Master's Thesis

Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives

1450207 - Dang Tran Thai

Supervisor: Prof. Ho Tu Bao
Main Examiner: Prof. Ho Tu Bao
Examiners: Prof. Mitsuru Ikeda
Prof. Tsutomu Fujinami
Assoc. Prof. Dam Hieu Chi

School of Knowledge Science
Japan Advanced Institute of Science and Technology

July, 2016

Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives

Dang Tran Thai(1450207)

School of Knowledge Science
Japan Advanced Institute of Science and Technology

September 2016

Keywords: Electronic Medical Records, clinical text, sentiment classification, linear combination, language models, negation, short text.

In recent years, the emergence of Electronic Medical Records (EMRs) opens an opportunity to improve the quality of healthcare and reduce medical cost. EMRs are well known as digitized medical records mostly created by doctors and nurses in hospital. EMR data is a rich and valuable resource including various data types such as digitized images, laboratory test, clinical text, in which the clinical text that contains information of patient health status such as symptoms, observations, physician's assessments plays an important role in EMRs exploitation. Therefore, EMRs exploitation mostly refers to clinical text exploitation.

Clinical text exploitation is still in infancy stage and poses a lot of challenges in analyzing and mining. As the clinical text is almost spoken language that is also called informal text, so it makes processing and mining on such kind of text become more challenging. The challenges can come from several features of this text such as: ungrammatical text and disambiguation of abbreviations; the shortness of text that makes classification become ineffective; implicit and vague expressions make sentiment analysis on clinical text to evaluate patient health status become more challenging; the text is strongly related to time. Moreover, we also have to face with the lack of annotated data, lexicon resources for higher level analysis such as information extraction, sentiment analysis, adverse drug reactions detection.

One of perspectives in clinical text exploitation that our study concentrates on is to evaluate patient health status through symptoms, diseases, conditions observed

during the treatment, and doctor's assessments which are noted in the clinical text. Patient health status evaluation can support doctor diagnosis and treatment. Moreover, it can be exploited for researches related to drug usage such as adverse drug reactions detection, drug repositioning. Evaluating the health status means that we have to determine whether observed symptoms, conditions, and physician's assessments are positive or negative. That inspires us to pose a problem of doing sentiment analysis on clinical text to evaluate patient health status.

Sentiment analysis that is also called opinion mining is a study field aiming to build methods for automatically analyzing people's opinions, sentiments, attitudes, emotions towards entities. Sentiment analysis is specified through many concrete problems in which document/sentence-level sentiment classification and aspect-based sentiment analysis are backbone in this study field.

Relying on similar points and different points between patient health status evaluation and the original sentiment analysis, we extend sentiment analysis for medical domain. In our initial study, we focus on sentiment classification on clinical text at sentence level. Doing sentiment classification on clinical text is significantly different from on normal text due to some specific features of the clinical text. In order to solve this problem, we have to face with four main challenges which are lack of domain-specific sentiment lexicon resources, implicit sentiment, various forms of negation, shortness of text.

In this thesis, we present our study of using a mixture of language models for score-based sentiment classification on clinical narratives. Our proposed method is a score-based classification method that can deal with the lack of sentiment lexicon resources, the variety of negation forms in clinical text, the shortness of text. The key idea is to use a linear combination of terms extracted from different language models to estimate an overall sentiment score of a sentence. Additionally, through using the linear combination, we derive a new vector representation called language-model-based representation that can help classification method work more effective on short text.

In conclusion, our study aims to build a groundwork for sentiment analysis on clinical text. We start with a backbone problem in sentiment analysis – sentiment classification. In order to effectively do sentiment classification on clinical text, we propose a score-based classification method that can deal with several challenges in this problem. This study initially reaches our proposed objectives.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor, professor Ho Tu Bao for his encouragement and support me during my master period. His advices and comments help me become more mature in both academic research and daily life.

Secondly, I would like to thank to the committee members, professor Dam Hieu Chi, professor Mitsuru Ikeda, professor Tsutomu Fujinami for your kind and useful comments on my research.

Thirdly, I would like to thank to my friends, especially all members in Ho Laboratory for your support and encouragement in my study and daily life. I have studied a lot through discussion with them.

Finally, I would like to express my gratitude to my family members who always stay by my side, encourage me during my study period at JAIST.

Dang Tran Thai

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem and objectives of our study

Electronic Medical Records (EMRs) are a valuable resource that contains rich and proper and believable medical knowledge ensured by doctors and nurses who are professional and have much experience in patient treatment. Exploiting EMRs is a new problem and poses many challenges, but is promising to make an innovation in heathcare. Although EMR data is in various types such as digitized image, text, signal, etc, the EMR exploitation mostly refers to analyzing clinical text. Clinical text almost comes from notes or narratives created by doctors and nurses during their patient treatment that contain information of symptoms, conditions, observations, assessments. Clinical text analysis demands to develop specific Natural Language Processing and Text Mining methods to adapt with such domain of text due to several characteristics of this text.

One of perspectives in clinical text exploitation is to evaluate patient health status through symptoms, diseases, conditions observed during the treatment, and doctor's assessments which are noted in the clinical text. This work plays an important role in supporting doctor diagnosis, treatment, and drug usage. To evaluate the health status, the essential point is to determine whether such symptoms, observations, or assessments are positive or negative, which inspires us to raise a problem of doing sentiment classification on clinical text to evaluate patient health status.

Doing sentiment classification on clinical text is significantly different from on normal text such as product-review data, etc due to several specific characteristics of the clinical text. The sentiments in clinical text are often implicitly expressed that requires medical knowledge to infer. Additionally, we often lack sentiment lexicon

resources for medical domain. Moreover, we also have to face with other problems such as various forms of negation used in clinical text, and the shortness of such kind of text. These problems make sentiment classification on the clinical text become more difficult that motivates our study.

Our study aims to propose a sentiment classification method that can work effectively on clinical text. This study requires us to overcome challenges in this task. More concretely, the target of our study is specified through three following objectives:

- Developing a method that can learn sentiment lexicon resources for medical domain.
- Constructing an effective classification method that can deal with problem of variety of negation forms.
- Finding out an effective representation to achieve high performance when classifying on short clinical text.

1.2 Thesis structure

The thesis consists of 4 chapters. The first chapter introduces about the problem and objectives of our research.

In chapter 2, we give an introduction of EMRs and clinical text, benefits of clinical text exploitation, and challenges of this work.

Chapter 3 gives a definition of original sentiment analysis or opinion mining, and two key tasks in this study field that are often performed on product-review data, one is document/sentence-level sentiment classification, and the other is aspect-based sentiment analysis. After that, relying on the original sentiment analysis on normal text, we introduce a new direction of sentiment analysis that performs on clinical text, and then make a comparison between the new direction and the old one. The comparison helps us raise problems for our study. In addition, we also present challenges in carrying out sentiment classification on clinical text because of specific features of such text.

In last chapter, we present our study of using mixture of language models for score-based sentiment classification on clinical narratives in detail. That focuses on dealing with three main problems, the first one is lack of domain-specific lexicon

resources, the second one is a variety of negation forms used in clinical text, and the third is short length of text. To solve these problems, we propose to use a linear combination of terms extracted from different language models to estimate the overall sentiment score of a sentence. Additionally, through using the linear combination, we derive a novel vector representation for short text which is called language-model-based representation.

Chapter 2

Electronic Medical Records and Clinical Text

Recently, the emergence of *Electronic Medical Records (EMRs)* opens an opportunity to improve the quality of health care such as diagnosis support, and post-market drug safety, and reduce medical cost. EMR data is a valuable and potential resource for exploitation, moreover, such data is created by doctors and nurses who have much medical knowledge and treatment experience, so the information has high quality and is more believable. EMRs consist of digitized images, laboratory tests, clinical text, in which clinical text is mostly focused on. The clinical text contains information of patient health status such as symptoms, diseases, conditions, and physician's assessments that provides a rich material for medical researches. Therefore, EMR exploitation mostly refers to clinical exploitation.

Clinical text exploitation is still in early stage and poses a lot of challenges in analyzing and mining. The clinical text is almost spoken language text that comes from notes of doctors and nurses, so it is almost informal. In addition, the content of clinical text is not explicitly expressed.

This chapter gives a brief introduction of EMRs, clinical text, and challenges in clinical text analytics in general.

2.1 What are Electronic Medical Records and Clinical Text?

Electronic Medical Records (EMRs) are well known as digitized medical records cre-

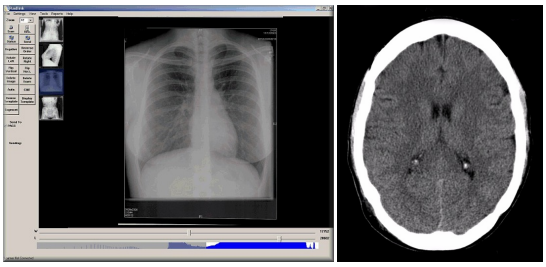
ated by medical organizations that deliver health care. Mostly, EMRs are created by doctors and nurses in hospitals during their patient's treatment.

