECHO* done with results pending.", abbreviated terms are italic.

- **Different meaning of general terms in specific domain**: Some terms have different meanings or attitudes in different context or domain of interest. Considers the sentence "2210 gram 35 week male born to a 24 yo multip with *negative*

PNS except GBS unknown" from medical doctor note (Figure 2.2c), term "negative" may not make the sentence negative, according to that the term is a result of which laboratory test.

- **Usage of domain-specific technical terms**: Domain-specific term is a main problem when do sentiment analysis on other domains. Some terms contain implicit senses, e.g., disease names are considered negative in medical domain. General-purpose lexicon do not cover the sentiment of those terms. For example, in the sentence "There was *jugulovenous distention* to the level of the jaw" from discharge summary (Figure 2.2b), if a semantic type of the term "jugulovenous distention" is not known, the sentiment of this sentence is neutral.

This dissertation focuses on the last mentioned issue since the domain knowledge for determining semantic type of terms is already existed, i.e., UMLS. The Unified Medical Language System or UMLS[1] is a corpus containing details of vocabularies in biomedical science. Terms is UMLS is annotated with their semantic types, e.g., Body Location or Region (blor), or Disease or Syndrome (dsyn). There are 133 semantic types in total, categorized into 15 groups, i.e., Activities and Behaviors (ACTI), Anatomy (ANAT), Chemicals and Drugs (CHEM), Concepts and Ideas (CONC), Devices (DEVI), Disorders (DISO), Genes and Molecular Sequences (GENE), Geographic Areas (GEOG), Living Beings (LIVB), Objects (OBJC), Occupations (OCCU), Organizations (ORGA), Phenomena (PHEN), Physiology (PHYS) and Procedures (PROC).

2.4 Sentiment Analysis in Medical Domain Using UMLS

UMLS was used for lexicon-based sentiment analysis in [45] and [46]. In [45], Na et al. proposed a rule-based linguistic method for sentiment classification and applied it to a dataset containing 1,000 clauses extracted from drug reviews. Each classification rule was manually defined considering grammatical relations, part-of-speech, and polarity scores of terms in a clause. The polarity score of a term was obtained from a generic

[1]https://www.nlm.nih.gov/research/umls/

lexicon (9,630 terms) and a domain-specific lexicon (10 terms). To compensate for the small domain-specific lexicon, disorder terms in a clause were identified using the UMLS Disorders semantic group and the polarity score of -1 was assigned to them. Compared to polarity classification using Support Vector Machine (SVM) with the bag-of-word and negation features, the proposed rule-based method improved the classification accuracy by 9%.

Asghar et al. [46] generated a health-related sentiment lexicon, called SentiHealth. A list of terms was first constructed by expanding an initial seed list of health-related words over a set of web repositories using Boot-strapping. UMLS was then employed for checking whether a term in the list was a valid medical term. A polarity score of each medical term in the list was calculated by applying a probability-based scoring method to a dataset consisting of 8,230 patient-authored drug reviews, collected from on-line forums. To assess the effectiveness of SentiHealth, the Vote & Flip algorithm [47] was applied on a dataset consisting of 17,830 drug reviews. Compared to three lexicons generated by Delta Scoring [48], Lexicon-based + Information Gain [49], and Revised Mutual Information [50], SentiHealth improved the classification accuracy between 9% and 23%.

UMLS has also been used to generate features for machine-learning-based sentiment analysis, e.g., in [11, 13, 51]. In [11], Niu et al. performed polarity classification using Support Vector Machine (SVM) with five feature sets, i.e., unigrams, bigrams, change phrases, negations, and categories. Each feature in the 'categories' set was the number of occurrences of one UMLS semantic type (e.g., *dsyn* or *patf*). Experiments were conducted to classify 1,509 sentences, obtained from articles in Clinical Evidence,[2] into four classes, i.e., positive, negative, neutral, and no outcome. Compared to the use of the first four feature sets alone, the incorporation of the 'categories' features decreased the error rate from 21.38% to 20.58%.

In [13], medical terms occurring in medical article abstracts were replaced with their UMLS semantic types. Nine semantic types from the Disorders group, i.e., acquired abnormality (*acab*), anatomical abnormality (*anab*), cell or molecular dysfunction(*comd*), congenital abnormality (*cgab*), disease or syndrome (*dsyn*), injury or poisoning (*inpo*),

[2]https://www.bmj.com/specialties/clinical-evidence

mental or behavioural dysfunction (*mobd*), neoplastic process (*neop*), and pathologic function (*patf*), and one type from the Living Beings group, i.e., virus (*virs*), were considered in the replacement. Four classification methods, i.e., Bayesian Net (BN), Naïve Bayes (NB), SVM, and C4.5 Decision Tree (C4.5), were applied to classify 520 medical article abstracts, collected from Journal of Family Practice[3] (JFP) and MEDLINE,[4] into three classes, i.e., positive, negative, and no outcome. Using unigrams with the replacement of medical terms, the classification accuracies were improved by 0.5% on average.

In [51], the TF-IDF values of UMLS semantic types appearing in a dataset were considered as domain-specific features, called ST features. The ST features were used in combination with content-based features, i.e., bag-of-word, word embeddings, and concept embeddings. Four classification methods, i.e., NB, Random Forest (RF), Sequential Minimal Optimization (SMO), and Vote, were applied to 3,747 patient-authored sentences concerning three diseases, i.e., allergic diseases, Crohn's diseases, and breast cancer, collected from the MedHelp health site[5]. Compared to the use of only content-based features, the incorporation of ST features improved the classification accuracies by 2.4% on average.

Table 2.2 summarizes the related works described above. Compared to these works, this study focuses more on how to assign an appropriate polarity score to an individual semantic type in the Disorders group, and also how to use information about semantic types in an input sentence to select an appropriate classification method. In terms of document types, our study considers sentences taken from clinical narratives, which contain more medical terms and involve less subjective use of language, compared to patient-authored documents (considered in [45, 46, 51]) and well-written documents (considered in [11, 13]).

[3]https://www.mdedge.com/familymedicine

[4]https://www.nlm.nih.gov/bsd/medline.html

[5]https://www.medhelp.org

Table 2.2: Related works with exploitation of UMLS

Publication	Classification Method	Dataset	Data Source	Uses of UMLS
Na et al. [45]	Rule-based	1,000 clauses	On-line drug reviews	Tag disorder terms and set them to polarity score of -1
Asghar et al. [46]	Vote-switch (lexicon-based)	26,060 drug reviews	On-line forums and public dataset	Check whether a term listed in candidate lexicon is a medical term
Niu et al. [11]	SVM	1,509 sentences	Clinical Evidence, publication	Count the number of occurrences of each UMLS semantic type
Sarker et al. [13]	BN, C4.5, NB, SVM	520 medical article abstracts	JFP and MEDLINE	Replace specific medical terms with their semantic types
Carrillo et al. [51]	NB, RF, SMO, Vote	3,747 sentences	On-line health forum	Compute TF-IDF values of UMLS semantic types

Chapter 3

Aspect-based Sentiment Analysis Exploiting Linguistic Structure

The primary aim of aspect-based sentiment analysis is to extract the sentiment of a specific aspect from an opinion textual document. Generally, text segments relevant to an aspect are first localized. Using lexicon-based approaches, polarity scores from terms in the segments are subsequently combined into an overall score. In this chapter, we propose a new score aggregation strategy that utilizes linguistic structures in several ways. A given textual document is segmented into elementary discourse units (EDUs) with relations between them. Polarity scores for EDUs are then computed from all aspect-related terms, identified using term dependency structure. The EDU scores are hierarchically combined into the scores for their local aspect segments and subsequently into those for aspect segments. Experimental results on on-line product reviews demonstrate that our new score aggregation method and EDU-level score calculation, exploiting term dependencies, improve sentiment classification at the level of local aspect segments.

3.1 Aspect-based Text Segmentation

A review text may contain many aspects, and each aspect may be expressed in many segments that are not necessarily adjacent to each other and may in general have individually independent polarity about the aspect. For example, as shown in Figure 3.1,

the segment "Switching from another operating system" is classified as neutral for the 'system' aspect, while the segment "I love not only the camera, but the Windows Phone operating system" holds a positive polarity for the same aspect. Consequently, to analyse the sentiment of an aspect, sentiment analysis should first be applied to these individual text segments, which are called *local aspect segments*.

> "The complete mobile package, great hardware, apps, more". Switching from another operating system, I wasn't sure what to expect (other than a good camera). I love not only the camera, but the Windows Phone operating system. Wish it were less expensive, but with this kind of hardware, I'm willing to pay for the extras.

Figure 3.1: A sample review showing non-adjacent text segments concerning the 'system' aspect

A dataset consisting of mobile phone reviews collected from CNET[1] is considered in this chapter. Each review is concerned with one or more aspects in a predetermined collection of 13 manually defined aspects, i.e., screen, application, network, system, camera, capacity, power/battery, sensor, accessory, size, hardware/body, sound, and price. An RST parser [52] is employed to segment a textual review into EDUs and connect them using RST relations.

To define a local aspect segment, we introduce the concept of a *key EDU* and that of an *aspect segment*. Given EDUs connected by RST relations, aspects relevant to a review are identified by matching their predetermined keywords with terms in the EDUs. An EDU containing at least one keyword of an aspect is called a *key EDU* for the aspect. The key EDUs for a given aspect and their adjacent EDUs with respect to RST relations are grouped into an *aspect segment* for that aspect.

We define a *local aspect segment* as either (1) a span of RST elements that has a key EDU as its nucleus or (2) a key EDU that appears as a satellite of an RST relation. Figure 3.2 illustrates local aspect segments in a product review, where the squares labelled with 0-9 represent EDUs and the bold-border squares (i.e., the squares with the labels 0, 1, 3, 4 and 5) represent key EDUs. The figure contains six local

[1]http://www.cnet.com/topics/phones/products

Table 3.1: Local aspect segments extracted from the sample review in Figure 3.2

Aspect	**Local aspect segment**	**EDU**
'hardware'	*Hardware*-1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
'application'	*Application*-1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
'camera'	*Camera*-1	2, 3
	Camera-2	4, 5, 6, 7, 8, 9
'system'	*System*-1	1
	System-2	5, 6, 7, 8, 9

aspect segments relevant to four aspects, which are detailed in Table 3.1. The aspect segment for the 'camera' aspect, as well as that for the 'system' aspect, consists of two local aspect segments. Local aspect segments for different aspects may involve the same group of EDUs, e.g., the local aspect segments *Hardware*-1 and *Application*-1 in Table 3.1.

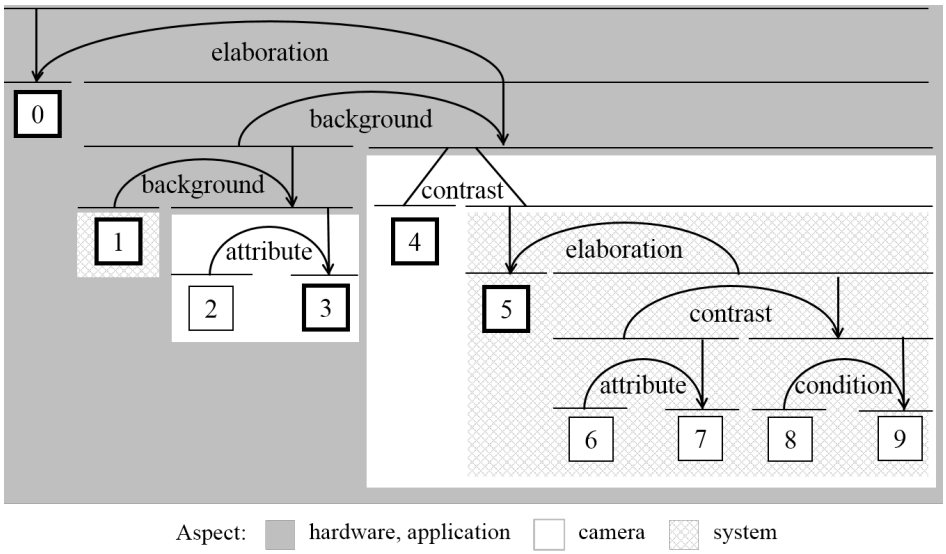


Figure 3.2: An RST relation tree depicting aspect segments and their components

3.2 EDU-level Score Calculation

For automatic calculation of EDU-level polarity scores, a lexicon-based method is applied in this study. The method uses SentiWordNet [53] to determine the polarity scores of all individual terms appearing in an EDU and takes their average score as the polarity score of the EDU. The resulting score is flipped to the opposite sign (i.e., negative or positive) if at least one negation term appears in the EDU. More precisely, given an EDU E, let $term(E)$ be the bag of all terms appearing in E and then let the polarity score of the EDU E, denoted by $score(E)$, be defined by

$$score(E) = (-1)^{neg(E)} \cdot \frac{1}{N} \cdot \sum_{t \in term(E)} \Big(score(t) \Big), \tag{3.1}$$

where for any term t, $score(t)$ is the polarity score of t obtained from SentiWordNet, N is the number of terms appearing in E and

$$neg(E) = \begin{cases} 1, & \text{if there is a negation term appearing in } E, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

3.3 Aspect-Based Polarity Score Aggregation Using RST Structure

We propose a method for calculating a sentiment score of a certain aspect by hierarchically aggregating polarity scores from its EDUs. In particular, the aggregation method operates first at the lower level (the level of local aspect segments) and then at the upper level (the level of aspect segments) consecutively. At the level of local aspect segments, two calculation schemes, called the *All-EDU averaging* (*AEA*) scheme and the *Top-RST-level weighted averaging* (*TWA*) scheme, which are described below, are used. At the level of aspect segments, the overall polarity score of an aspect is calculated by averaging all the scores computed from the lower level.

3.3.1 Method Description

The AEA scheme is a simple approach for aggregating the polarity scores from EDUs in a local aspect segment. It simply takes the average value of polarity scores and ignores the linguistic structure between individual EDUs.

To exploit the RST structure of a local aspect segment, the TWA scheme considers the topmost RST relation in a local segment. Specifically, when a local aspect segment contains multiple EDUs, we first divide them into two parts, a nucleus and a satellite according to the RST relation, and determine their individual polarity scores. The score of the nucleus is the score of the key EDU (i.e., the nucleus of the topmost relation). The score of the satellite is the average score of all the remaining EDUs in the local aspect segment. Consider Figure 3.2 and Table 3.1, for example, the polarity score of the nucleus of the *System*-1 segment is the score of EDU 5, and the score of the satellite is obtained by averaging the scores of EDUs 6, 7, 8 and 9. Let L be a given local aspect segment. Let $score_{nuc}(L)$ denote the score of the nucleus part of L and $score_{sat}(L)$ denote that of the satellite part. Using a predefined weight w of the satellite part, the score of L, denoted by $score(L)$, is calculated as follows:

$$score(L) = \frac{score_{nuc}(L) + w \cdot score_{sat}(L)}{1 + |w|} \tag{3.3}$$

In this study, a predefined satellite weight is chosen from the set -2, -1, -0.5, 0, 0.5, 1, 2, regardless of RST relation types. A satellite weight is utilized for describing the contribution strength of the polarity score of a satellite part to the polarity of its local aspect segment as depicted in Table 3.2. The negative weights (i.e., -2, -1, and -0.5) are used if the polarity of a satellite is opposite to the polarity of the nucleus part.