EMRs include digitized images such as X-ray images and CT scan, medical laboratory test, and daily clinical notes/narratives of doctors and nurses, in which the clinical notes that contain *clinical text* are most valuable for exploitation. Figure 2.1 shows several examples of EMRs, in which Figure 2.1a shows X-ray images and CT scan, Figure 2.1b shows laboratory test, and Figure 2.1c shows daily clinical notes of doctors and nurses.

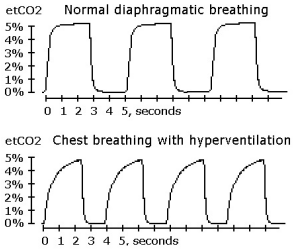
2.2 Clinical text – a valuable resource for health care innovation

EMR data exploitation, especially clinical text exploitation, is potential to open an opportunity for health care innovation. Clinical text is a large and worthy resource that contains information of patient's symptoms, observations known as phenotype information, and physician's assessments. This text is noted by doctors and nurses who have medical knowledge and experience in patient treatment, so the information in clinical text is more sufficient and believable. That makes the EMR data become more valuable than other medical data sources such as patient spontaneous reports or medical social data. Clinical text exploitation plays an important role in improving diagnosis support, and post-market drug safety, and drug repositioning. For the diagnosis support, the clinical text is collected from a large amount of patients that provides a rich amount of evidences to support physicians in making their decisions. In post-market drug safety, symptoms, observations of adverse drug reactions, and new indications of drugs noted in detail in clinical text help us recognize abnormalities in drug usage and new uses of drugs.

In order to support clinical data mining, several databases were created that collect EMRs from hospitals or medical centers. One of well-known databases is the *MIMIC II database* [40]. Data of more than 30,000 patients in the MIMIC II database was collected from Beth Israel Deaconess Medical Center in Boston, USA from 2001 to 2008. This data is organized into tables that store apart of personal information of patients and treatment information of each patient related to diseases patients got, drug usage, symptoms, doctor's assessments, laboratory test results, etc. For privacy protection, the data in this database was de-identified that removes sensitive personal information of patients. The MIMIC II database has been used in many data-driven medical studies.



a)



b)

3,2075,,,2682-09-07 00:00:00 EST,,,,,"DISCHARGE_SUMMARY",,"

Admission Date: [**2682-9-7**]
Discharge Date: [**2682-9-18**]
Date of Birth: [**2606-2-28**] Sex: M

Service: Medicine

CHIEF COMPLAINT: Admitted from rehabilitation for hypotension (systolic blood pressure to the 70s) and decreased urine output.

HISTORY OF PRESENT ILLNESS: The patient is a 76-year-old male who had been hospitalized at the [**Hospital1 3007**] from [**8-29**] through [**9-6**] of 2002 after undergoing a left femoral-AT bypass graft and was subsequently discharged to a rehabilitation facility.

On [**2682-9-7**], he presented again to the [**Hospital1 3087**] after being found to have a systolic blood pressure in the 70s and no urine output for 17 hours. A Foley catheter placed at the rehabilitation facility yielded 100 cc of murky/brown urine. There may also have been purulent discharge at the penile meatus at this time.

DATE: [**2682-8-24**] 6:02 PM
CHEST (PORTABLE AP)
Reason: please check picc tip placement. #4f, sl, v-cath for abx. pl

---

UNDERLYING MEDICAL CONDITION:
76 year old man with cellulitis lower extremity.
REASON FOR THIS EXAMINATION:
please check picc tip placement. #4f, sl, v-cath for abx. please page with stat wet read to beeper # [**Pager number 1698**]. thanks

---

FINAL REPORT
INDICATIONS: 76 y/o male with lower extremity cellulitis, check PICC line placement. Being presented for evaluation on [**2682-8-25**]. This study is compared with the prior exam of [**2682-7-10**].

AP PORTABLE CHEST: A left sided PICC line is seen with its tip in the atriocaval junction. There is no evidence of pneumothorax. The heart is normal in size. The mediastinal and hilar contours are unremarkable. The pulmonary vascularity is normal. There are no pleural effusions. The lung fields are clear. The soft tissue and osseous structures are unremarkable.

IMPRESSION: Satisfactory placement of a left PICC line. No evidence of pneumothorax.

c)

Figure 2.1: Examples of Electronic Medical Records

Clinical text exploitation poses several challenges for Natural Language Processing (NLP) and Text Mining due to some characteristics of such kind text that is discussed in next section.

2.3 Challenges in clinical text analytics

As mentioned in previous section, EMRs exploitation concentrates on analyzing clinical text including various NLP and Text Mining tasks as follows:

- Medical phrases identification.
- Spelling and grammatical errors correction.
- Word/abbreviation disambiguation.
- Medical concepts recognition-Name Entity Recognition (NER).
- Clinical text representation.
- Relation, temporal information extraction.
- Sentiment analysis for evaluating patient's health status.

Several specific characteristics of clinical text make the the tasks mentioned above become more challenging. To promote the development of NLP methods to solve several tasks in analyzing clinical text, challenges such as I2B2 Challenges for English since 2006 and NTCIR Challenges for Japanese since 2013 have been given for researchers. Several features of clinical text that pose challenges in analyzing are as below.

Clinical text is not grammatical and contains lot of abbreviations that make syntactic parsing and concepts understanding more difficult. As clinical text often comes from quick notes of doctors and nurses, so the grammar is not strictly care. Additionally, due to the quick notes, the doctors and nurses usually use abbreviations. In clinical text analysis, abbreviation restoration is also a big problem because of the disambiguation of concepts corresponding to the abbreviation. For example, the abbreviation "BPS" can stand for several concepts "Blood Pressure", "Beats Per Second", "Bisphosphonates", and "Behavioral Pain Scale". Methods implemented in existing medical ontologies such as MetaMap, MedLEE, cTAKEs are not effective enough to deal with the abbreviation disambiguation [68].

We also have to face with lack of annotated data, lexicon resources for higher level analysis such as temporal information extraction, sentiment analysis on clinical text to evaluate patient health status, adverse drug reactions detection, drug repositioning, etc. EMRs analytics is a new problem and still in infancy stage, so the annotated sets are not available and have been separately built by research groups. Therefore, we lack official benchmark annotated sets for fairly evaluating methods. Moreover, building annotated data sets is also costly and time-consuming and requires lots of involvement of medical experts.

Beside two challenges mentioned above, the shortness of clinical text also makes a difficulty in text classification. The shortness causes a problem of text representation that does not provide enough word co-occurrence for good similarity measures [58]. In our work of sentiment classification on clinical text, the shortness of text is an essential challenge we need to overcome to achieve high performance that will be discussed in detail in next section.

Another feature of clinical text is to contain lots of implicit and vague expressions, various negation forms that pose some problems of doing sentiment analysis on clinical text to evaluate patient's health status. Different from evaluating a product that customers can directly give their positive or negative comments, in clinical notes, doctors just note observed symptoms, and several preliminary assessments instead of immediately make a conclusion of patient's health status as positive or negative. That makes content of clinical text are almost descriptions instead of opinions. However, when doing sentiment analysis on such kind of text, we must infer the patient status as positive or negative based on such descriptions and medical knowledge. Moreover the problem of various forms of negation used in this text that is discussed in detail in chapter 3 also makes the sentiment classification on such text become more challenging.

The last characteristic of clinical text is strongly related to time. As doctors and nurses create clinical note each time they visit patients, so clinical notes keep information of patient's health status over time. Therefore, clinical notes are considered time series data that poses some challenges in time-series data mining. Moreover, it also poses challenges in temporal events and time extraction [35].

Chapter 3

Sentiment analysis

Sentiment analysis, also called *opinion mining*, is a study field that aims to build a class of methods for automatically analyzing people's opinions, sentiments, evaluations, attitudes, emotions towards entities such as products, services, organizations, events, topics, etc. Therefore, originally, most of methods for sentiment analysis have been developed to work on product-review data that is almost comments/posts expressing evaluations and opinions related to movies, mobiles, computers, etc collected from websites, social network.

Sentiment analysis can be extended for medical domain instead of only focusing on product-review domain. The original sentiment analysis on product-review data inspires our research that aims to evaluate patient's health status in clinical text to support doctor treatment or researches regarding drug side effects detection and drug repositioning. However, doing sentiment analysis on clinical text is not as same as that on the product-review data due to several differences between two kinds of text. Moreover, this topic is a new direction in sentiment analysis and still in early stage, several specific characteristics of clinical text pose some challenges in finding appropriate solutions adapting with the clinical text.

This chapter includes three sections, the first one is to make a brief review of the original sentiment analysis on product-review data, the second one is to introduce about a new direction of sentiment analysis that works on clinical text, and the last one shows challenges we have to face with when carrying out this task. Moreover, through these sections, people can see the similar points and different points of sentiment analysis on product-review domain and medical domain that is the inspiration for our study.

3.1 Sentiment analysis on product-review text

As a demand in industry, business, politics that companies or organizations would like to gather a large mount of people opinions regarding their products or events, political strategies, etc then analyze such data to discover valuable information for offering their decision making. That partially promotes the rapid development of sentiment analysis (opinion mining), especially focusing on product-review data. In this section, we present a definition of sentiment analysis, and several typical tasks in this study field with existing related studies.

3.1.1 Definition

As the definition in [44], sentiment analysis that is also called opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. In other word, the essential target of sentiment analysis is to automatically determine whether people's opinions are positive or negative.

Sentiment analysis is a challenging problem in Natural Language Processing (NLP) field. In general, it is mostly carried out in 3 levels of text as follows:

- Document level: That investigates whether the general opinion in a document is positive or negative.
- Sentence level: That investigates the sentiment orientation of each sentence.
- Aspect/feature level: That evaluates each aspect of a product.

Sentiment analysis touches many problems in NLP such as: text classification, name entities recognition, negation handling, word disambiguation. Those are challenging NLP tasks.

For text classification, as mentioned above, it investigates people's opinions, so it is necessary to classify such opinions into two classes positive and negative. This task is also called *sentiment classification* which is the backbone of sentiment analysis.

In aspect level analysis, before evaluating aspects of a product, we have to recognize such aspects mentioned in text that forms the problem of Name Entities Recognition (NER). NER is not a completely solved problem, there still has several challenges to overcome for achieving high performance.

As sentiment words such as “good”, “interesting”, “beautiful”, “like”, etc are essential evidences for identifying the sentiment orientation of documents or sentences, however the opinion is also indirectly expressed or there are negation words in expressions that make the sentiment orientation inverse. Those pose two challenging problems of negation handling and word disambiguation.

Sentiment analysis has many real-life applications, particularly in business and industry. Along with the rapid growth of social media that provides a huge amount of data, it is a power tool for supporting decision making. For example, in industry, for customers, sentiment analysis processes and synthesizes feedbacks of products from previous users to discover the opinion trend related to those products that they can base on to decide if they should buy or not. These feedbacks also help producers improve the quality of their products and services.

3.1.2 Typical tasks in sentiment analysis

Sentiment analysis is specified in various tasks such as document-level and sentence-level sentiment classification, aspect-based sentiment analysis, opinion summarization, opinion spam detection. In such tasks, *document/sentence-level sentiment classification*, and *aspect-based sentiment analysis* are two key problems which play a role of groundworks for other problems in sentiment analysis.

Document/sentence-level sentiment Classification

Sentiment classification is the backbone of sentiment analysis that is carried out for both document and sentence level to determine whether the documents/sentences express positive or negative opinion. Different from document-level, in sentence-level classification, before doing sentiment classification, people have to determine if a sentence contains opinions or not that is called subjectivity classification. After that, sentences identified as containing opinions will be classified into positive or negative ones.

In general, sentiment classification can be described as Figure 3.1. In this framework, the sentiment classification includes two main components: Document/sentence representation, and using a predict model to assign the sentiment labels. The labels are mainly positive and negative, sometimes, people consider more neutral one.

Regarding methods for sentiment classification, in [45] and [12], the authors made a review of sentiment classification techniques that follow three main approaches:

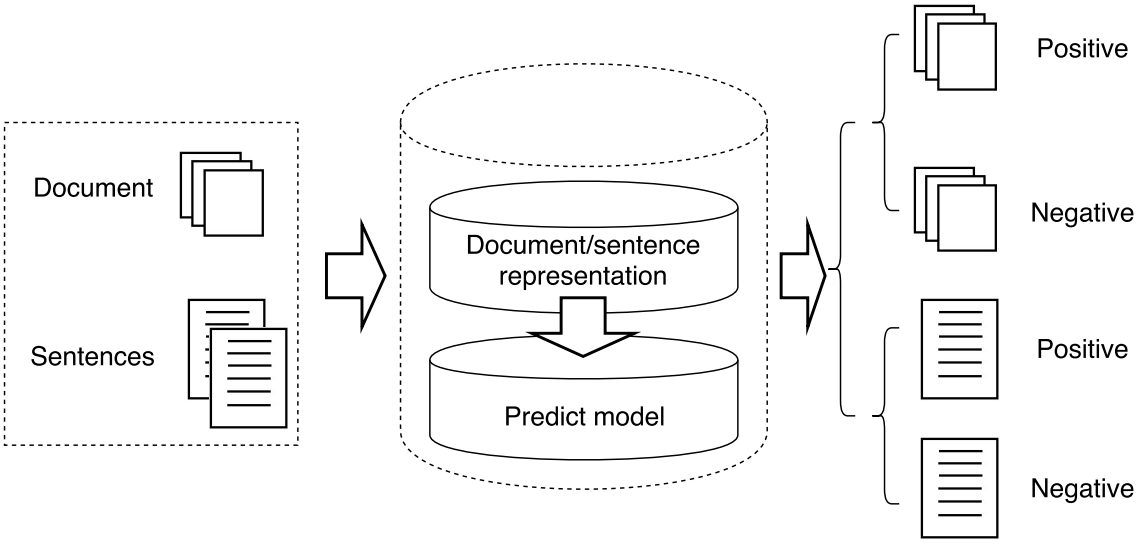


Figure 3.1: General framework of sentiment classification.

lexicon-based approach, *machine learning-based approach*, *hybrid approach*. In these approaches, machine learning approach including supervised learning, unsupervised learning, and semi-supervised learning is more popular and has many related works.

The *lexicon-based approach* essentially bases on an available dictionaries of sentiment words/phrases with a score measuring strength of association between such words/phrases and the sentiment label of sentence/document, which is called opinion lexicon. The overall sentiment score of sentence/document that is used to decide the label is aggregated from score of words and phrases. For example, in [60], Taboada *et al.* estimated sentiment score of a document by using sentiment word dictionary incorporating with negations. In addition, Turney [64] predicted sentiment labels by average semantic orientation of the phrases containing adjectives and adverbs. The semantic orientation of a phrase is measured by Point Mutual Information (PMI). In a similar work, Dave *et al.* [17] also summed up the scores of all terms belonging to the document to determine the label.

Several lexicon resources were built to offer sentiment classification such as Harvard Inquirer [59], Micro-WNOp [13], and SentiWordNet [4]. Harvard Inquirer source consists of words and their corresponding categories in which two categories "Positiv" (positive) and "Negativ" (negative) account for a majority of words, besides, there are additional categories used to give more semantic information of the words. In contrast, Micro-WNOp and SentiWordNet resources give both positive and neg-

ative scores for each word instead of assigning an unique categories for the word because the word can appear in different context in both positive and negative document/sentence. The score estimation essentially bases on the frequency of words in positive and negative documents/sentences in an annotated dataset. Therefore, a word can have both non-zero positive score and negative score because it can appear in both positive and negative sentences/documents.

The lexicon resources just contain a limited number of sentiment words while we have to do sentiment analysis on various datasets belonging to many domains that can contains new sentiment words, thus to adapt with a new domain, the lexicon dictionaries must be updated by adding the new sentiment words or updating the sentiment scores. Several methods were proposed to automatically extract new sentiment words. In [31], to extract new sentiment words, Huang *et al.* proposed an unsupervised data-driven framework and design statistical measures to estimate the possibility of a word being a new sentiment word. In addition, Saif *et al.* [56] proposed a lexicon adaptation approach that uses the contextual semantics of words to capture their contexts in tweet messages and update their corresponding sentiment orientations and sentiment scores.

The *machine learning-based approach* is a popular approach for sentiment classification. It includes two key steps, one is feature extraction, the other is to build a predictive model. The feature set is mostly built by using lexicon resources and linguistic information.

Various predictive models were proposed to predict the sentiment label of a document or sentence in problems of subjectivity classification and sentiment classification. In [22], Ding *et al.* used words surrounding product feature to determine opinion orientation on the product feature. They also combined multiple opinion words to arrive at final decision and integrated negation rules to handle context-dependent opinions. In similar work, Kim et al. [37] used maximum entropy with lexicon features for this task. For the subjectivity classification problem, patterns associated with objectivity were used as features for Naive Bayes classifier [66]. Hidden Conditional Random Fields (HCRFs) was used for sentence level classification in [61]. In [62], semi-supervised latent variable models were utilized to combine coarse-grained and fine-grained supervision benefits for sentence-level classification. In other work, Agarwal *et al.* [1] represented a tweet in tree form by using tree kernel. Gamon used Support Vector Machine (SVM) with lexicon features in [24]. McDonald *et al.* [46] investigated the sequence of several techniques for classification. Benamara *et al.* [6] proposed new subjectivity classification at segment level that is more appro-

priate for discourse-based sentiment analysis. In addition, Hassan *et al.* [27] proposed a graphical model using lexical items, part-of-speech tags, dependency relations to determine the attitude of participants in an online discussion.