3.3.2 Experiments and Results

After describing our dataset and experimental settings, we compared the aggregation methods, which combine all relevant EDU scores at both the lower and higher levels (the local aspect segment level and the aspect level).

Table 3.2: The meaning of each predefined weight of a satellite

Weight value	**Meaning (effect to a local aspect segment)**
0	No effect
0.5	Effect less than the nucleus part
1	Effect equal to the nucleus part
2	Effect more than the nucleus part

Dataset

The approaches have been evaluated on an on-line product review dataset containing 465 aspect segments, with a total of 834 local aspect segments and 2,545 EDUs. Separated by 3 polarity classes, the dataset contains 498, 156 and 180 positive, neutral and negative local aspect segments respectively; and 315, 45 and 105 positive, neutral and negative aspect segments, respectively. To investigate score aggregation at the level of local aspect segments and that of aspect segments, three human annotators manually labelled each individual local aspect segment and each individual aspect in the dataset by selecting polarity scores from the choices -1, -0.5, 0, 0.5 and 1, which denote 'strongly negative', 'weakly negative', 'neutral', 'weakly positive' and 'strongly positive' respectively. The average annotated score was taken as its actual label.

Results

We evaluated the methods for aggregating the polarity scores at the level of local aspect segments, i.e., the AEA and TWA schemes. The scores of EDUs in the local aspect segments are obtained from the EDU scoring method in Section 3.2. Table 3.3 shows the evaluation results.

Considering score aggregation schemes, TWA achieves higher accuracy than AEA at certain values of the satellite weight. The results demonstrate that the employment of RST linguistic structure to separately consider a nucleus and its related satellite part yields better results, compared to ignoring it.

At the aspect level, we simply take the average score of all local aspect segments

Table 3.3: Performance comparison of the score aggregation methods, AEA and TWA, on 834 local aspect segments

Score aggregation method	**Number of correctly classified local aspect segments**
AEA	492 (59.0%)
TWA	
weight = 2	490 (58.8%)
weight = 1	493 (59.1%)
weight = 0.5	**499 (59.8%)**
weight = 0	496 (59.5%)
weight = -0.5	468 (56.1%)
weight = -1	441 (52.9%)
weight = -2	419 (50.2%)

related to a given aspect as the polarity score of the aspect. Table 3.4 shows the evaluation results at the aspect level using different aggregation schemes at local aspect segments, i.e., AEA and TWA.

The results obtained when the scores of local aspect segments are aggregated by AEA (cf. the first row of Table 3.4) are slightly higher than those obtained when TWA is used. When the result obtained using a certain experimental setting at the local aspect level (cf. Table 3.3) is better than the results obtained using other settings, the result at the higher level (the level of aspect segments) is not necessarily better than the rest using the same experimental setting. Possible reasons may be (1) inappropriateness of taking the average polarity score of all local aspect segments as the score aggregation result (under the assumption that all local aspect segments similarly contribute to the overall score), or (2) incorrectness of individual scores at the level of local aspect segments.

To validate whether the overall polarity of an aspect can potentially be averaged from its local aspect segments, the manually annotated scores of the local aspect segments are used to compute the score of the aspect. The result shows that 445 aspect segments (i.e., 95.7% of all aspect segments) are correctly classified. Consequently, the

Table 3.4: Performance comparison of the score aggregation methods, AEA and TWA, on 465 aspect segments

Score aggregation method	**Number of correctly classified aspect segments**
AEA	**325 (69.7%)**
TWA	
weight = 2	318 (68.4%)
weight = 1	323 (69.5%)
weight = 0.5	323 (69.5%)
weight = 0	315 (67.7%)
weight = -0.5	294 (63.2%)
weight = -1	273 (58.7%)
weight = -2	253 (54.4%)

actual cause of inconsistency between the results in Table 3.3 and those in Table 3.4 is the classification incorrectness at the level of local aspect segments, which is outside the scope of this work.

3.4 Rule-Based Polarity Aggregation Using Rhetorical Structure

In this study, topmost RST structures are taken into consideration for sentiment classification. After local aspect segments for a certain aspect are extracted and the polarity scores of all EDUs are calculated, local aspect segments are transformed to feature vectors representing their topmost structures. These feature vectors are used for training a classification model. Although several classification algorithms are applicable, we apply rule-based algorithms, with an expectation that how rhetorical relation types are utilized can be clarified in a human readable manner through the content of the resulting classification rules.

3.4.1 Feature Vector Transformation

Figure 3.3 shows the number of EDUs in a local aspect segment extracted from dataset of product reviews that was used in this section. Less than 25 percent of the local aspect segments contain three or more EDUs, i.e., most local aspect segments have only topmost RST relations. Deeper relations are thus neglected since they may not have much effect on the dataset.

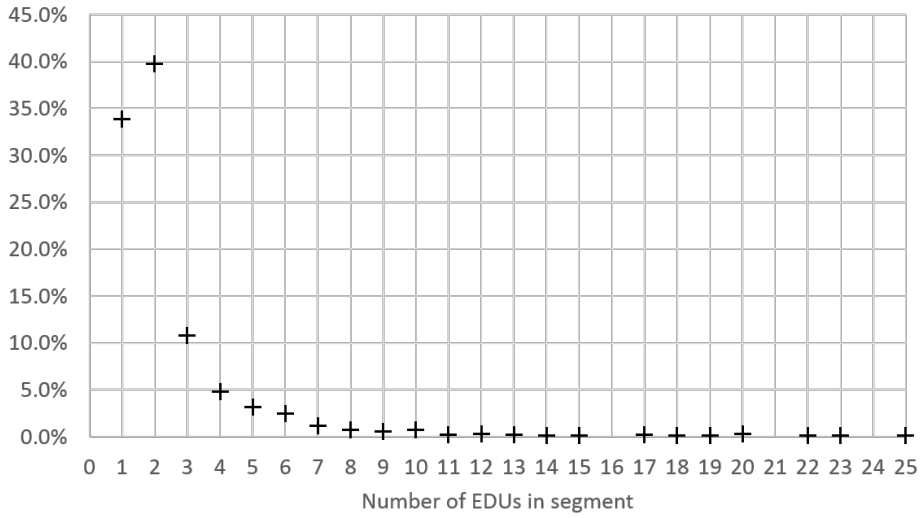


Figure 3.3: Distribution of number of EDUs in aspect-based segments from product reviews

Before applying a rule induction algorithm to the local aspect segments, they were transformed into feature vectors containing the topmost relations of their key EDUs. To compare the effect of RST relation types, two types of feature vectors are considered. A vector of the first type, referred to as a NRT vector (a vector "with no relation type"). Given a local aspect segment E, the NRT vector of E is constructed as follows:

1. If E consists only of a key EDU, say K, and its satellite EDU, say S, with no other EDU, then the NRT vector of E is $[P_K, P_S]$, where P_K and P_S are the polarity of K and that of S, respectively.

2. If E consists only of a key EDU, say K, with no other EDU, then the NRT vector of E is [outside, P_K], where P_K is the polarity of K. In this case, K is always a satellite EDU. The term 'outside' is used to indicate that the nucleus

of the RST relation connected with K is outside E, and is ignored for sentiment consideration.

3. If E consists of a key EDU, say K, and a subtree, say ST, as the satellite of its topmost RST relation, then the NRT vector of E is $[P_K, P_{ST}]$, where P_K is the polarity of K and P_{ST} represents the average polarity of the EDUs in ST. P_{ST} takes the values 'positiveTree', 'neutralTree', and 'negativeTree' when the average polarity of the EDUs in ST is positive, neutral, and negative, respectively.

A vector of the second type, referred to as a WRT vector (a vector "with a relation type"), extends that of the first type by including the topmost relation type of a local aspect segment as an additional feature.

To illustrate, assume that

- E_1 is a local aspect segment that consists of two EDUs, which are a nucleus (key) EDU with positive polarity and a satellite EDU with negative polarity, connected by an "elaboration" relation,
- E_2 is the left local aspect segment in Fig. 3.4, consisting solely of a key EDU with negative polarity, which is a satellite EDU of a "background" relation, and
- E_3 is the right local aspect segment in Fig. 3.4, containing a nucleus (key) EDU with neutral polarity and a satellite subtree of a "condition" relation with the average polarity score being positive.

The NRT feature vectors and the WRT feature vectors obtained from E_1, E_2, and E_3 are shown in Table 3.5.

3.4.2 Classification Rule Induction

Two rule induction methods, PRISM and PART, are used in this study. They represent two major rule generation paradigms, i.e., rule generation based on sequential covering and that based on a decision tree. WEKA API [54] is used for our implementation.

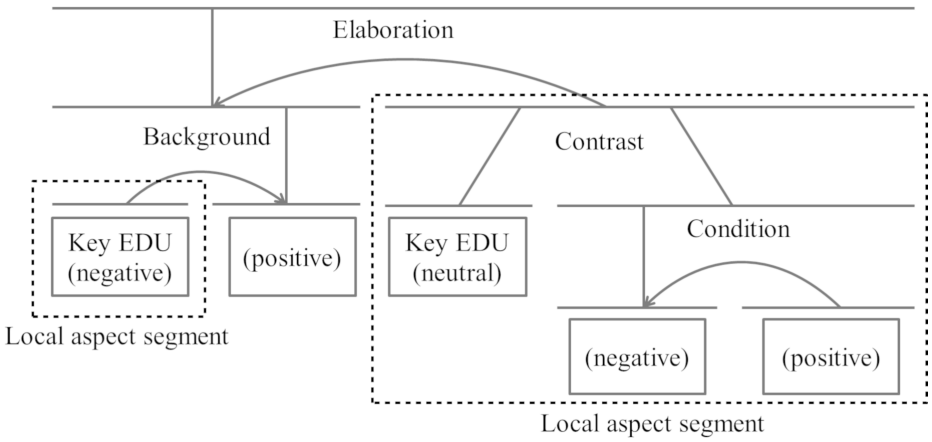


Figure 3.4: Examples of local aspect segments expanded from key EDUs.

Table 3.5: Examples of feature vectors

Example	Feature vector	
	NRT	WRT
E_1	[positive, negative]	[elaboration, positive, negative]
E_2	[outside, negative]	[background, outside, negative]
E_3	[neutral, positiveTree]	[contrast, neutral, positiveTree]

PRISM

PRISM [55] employs a sequential covering technique. It produces rules to cover all training instances of each classification class (e.g., positive polarity or negative polarity) separately, one by one. When dealing with a class, PRISM starts by considering rules with only one attribute-value pair in their conditions, e.g., (*relation* = *elaboration*) or (*nucleus* = *positive*). The rule with the highest confidence is first considered. If its confidence value is equal to 1.0, the rule is included in the resulting rule set. If its confidence value is less than 1.0, a set of new rule candidates is generated by adding another attribute-value pair into the condition part. A new rule candidate is selected and proceeded by the same criteria, but only the group of instances that satisfy the previous condition are considered for determining the confidence value. After the selected rule candidate is added to the rule set, all instances that satisfy the

candidate rule are removed. The process is repeated to produce another new rule until all instances of the class being considered are covered. The training set is then restored and the process is repeated for another class.

PART

PART [56] constructs a decision tree from training data, and uses the resulting tree to generate classification rules. The main difference between PART and other rule induction methods based on decision trees, e.g., C4.5 [57] and RIPPER [58], is that PART creates a "partial" tree and generates only one rule at a time. After a rule has been generated, the instances covered by the rule are removed from the training set, and the generation process is continued to produce another rule. To avoid a situation in which some significant rules with low coverage are swallowed by rules with higher coverage during a rule prunning process, PART does not perform global optimization on a rule set.

3.4.3 Experiments and Results

Datasets

In this study, experiments are conducted on two datasets of customers' product reviews, which are described below.

- *CNET Mobile Phone Reviews*:

  The first dataset, referred to as D1, consists of 139 reviews concerning mobile phone models from CNET.com[2]. 664 local aspect segments were extracted and manually annotated with positive or negative labels, resulting in 492 positive and 172 negative local aspect segments.

- *Liu Bing's Dataset*:

  The second dataset, referred to as D2, contains 640 product reviews from Amazon.com[3] and CNET.com. It was collected by Liu Bing's research group and was

[2]https://www.cnet.com/

[3]https://www.amazon.com/

used in [4, 59, 60]. The product reviews in D2 belong to six domains, e.g., cameras, media players, mobile phones, networking devices, software and miscellanea. Sentences in these reviews were originally annotated with aspect keywords and their polarity scores. To analyse them at the level of local aspect segments, we manually grouped all annotated keywords into aspects and formed local aspect segments for these aspects using rhetorical structures. 2,638 local aspect segments were obtained, 1,807 and 831 of which are positive and negative, respectively.

Experimental Settings

- *Discourse Relation Type Exploitation*:

  All experiments are conducted separately for each feature vector type (cf. Section 3.4.1) to study the influence of RST relation types on classification rule induction. The NRT type represents the setting in which the information about RST relation types is not used. The WRT type takes RST relation types into the consideration.

- *Evaluation Schemes and Baseline Methods*:

  Two evaluation schemes are used in our experiments.

  - *10-fold cross validation*: This scheme applies 10-fold cross validation on each dataset. When the cross validation is applied to the D1 and D2 datasets, the scheme is referred to as D1-CV and D2-CV, respectively. The 10-fold cross validation is also applied on the mixture of the D1 and D2 datasets and is referred as D3-CV.
  - *Cross-dataset rule application*: In this scheme, one dataset is used to generate a set of rules and the other dataset is then used to test the obtained rules. The scheme that applies the rules learned from D2 to D1 is referred to as D1-XR. On the other hand, when the rules learned from D1 are applied to D2, the scheme is referred to as D2-XR.

  The following two score aggregation methods for local aspect segments are used as baselines in this study:

- *Baseline-I*: Only the polarity score of a key EDU is used to determine the polarity of a local aspect segment.
- *Baseline-II*: The average value of the polarity scores of all EDUs in a local aspect segment is used to determine the polarity of the segment.

These two methods were employed for score aggregation in [39]. Neither of them employs RST relation information.

- *Rule Ordering*:

  The PRISM method produces rules based on a predetermined class ordering. When applied to our datasets, it produces rules for the positive polarity class before producing those for the negative one. The resulting rules are applied in the same ordering. Consequently, when two applicable rules are in conflict, a data instance will be classified depending on the class ordering; for example, when two rules with the same condition but different polarity predictions are both applicable to a local aspect segment, the segment will be classified as positive. To prevent unfair class polarity setting, two measurements, confidence and coverage, are used for rule ordering in this study. The first measurement indicates the precision of a rule on training instances. The second one indicates the proportion of training instances covered by a rule to those belonging to the class predicted by the rule. Let r be a rule $(P \Rightarrow Q)$. The confidence of r and the coverage of r, denoted by $conf(r)$ and $cov(r)$, respectively, are defined by

  - $conf(r) = n(P \wedge Q)/n(P)$, and
  - $cov(r) = n(P \wedge Q)/n(Q)$,

  where for any given condition C, $n(C)$ denotes the number of training instances that satisfy the condition C.

Results

Table 3.6 shows the resulting f-measure for the positive and negative polarity classes (f_{pos} and f_{neg}) and the accuracy obtained from each experimental setting, where the datasets and evaluation schemes (cf. Section 3.4.3) are shown as columns and the

Table 3.6: Accuracy (acc) and f-measure for the positive and negative polarity classes (f_{pos} and f_{neg}) when applying the proposed rule-based methods

Feature	Method	D1-CV			D2-CV			D3-CV			D1-XR			D2-XR		
vector type		f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc
-	Baseline-I	.807	.517	.714	.772	.544	.664	.779	.539	.674	.807	.517	.714	.772	.544	.664
	Baseline-II	.811	.463	.717	.775	.502	.669	.783	.495	.678	.811	.463	.717	.775	.502	.669
NRT	PRISM															
	default	.849	n/a	.738	.813	n/a	.685	.821	n/a	.696	.851	n/a	.741	.813	n/a	.685
	confidence-based	.837	.358	.739	.813	.525	.732	.820	.488	.733	.836	.500	.753	.825	.491	.740
	coverage-based	.789	.478	.699	.795	.560	.721	.802	.550	.725	.821	.499	.736	.793	.534	.713
	PART	.847	.446	.761	.823	.504	.740	.829	.493	.744	.851	.467	.767	.824	.510	.741
WRT	PRISM															
	default	.834	.042	.708	.811	.066	.683	.819	.037	.694	.849	.065	.739	.802	.102	.667
	confidence-based	.818	.342	.706	.810	.480	.719	.812	.451	.718	.841	.451	.753	.795	.324	.676
	coverage-based	.769	.512	.678	.761	.534	.682	.758	.527	.678	.798	.520	.715	.720	.515	.636
	PART	.825	.386	.727	.814	.507	.730	.826	.489	.741	.855	.457	.771	.794	.291	.680

classification methods are divided into three groups by rows. The first group shows the performance of the two baseline methods, Baseline-I and Baseline-II. The second and the third groups show the performance of the rule-based classification methods when NRT and WRT feature vectors, respectively, are used. Each of them consists of the results obtained using PRISM and PART. The results obtained from PRISM are divided into three rows, with different rule ordering schemes, where "default" denotes the original rule ordering sequence produced by PRISM (starting with the rules for the positive class, followed by those for the negative class), and "confidence-based" and "coverage-based" denote the descending sequences of rules ordered by confidence values and coverage values, respectively (cf. Section 3.4.3).

In terms of accuracy, when NRT vectors are used, both PRISM and PART improve the classification performance, compared to the two baseline methods, on all evaluation schemes, except for the case when PRISM is applied on D1-CV with the coverage-based rule ordering. Using WRT vectors, both of them also improve the classification performance on most schemes, except for the application of PRISM to D1-CV. PART yields slightly higher accuracy compared to PRISM with its best setting, i.e., PRISM with the confidence-based rule ordering. Compared to Baseline-II, PART improves

Table 3.7: Top three rules, ranked by confidence and then by coverage, obtained from D1-XR using PRISM and PART

Feature vector type	Rule induction	Example rule	Confidence	Coverage
NRT	PRISM	$((nucleus = neutral) \wedge (satellite = negativeTree)) \Rightarrow (segment = positive)$	1.000	0.008
		$((nucleus = neutral) \wedge (satellite = neutralTree)) \Rightarrow (segment = positive)$	1.000	0.008
		$((nucleus = neutral) \wedge (satellite = neutral)) \Rightarrow (segment = positive)$	1.000	0.006
	PART	$(nucleus = positive) \Rightarrow (segment = positive)$	0.831	0.549
		$(satellite = positive) \Rightarrow (segment = positive)$	0.803	0.514
		$(nucleus = negative) \Rightarrow (segment = negative)$	0.518	0.337
WRT	PRISM	$((relation = evalution) \wedge (nucleus = positive) \wedge (satellite = positive)) \Rightarrow (segment = negative)$	1.000	0.023
		$((relation = attribution) \wedge (nucleus = positive) \wedge (satellite = positive)) \Rightarrow (segment = positive)$	1.000	0.018
		$((relation = enablement) \wedge (satellite = neutral)) \Rightarrow (segment = negative)$	1.000	0.017
	PART	$((relation = elaboration) \wedge (nucleus = positive)) \Rightarrow (segment = positive)$	0.908	0.219
		$((relation = joint) \wedge (nucleus = positive)) \Rightarrow (segment = positive)$	0.917	0.134
		$((relation = elaboration) \wedge (satellite = positive)) \Rightarrow (segment = positive)$	0.871	0.248

the accuracy from 71.7% to 76.7% on D1-XR, and improves approximately 4-7% of accuracy on the other schemes. The accuracy values are not improved when WRT feature vectors are used instead of NRT feature vectors. The top three rules, ranked by confidence and then by coverage, produced by PRISM and PART on D1-XR are shown in Table 3.7.

A closer examination on each polarity class reveals that the rule-based classification methods improve the f-measure for the positive polarity class; however, they worsen the f-measure for the negative polarity class. PRISM with the coverage-based rule ordering is the only method with the resulting f-measure for the negative polarity class being comparable to that obtained from the baseline methods. However, its accuracy is lower than PART and PRISM with other schemes. To address the issue of the negative polarity class, recall and precision values for the class are examined. Compared to the baseline methods, the recall values for the negative class are obviously decreased on each rule-based scheme, except for PRISM with the coverage-based rule ordering. Some attempts to remedy the situation are made, i.e., a combination with a baseline method, setting a confidence threshold, or removal of conflicting rules. They are detailed in the next subsection.

Table 3.8: The classification performance obtained from PRISM+BL

Feature vector type	Method	D1-CV			D2-CV			D3-CV			D1-XR			D2-XR		
		f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc
NRT	PRISM+BL															
	default	.849	n/a	.738	.813	n/a	.685	.821	n/a	.696	.851	n/a	.741	.813	n/a	.685
	confidence-based	.837	.358	.739	.813	.525	.732	.820	.488	.733	.836	.500	.753	.825	.491	.740
	coverage-based	.789	.478	.699	.795	.560	.721	.802	.550	.725	.821	.499	.736	.793	.534	.713
WRT	PRISM+BL															
	default	.837	.090	.723	.811	.074	.686	.821	.046	.697	.850	.065	.741	.806	.147	.682
	confidence-based	.822	.368	.721	.811	.483	.723	.814	.455	.721	.842	.451	.755	.799	.353	.691
	coverage-based	.773	.527	.693	.762	.536	.685	.760	.530	.681	.799	.520	.717	.725	.531	.651

3.4.4 Classification Rules with Heuristics

A Combination with a Baseline Method

Unlike PART, PRISM does not produce a complement rule to classify a local aspect segment when no rule is applicable to it. To complement the PRISM rule-based classifier, the Baseline-I method is applied when there is no applicable rule. We refer to this combination as PRISM+BL. The obtained results are shown in Table 3.8. PRISM+BL yields better accuracy compared to PRISM with WRT feature vectors (cf. Table 3.6). When NRT feature vectors are used, both PRISM and PRISM+BL yield the same results.

Setting Rule Confidence Thresholds

The confidence value of a rule indicates its precision on training instances. Rules with low confidence values tend to yield low classification accuracy on test sets. One possible strategy to improve classification performance is to set a confidence threshold for discarding such rules.

We first conduct experiments by varying confidence threshold values in the range of 0.0-1.0 and applied them to the rules obtained from PRISM. Assuming that a threshold value t is used, a local aspect segment E is classified if and only if some rule with the confidence value greater than t is applicable to E. Figure 3.5 shows the average percentage of classified local aspect segments, calculated from all evaluation schemes,

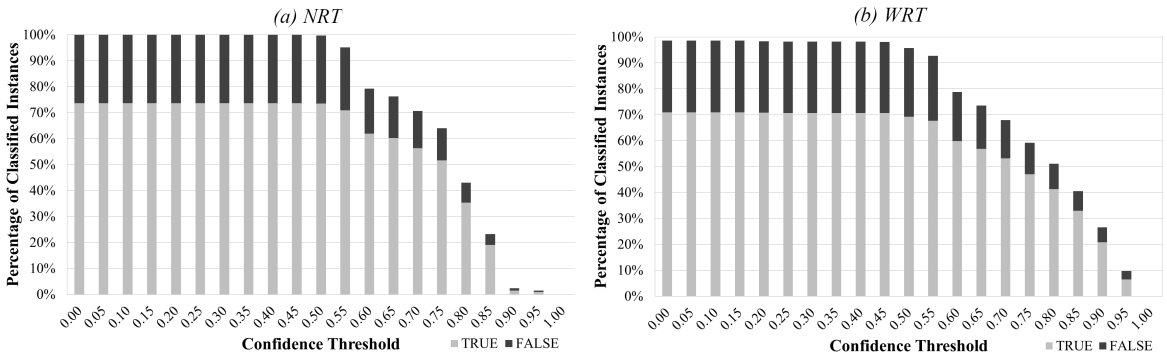


Figure 3.5: Classification results obtained from PRISM using the confidence-based rule ordering at confidence threshold values between 0.0-1.0

at each threshold value. Each bar in the figure is divided into two parts representing the proportion of correctly (TRUE) classified segments to incorrectly (FALSE) classified segments. As the confidence threshold value increases, the proportion of correctly classified segments tends to increase, however, the percentage of classified segments continuously decreases due to the reduction of the overall rule coverage.

To address the rule coverage issue, we next use PRISM+BL, instead of PRISM, in the same experiment setting. Figure 3.6 shows the obtained results. PRISM+BL with confidence threshold setting slightly improves the accuracy and decreses the incorrect-classification rate, compared to that without threshold setting (which can be seen from the bars at the threshold value zero). For WRT feature vectors, the highest accuracy of 72.0% is obtained at the threshold value 0.6, and the incorrect-classification rate is 26.9% at this threshold value. (At the threshold value zero with WRT vectors, the accuracy is 71.6% and the incorrect-classification rate is 28.1%.) For NRT feature vectors, when the threshold value is higher than 0.6, the accuracy slightly decreases, compared to that without threshold setting.

We extend PRISM+BL by setting a threshold value based on the Baseline-I method. More precisely, the threshold value is set to be the minimum confidence value on training instances of the four rules in Figure 3.7, the application of which corresponds to the application of the Baseline-I method. The method obtained by this extension is referred to as PRISM+BL+TH_{BL}. PART is also extended in the same way into PART+BL+$TH_{BL.}$ Table 3.9 shows the results obtained from these extended methods.

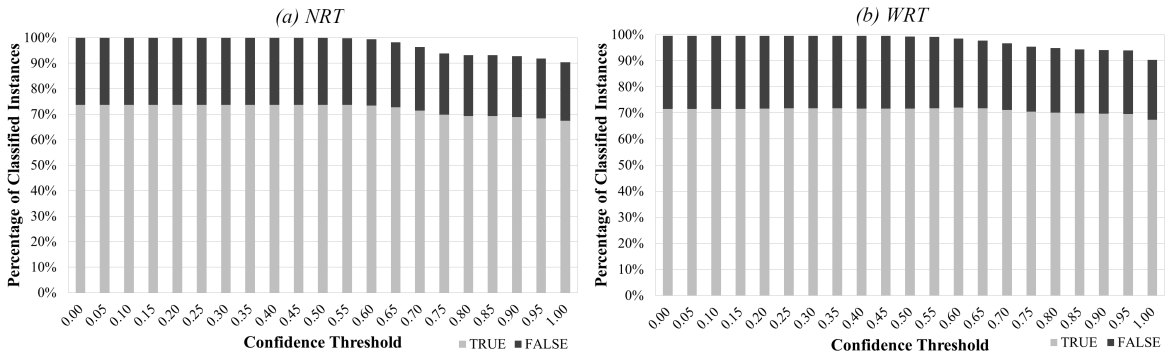


Figure 3.6: Classification results obtained from PRISM+BL using the confidence-based rule ordering at confidence threshold values between 0.0-1.0

$(nucleus = positive) \Rightarrow (segment = positive)$
$((nucleus = outside) \wedge (satellite = positive)) \Rightarrow (segment = positive)$
$(nucleus = negative) \Rightarrow (segment = negative)$
$((nucleus = outside) \wedge (satellite = negative)) \Rightarrow (segment = negative)$

Figure 3.7: Rules corresponding to the Baseline-I method

Compared to PRISM+BL (cf. Table 3.8), PRISM+BL+TH_{BL} improves the classification performance, especially the f-measure values for the negative class on all evaluation schemes. PART+BL+TH_{BL}, however, does not improve the performance of PART (cf. Table 3.6).

Handling Conflicting Rules

Conflicting rules are rules that have the same condition part but different polarity predictions. Basic conflict resolution by rule ordering tends to improve the classification performance, e.g., the results obtained using the confidence-based rule ordering tend to be better than those obtained using the default rule ordering (cf. Table 3.6). To further analyse the effect of rule conflicts, PRISM+BL and PRISM+BL+TH_{BL} are extended by removing all conflicting rules before a rule application process. The two resulting methods are referred to as PRISM+BL+RM_{all} and PRISM+BL+TH_{BL}+RM_{all}, respectively. Their performance is shown by the first six rows for each feature vector type in Table 3.10. On most evaluation schemes, these methods yield higher accuracy compared to the baseline methods (cf. the first two rows of Table 3.6), while their obtained

Table 3.9: The classification performance obtained from PRISM+BL+TH_{BL} and PART+BL+TH_{BL}

Feature	Method	D1-CV			D2-CV			D3-CV			D1-XR			D2-XR		
vector type		f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc
NRT	PRISM+BL+TH_{BL}															
	default	.842	.432	.753	.813	.544	.732	.816	.534	.734	.834	.517	.752	.825	.491	.740
	confidence-based	.842	.432	.753	.813	.550	.735	.816	.534	.735	.834	.513	.752	.825	.491	.740
	coverage-based	.830	.518	.747	.795	.545	.702	.802	.539	.710	.821	.517	.735	.814	.544	.735
	PART+BL+TH_{BL}	.847	.451	.761	.823	.497	.737	.829	.488	.742	.851	.471	.767	.824	.503	.738
WRT	PRISM+BL+TH_{BL}															
	default	.830	.365	.732	.804	.518	.713	.810	.504	.720	.833	.471	.744	.815	.352	.709
	confidence-based	.820	.404	.723	.804	.516	.716	.809	.503	.720	.833	.466	.744	.805	.360	.699
	coverage-based	.821	.528	.738	.791	.532	.697	.801	.530	.710	.822	.511	.735	.784	.520	.688
	PART+BL+TH_{BL}	.825	.365	.726	.816	.504	.730	.826	.491	.740	.855	.457	.771	.800	.271	.687

f-measure values for the negative class are comparable. Compared to PRISM+BL and PRISM+BL+TH_{BL} (cf. Tables 3.8 and 3.9, respectively), although the overall resulting accuracy is higher only on a few evaluation schemes, PRISM+BL+RM_{all} and PRISM+BL+TH_{BL}+RM_{all} yield higher f-measure values for the negative class on most schemes. When all conflicting rules are removed, the methods with WRT vectors yield higher accuracy compared to those with NRT vectors on all evaluation schemes, except for D1-CV.