Feature extraction, which is known as text representation, plays an important role to obtain high performance in sentiment classification. Mostly, documents/sentences are represented by vectors based on several linguistic features as follows:

- Lexicon dictionaries: that consists of sentiment words and phrases such as "good", "wonderful", "bad", "poor", etc [54], [28], [30].
- Term and their frequency: This kind of feature is related to n-grams model such unigram (individual words), bigram (sequence of two adjacent words), etc [50].
- Part-of-speech: As adjectives and verbs or several nouns often used to express opinions, so part-of-speech is also important in feature extraction [30], [5].
- Syntactic dependency: Words dependency-based features are generated through syntactic parsing process [67], [5].

In addition, other information such as hashtags and smileys is also used [18]. Moreover, algorithms for automatically learning (includes determining and extracting) such kinds of feature were proposed. In [55], Riloff *et al.* utilized a bootstrapping process for learning and extracting linguistic rich patterns for subjective expressions. In another work, Wiebe [65] used a method for word clustering according to distributional similarity to identify strong clues of subjectivity.

As the strategy of supervised method is to assume that the training data set and the test set share the similar distribution that means the difference between two these sets will make the performance decrease. However, in fact, test set or training set can belong to different domain or come from different sources that leads to their different distribution. To overcome this drawback, the training should be updated to adapt with the test set that demands to develop *hybrid methods* also called semi-supervised methods or cross-domain sentiment classification [3], [9].

Aspect-based sentiment analysis

Document-level and sentence-level sentiment classification mentioned above are not sufficient for applications because they just analyze general opinions of a product instead of each attribute of such product. For example, a document can evaluate a

product as positive, but it does not means all aspects of this product are also positive. Therefore, to completely evaluate, we need to determine aspects and do sentiment classification according to each aspect. That is called *aspect-based sentiment analysis*.

In general, aspect-based sentiment analysis consists of two steps, one is *aspect extraction*, and the other is *aspect sentiment classification*. As aspect sentiment classification is similar sentence-level or clause-level classification, thus we concentrate on presenting aspect extraction step.

Aspect extraction is well-known as Information Extraction (IE) or Name Entity Recognition (NER). The aspect can be explicitly or implicitly expressed. For example, we consider two following sentences:

1. The battery life of this iphone is long.
2. This car is expensive.

In sentence 1, the aspect "battery life" is expressed explicitly while in sentence 2, "expensive" is a sentiment word indicating the aspect "price", but this aspect is not directly mentioned in the sentence. Most of existing works just focus on explicit aspect.

In [44], the authors mentioned 4 main approaches for aspect extraction as follows:

- Extraction based on frequent nouns and noun phrases
- Extraction by exploiting opinion and target relations
- Extraction using supervised learning
- Extraction using topic modeling

For Extraction based on frequent nouns and noun phrases, Hu *et al.* [30] extracted nouns and noun phrases by using part-of-speech tagger, their frequencies are counted, then a frequency threshold is used to select nouns and noun phrases having high frequency. This algorithm was improved in the work [52] by removing imprecise discovered noun phrases. The retrieved noun phrases will be evaluated by estimating Pointwise Mutual Information (PMI) between those and known words indicating predefined aspects. In addition, in [7], Blair-Goldensohn *et al.* obtained frequent nouns and noun phrases that are inside syntactic patterns indicating sentiments.

The key idea of extraction by using opinion and target relations is that as opinions usually target to an aspect, and opinions are indicated through sentiment words

that we have already known, so the aspect detection can be based on the sentiment words. For example, in the sentence "The battery life is long", the word "long" is a sentiment word that targets to the aspect "battery life". Most of methods following this approach utilized syntactic parser to discover dependencies between sentiment words and their corresponding aspects [69], [57].

Since aspect extraction is considered a NER or IE problem, various supervised learning methods can be applied to solve. Several methods based on sequential labeling (sequential learning). Jin *et al.* [34] used a lexical Hidden Markov Model (HMM) model to learn patterns to extract aspects and opinion expressions. Jakob *et al.* also used Conditional Random Field (CRF) for the same target. Additionally, other supervised methods were also utilized. In [39], the authors first found candidates being pairs between aspects and opinions word using dependency tree, and then employed a tree-structure classification method to identify whether each candidate is a relation between an aspect and its corresponding evaluation or not.

The recent approach for aspect extraction is to utilize statistical topic models. Topic modeling is an unsupervised learning method that assumes a document is a mixture of topics, and each topic is a distribution of words. The output topic is a cluster of words that are grouped by considering the words co-occurrence probability. Technically, topic models are Bayesian network.

In aspect-based sentiment analysis, topics can be considered as aspects. However, when using topic models to generate topics, the output topics include both aspect an sentiment words, and we need to separate it. Therefore, general topic models such as Probabilistic Latent Semantic Analysis (pLSA) [29] and Latent Dirichlet allocation (LDA) [8] are not enough for sentiment analysis. Several works tried using topic models for aspect extraction. Mei *et al.* [47] proposed a joint model for sentiment analysis that was based on pLSA. In [63], the authors showed that general topic models such as LDA may not be appropriate for detecting aspects. The reason is that LDA depends on topics distribution differences and co-occurrence of words in a document, but topics in opinion documents are homogenous that makes LDA ineffective, thus they proposed multigrain topic models. In addition, Lin *et al.* [43] proposed a joint-sentiment model by extending LDA, however, aspect words and sentiment words were still not explicitly separated. In [48], a semi-supervised joint model was proposed which allows the user to provide some seed aspect terms for some topics to guide the inference.

Table 3.1: Examples of patient health status evaluation based on clinical text

	Sentence	label
1	There is moderate cardiomegaly.	Negative
2	Restless and agitated most of the night	Negative
3	He was also complains of shortness of breath.	Negative
4	There has significant improvement in pleural effusion.	Positive
5	There is no evidence of pleural effusion.	Positive
6	There has been marked decrease in right pleural effusion.	Positive
7	Less nauseous than previous	Positive

3.2 Sentiment analysis on clinical text

The key point to answer the question that why we do sentiment analysis on clinical text? is to find similar points between the original sentiment analysis and the extended one for clinical text. These points can be in the purpose, and the problems posed.

As mentioned in chapter 2, clinical text is a valuable resource that reflects patient health status through observations of symptoms, progress in treatment, abnormalities, and physician's assessments which are called phenotype information. Therefore, determining such observations and assessments as positive or negative or neutral towards a disease or a drug or combination of drugs plays an important role in supporting doctor treatment. Moreover, in researches related to drug producing and using, evaluation of phenotype information gives significant evidences for adverse drug reaction detection, drug repositioning. Table 3.1 shows some examples of clinical text that gives information of patient health status with its corresponding evaluation as positive or negative. Thus, the purpose of patient health status evaluation is similar to sentiment analysis.

Classifying patient health status can be done on both documents (notes) or sentences that is equivalent to document-level and sentence-level classification. Moreover, when evaluating patient status towards a disease or after using a drug or combination of drugs, doctors often base on many criteria (aspects). For examples, to make a conclusion of health status of a patient who got heart disease, several aspects such as heart rate, blood pressure, characteristics of ventricles, etc are considered. That also forms a aspect-based analysis problem like the original sentiment analysis.

Although there are many similar points between sentiment analysis on product-

review data and clinical text, the methods for to solve them are different because of several specific characteristics of clinical text. That poses some challenges which will be discussed in section 3.3.

3.3 Challenges in sentiment analysis on clinical text

Sentiment analysis on clinical text is a new problem and still in early stage. There are several works attempting to do that on text come from doctor/nurse narratives and medical forums. Ali *et al.* [2] applied the methods such as Naive Bayes, Support Vector Machine (SVM), Logistic-R to classify the posts in medical forums. Additionally, SVM and Naive Bayes were also used in [14] to determine the watchlist of drugs as positive or negative in drug surveillance. In [21], Deng *et al.* applied dictionary-based method to classify nurse letters, radiology reports in the MIMIC II database. Moreover, they also presented some difficulties when doing classification on such data set due to the nature of clinical text such as implicit description of health status, small number of sentiment words used, etc. Besides, Na *et al.* [49] did clause-level sentiment classification using pure linguistic approach.

Most previous works mentioned above are not good enough to achieve high performance because their methods are just simple methods for general sentiment analysis that are not appropriate for sentiment classification on clinical text. Sentiment analysis on clinical text is significantly different from that on general text due to some specific features of clinical text. Therefore, to create suitable methods for sentiment classification on clinical text, it is necessary to investigate the nature of such kind of text that poses challenges in analyzing. These challenges will be discussed in detail in subsections 3.3.1, 3.3.2, 3.3.3, 3.3.4.

3.3.1 Lack of domain-specific lexicon resources

Since sentiment classification analysis on clinical text is still in early stage, it lacks specific sentiment lexicon resources for medical domain. For product-review data, people have already created several lexicon resources such Micro-WNOp and SentiWordNet to offer analyzing. However, to adapt with the medical domain, these general lexicon resources must be extended by updating sentiment words from medical literature and clinical documents.