A detailed investigation shows that many rules that are in conflict have different confidence values. The conflicts created by such rules can be resolved by the confidence-based rule ordering. This rule ordering, however, does not resolve conflicts between a pair of rules having the same confidence value. We consider another extension of PRISM+BL and PRISM+BL+TH_{BL} by removing every pair of conflicting rules that have the same confidence value. These two methods are referred to as PRISM+BL+RM and PRISM+BL+TH_{BL}+RM, respectively. Their performance is shown by the third and fourth row groups (rows 7–12) for each feature vector type in Table 3.10. These methods slightly improve f-measure values for the negative class compared to PRISM+BL and PRISM+BL+TH_{BL}, while their accuracy values are quite similar. Compared to their corresponding methods in which all conflicting rules

Table 3.10: The classification performance obtained from additional experimental settings for eliminating conflicting rules

Feature	Method	D1-CV			D2-CV			D3-CV			D1-XR			D2-XR		
vector type		f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc	f_{pos}	f_{neg}	acc
NRT	PRISM+BL+RM_{all}															
	default	.818	.524	.733	.772	.544	.664	.779	.539	.674	.807	.517	.714	.791	.544	.696
	confidence-based	.818	.524	.733	.772	.544	.664	.779	.539	.674	.807	.517	.714	.791	.544	.696
	coverage-based	.818	.524	.733	.772	.544	.664	.779	.539	.674	.807	.517	.714	.791	.544	.696
	PRISM+BL+TH_{BL}+RM_{all}															
	default	.819	.518	.733	.772	.544	.664	.779	.539	.674	.807	.517	.714	.791	.544	.696
	confidence-based	.819	.518	.733	.772	.544	.664	.779	.539	.674	.807	.517	.714	.791	.544	.696
	coverage-based	.819	.518	.733	.772	.544	.664	.779	.539	.674	.807	.517	.714	.791	.544	.696
	PRISM+BL+RM															
	default	.850	.086	.742	.812	.005	.684	.820	.006	.695	.851	n/a	.741	.813	n/a	.685
	confidence-based	.837	.399	.744	.812	.525	.731	.818	.489	.732	.836	.500	.753	.825	.491	.740
	coverage-based	.789	.478	.699	.795	.560	.721	.802	.550	.725	.821	.499	.736	.793	.534	.713
	PRISM+BL+TH_{BL}+RM															
	default	.842	.432	.753	.813	.544	.732	.816	.534	.734	.834	.517	.752	.825	.491	.740
	confidence-based	.842	.432	.753	.813	.550	.735	.816	.534	.735	.834	.513	.752	.825	.491	.740
	coverage-based	.830	.518	.747	.795	.545	.702	.802	.539	.710	.821	.517	.735	.814	.544	.735
WRT	PRISM+BL+RM_{all}															
	default	.811	.449	.718	.780	.536	.682	.780	.527	.682	.816	.524	.730	.791	.509	.704
	confidence-based	.812	.475	.723	.779	.535	.680	.779	.527	.681	.816	.524	.730	.788	.515	.702
	coverage-based	.811	.487	.723	.779	.536	.681	.779	.527	.681	.816	.524	.730	.777	.520	.693
	PRISM+BL+TH_{BL}+RM_{all}															
	default	.831	.481	.744	.782	.539	.682	.786	.532	.686	.818	.515	.730	.811	.524	.727
	confidence-based	.831	.500	.747	.781	.538	.681	.786	.532	.686	.818	.515	.730	.808	.522	.724
	coverage-based	.830	.520	.747	.782	.538	.682	.785	.533	.685	.818	.515	.730	.798	.521	.704
	PRISM+BL+RM															
	default	.829	.138	.714	.809	.104	.684	.821	.099	.700	.843	.129	.733	.806	.170	.683
	confidence-based	.814	.387	.714	.809	.493	.721	.815	.477	.725	.835	.471	.747	.799	.364	.692
	coverage-based	.786	.536	.706	.764	.536	.685	.761	.530	.682	.801	.518	.717	.738	.531	.662
	PRISM+BL+TH_{BL}+RM															
	default	.829	.364	.730	.804	.518	.713	.810	.503	.720	.833	.471	.744	.815	.352	.709
	confidence-based	.819	.403	.721	.804	.516	.716	.809	.503	.720	.833	.466	.744	.805	.360	.699
	coverage-based	.823	.528	.741	.791	.533	.698	.802	.531	.711	.822	.507	.735	.784	.520	.688

are removed, PRISM+BL+RM and PRISM+BL+TH_{BL}+RM yield higher accuracy on most evaluation schemes.

3.4.5 Conclusions

By using rules induced from feature vectors representing RST structures, the accuracy of aspect-based sentiment classification has been shown to be improved by approximately 4–7% on our datasets, compared to a simpler classification method that relies solely on the average polarity score of relevant EDUs. Although the overall accuracy values are improved by using rules induced by the PRISM and PART algorithms, the f-measure values for the negative polarity class are decreased. To address this issue, three heuristic approaches, i.e., a combination with a baseline method, confidence threshold setting, and removal of conflicting rules, are applied. The combination of these three heuristic approaches yields satisfactory classification results, without sacrificing the f-measure values for the negative class. Further work includes a modification/extension of the representation of local aspect segments, e.g., by incorporation of information about deeper levels of RST relations into feature vectors.

Chapter 4

Aspect-Based Sentiment Analysis on Clinical Text

4.1 Identification of Aspects in Clinical Narratives

In this preliminary study we examine only aspects of two entities, i.e. a health status and a medication. The health status entity is about a status of a patient's health. Aspects of this entity can be organs (e.g., heart), body parts (e.g., coronary vain), and body functions (e.g., pulmonary circulation). For the medication entity, aspects can be either medicines (e.g., aspirin) or procedure (e.g., surgery).

Consider 15 groups of semantic types in UMLS, i.e., activities and behaviors (ACTI), anatomy (ANAT), chemicals and drugs (CHEM), concepts and ideas (CONC), devices (DEVI), disorders (DISO), genes and molecular sequences (GENE), geographic areas (GEOG), living beings (LIVB), objects (OBJC), occupations (OCCU), organizations (ORGA), phenomena (PHEN), physiology (PHYS), and procedures (PROC), listed in Table 4.1. Some groups can identify aspects of the targeted entities. We test that assumption by listing terms in a clinical text, which are in specific semantic-type groups, and manually annotating them whether they are binding to a sentiment, i.e., negative, neutral, and positive. The specific semantic-type groups are selected manually by considering their meaning. ANAT, DISO, PHEN, and PHYS groups are selected for the health status entity, while CHEM and PROC groups are selected for the medication

entity.

Table 4.1: UMLS semantic types separated by their groups

Semantic groups	Abbrev.	Semantic types
ACTI: Activities & Behaviors	acty	Activity
	bhvr	Behavior
	dora	Daily or Recreational Activity
	evnt	Event
	gora	Governmental or Regulatory Activity
	inbe	Individual Behavior
	mcha	Machine Activity
	ocac	Occupational Activity
	socb	Social Behavior
ANAT: Anatomy	anst	Anatomical Structure
	blor	Body Location or Region
	bpoc	Body Part, Organ, or Organ Component
	bsoj	Body Space or Junction
	bdsu	Body Substance
	bdsy	Body System
	cell	Cell
	celc	Cell Component
	emst	Embryonic Structure
	ffas	Fully Formed Anatomical Structure
	tisu	Tissue
CHEM: Chemicals & Drugs	aapp	Amino Acid, Peptide, or Protein
	antb	Antibiotic
	bacs	Biologically Active Substance
	bodm	Biomedical or Dental Material
	carb	Carbohydrate
	chem	Chemical
	chvf	Chemical Viewed Functionally
	chvs	Chemical Viewed Structurally
	clnd	Clinical Drug
	eico	Eicosanoid
	elii	Element, Ion, or Isotope

Continued on next page

Table 4.1 – *Continued from previous page*

Semantic groups	Abbrev.	Semantic types
	enzy	Enzyme
	hops	Hazardous or Poisonous Substance
	horm	Hormone
	imft	Immunologic Factor
	irda	Indicator, Reagent, or Diagnostic Aid
	inch	Inorganic Chemical
	lipd	Lipid
	nsba	Neuroreactive Substance or Biogenic Amine
	nnon	Nucleic Acid, Nucleoside, or Nucleotide
	orch	Organic Chemical
	opco	Organophosphorus Compound
	phsu	Pharmacologic Substance
	rcpt	Receptor
	strd	Steroid
	vita	Vitamin
CONC: Concepts & Ideas	clas	Classification
	cnce	Conceptual Entity
	ftcn	Functional Concept
	grpa	Group Attribute
	idcn	Idea or Concept
	inpr	Intellectual Product
	lang	Language
	qlco	Qualitative Concept
	qnco	Quantitative Concept
	rnlw	Regulation or Law
	spco	Spatial Concept
	tmco	Temporal Concept
DEVI: Devices	drdd	Drug Delivery Device
	medd	Medical Device
	resd	Research Device
DISO: Disorders	acab	Acquired Abnormality
	anab	Anatomical Abnormality
	comd	Cell or Molecular Dysfunction
	cgab	Congenital Abnormality

Continued on next page

Table 4.1 – *Continued from previous page*

Semantic groups	Abbrev.	Semantic types
	dsyn	Disease or Syndrome
	emod	Experimental Model of Disease
	fndg	Finding
	inop	Injury or Poisoning
	mobd	Mental or Behavioral Dysfunction
	neop	Neoplastic Process
	patf	Pathologic Function
	sosy	Sign or Symptom
GENE: Genes & Molecular Sequences	amas	Amino Acid Sequence
	crbs	Carbohydrate Sequence
	gngm	Gene or Genome
	mosq	Molecular Sequence
	nusq	Nucleotide Sequence
GEOG: Geographic Areas	geoa	Geographic Area
LIVB: Living Beings	aggp	Age Group
	amph	Amphibian
	anim	Animal
	arch	Archaeon
	bact	Bacterium
	bird	Bird
	euka	Eukaryote
	famg	Family Group
	fish	Fish
	fngs	Fungus
	grup	Group
	humn	Human
	mamm	Mammal
	orgm	Organism
	podg	Patient or Disabled Group
	plnt	Plant
	popg	Population Group
	prog	Professional or Occupational Group
	rept	Reptile
	vtbt	Vertebrate

Continued on next page

Table 4.1 – *Continued from previous page*

Semantic groups	Abbrev.	Semantic types
	virs	Virus
OBJC: Objects	enty	Entity
	food	Food
	mnob	Manufactured Object
	phob	Physical Object
	sbst	Substance
OCCU: Occupations	bmod	Biomedical Occupation or Discipline
	ocdi	Occupation or Discipline
ORGA: Organizations	hcro	Health Care Related Organization
	orgt	Organization
	pros	Professional Society
	shro	Self-help or Relief Organization
PHEN: Phenomena	biof	Biologic Function
	eehu	Environmental Effect of Humans
	hcpp	Human-caused Phenomenon or Process
	lbtr	Laboratory or Test Result
	npop	Natural Phenomenon or Process
	phpr	Phenomenon or Process
PHYS: Physiology	celf	Cell Function
	clna	Clinical Attribute
	genf	Genetic Function
	menp	Mental Process
	moft	Molecular Function
	orga	Organism Attribute
	orgf	Organism Function
	ortf	Organ or Tissue Function
	phsf	Physiologic Function
PROC: Procedures	diap	Diagnostic Procedure
	edac	Educational Activity
	hlca	Health Care Activity
	lbpr	Laboratory Procedure
	mbrt	Molecular Biology Research Technique
	resa	Research Activity
	topp	Therapeutic or Preventive Procedure

Experimental results tested on 30 example clinical narratives show that 237 terms are binding to a polarity while 523 remaining terms are not binding to any polarity. We roughly infer that only some semantic types in the selected groups can identify aspects of the targeted entity. In order to analyse the effect of the suspect semantic types, we calculate confidence and support values for each type with respect to the results of the aspect identification. Table 4.2 shows the calculated results.

From Table 4.2, many semantic types can accurately identify aspects with confidence values of 1.0 but they rarely appear in the example documents. The semantic type with the most support value is body part, organ, or organ component (*bpoc*) that can identifies aspects with confidence value more than 0.8.

4.2 Text Segmentation with Respect to Aspect

Aim of this section is to determine difference of clinical and general text, in terms of aspect-based segmentation. Aspects in clinical text are mainly concerning parts or systems of patient's body. The Anatomy semantic group of UMLS is used for identifying aspect terms. After aspects are identified, the text is segmented using the same method in Section 3.1. Figure 4.1 shows a segmentation example from a clinical narrative to three aspect-based segments. Aspect keywords are in blue color, i.e., 'heart', 'lungs', and 'posterior tibialis'. The segments consist of four, nine, and two clauses (EDUs, denoted by Ex where x is a counting number) corresponding to 'heart', 'lungs', and 'posterior tibialis', respectively. Rhetorical structure is shown in the second line of each segment.