Several works attempt to make an extension of existing sentiment lexicon resources to adapt with a specific domain. In [25], the authors merged terms from

SentiWordNet and Subjectivity Lexicon. After that, they extracted opinion terms from drug reviews then updated to the merged lexicon resource. In similar work, Deng *et al.* [20] also created an extension of the well-known subjectivity lexicon. In [23], Du *et al.* proposed an adapted information bottleneck method for constructing a domain-orientated sentiment lexicon.

As mentioned in [19], different from sentiment words used in general text that are almost adjectives, the sentiment words in clinical text are nouns that indicate concepts of symptoms, medical conditions, diseases (ex. "sick", "cough", "pleural effusion", etc). Therefore, in order to address this problem, lexicon resources for medical domain need to link these concepts to sentiment. To do so, we need to add sentiment information for each medical concepts in fundamental medical ontologies such as Unified Medical Language System (UMLS) [11].

3.3.2 Implicit sentiment

Relying on the analysis presented in [19], Denecke *et al.* confirmed that clinical text differs from social media data in term of word usage. In social media data, most of sentiments, opinions, or evaluations are explicitly expressed through adjectives (ex. "good", "bad", etc) and verbs (ex. "like", "hate", etc). In contrast, in clinical text, sentiment is often implicit that requires a inference of concepts based on medical knowledge. The implicit sentiment is descriptions of patient's health status, symptoms such as "severe pain", "high blood pressure". Besides, clinical text also contains explicit sentiment in doctor/nurse assessment about the patient health status such as "well", "normal".

In [21], the authors showed that one of difficulties in sentiment classification on clinical narratives is implicit sentiment detection. A simple solution is to build sentiment medical lexicon resources mentioned in subsection 3.3.1. Besides, we also proposed a supervised method that learns implicit sentiments in clinical text. This method is described in chapter 4.

3.3.3 Various forms of negation in clinical text

The negation is also called *sentiment shifter* that is known as expressions used to change the sentiment orientation of a sentence. For example, in the sentence:

"This car is not expensive."

If this sentence does not contain the word "not" it will be negative, however this word make the sentence become positive. That means the orientation of sentence is changed.

The sentiment shifters in product-review data are explicit, which are just negation words such as "not", "no", "don't", etc. In contrast, the clinical text contains descriptions of patient's health status, to express the improvement of patient status, nurses or doctors often use the negation of symptoms or negative observation, however, the negation is in various forms not only negation words. For instance, we consider the following sentences and clauses:

1. There has significant improvement in pleural effusion.
2. There is no evidence of pleural effusion.
3. There has been marked decrease in right pleural effusion.
4. Less nauseous than previous.

The example shows various forms of sentiment shifters in clinical text. In sentence 1, 2 the sentiment shifters are a negation word ("no") and a strongly positive word ("improvement") respectively. Moreover, in sentence 4, the sentiment shifter is a phrase "less nauseous". In this case, "less" can be positive or negative in different contexts while "nauseous" is negative, and the sentiment of the sentence that is positive strongly depends on the phrase instead of each individual word. Besides, sentiment shifters are not only phrases but also sequences of non-adjacent words as "decrease...pleural effusion" like sentence 3.

The problem of sentiment shifters was mentioned and solved by several methods on product-review domain. Such methods follow one of two main approaches, one is negation words and scope of the negation detection, the other is simple voting for overall sentence's sentiment score by word/phrase scores.

For the first approach, as sentiment shifters in product-review are almost negation terms, so several works attempt to detect such terms, and the scope of negation in the sentence. In [51], Polanyi *et al.* described how the base attitudinal valence of lexical item can be modified by context and proposed a simple "proof of concept" implication for some context shifters. In other work, Li *et al.* [41] presented a shallow semantic parsing approach to learn the scope of negation. The effect of valence shifters on classification was examined in [36]. The parser and some heuristic rules was used to identify the scope of negation [33]. In [42], Li *et al.* proposed a feature

selection method to generate scale polarity shifting training data, and a combination of classifiers to improve the performance.

There are few works following the second approach. Dave *et al.* [17] made a simple voting for deciding sentiment label by summing scores of words and phrases. Ikeda *et al.* [32] proposed a method that models polarity shifters better than simple voting by sentiment word method. In [38], Kiritchenko *et al.* determined the sentiment scores of words in the presence of negation by detecting negation context via computing two scores of term in two parts: affirmative context, and negated context.

The first approach often gives a better performance than the second one due to the intensive analysis of word contexts while the second one is more flexible because of the specific language independence. However, to deal with the problem of various negation forms, the first approach seems to be not effective because it is difficult to exactly capture all variants of sentiment shifters. Therefore, the second one is more appropriate, but it requires some modifications to enhance word's contexts considering instead of individually aggregating scores at word-level or phrase-level. For example, the word "improvement" is a strong positive word, so its score can dominate the other and rule the sentence's score while the word "less" may not due to a weaker positive sense. However, the phrase "less nauseous" with more positive purity volume can make a bigger influence on the sentence's score. It helps to raise an hypothesis that the sentence score does not separately depends on word or phrase scores. Thus we can simultaneously sum up word and phrase scores by a linear combination in which the coefficients characterize how words and phrases affect the sentence sentiment orientation. That is presented in detail in next chapter about our proposed method for sentiment classification on clinical text.

3.3.4 Shortness of clinical text

The short text causes a problem of text representation that requires a particular representation method instead of normal methods such as bag-of-words, or bag-of-n-grams. The reason is that the short length of text does not provide enough word co-occurrence or shared context for good similarity measures [58]. That means sentences have similar meaning but contain different words. For example, consider two following sentences s_1, s_2:

1. s_1: There has been marked decrease in right pleural effusion.

2. s_2: Free fluid volume in right lung reduces.

Two sentence contain 14 words in total as {"there", "has", "been", "marked", "decrease", "in", "right", "pleural", "effusion", "free", "fluid", "volume", "lung", "reduces"}, in which there are some words are synonyms or have similar meaning (belonging to the same topic). In this example, we can group similar meaning words into topics as follows:

- Topic 1: {decrease, reduces} (Those are synonyms)
- Topic 2: {pleural, lung}
- Topic 3: {effusion, fluid}

Two sentences are represented by two vectors, and their size is the number of words in vocabulary, in this case the size is 14. Each element of the vector indicates a word in the vocabulary, if this word appear in the sentence, the corresponding value of this element will be 1 and vice versa. Therefore, two sentences in the example have the representation as follows:

- $s_1 = (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0)$
- $s_2 = (0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 1)$

Assume that we measure the similarity of meaning of two sentences by considering the Euclid distance between two vectors by following equation.

$$d(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + ... + (u_n - v_n)^2} \tag{3.1}$$

Thus $d(s_1, s_2) = \sqrt{12}$. However, two sentences contain words that have similar meaning, so we can consider that such words appear in both sentences. For example, we consider three words "reduces", "lung", "fluid" appearing in sentence s_1 because their meanings are similar to "decrease", "pleural", "effusion" respectively. Similarly, "decrease", "pleural", "effusion" are considered to appear in sentence s_2. Therefore, two sentences are re-represented as the following:

- $s_1 = (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1)$
- $s_2 = (0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$

The similarity between two sentences is also re-computed as $d(s_1, s_2) = \sqrt{6}$ which shows that these sentences are more similar. Thus, the new representation reflects the sentence meanings better than the previous one.

Through this example, we see that although sentences s_1, s_2 contain many different words, their meanings are similar because several words belong to the same topics and can replace each other. Thus, the representation of short text often bases on the appearance of topics to enhance the word co-occurrence in such sentences instead of just considering the appearance of words in the sentences. To discover latent topics of words, probabilistic topic models are commonly utilized. Therefore, LSA, pLSA, LDA have been widely applied in short text representation [58], [53], [15]. Besides, PMM-based classifier based on conditional probabilities of upcoming symbol given several previous symbols was applied for topic and non-topic classification [10]. Dai *et al.* [16] proposed cluster-based representation method named CREST to deal with the shortness and sparsity of text.

Chapter 4

Mixture of language models for sentiment classification on clinical narratives

As sentiment classification is a backbone of sentiment analysis, on clinical narratives, it plays a role of groundwork to analyze patient's health status, medical condition and treatment. The work posed challenges due to the shortness, and implicit sentiment of the clinical text.

In this chapter, we show our study of sentiment classification on clinical text that focuses on dealing with three problems, the first one is lack of domain-specific lexicon resources, the second one is various forms of negation in clinical text, and the last one is shortness of clinical text. Our study shows that a sentiment score of a sentence simultaneously depends on scores of its terms including words, phrases, sequences of non-adjacent words, thus we propose to use a *linear combination* which can incorporate the scores of terms extracted by various language models with the corresponding coefficients for estimating the sentence's score. Through utilizing the linear combination, we derive a novel vector representation of a sentence called *language-model-based representation* that is based on average scores of kinds of term in the sentence to help supervised classifiers work more effectively on the clinical narratives.