All 523 obtained aspect-based segments are next manually annotated EDUs being actually related to the aspect of the segment. For example, considers segments in Figure 4.1, the 'heart' and 'posterior tibialis' segments contain EDUs that are entirely related to the aspects. On the contrary, in the 'lungs' segment, only first EDU is related to lungs. Distribution of the number of EDUs in an aspect-based segment from manual annotation is slightly different compared to those obtained from automatic segmentation, as shown in Figure 4.2. The most number of EDUs in a segment is five when manual segmentation is applied, but it can be up to 25 EDUs by using automatic

Table 4.2: Confidence and support values of semantic types for aspect identification

Confidence value	Semantic types of manually-labelled aspect terms (support value)	
	With polarity	Without polarity
=1.0	*ortf* (0.038)	*hlca* (0.075)
	tisu (0.034)	*clnd* (0.046)
	bsoj (0.013)	*bodm* (0.032)
	aapp|*phsu* (0.008)	*antb*|*orch* (0.021)
	nsba|*orch*|*phsu* (0.004)	*npop* (0.018)
	cgab (0.004)	*elii*|*phsu* (0.018)
	aapp (0.004)	*orga* (0.014)
		elii (0.014)
≥0.8 and <1.0	*blor* (0.165)	*orch*|*phsu* (0.068)
	bpoc (0.498)	*lbpr* (0.043)
		diap (0.057)
		orgf (0.029)
≥0.6 and <0.8	*clna* (0.059)	*topp* (0.107)
	bdsu (0.034)	*phsu* (0.032)
	bacs|*carb*|*phsu* (0.008)	*phsf* (0.011)
		lbtr (0.011)
≥0.5 and <0.6	*aapp*|*bacs* (0.008)	*qlco* (0.014)
	antb (0.004)	*aapp*|*bacs* (0.007)
	inch|*phsu* (0.004)	*antb* (0.004)
		inch|*phsu* (0.004)

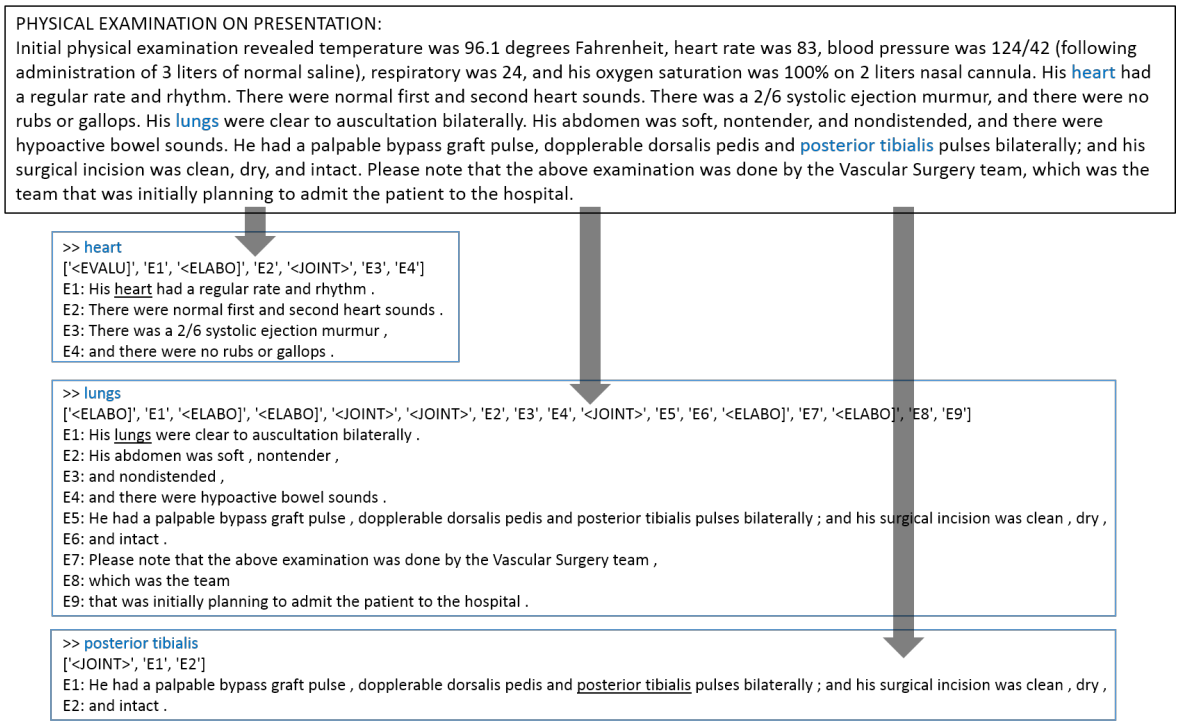


Figure 4.1: Example of aspect-based segmentation on a clinical narrative using the rhetorical structure

segmentation. However most segments from clinical narratives consist of only one EDU. Their portions are around 60 percent both manual and automatic segmentation. For product reviews, from Figure 3.3, local aspect segments containing one EDU and two EDUs are approximately 35 and 40 percent, respectively, of all segments.

In summary, an expression about an aspect in clinical text is usually in one main clause, without supplementary clauses, while that in product reviews commonly uses one or two clauses to describe opinion about an aspect. Therefore the methods for aggregation sentiment across EDUs introduced in Section 3.3 may not be influential on clinical text.

4.3 Sentiment Analysis on Clinical Narrative

A *clinical narrative* is a text containing details of a patient while staying in a hospital. It is a plain text recorded freely by clinical staffs. The contents of the narrative are about health status of a patient, the medicine which a patient took, or other details

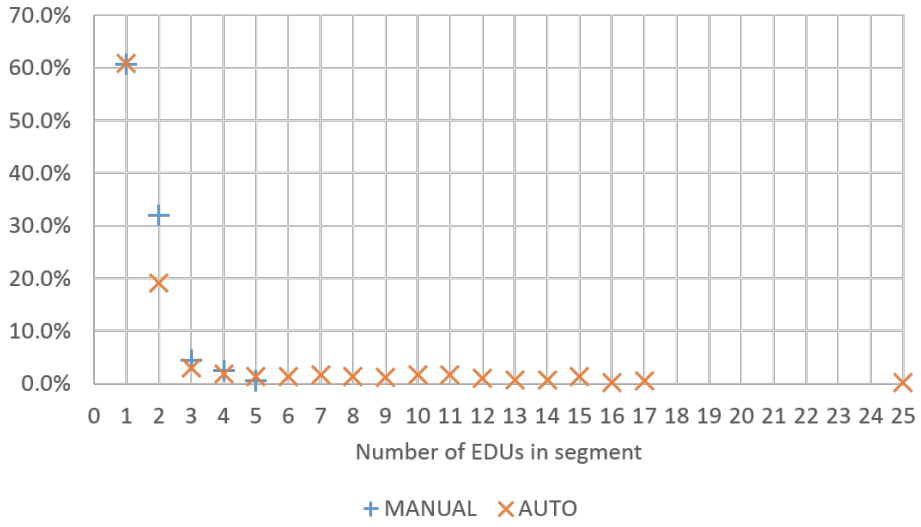


Figure 4.2: Distribution of number of EDUs in aspect-based segments from clinical narratives

the staffs found in a patient. In the MIMIC II database [19], the clinical narrative is divided into four groups, i.e., a radiology report, a nurse note, a medical doctor note, and a discharge summary. The narratives are useful in case of tracing what happen to a patient. Physicians can check results or effects of medications or treatments, applied to a patient, by reading the narratives, noted periodically by nurses. A suitable procedure would be next processed if the content, written in the narrative, gave enough information about the patient. Reading all narratives concerning each patient consumes a large amount of time. Sentiment analysis is applied to help filtering out less important parts of the narratives. An assumption is that important parts expresses a sentiment orientation, i.e., positive or especially negative.

An issue that makes the sentiment on a clinical narrative hard to analyse is the large number of medicine-technical terms, e.g., disease names or treatment processes, in the narrative. For example, a sentiment of a text '*Appears to have premature atrial contraction with bundle showing*' depends greatly on the meaning of a term '*premature atrial contraction*' was known. The Unified Medical Language System (UMLS) is an important knowledge base in the medical domain. It contains the technical terms which are mapped with their semantic types and semantic groups. For example, the term "premature atrial contraction" is mapped to the "disease or syndrome" (dsyn) semantic

type, which belongs to the Disorders semantic group. The Disorders semantic group, have potential to indicate the sentiment of their corresponding words, which frequently do not exist in a general-purpose lexicon. The Disorders group consists of 12 semantic types, i.e., acquired abnormality (*acab*), anatomical abnormality (*anab*), cell or molecular dysfunction (*comd*), congenital abnormality (*cgab*), disease or syndrome (*dsyn*), experimental model of disease (*emod*), finding (*fndg*), injury or poisoning (*inpo*), mental or behavioural dysfunction (*mobd*), neoplastic process (*neop*), pathologic function (*patf*),and sign or symptom (*sosy*). Most types in the Disorders group, e.g., *anab* or *dsyn*, are recognized as having a negative sense. So terms of those types should be treated as same as their types.

Lexicon-based sentiment analysis determines the sentiment score of a target text portion by aggregating the scores of terms in the portion that are obtained from a predetermined sentiment lexicon. Two types of sentiment lexicons are usually used [49]: (i) a generic sentiment lexicon, e.g., SentiWordNet [53], and (ii) a trained sentiment lexicon, i.e., a lexicon specifically learned from labelled samples of the targeted dataset [49, 61].

This study aims to exploit UMLS semantic types, especially in the Disorders group, to lexicon-based sentiment analysis both using a generic lexicon and a trained lexicon. Experiments focuses on investigating how semantic types in the Disorders group affect sentiment analysis on the text portion. The first objective is to investigate which sentiment is actually expressed by a term with a certain semantic type in Disorders group, e.g., *fndg* and *sosy*. The second objective is to investigate whether only one classification method is sufficient for text portions containing terms with different semantic types. Classifier combination is employed to solve the latter objective.

4.3.1 Methods Based on a Generic Lexicon

A generic lexicon is a lexicon generated from data sources that are not specific to a particular domain. The generic lexicon used in this paper is SentiWordNet [53]. Given a term t, let $score_{swn}(t)$ denote the polarity score of t obtained from SentiWordNet. The polarity scores from SentiWordNet are used for sentiment classification by three

methods, referred to as SWN, SWN.DE and SWN.LR, which are described below.

SWN

The first method, SWN, is a basic lexicon-based method used as a baseline for efficiency comparison in this section. The method determines the polarity score of an input sentence by calculating the sum of the polarity scores of all terms appearing in the sentence, with the score of a term being flipped to the opposite sign (i.e., positive or negative) if the term appears after a negation term, e.g., 'no' or 'not'. More precisely, given a sentence S, let $term(S)$ be the bag of all terms appearing in S and then let the polarity score of the sentence S, denoted by $score(S)$, be defined by

$$score(S) = \sum_{t \in term(S)} \Big((-1)^{neg(t)} \cdot score(t) \Big), \tag{4.1}$$

where for any term t, $score(t) = score_{swn}(t)$ and

$$neg(t) = \begin{cases} 1, & \text{if } t \text{ appears in } S \text{ after a negation term,} \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

SWN.DE

A *DISO type* is a semantic type in the Disorders group. Let *Dtype* be the set of all DISO types, i.e., $Dtype = \{acab, anab, comd, cgab, dsyn, emod, fndg, inpo, mobd, neop, patf, sosy\}$. Given a type d in *Dtype*, let a term that is mapped by UMLS to the type d be called a d-term. For each type d in *Dtype*, a d-term is also simply called a *DISO term*. The second method, SWN.DE, assumes that a DISO term expresses a negative sentiment, with the polarity score of -1. SWN.DE determines the polarity score of a sentence S in the same way as SWN except that the polarity score of -1 is assigned to each DISO term. That is, for each term t in $term(S)$, SWN.DE determines $score(t)$ in Equation 4.1 as follows:

$$score(t) = \begin{cases} -1, & \text{if } t \text{ is a DISO term,} \\ score_{swn}(t), & \text{otherwise.} \end{cases} \tag{4.3}$$

SWN.LR

Unlike SWN.DE, which assigns the polarity score of -1 to every DISO term, the third method, SWN.LR, tries to assign polarity scores to DISO terms based on their types. To find proper polarity scores, logistic regression (LR) is employed. LR finds the best fitting regression coefficient β_i for each independent variable x_i in a linear combination

$$l(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n. \tag{4.4}$$

Let S be a given sentence. Let S' be the text portion obtained from S by removing all DISO terms, and let $score(S')$ be the polarity score of S' calculated by the SWN method. For each DISO type d in $Dtype$, let $N_{d,dir}$ (respectively, $N_{d,neg}$) denote the number of all d-terms that appear not after (respectively, after) a negation term. (Intuitively, "dir" stands for "directly used", while "neg" stands for "negated".) Following Equation 4.4, the linear combination obtained from S is

$$l(S) = \beta_0 + \beta_{swn} \cdot score(S') + \sum_{d \in Dtype} \Big(\beta_{d,dir} \cdot N_{d,dir} + \beta_{d,neg} \cdot N_{d,neg} \Big). \tag{4.5}$$

For each d in $Dtype$, the coefficient $\beta_{d,dir}$ (respectively, $\beta_{d,neg}$) represents the polarity score of a d-term that appears not after (respectively, after) a negation term. The coefficient β_{swn} gives the weight of a score obtained from SentiWordNet. The coefficient β_0 refers to an intercept, representing a bias of the classifier. The best fitting regression coefficients are learned from a training set.

Experimental Setting and Results

A dataset used in this paper was taken from the clinical narrative part in MIMIC II database [19]. Each sentence in the dataset was annotated with a polarity orientation, i.e., positive or negative. The dataset consists of 2,504 sentences, which are divided into 1,237 positive sentences and 1,267 negative sentences.

One hundred iterations of experiments were conducted. In each iteration, the dataset was randomly separated into a training set and a test set with a ratio of 60/40. The training set was used for training the regression coefficients in SWN.LR,

while the test set was used for performance evaluation of all the three methods. If the polarity score of an input sentence, obtained from the classification method, is less than zero, it is classified as negative; otherwise, it is classified as positive. The average accuracy values obtained from the test sets in the 100 iterations were taken as the experimental results.

Table 4.3 shows the experimental results in different segmentations. The rows in the table are divided into three row groups. The first row of the first row group shows the results obtained from all sentences. The second row and the third row of the first row group show the results obtained from sentences that contain and do not contain DISO terms, respectively. Each row in the second row group (respectively, the third row group) shows the results obtained from sentences containing DISO terms having one specific DISO type without other DISO terms (respectively, possibly with other DISO terms). For example, the row with label 'with only *acab*' (respectively, 'with *acab*') shows the results obtained from sentences containing *acab*-terms[1] without any other DISO term (respectively, possibly with other DISO terms). Since neither *comd*-term nor *emod*-term appears in the dataset, the DISO types *comd* and *emod* are neglected.

From Table 4.3, SWN.LR yields the highest overall accuracy of 0.710 when all sentences in the test sets are considered. When only sentences containing DISO terms (respectively, not containing DISO terms) are considered, the highest average accuracy value is obtained from SWN.DE (respectively, SWN.LR). When considering sentences containing only one specific DISO type, SWN.DE and SWN.LR yield higher accuracy than SWN with a 99% level of confidence for the DISO types *acab*, *anab*, *dsyn*, *neop*, *patf*, and *sosy*. For the DISO types *fndg* and *inpo*, SWN.DE also yield higher accuracy than SWN with a 95% level of confidence. When considering sentences containing one specific DISO type possibly with some other types, SWN.DE and SWN.LR yield higher accuracy than SWN with a 99% level of confidence for all DISO types except *mobd*.

Table 4.4 shows the average values of regression coefficients, which are learned from the training sets generated in the 100 iterations of the experiments. Since $\beta_{swn} = 0.744$, all scores from SentiWordNet are 24.6 percent less significant in SWN.LR, compared to

[1]Using the notation introduced in Section 4.3.1, an *acab*-term is a DISO term that is mapped by UMLS to the DISO type *acab*.

Table 4.3: Accuracies of methods based on SentiWordNet, averaged from 100 iterations

Segment Type	#Sentences	SWN	SWN.DE	SWN.LR
all sentences	1,033	0.582	0.676**	**0.710****
with DISO terms	660	0.599	**0.745****	0.740**
w/o DISO terms	373	0.553	0.553	**0.658****
with only *acab*	3	0.333	**0.667****	0.617**
with only *anab*	10	0.566	**0.781****	0.770**
with only *cgab*	6	**0.847**	0.740	0.693
with only *dsyn*	93	0.586	0.863**	**0.872****
with only *fndg*	206	0.562	**0.569***	0.545
with only *inpo*	21	0.660	**0.687***	0.660
with only *mobd*	6	**0.705**	0.490	0.400
with only *neop*	3	0.593	**1.000****	0.933**
with only *patf*	85	0.628	0.904**	**0.907****
with only *sosy*	54	0.538	**0.779****	0.771**
with *acab*	13	0.603	0.728**	**0.736****
with *anab*	20	0.624	**0.841****	0.835**
with *cgab*	10	0.708	**0.844****	0.809**
with *dsyn*	172	0.617	0.867**	**0.877****
with *fndg*	333	0.592	**0.654****	0.644**
with *inpo*	36	0.672	**0.749****	0.717**
with *mobd*	19	**0.706**	0.680	0.681
with *neop*	13	0.321	0.820**	**0.842****
with *patf*	151	0.647	**0.896****	**0.896****
with *sosy*	105	0.607	0.778**	**0.791****

* and ** indicate significant improvement compared to SWN with $p < 0.05$ and $p < 0.01$, respectively.

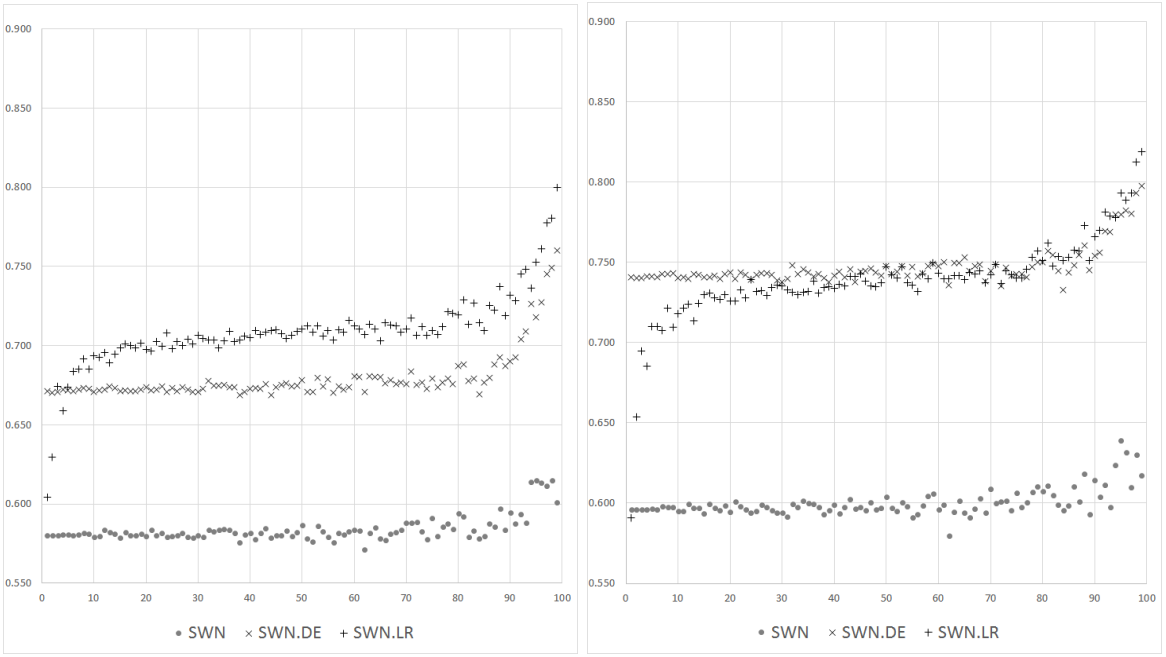
Table 4.4: Regression coefficients of SWN.LR, averaged from 100 iterations

Parameter	Value	Parameter	Value
β_0	-0.380	β_{swn}	0.744
$\beta_{acab,dir}$	0.095	$\beta_{acab,neg}$	0.273
$\beta_{anab,dir}$	-1.386	$\beta_{anab,neg}$	-0.157
$\beta_{cgab,dir}$	-0.485	$\beta_{cgab,neg}$	0.555
$\beta_{dsyn,dir}$	-1.604	$\beta_{dsyn,neg}$	1.871
$\beta_{fndg,dir}$	-0.273	$\beta_{fndg,neg}$	0.961
$\beta_{inpo,dir}$	-0.686	$\beta_{inpo,neg}$	1.127
$\beta_{mobd,dir}$	-0.169	$\beta_{mobd,neg}$	-0.398
$\beta_{neop,dir}$	-1.003	$\beta_{neop,neg}$	0.529
$\beta_{patf,dir}$	-1.868	$\beta_{patf,neg}$	1.430
$\beta_{sosy,dir}$	-1.036	$\beta_{sosy,neg}$	1.036

SWN. Since β_0 = -0.380, a sentence containing no sentiment term and no DISO term is classified as negative. Due to β_0 and β_{swn}, even when only sentences without DISO terms are considered (cf. the third row of the first row group in Table 4.3), SWN.LR yields higher performance than SWN. For each type $d \in \{anab, dsyn, neop, patf, sosy\}$, $\beta_{d,dir}$ gives a strong negative polarity score of less than -1. For each type $d \in Dtype - \{anab, mobd\}$, $\beta_{d,neg}$ gives a positive polarity score. Considering the DISO type *dsyn*, for example, on average, the polarity score of $\beta_{dsyn,dir}$ = -1.604 is assigned to a *dsyn*-term that appears not after a negation term, when that of $\beta_{dsyn,neg}$ = 1.871 is assigned to a *dsyn*-term that appears after a negation term.

The Effect of Training Set Size

To investigate the effect of the size of a training set on the performance of SWN.LR, an addition set of experiments was considered with the proportion of a training set to the whole dataset being varied from 1 to 99 percent. With the proportion value of 20 percent, for example, the dataset was divided into a training set and a test set



(a) All sentences

(b) Sentences containing DISO terms

Figure 4.3: Accuracy of methods using generic lexicon at varied train set proportion

with a ratio of 20/80. For each proportion value, ten iterations of experiments were conducted, in each of which a training set was randomly selected. The average accuracy values obtained from all sentences are shown in Fig.4.3a, while those obtained from only sentences containing DISO terms are shown in Fig.4.3b. In Fig.4.3a, SWN.LR yields the highest performance when the proportion of the training set is higher than 5 percent. In Fig.4.3b, when the proportion of a training set is greater than 35 percent, the performance of SWN.LR is comparable to that of SWN.DE.

4.3.2 Methods Based on Trained Lexicons

From a given training set, polarity scores are assigned to all terms appearing in the training set to generate a trained sentiment lexicon. Two sentiment classification methods using trained lexicons, referred to as TL and TL.R (TL with term replacement), are considered in this study. The former uses a lexicon trained without considering DISO types, while the latter uses that trained by considering DISO types.

TL

The first method determines a polarity score for each term appearing in a training set without considering DISO types. As in [62], the polarity score of a term t appearing in the training set, denoted by $score_{tr}(t)$, is calculated by

$$score_{tr}(t) = \frac{p(t|positive) - p(t|negative)}{p(t|positive) + p(t|negative)}, \tag{4.6}$$

where $p(t|positive)$ (respectively, $p(t|negative)$) is the number of occurrences of t in all positive (respectively, negative) sentences of the training set divided by the total number of term occurrences in the positive (respectively, negative) sentences.

Using the trained lexicon, the polarity score of a text portion S, denoted by $score(S)$, is defined by

$$score(S) = \sum_{t \in term(S)} score(t), \tag{4.7}$$

where $term(S)$ is the bag of all terms appearing in S and for any term t,

$$score(t) = \begin{cases} score_{tr}(t), & \text{if } t \text{ appears in the training set,} \\ 0, & \text{otherwise.} \end{cases} \tag{4.8}$$

TL.R

For any $D \subseteq Dtype$, *TL.R with respect to* D is a method that generates a lexicon by considering the DISO types in D as follows: First, for each DISO type $d \in D$, each occurrence of a d-term is replaced with $\langle d \rangle$. For example, the text portion "*Appears to have premature atrial contraction with bundle showing*" is changed to "*Appears to have* $\langle dsyn \rangle$ *with bundle showing*" since "premature atrial contraction" is a *dsyn*-term. Consequently, Equation 4.6 is applied to the modified training set. The polarity scores of all terms in the training set, including those of the term $\langle d \rangle$ where $d \in D$, are retrieved.

When the trained lexicon is applied to classification, TL.R with respect to D determines $score(t)$ in Equation 4.7 as follows:

$$score(t) = \begin{cases} score_{tr}(\langle d \rangle), & \text{if } t \text{ is a } d\text{-term and } d \in D, \\ score_{tr}(t), & \text{if } t \text{ is in training set and does not have any types in } D, \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

Experimental Setting and Results

Two consideration schemes about DISO types are proposed. First, considering all DISO types at once, TL.R with respect to *Dtype* is proposed and referred to as TL.R(all). Second, considering only a DISO type $d \in Dtype$, TL.R with respect to $\{d\}$ is proposed and referred to as TL.R(d). For example, TL.R($dsyn$) is TL.R with respect to $\{dsyn\}$.

TL and the TL.R methods were evaluated using the dataset and experimental setting described in Section 4.3.1 (i.e., 100 iterations of trials, in each of which the dataset was randomly separated into a training set and a test set with a ratio of 60/40). Table 4.5 shows the average accuracies obtained from the experiments. From the row 'with DISO terms', for example, the methods TL, TL.R($anab$), TL.R($cgab$), TL.R($dsyn$), and TL.R($neop$) yield the same average accuracy of 0.830.

From Table 4.5, TL.R(all) and TL.R($dsyn$) yield the highest average accuracies on 9 and 6 segment types, respectively, while the other methods give the highest average accuracies on equal or less than 4 segment types. TL.R(all) and TL.R($dsyn$) yield higher performance than TL on 11 and 12 segment types, respectively. From those segment types, there are 4 (respectively, 2) segment types on which TL.R(all) (respectively, TL.R($dsyn$)) yield higher performance than TL with the level of confidence being at least 95%. Comparing between TL.R(all) and TL.R($dsyn$), TL.R($dsyn$) performs better on 14 segment types, while TL.R(all) performs better on 8 segment types.

For each DISO type $d \in Dtype$, Table 4.6 shows the average polarity score of the term $\langle d \rangle$ in the trained lexicons generated by TL.R(all) in the experiments. For each $d \in \{anab, dsyn, neop, patf, sosy\}$, the average polarity score of $\langle d \rangle$ in the generated trained lexicons is negative.

Table 4.5: Accuracies of methods based on trained lexicons, averaged from 100 iterations

Segment Type	TL	TL.R										
		all	*acab*	*anab*	*cgab*	*dsyn*	*fndg*	*inpo*	*mobd*	*neop*	*patf*	*sosy*
all sentences	0.812	0.806	0.811	0.812	0.812	**0.813**	0.806	0.811	0.811	0.812	0.810	0.810
with DISO terms	**0.830**	0.820	0.829	**0.830**	**0.830**	**0.830**	0.820	0.829	0.828	**0.830**	0.826	0.827
w/o DISO terms	0.781	**0.782**	0.780	0.780	0.780	**0.782**	0.780	0.780	0.780	0.781	0.781	0.781
with only *acab*	0.817	0.750	0.710	0.823	0.817	**0.850**	0.820	0.817	0.817	0.817	0.820	0.817
with only *anab*	0.769	**0.777**	0.768	0.769	0.769	0.756	0.770	0.766	0.768	0.766	0.772	0.771
with only *cgab*	0.883	**0.912**	0.883	0.885	0.900	0.845	0.885	0.883	0.883	0.885	0.883	0.885
with only *dsyn*	0.849	**0.868****	0.849	0.848	0.848	0.865**	0.852	0.849	0.849	0.849	0.849	0.847
with only *fndg*	0.813	0.785	0.813	0.813	0.813	0.812	0.786	0.813	0.812	**0.814**	0.812	0.812
with only *inpo*	0.757	0.758	0.756	0.758	0.760	**0.761**	0.756	0.749	0.756	0.757	0.757	0.754
with only *mobd*	0.718	0.652	0.718	0.718	0.718	0.717	0.713	0.722	0.643	0.718	0.718	**0.723**
with only *neop*	0.687	**0.810****	0.687	0.687	0.687	0.690	0.687	0.687	0.687	0.800**	0.690	0.687
with only *patf*	**0.903**	0.879	0.902	0.902	0.902	0.900	0.901	**0.903**	0.902	0.902	0.880	0.901
with only *sosy*	0.739	0.734	0.739	0.740	0.739	0.740	**0.742**	0.740	0.739	0.740	0.741	0.732
with *acab*	0.682	0.625	0.657	**0.687**	0.681	0.669	0.674	0.664	0.681	0.667	0.634	0.682
with *anab*	0.840	**0.878****	0.835	0.872**	0.840	0.845	0.832	0.836	0.839	0.838	0.841	0.841
with *cgab*	0.878	**0.899**	0.878	0.886	0.886	0.866	0.849	0.878	0.878	0.879	0.884	0.878
with *dsyn*	0.855	**0.866****	0.858	0.857	0.854	0.860	0.853	0.855	0.855	0.854	0.857	0.855
with *fndg*	0.830	0.814	0.829	**0.831**	0.830	0.830	0.812	0.829	0.829	0.830	0.828	0.826
with *inpo*	0.773	0.774	0.772	0.774	0.774	**0.777**	0.774	0.760	0.772	0.773	0.773	**0.777**
with *mobd*	0.821	0.781	0.821	0.821	0.821	0.797	0.815	0.821	0.788	0.821	0.820	**0.822**
with *neop*	0.829	0.829	0.811	0.828	0.829	0.852*	**0.856***	0.829	0.829	0.818	0.835	0.831
with *patf*	0.902	0.894	0.901	0.902	0.902	0.901	0.898	0.902	0.902	0.901	0.890	**0.903**
with *sosy*	0.773	**0.779**	0.772	0.773	0.773	0.774	0.777	0.771	0.772	0.773	0.777	0.764

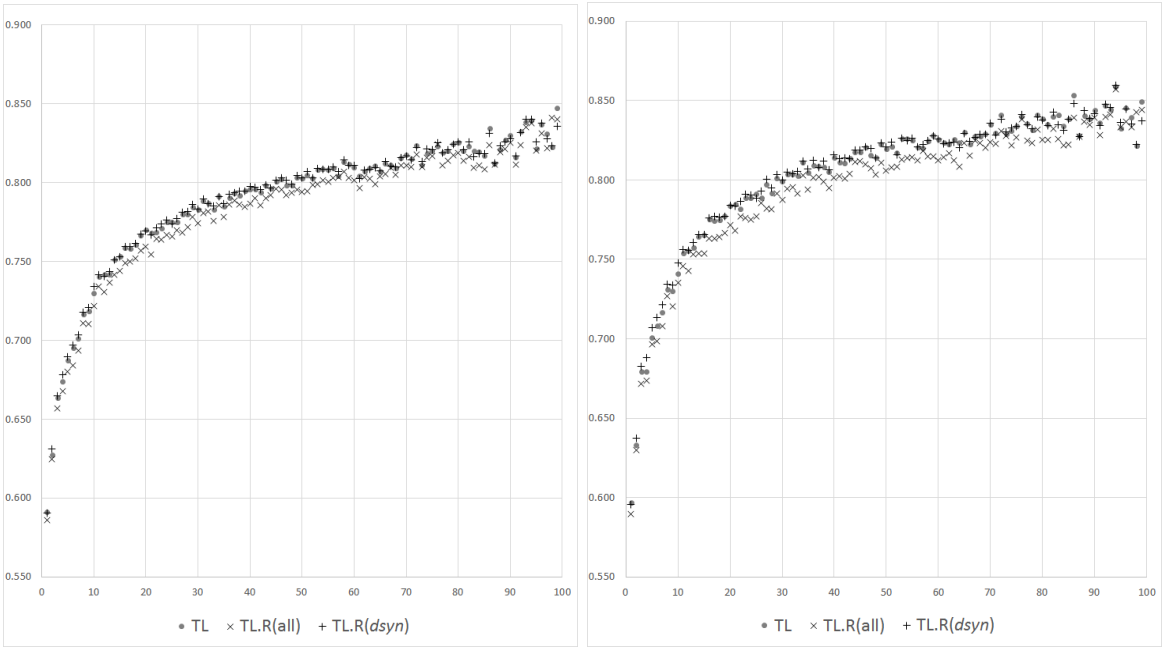
* and ** indicate significant improvement compared to TL with $p < 0.05$ and $p < 0.01$, respectively.