4.1 Problem and challenges of sentiment classification on clinical narratives

As clinical narratives reflect the patient's health status through observations of symptoms, progress in treatment, and physician's assessment, determining such observations and assessments as positive or negative or neutral towards a disease plays an important role in therapeutic assistance and abnormality recognition.

The text in clinical narratives has several particular characteristics that pose some challenges for sentiment classification on such kind of text mentioned in chapter 3. In our work, we have to face with 3 main challenges as follows:

- Lack of medical-domain sentiment lexicon resources.
- The diversity of sentiment shifters used in clinical text.
- The shortness of clinical text.

As a fact that symptoms, observations reflecting patient's health status are often expressed in a sentence or clause, so we do sentiment classification at sentence level. In this study, we just only focus on the problem of sentiment classification and skip subjectivity classification.

4.2 Our idea

The first challenge mentioned in section 4.1, which is lack of sentiment lexicon for medical domain, requires a method that helps to build the lexicon resources. Thus, we proposed a supervised method that can extract sentiment terms with their corresponding sentiment score from annotated data set.

For the diversity of sentiment shifters used in clinical text, recall the example mentioned in subsection 3.3.3 in chapter 3, we see that the sentiment of a sentence can depend on negation words and strongly positive words like "no" and "positive" or phrases such as "less nauseous" or sequences of non-adjacent words like "decrease...pleural effusion". We generalize this observation by a hypothesis that the sentiment score of a sentence simultaneously depends on scores of its words, phrases, sequences of non-adjacent words with different weights.

Relying on the hypothesis mentioned above, we proposed a method that simultaneously sums up the score of words, phrases, sequences of words extracted by

different language models by a linear combination. The linear combination is a simple and efficient model for voting sentiment score of sentence with low computational cost that characterizes the importance of its components via the corresponding coefficients. Moreover, basing on such linear combination, we are able to derive a novel vector representation of a sentence called language-model-based representation. Our proposed idea is formulated as the following:

Assume that:

- $\mathbf{L} = \{L_1, L_2, ..., L_m\}$ is a set of m language models used to extract terms.
- $\mathbf{T} = \{L_1(s), L_2(s), ..., L_m(s)\}$ where $L_i(s), i = 1, 2, .., m$ is a set of terms extracted from the sentence s according to the language model L_i.

For each term $t \in L_i(s)$ compute $Score(t)$. An average score over all terms belonging to $L_i(s)$ is computed by the following equation:

$$Score(L_i(s)) = \frac{\sum_{t \in L_i(s)} Score(t)}{N_i} \tag{4.1}$$

where N_i is the number of terms in $L_i(s)$.

The sentiment score of the sentence s is defined as a linear combination over $Score(L_i(s))$ as the following:

$$Score(s) = \sum_{i=1}^{m} w_i \times Score(L_i(s)) \tag{4.2}$$

$$\begin{cases} Score(s) > 0 \Rightarrow positive \\ Score(s) < 0 \Rightarrow negative \end{cases}$$

In the linear combination, the coefficients $(w_1, w_2, ..., w_n)$ characterize how the sentence's score depends on each $Score(L_i(s))$. If the sentence's score is strongly related to a kind of term, its coefficient is larger that means there is a bias for such kind of term. Besides, some kinds of term contribute to sentence's score identification with equal roles. Therefore, we pose three assumptions regarding the coefficient's values as follows:

- Assumption 1: The values of coefficients $(w_1, w_2, ..., w_m)$ are different. That means there is a bias in the voting process.
- Assumption 2: The values of coefficients are equal, and set as 1.

- Assumption 3: That incorporates assumption 1 and assumption 2. There exists a subset of language models following assumption 1, and the rest is appropriate with assumption 2. In this case, the sentence's score is computed as the following:

$$Score(s) = \sum_{i=1}^{k} w_i \times Score(L_i(s)) + \sum_{i=k+1}^{m} Score(L_i(s)) \tag{4.3}$$

  where k, $m - k$ are the number of language models following assumption 1, assumption 2 respectively.

Through the experiments and interpretations, we assess that if the components $Score(l_i(s))$ have a weak linear relationship, assumption 1 is more appropriate to obtain a better performance because in this case, there will has a conflict when aggregating such components, so we need to adjust the aggregation by a priority setting via adding the different weights for the components. Otherwise, in case such components have a strong linear relation that means we can use one of them to make the aggregation to make the decision, and we do not need to adjust them, thus assumption 2 is more appropriate. The detail and explanation are presented in section 4.4.

Equation 4.2 gives an idea of a vector representation for a sentence that is different from most of previous works using topics of words. In this equation, the sentence's score depends on the concurrent contribution of the components $Score(l_i(s))$, thus the set $\mathbf{S} = \{Score(L_1(s)), Score(L_2(s)), ..., Score(L_m(s))\}$ could be considered a feature set to represent the sentence that is called language-model-based representation. Through such method, the similarity measure of two sentences is based on the comparison between the sentence's scores which are decomposed into the components $Score(L_i(s))$ instead of enhancing the co-occurrence of common words like using topic models. Figure 4.1 shows our idea of language-model-based representation to deal with the shortness of text.

4.3 Proposed method

Relying on the proposed idea mentioned in previous section, we propose a method that includes three main steps: language-model-based terms extraction, sentiment score measure, and feature derivation and linear combination coefficients estimation.

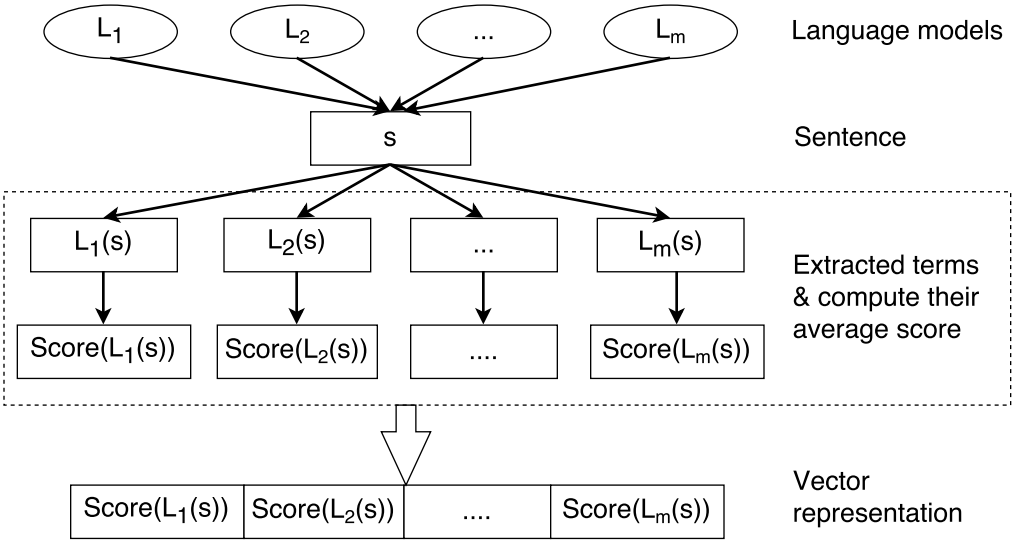


Figure 4.1: Language-model-based representation method

4.3.1 Language-model-based terms extraction

Language model is a statistical model that reflects a probability distribution over sequences of words. It includes two components, one is sequences of words, the other is a corresponding probability of each sequence that is estimated through frequency of this sequence in a corpus.

In this step, we use the first component of language models including n-gram and skip-gram models which play a role as templates in terms extraction. For example, we can use unigram model for extracting words, bigram model for extracting two adjacent words, trigram model for extracting three adjacent words. More general then n-gram models that help to extract sequences of adjacent words, skip-gram [26] models can capture not only sequences of adjacent words by also sequences of non-adjacent words. That help us to extend the context of words consideration. For example, we consider the following sentence:

"There is no evidence of pleural effusion."

Various language models such as unigram, bigram, trigram, 1-skip-bigram, 2-skip-bigram, 3-skip-bigram, 4-skip-bigram, 1-skip-trigram, 2-skip-trigram are used in this step. Table 4.1 shows an example of language models utilization for term extraction.

Table 4.1: Terms extraction based on language models

Language model	**Extracted terms**
unigram	there, is, no, evidence, of, pleural, effusion
bigram	there is, is no, no evidence, evidence of, of pleural, pleural effusion
1-skip-bigram	there is, there no, is no, is evidence, no evidence, no of, evidence of, evidence pleural, etc.
2-skip-bigram	there is, there no, there evidence, is no, is evidence, is of, no evidence, no of, no pleural, etc.
trigram	there is no, is no evidence, no evidence of, evidence of pleural, of pleural effusion.
1-skip-trigram	there is no, there is evidence, there no evidence, is no evidence, is evidence of, is no of, etc.

As the definition in [26], k-skip-n-grams consider k or lest skips to construct n-gram. For example, 3-skip-bigram includes 3 skips, 2 skips, 1 skip, 0 skips (bigram). Relying on number of tokens in terms, the language models are divided into three groups as the following:

- Group 1: Occurrence of words individually (unigram).
- Group 2: Co-occurrence of two words (bigram, 1-skip-bigram, 2-skip-bigram, 3-skip-bigram, 4-skip-bigram).
- Group 3: Co-occurrence of three words (trigram, 1-skip-trigram, 2-skip-trigram).