Table 4.6: Polarity scores of DISO types generated in TL.R(all), averaged from 100 iterations

Term	Polarity Score
⟨*acab*⟩	0.205
⟨*anab*⟩	-0.520
⟨*cgab*⟩	0.286
⟨*dsyn*⟩	-0.262
⟨*fndg*⟩	0.027
⟨*inpo*⟩	0.010
⟨*mobd*⟩	0.125
⟨*neop*⟩	-0.383
⟨*patf*⟩	-0.290
⟨*sosy*⟩	-0.233

The Effect of Training Set Size

To investigate the effect of the size of a training set on the performance of TL, TL.R(all), and TL.R(*dsyn*), the experimental setting in Section 4.3.1 was applied (i.e., the proportion of a training set to the whole dataset being varied from 1 percent to 99 percent, with 10 iterations being conducted for each proportion value). Figure 4.4a and Figure 4.4b show the average accuracies of TL, TL.R(all), and TL.R(*dsyn*) on the segment types 'all sentences' and 'with DISO terms', respectively. In both figures, regardless of the proportion of a training set, TL and TL.R(*dsyn*) yield almost the same average accuracy, and both of them perform slightly better than TL.R(all). In Figure 4.4a, to achieve the average accuracy of 0.800, the methods TL, TL.R(all) and TL.R(*dsyn*) require the proportion of at least 49 percent, 65 percent, and 49 percent, respectively. To achieve the same average accuracy (i.e., 0.800) in Figure 4.4b, the three methods require the proportion of at least 33 percent, 40 percent, and 31 percent, respectively.



(a) All sentences (b) Sentences containing DISO terms

Figure 4.4: Accuracy of methods using trained lexicon at various training set proportion

4.3.3 Classifier Selection

From the results of the 100 iterations of the experiments in Sections 4.3.1 and 4.3.2, different classification methods yield different performance on each segment type of the dataset (cf. each row in Tables 4.3 and 4.5). A classifier selection, described by Algorithm 1, is proposed to select the method appropriate for an input sentence, based on DISO types occurring in the sentence and performance of the method in Sections 4.3.1 and 4.3.2.

Preliminary Notation

To describe Algorithm 1, the following notations are used:

- $avgAcc(m, t)$ is the average accuracy obtained by using the method m on the segment type t, shown in Tables 4.3 and 4.5.
- $numWin(m_1, m_2, t)$ is the number of iterations of the experiments (cf. Sections 4.3.1 and 4.3.2) in which the method m_1 yields higher accuracy than the method m_2 on the segment type t.

Algorithm 1: Classifier Selection MethSel(s,G,cr)

Input:

- s, a given sentence
- G, a set of candidate methods
- cr, a measurement criterion for classifier selection

Output: x, the method selected for classifying the input sentence s

```
1  let D be the set of all DISO types appearing in s
2  if D is the empty set then
3  |  P := PerfQuads(G, 'w/o DISO terms', cr)
4  else if D is a singleton set {d} then
5  |  P := PerfQuads(G, withOnly(d), cr)
6  else
7  |  P := ⋃_{d∈D} PerfQuads(G, with(d), cr)
8  end
9  p := MaxPerfQuad(P)
10 x := method(p)
11 return x
```

- Let m be a classification method,
  t be a data segment type,
  and cr be a measurement criterion for classifier selection.

  The performance value of m on a segment of type t with respect to cr is denoted by $val(m, t, cr)$.

  Two criteria 'Ac' and 'Fq' are considered.
  The former criterion compares the average accuracy of m with that of TL.
  The latter criterion compares the number of iterations in which m performs better than TL with the number of those in which m performs worse than TL.

  More precisely, based on the results in Sections 4.3.1 and 4.3.2, $val(m, t, \text{`}Ac\text{'})$ and $val(m, t, \text{`}Fq\text{'})$ are defined as follows:

  - $val(m, t, \text{`}Ac\text{'}) = avgAcc(m, t) - avgAcc(\text{TL}, t)$
  - $val(m, t, \text{`}Fq\text{'}) = numWin(m, \text{TL}, t) - numWin(\text{TL}, m, t)$

- Types of data segments (cf. Tables 4.3 and 4.5) are referred to as follows.
  - $withOnly(d)$ is the segment type with label 'with only d'.
  - $with(d)$ is the segment type with label 'with d'.

- The performance quadruple p is the quadruple $\langle m, t, cr, v \rangle$, where $v = val(m, t, cr)$. v is called the performance value of p. The method m is also denoted by $method(p)$.

- Given a set G of classification methods, a segment type t, and a criterion cr, let a set $\textsc{PerfQuads}(G, t, cr)$ of performance quadruples be defined by

$$\textsc{PerfQuads}(G, t, cr) = \{\langle m, t, cr, v \rangle \mid m \in G, v = val(m, t, cr)\}.$$

- Given a set P of performance quadruples, let $\textsc{MaxPerfQuad}(P)$ denote a performance quadruple with the maximal performance value in P.

Selection Algorithm Description

Algorithm 1 selects a classification method x for a given input sentence s from a set G of methods with respect to a criterion cr. Suppose that G contains the 12 methods considered in Table 4.5 and the criterion cr is '*Ac*'. First, a set D containing all DISO types in the input sentence s is determined at Line 1. We explain how the output method x is selected by using the following three examples:

Example 1

Suppose that $D = \emptyset$. A set $P = \textsc{PerfQuads}(G, \text{'w/o DISO terms'}, \text{'}Ac\text{'})$ constructed at Line 3 contains 12 performance quadruples (one for each method in G), determined by the average accuracy values in the row 'w/o DISO terms' in Table 4.5. For example, since

$$\begin{aligned} & val(\text{TL.R(all)}, \text{'w/o DISO terms'}, \text{'}Ac\text{'}) \\ = \ & avgAcc(\text{TL.R(all)}, \text{'w/o DISO terms'}) - avgAcc(\text{TL}, \text{'w/o DISO terms'}) \\ = \ & 0.782 - 0.781 \\ = \ & 0.001, \end{aligned}$$

the performance quadruple for the method TL.R(all) is

$$\langle \text{TL.R(all)}, \text{'w/o DISO terms'}, \text{'}Ac\text{'}, 0.001 \rangle.$$

Since TL.R(all) and TL.R(*dsyn*) yield the highest average accuracy (0.782) in that row, the method x obtained at Line 10 is either TL.R(all) or TL.R(*dsyn*).

Example 2

Suppose that $D = \{inpo\}$. A set $P = \textsc{PerfQuads}(G, withOnly(inpo), \text{'}Ac\text{'})$, containing 12 performance quadruples, is constructed at Line 5 from the average accuracy values in the row 'with only *inpo*' of Table 4.5. Since TL.R(*dsyn*) yields the highest average accuracy in that row, the method x obtained at Line 10 is TL.R(*dsyn*).

Example 3

Suppose that $D = \{dsyn, patf\}$. A set P constructed at Line 7 is the union of P_1 and P_2, where $P_1 = \textsc{PerfQuads}(G, with(dsyn), \text{'}Ac\text{'})$ and $P_2 = \textsc{PerfQuads}(G, with(patf), \text{'}Ac\text{'})$,

which are determined by the average accuracy values in the row 'with *dsyn*' and the row 'with *patf*' of Table 4.5, respectively. Since TL.R(all) yields the highest average accuracy in the row 'with *dsyn*', the best method in P_1 is TL.R(all) with the performance value being

$$\begin{aligned} & val(\text{TL.R(all)}, \text{`with } dsyn\text{'}, \text{`}Ac\text{'}) \\ = \ & avgAcc(\text{TL.R}(all), \text{`with } dsyn\text{'}) - avgAcc(\text{TL}, \text{`with } dsyn\text{'}) \\ = \ & 0.866 - 0.855 \\ = \ & 0.011. \end{aligned}$$

Similarly, the best method in P_2 is TL.R($sosy$) with the performance value being

$$\begin{aligned} & val(\text{TL.R}(sosy), \text{`with } patf\text{'}, \text{`}Ac\text{'}) \\ = \ & avgAcc(\text{TL.R}(sosy), \text{`with } patf\text{'}) - avgAcc(\text{TL}, \text{`with } patf\text{'}) \\ = \ & 0.903 - 0.902 \\ = \ & 0.001. \end{aligned}$$

Since $val(\text{TL.R(all)}, \text{`with } dsyn\text{'}, \text{`}Ac\text{'}) > val(\text{TL.R}(sosy), \text{`with } patf\text{'}, \text{`}Ac\text{'})$, the method x obtained at Line 10 is TL.R(all).

When the criterion 'Fq' is considered, the performance value $val(m, t, \text{`}Fq\text{'}, v)$ is used instead of $val(m, t, \text{`}Ac\text{'}, v)$.

Experimental Settings and Results

Given a set G of classification methods and a criterion cr, let CS(G,cr) denote a method that uses the classification method MethSel(s,G,cr) to classify each sentence s in a dataset. Two sets of classification methods,

- $G_1 = \{\text{TL, TL.R(all)}\} \cup \{\text{TL.R}(d) \mid d \in Dtype - \{comd, emod\}\}$ and
- $G_2 = G_1 \cup \{\text{SWN, SWN.DE, SWN.LR}\}$,

are considered. Ten iterations of experiments using the methods TL and CS(G,cr), where $G \in \{G_1, G_2\}$ and $cr \in \{\text{`}Ac\text{'}, \text{`}Fq\text{'}\}$, are conducted on the same experimental setting described in Section 4.3.1. Table 4.7 shows the average accuracy values obtained from those 10 iterations of the experiments.

From Table 4.7, CS(G_2,'Fq') yields the highest average accuracy of 0.820 when all sentences in the test sets are considered. The method yields higher performance than TL on all segment types except the segment type 'w/o DISO terms'. Among those 12 segment types, CS(G_2,'Fq') improve the performance with the level of confidence being at least 95% on eight of them. Other methods using classifier selection also yield higher performance compared to TL on most segment types. Considering sets of classification methods (G_1 and G_2), the methods choosing the classifier from G_2, i.e., CS(G_2,'Ac') and CS(G_2,'Fq'), give higher performance than those choosing from G_1 on most segment types. Considering criteria for classifier selection ('Ac' and 'Fq'), the performance of the methods with respect to 'Ac' and that with respect to 'Fq' are quite not different on the same segment types.

The Effect of Training Set Size

From Table 4.7, the classification results are improved when the method appropriate for an input sentence are selected and used for its sentiment classification. The DISO types containing in the sentence are the important keys to select the method.

When the investigation for the effect of the size of a training set is conducted as in Section 4.3.1, the average accuracy values are shown in Fig. 4.5a and 4.5b. In both figures, CS(G_2,'Ac') and CS(G_2,'Fq') yields the highest performance. Those methods require a smaller training set to yield the same performance compared to TL. To achieve the average accuracy value of 0.80, for example, the proportion value of 30 percent is required by CS(G_2,'Ac') and CS(G_2,'Fq') while that of 40 percent is required by TL in Fig. 4.5a. Similarly, in Fig. 4.5b, the proportion value of 10 percent is required by CS(G_2,'Ac') and CS(G_2,'Fq') while that of 30 percent is required by TL.

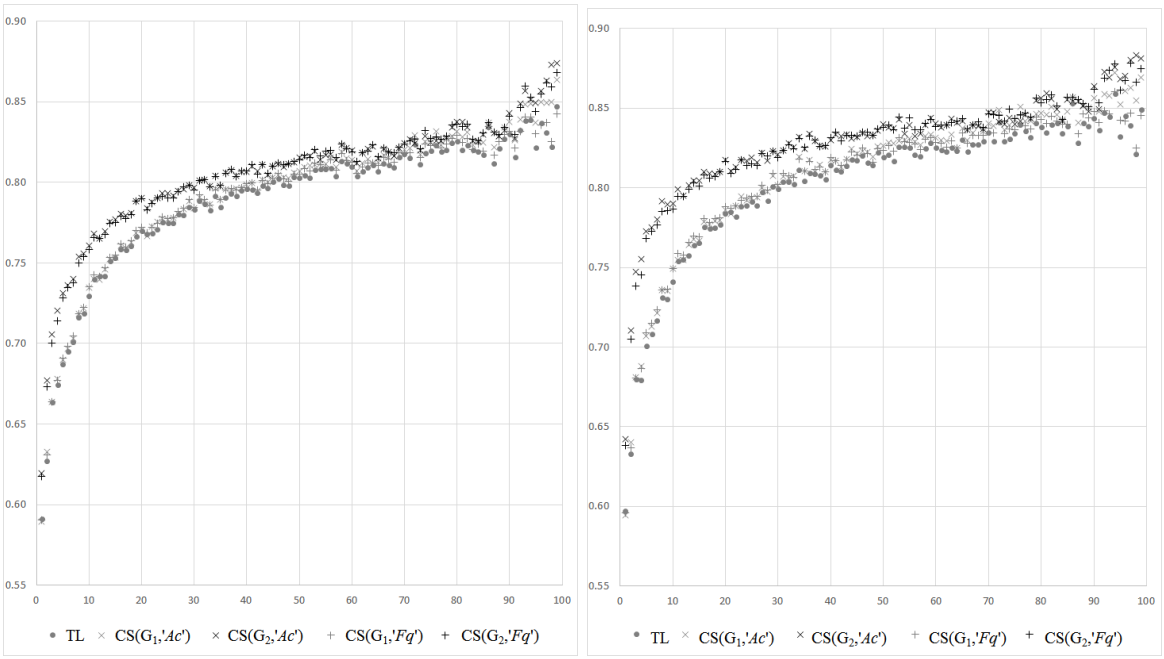
4.3.4 Conclusions

Sentiment analysis on clinical narrative can be improved using a domain-specific knowledge corpus, i.e., UMLS. In this study, UMLS is exploited to improve lexicon-based sentiment analysis methods both using the generic lexicon and using the trained lexicon. For the former, the polarity score of a term with a semantic type in the Disorders

Table 4.7: Accuracy of classifier selection methods compared to TL, running on 10 iterations.

Segment Type	#Sentence	TL	CS(G_1,'Ac')	CS(G_2,'Ac')	CS(G_1,'Fq')	CS(G_2,'Fq')
all sentences	1033	0.810	0.815	0.817	0.813	**0.820***
with DISO terms	660	0.828	0.836*	0.839**	0.833	**0.843****
w/o DISO terms	373	**0.779**	0.777	0.777	0.777	0.777
with only *acab*	3	0.767	**0.800**	**0.800**	**0.800**	**0.800**
with only *anab*	10	**0.820**	**0.820**	0.760	**0.820**	**0.820**
with only *cgab*	6	0.867	**0.917**	**0.917**	**0.917**	**0.917**
with only *dsyn*	93	0.852	0.882**	**0.889****	0.882**	**0.889****
with only *fndg*	206	0.820	**0.822**	**0.822**	**0.822**	**0.822**
with only *inpo*	21	0.733	**0.738**	**0.738**	**0.738**	**0.738**
with only *mobd*	6	0.683	**0.700**	0.683	**0.700**	**0.700**
with only *neop*	3	0.767	0.933**	**1.000****	0.933**	**1.000****
with only *patf*	85	0.891	0.891	**0.905**	0.891	**0.905**
with only *sosy*	54	0.724	0.728	**0.750**	0.728	**0.750**
with *acab*	13	0.654	0.692	**0.792****	0.654	0.754**
with *anab*	20	0.865	0.895	0.815	0.895	**0.900***
with *cgab*	10	0.880	0.910	**0.940**	0.910	**0.940**
with *dsyn*	172	0.857	0.869	0.890**	0.866	**0.892****
with *fndg*	333	0.829	**0.834**	0.825	0.831	0.832
with *inpo*	36	0.769	0.786	**0.789**	0.772	0.769
with *mobd*	19	0.826	0.821	0.832	0.826	**0.837**
with *neop*	13	0.838	0.892	**0.938***	0.838	0.923*
with *patf*	151	0.893	0.896	**0.907**	0.891	0.902
with *sosy*	105	0.762	**0.781**	0.764	0.766	0.777

* and ** indicate significant improvement compared to TL with $p < 0.05$ and $p < 0.01$, respectively.



(a) Whole dataset

(b) Data only with DISO terms

Figure 4.5: Accuracy of classifier selection methods at various training set proportion

group is modified to the score suitable for its type. For the latter, the semantic types are considered in the process of a lexicon construction. Compared to the classification methods that do not use UMLS, i.e., SWN and TL, the use of UMLS improved accuracy between 0.001 and 0.128. When appropriate classifiers are chosen by classification selection, the classification accuracy on most segment types are improved between 0.001 and 0.200. Using a composite classifier, the amount of a training set required to achieve a specific value of classification accuracy is less than using only one method on all segment types. To achieve the accuracy of 0.800, for example, CS(G_2,'Ac') and CS(G_2,'Fq') requires around 10 percent less than TL when the experiment is conducted on the segment type 'all sentences'.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This dissertation proposes methods for sentiment analysis at the aspect level It is divided into two parts. The first part develops aspect-based sentiment classification methods using the rhetorical structure theory (RST), and tests the methods on product reviews collected from on-line sources. The second part applies the classification methods to clinical text in electronic medical records (EMRs).

In the first part (Chapter 3), linguistic structural information is employed to improve classification accuracy of aspect-based sentiment analysis. A discourse structure is used for segmenting a set of candidate clauses relevant to a particular aspect, called a local aspect segment. The polarity score of the local aspect segment is obtained by aggregating scores of all clauses, according to their discourse relations. The results demonstrate that our approach outperforms an approach that ignores the linguistic structure. The score aggregation method considering the hierarchical structure of the segment, i.e., TWA, yields a better accuracy of 0.8% at most, compared to the method that ignores the structure, i.e., AEA.

By using rules induced from feature vectors representing RST structures, the classification accuracy is improved by approximately 4-7%, compared to a method using the average polarity score of relevant EDUs. Although the overall accuracy is improved by using the rule-based methods, the f-measure values for the negative polarity class

are decreased. To address the issue, three heuristic methods, i.e., combination with a baseline method, confidence threshold setting, and removal of conflicting rules, are applied. The combination with heuristics yields satisfactory classification results, without decreasing the f-measure values for the negative class.

In the second part (Chapter 4), clinical text from EMRs is studied. There are few issues making text analysis on clinical text more challenging, e.g., uncommon abbreviations or medical-specific meanings of terms. This dissertation focuses on a domain-specific meaning of a medical term. The Unified Medical Language System (UMLS), containing semantic types of terms in clinical text, is promising to solve the issue of domain-specific meanings. UMLS is exploited to improve lexicon-based sentiment analysis methods both using generic lexicons and using trained lexicons. For methods using generic lexicons, the polarity score of a term with a Disorders semantic type is modified to the score suitable for its type. For those using trained lexicons, the semantic types are considered in the process of a lexicon construction. Compared to the classification methods that do not use UMLS, the use of UMLS improved accuracy between 0.1% and 12.8%. When appropriate classifiers are chosen by classification selection, the classification accuracy values on most segment types are improved between 0.1% and 20.0%.

5.2 Future Work

Aspect-based polarity score aggregation presented in Chapter 3 considers only the topmost relation in a local aspect segment. Further work includes investigation of the usage of information about the deeper-level relations of the local aspect segment. A feature vector for rule-based classification, for example, may include deeper-level relation types.

To apply composite classification on clinical text in Chapter 4, an appropriate classification method is selected by considering the presence of an individual semantic type. Method selection by considering a combination of more than one semantic type may be studied.

Other representations in natural language processing, e.g., word embeddings [63,

64, 65, 66, 67], and classification methods based on artificial neural network [68] are promising to improve the sentiment classification accuracy. They may incorporate UMLS semantic types when they are applied on clinical text.

Bibliography

[1] B. Pang, L. Lee, *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[3] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 427–434, IEEE, 2003.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, ACM, 2004.

[5] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*, pp. 9–28, Springer, 2007.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[7] S. Kim, J. Zhang, Z. Chen, A. H. Oh, and S. Liu, "A hierarchical aspect-sentiment model for online reviews.," in *AAAI*, 2013.

[8] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 815–824, ACM, 2011.

[9] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based twitter sentiment for stock prediction.," in *ACL (2)*, pp. 24–29, 2013.

[10] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial intelligence in medicine*, 2015.

[11] Y. Niu, X. Zhu, J. Li, and G. Hirst, "Analysis of polarity information in medical text," in *AMIA Annual Symposium Proceedings*, vol. 2005, p. 570, American Medical Informatics Association, 2005.

[12] A. Sarker, D. Mollá-Aliod, and C. Paris, "Automatic prediction of evidence-based recommendations via sentence-level polarity classification," in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Nagoya, Japan*, pp. 712–718, 2013.

[13] A. Sarker, D. Mollá-Aliod, C. Paris, *et al.*, "Outcome polarity identification of medical papers," 2011.

[14] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health.," in *ICWSM*, pp. 265–272, 2011.

[15] M. Sokolova and V. Bobicev, "What sentiments can be found in medical forums?," in *RANLP*, pp. 633–639, 2013.

[16] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, *et al.*, "Psychological language on twitter predicts county-level heart disease mortality," *Psychological science*, vol. 26, no. 2, pp. 159–169, 2015.

[17] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Use of sentiment analysis for capturing patient experience from free-text comments posted online," *Journal of medical Internet research*, vol. 15, no. 11, 2013.

[18] Y. Deng, M. Stoehr, and K. Denecke, "Retrieving attitudes: Sentiment analysis from clinical narratives," in *Medical Information Retrieval Workshop at SIGIR 2014*, p. 12, 2014.

[19] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[20] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri, and C. Haruechaiyasak, "Discovering consumer insight from twitter via sentiment analysis.," *J. UCS*, vol. 18, no. 8, pp. 973–992, 2012.

[21] S. Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proceedings of the 25th International Conference on World Wide Web*, pp. 167–176, International World Wide Web Conferences Steering Committee, 2016.

[22] Y. Choi and C. Cardie, "Learning with compositional semantics as structural inference for subsentential sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 793–801, Association for Computational Linguistics, 2008.

[23] K. Sadamitsu, S. Sekine, and M. Yamamoto, "Sentiment analysis based on probabilistic models using inter-sentence information.," in *LREC*, 2008.

[24] S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor, "Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 170–179, Association for Computational Linguistics, 2009.

[25] R. Trnavac and M. Taboada, "Discourse structure and attitudinal valence of opinion words in sentiment extraction," in *LSA Annual Meeting Extended Abstracts*, 2014.

[26] K. Denecke and W. Nejdl, "How valuable is medical social media data? content analysis of the medical web," *Information Sciences*, vol. 179, no. 12, pp. 1870–1880, 2009.

[27] P. Sondhi, M. Gupta, C. Zhai, and J. Hockenmaier, "Shallow information extraction from medical forum data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 1158–1166, Association for Computational Linguistics, 2010.

[28] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, pp. 415–463, Springer, 2012.

[29] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[30] S. Moghaddam and M. Ester, "On the design of lda models for aspect-based opinion mining," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 803–812, ACM, 2012.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[32] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[33] P. Casoto, A. Dattolo, and C. Tasso, "Sentiment classification for the italian language: A case study on movie reviews," *Journal of Internet Technology*, vol. 9, no. 4, pp. 365–373, 2008.

[34] S. Rustamov, E. Mustafayev, and M. A. Clements, "Sentiment analysis using neuro-fuzzy and hidden markov models of text," in *2013 Proceedings of IEEE Southeastcon*, pp. 1–6, IEEE, 2013.

[35] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, vol. 6, pp. 417–422, Citeseer, 2006.

[36] C. Zirn, M. Niepert, H. Stuckenschmidt, and M. Strube, "Fine-grained sentiment analysis with structural features.," in *IJCNLP*, pp. 336–344, 2011.

[37] L. Polanyi and M. van den Berg, "Discourse structure and sentiment," in *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 97–102, IEEE, 2011.

[38] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[39] N. Sanglerdsinlapachai, A. Plangprasopchok, and E. Nantajeewarawat, "Exploring linguistic structure for aspect-based sentiment analysis," *Maejo International Journal of Science and Technology*, vol. 10, no. 2, p. 142, 2016.

[40] J. M. Chenlo, A. Hogenboom, and D. E. Losada, "Rhetorical structure theory for polarity estimation: An experimental study," *Data & Knowledge Engineering*, vol. 94, pp. 135–147, 2014.

[41] J. M. Chenlo and D. E. Losada, "An empirical study of sentence features for subjectivity and polarity classification," *Information Sciences*, vol. 280, pp. 275–288, 2014.

[42] A. Hogenboom, F. Frasincar, F. De Jong, and U. Kaymak, "Using rhetorical structure in sentiment analysis," *Communications of the ACM*, vol. 58, no. 7, pp. 69–77, 2015.

[43] H. Wachsmuth, M. Trenkmann, B. Stein, and G. Engels, "Modeling review argumentation for robust sentiment analysis.," in *COLING*, pp. 553–564, 2014.

[44] F. Wang and Y. Wu, "Exploiting hierarchical discourse structure for review sentiment analysis," in *2013 International Conference on Asian Language Processing*, 2013.

[45] J.-C. Na, W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, and Y.-L. Theng, "Sentiment classification of drug reviews using a rule-based linguistic approach," in *The Outreach of Digital Libraries: A Globalized Resource Network*, pp. 189–198, Springer, 2012.

[46] M. Z. Asghar, S. Ahmad, M. Qasim, S. R. Zahra, and F. M. Kundi, "Sentihealth: creating health-related sentiment lexicon using hybrid approach," *SpringerPlus*, vol. 5, no. 1, p. 1139, 2016.

[47] Y. Choi and C. Cardie, "Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 590–598, Association for Computational Linguistics, 2009.

[48] G. Demiroz, B. Yanikoglu, D. Tapucu, and Y. Saygin, "Learning domain-specific polarity lexicons," in *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 674–679, IEEE, 2012.

[49] L. Goeuriot, J.-C. Na, W. Y. Min Kyaing, C. Khoo, Y.-K. Chang, Y.-L. Theng, and J.-J. Kim, "Sentiment lexicons for health-related opinion mining," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 219–226, ACM, 2012.

[50] M. Z. Asghar, A. Khan, S. Ahmad, I. A. Khan, and F. M. Kundi, "A unified framework for creating domain dependent polarity lexicons from user generated reviews," *PloS one*, vol. 10, no. 10, p. e0140204, 2015.

[51] J. Carrillo-de Albornoz, J. R. Vidal, and L. Plaza, "Feature engineering for sentiment analysis in e-health forums," *PloS one*, vol. 13, no. 11, p. e0207996, 2018.

[52] D. A. Duverle and H. Prendinger, "A novel discourse parser based on support vector machine classification," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 665–673, Association for Computational Linguistics, 2009.

[53] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.," in *LREC*, vol. 10, pp. 2200–2204, 2010.

[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[55] J. Cendrowska, "Prism: An algorithm for inducing modular rules," *International Journal of Man-Machine Studies*, vol. 27, no. 4, pp. 349–370, 1987.

[56] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," 1998.

[57] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, vol. 1. Morgan Kaufmann, 1993.

[58] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning*, pp. 115–123, 1995.

[59] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*, pp. 231–240, ACM, 2008.

[60] Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "Automated rule selection for aspect extraction in opinion mining," in *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 1291–1297, AAAI Press, 2015.

[61] K. Labille, S. Gauch, and S. Alfarhood, "Creating domain-specific sentiment lexicons via text mining," in *Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, 2017.

[62] T.-T. Dang and T.-B. Ho, "Mixture of language models utilization in score-based sentiment classification on clinical narratives," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 255–268, Springer, 2016.

[63] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations.," in *hlt-Naacl*, vol. 13, pp. 746–751, 2013.

[64] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification.," in *ACL (1)*, pp. 1555–1565, 2014.

[65] Z. Zhang and M. Lan, "Learning sentiment-inherent word embedding for word-level and sentence-level sentiment analysis," in *Asian Language Processing (IALP), 2015 International Conference on*, pp. 94–97, IEEE, 2015.

[66] H. Du, X. Xu, X. Cheng, D. Wu, Y. Liu, and Z. Yu, "Aspect-specific sentimental word embedding for sentiment analysis of online reviews," in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 29–30, International World Wide Web Conferences Steering Committee, 2016.

[67] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214–224, 2017.

[68] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

Publications

International Journals

[1] Nuttapong Sanglerdsinlapachai, Anon Plangprasopchok, Ho Tu Bao and Ekawit Nantajeewarawat, "Rule-Based Polarity Aggregation Using Rhetorical Structures for Aspect-Based Sentiment Analysis", International Journal of Knowledge and Systems Science, vol. 10(3), pp. 44–60, 2019.

[2] Nuttapong Sanglerdsinlapachai, Anon Plangprasopchok and Ekawit Nantajeewarawat, "Exploring Hierarchical Linguistic Structure for Aspect-Based Sentiment Analysis", Journal of Internet Technology, vol. 18(4), pp. 945–952, 2017.

International Conference Proceeding

[3] Nuttapong Sanglerdsinlapachai, Anon Plangprasopchok and Ekawit Nantajeewarawat, "Exploiting Rhetorical Structures to Improve Feature-Based Sentiment Analysis", Proceedings of 18th International Computer Science and Engineering Conference (ICSEC 2014), pp. 180–185, 30 July–1 August 2014, Khon Kaen, Thailand.