4.3.2 Term's sentiment score measure

Sentiment score of a term measure the related volume between the term and the sentence's sentiment label. We use the following equation to compute the term's sentiment score as in [17]:

$$Score(t) = \frac{p(t|positive) - p(t|negative)}{p(t|positive) + p(t|negative)} \tag{4.4}$$

$p(t|positive)$ is computed by taking number of times term t appears in positive sentences then dividing it by the total number of terms in the positive sentences. $p(p|negative)$ is also computed in the similar way. The term's score $Score(t)$ ranges

from -1 to 1. If $Score(t) > 0$ the sentiment orientation of the term is likely positive, and vice versa.

4.3.3 Language-model-based feature derivation and coefficient estimation

As we mentioned in section 4.2, the simultaneous contribution of various kinds of term to the sentence sentiment orientation is characterized by a linear combination of their score as equation 4.2, in which each coefficient indicates how each kind of term gives its influence on the sentence score. Therefore, identifying such influence is equivalent to estimating such coefficient. We need to estimate coefficients in case of assumption 1 and 3.

Algorithm 1: Linear combination coefficients learning

```
L = {L_1, L_2, ..., L_m} is a set of language models used.
for each sentence s in training set do
    vector := empty
    for each L_i ∈ L do
        Extracting a set of terms L_i(s) in the sentence s according to L_i
        for each term t in L_i(s) do
            Compute Score(t) by equation 4.4
        Compute score average Score(L_i(s)) by equation 4.1
        Append Score(L_i(s)) to vector
if L follows assumption 1 then
    Train with Support Vector Machine to to identify (w_1, w_2, ..., w_m)
if L follows assumption 2 then
    Set w_1 = w_2 = ... = w_m = 1
if L follows assumption 3 then
    if L1 ⊂ L follows assumption 1 then
        Train with Support Vector Machine to identify coefficients
    if L2 ⊂ L follows assumption 2 then
        Set the coefficients as 1
```

The most likely coefficients estimation is based on the training data. Each sentence in the training set is converted into the corresponding linear combination like equation 4.2, and then if the label of the sentence is positive the linear combination is greater than 0, and if it is negative, the combination is smaller than 0. For example,

we assume that we convert n sentences in the training data into a set of inequalities as the following:

$$\begin{cases} s_1 : \sum_{i=1}^{m} w_i \times Score(L_i(s_1)) < 0 \\ s_2 : \sum_{i=1}^{m} w_i \times Score(L_i(s_2)) > 0 \\ ... \\ s_n : \sum_{i=1}^{m} w_i \times Score(L_i(s_n)) > 0 \end{cases}$$

We see that determining the most likely $(w_1, w_2, ..., w_m)$ is equivalent to finding a hyperplane as a linear boundary of a data set represented by the set of vectors $\{Score(L_1(s_k)), Score(L_2(s_k)), ..., Score(L_m(s_k)\}, k = 1, 2, ..., n$ that are language-model-based representation. Thus, this problem can be solved by using Support Vector Machine (SVM) technique.

We proposed algorithm 1 for coefficients learning. In algorithm 1, to determine which assumption language models **L** should follow, we based on assessments presented in detail in section 4.4.

4.4 Experimental evaluation and discussion

4.4.1 Experimental Objectives

We conduct experiments to evaluate our proposed methods through three main objectives as follows:

- For each assumption, which language models are appropriate?
- Is the proposed method better than summing up words or phrases separately?
- How does language-model-based perform?

Besides, we also investigate classification performance of our method in both cases of balance data and imbalance data.

4.4.2 Data preparation

In the experiment, the MIMIC II dataset [40] that contains the information of more than 32,000 patients are used for our method evaluation. 6000 sentences that are

manually annotated with two labels "1" (positive) and "-1" (negative) are obtained from "NOTEEVENTS" records.

For evaluation method, the annotated data is randomly divided into 10 parts then 6 parts are used for training, and the rest for testing. This process is repeated 10 times, then we take an average of precision.

We aim to build a classifier that can work well on clinical narratives in case sentiment lexicon resources for medical domain are not available, so the classification method should not depend on a specific domain. Therefore, to investigate whether our proposed method with the derived assessments is robust and can be applied on other data set or not, we additionally use movie review data [1] for evaluation due to some fairly similar points. The text in movie review data set is also separated into sentences/snippets (short text), and also contains some kinds of sentiment shifters like the MIMIC II dataset.

In case of assumption 1 and 3, we use scikit learn – a python package implementing SVM algorithm with kernel functions [2] to determine coefficients.

4.4.3 Experiment results and interpretation

For each assumption, which language models are appropriate?

The experiments aims to determine which assumption is appropriate to a given language model. In the experiments, we consider the features generated from the language models in three groups and in the combination of such groups. All sentences are represented according to the language-model-based representation method. The full classification results of three assumptions with three groups are showed in Table 4.2.

- A comparison between group 2 and group 3

We consider language models in the same group, and make a comparison between language models in group 2 and group 3. Lines 1, 2, 7, 8 in Table 4.2 show that the features of group 1 provide remarkably higher performance than those of group 3 with both assumption 1 and 2. To explain why there is a significant different between the features of group 2 and group 3, we visualize the training set and testing set in Figure 4.2, then observe the distribution of data points. In Figure 4.2, the features

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[2] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Table 4.2: Coefficients assumptions with groups of language models investigation

	Method	**MIMIC II**	**Movie-Review**
	Our method		
1	assumption 1 with group 2	**0.823**	**0.736**
2	assumption 1 with group 3	0.69	0.507
3	assumption 1 with group 1 + group 2	0.799	0.747
4	assumption 1 with group 1 + group 3	**0.827**	**0.754**
5	assumption 1 with group 2 + group 3	0.807	0.605
6	assumption 1 with group 1 + group 2 + group 3	0.811	0.723
7	assumption 2 with group 2	0.817	0.732
8	assumption 2 with group 3	0.68	0.594
9	assumption 2 with group 1 + group 2	**0.836**	**0.756**
10	assumption 2 with group 1 + group 3	0.823	0.738
11	assumption 2 with group 2 + group 3	0.813	0.723
12	assumption 2 with group 1 + group 2 + group 3	0.832	0.751
13	assumption 3 with group 1 + group 2 + group 3 (*)	**0.836**	**0.764**
	Separately sum up terms		
14	terms from unigram	0.827	0.747
15	terms from bigram	0.769	0.688
16	terms from trigram	0.579	0.464
17	terms from 1-skip-bigram	0.799	0.709
18	terms from 2-skip-bigram	0.81	0.717
19	terms from 3-skip-bigram	0.812	0.721
20	terms from 4-skip-bigram	0.818	0.727
21	terms from 1-skip-trigram	0.644	0.556
22	terms from 2-skip-trigram	0.678	0.599
	Bag-of-words		
23	SVM + bag-of-words	0.698	0.503

(*): The sentence's score is computed by the following equation:

$$Score(s) = \sum_{i=1}^{k} w_i \times Score(L_i(s)) + \sum_{j=1}^{h} Score(L_j(s))$$

where $L_i \in$ group 1 and group 3, $L_j \in$ group 2.

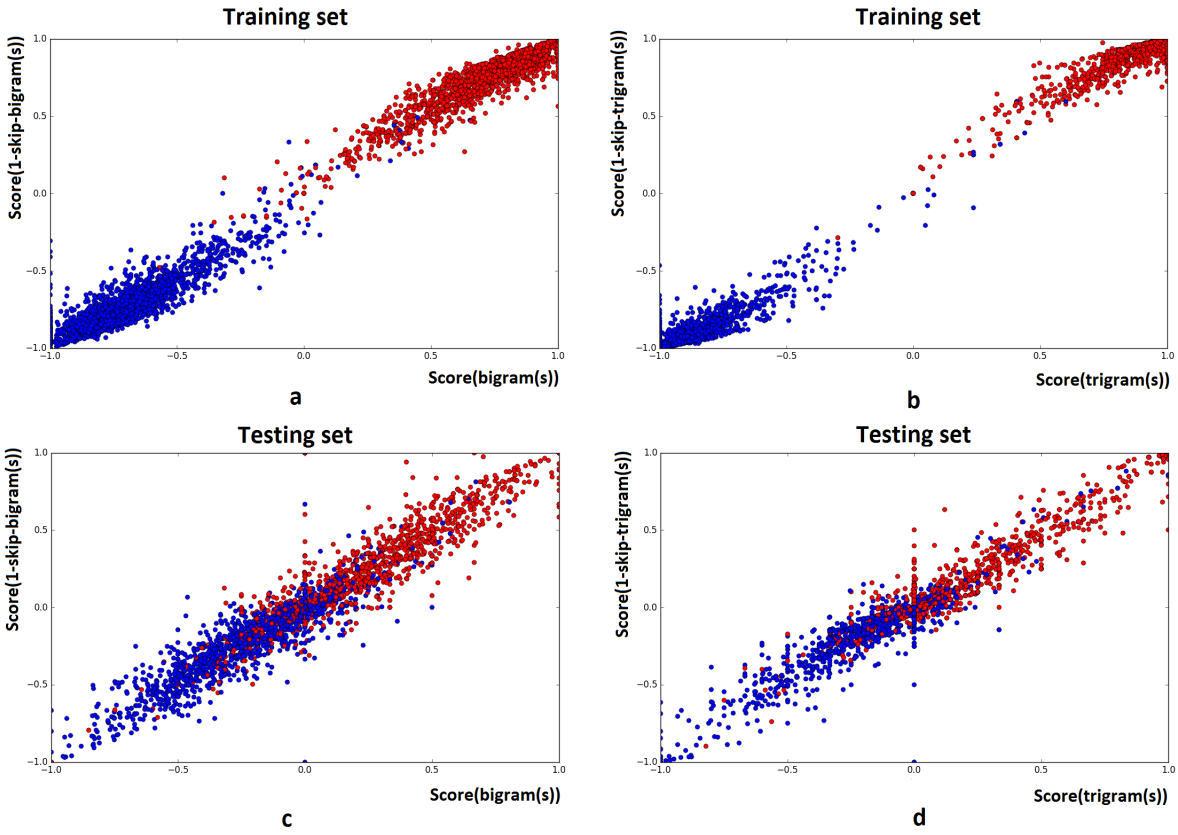


Figure 4.2: A visualization of training and testing set with features of group 2 and group 3.

of group 2 are generated by bigram and 1-skip-bigram models ($Score(bigram(s))$, $Score(1-skip-bigram(s))$), and the features of group 3 are generated by trigram and 1-skip-trigram models ($Score(trigram(s))$, $Score(1-skip-trigram(s))$). The blue, and red points indicate negative sentences and positive sentences respectively. Figure 4.2a and 4.2c show the data points with the features of group 2, and Figure 4.2b, 4.2d show the data points with the features of group 3.

We observe that the language models in a same group often generate their features with similar value, so the points in Figure 4.2 almost fluctuate around the bisector $y = x$ with close distance.

Figure 4.2a and 4.2b show a difference of the points distribution between group 2 and group 3. The data points of group 2 tend to spread along the bisector while the

Table 4.3: Correlation coefficient among features generated by different combinations of three groups

Group pair	correlation coefficient on MIMIC II	correlation coefficient on Movie Review
group 1 + group 2	0.901	0.893
group 1 + group 3	0.844	0.837
group 2 + group 3	0.977	0.983

data points of group 3 tend to converge at the corners. The reason is that sentiment orientations of terms extracted by language models of group 3 is almost pure with very high absolute value of score because the probability of co-occurrence of three words in a sentence is very small that gives poor information for prediction. In addition, the sentences in testing set are represented through the lexicon extracted from training set, so the terms of group 3 appearing together in a training sentence have a less chance to co-occur in the testing in the testing sentence that makes the testing set significantly different from the training set. In contrast to group 3, due to the higher probability of co-occurrence of two words, features of group 2 make our method get better accuracy. We also obtain a similar result when doing classification on movie-review data set. Therefore, we have an assessment of using language models in a same group as the following:

Assessment 1: When building the feature set by language models in a same group, the language models considering the co-occurrence of two words provide better performance than ones considering the co-occurrence of three or more words.

- A comparison among different combination of groups

Lines 3, 4, 5, 6 show the accuracy when using assumption 1 with different combinations of three groups. We obtained the highest precision by incorporating language models of group 1 and group 3 (line 4), and get lower accuracy on other combinations. The quality of features depends on the linear relationship among them. If the features have a strong linear relation, there is less information to make a decision because they are considered as duplicated features, and the decision is just based on one of them. The volume of linear relationship between two features can be measured via correlation coefficient. In case the correlation coefficient is close to 1 or -1, the linear relation is strong. Table 4.3 shows the correlation coefficient of features generated by incorporating groups. For each group, we take a language model to generate the feature because other ones also generate the similar feature.

From Table 4.3, we observe that the features generated by language models of group 1, and group 3 have lowest correlation coefficient on both MIMIC and movie-review data that explains why such features get high performance of classification with assumption 1.

Although the combinations of group 1 and group 2 or group 2 and group 3 do not produce the high performance with assumption 1 on the MIMIC dataset and movie-review dataset, they get better results with assumption 2 (showed lines 9, 11).

Through the results showed from lines 1 to 12, we have an assessment to select the appropriate assumption for language models as the following:

Assessment 2: Assumption 1 is appropriate for language models whose generated features have a weak linear relation. In case such features have a strong relation, assumption 2 is more appropriate.

There has an interesting meaning inside this assessment. In case the features have a weak linear relationship, it will raise a conflict when aggregating, so we need a referee to judge which features are important then give such features a priority. In our method, the priority is characterized through the coefficients. Otherwise, if such features strongly linearly depend on each other, no conflict happens, so the referee is not necessary.

Is the proposed method better than summing up words or phrases separately?

Line 13 shows our best result when we use assumption 3 with a combination of three groups, in which the feature of group 1 and group 3 are aggregated with the different coefficients. We obtained 83.6% on the MIMIC II dataset and 76.4% on movie-review data.

From line 14 to line 22, we show the results when using each language model to extract terms then make their score summing up. By this method, unigram has highest performance (82.7% on the MIMIC and 74.7% on the movie-review), but it is not better than our method with assumption 3 that considers the interaction among terms extracted from different language models in voting for sentence's score.

How does language-model-based perform?

In Table 4.2, methods showed from line 1 to line 22 use language-model-based representation. Those show the better performance than using bag-of-words (showed in

Table 4.4: Influence of balance and imbalance training set on classification performance

	Method		MIMIC II	Movie-Review
1	sum up score (unigram)	B	0.827	0.747
2	sum up score (unigram)	IB-P	0.805	0.72
3	sum up score (unigram)	IB-N	0.813	0.726
4	assumption 1 with group 2	B	0.823	0.731
5	assumption 1 with group 2	IB-P	0.715	0.585
6	assumption 1 with group 2	IB-N	0.783	0.582
7	assumption 2 with group 1 + group 2	B	0.836	0.756
8	assumption 2 with group 1 + group 2	IB-P	0.799	0.695
9	assumption 2 with group 1 + group 2	IB-N	0.82	0.711

- B: Balance data set
- IB-P: Imbalance data set with greater number of positive sentences.
- IB-N: Imbalance data set with greater number of negative sentences.

line 23). These results show that language-model-based representation provide fairly high performance when classifying on short text.

Influence of balance and imbalance training set on classification performance

The experiment aims to examine the influence of balance and imbalance training data on the classification performance. A balance set contains an equal number of positive and negative sentences while an imbalance set is in contrast. The proportion between positive sentences and negative sentences impacts the term's score measure in equation 4.4. Table 4.4 shows how the proportion affects the classification performance of our method.

Table 4.4 shows that imbalance sets make the accuracy reduce on both MIMIC

and movie-review dataset. The difference between number of positive sentences and negative sentences makes the term's score measure not fair, thus the scores are not precise.

4.5 Conclusion and future work

The paper presents our work on sentiment classification on clinical narratives. In this work, we proposed a classification method to deal with three challenges of such text: the lack of sentiment lexicon for medical domain, the diversity of sentiment shifters, and the shortness of text. Our method uses a mixture of language models to extract terms, then estimate the sentiment score of sentences by a linear combination of such term's scores. In addition, we also derive a novel vector representation according to the language models used to extract terms that can work better on short text. Moreover, this method is flexible and independent with a specific language. The experimental results show the improvement of classification performance by using our method.

Beside the advantages, our method still has some drawbacks. The exist of sentiment shifters in training data makes the estimation of term's score sometimes is not precise. We also have to face with the problem of sparse data when using language models in group 3. Therefore, we plan to overcome these drawbacks to improve the performance of our method in the future work.

Publications during master course

1. Dang, Tran-Thai, and Tu-Bao Ho. "Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives." International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer International Publishing, 2016.

2. Tran-Thai Dang, Phetnidda Ouankhamchan, Tu-Bao Ho. Detection of New Drug Indications from Electronic Medical Records. The 12th IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF) 2016. (under review)

3. Tu-Bao Ho, Ly Le, Dang Tran Thai, Siriwon Taewijit. Data-driven Approach to Detect and Predict Adverse Drug Reactions. Current Pharmaceutical Design Journal 2016, Vol. 22, No. 23.

Bibliography

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (2011), Association for Computational Linguistics, pp. 30–38.

[2] Ali, T., Schramm, D., Sokolova, M., and Inkpen, D. Can i hear you? sentiment analysis on medical forums. In *IJCNLP* (2013), pp. 667–673.

[3] Aue, A., and Gamon, M. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)* (2005), vol. 1, pp. 2–1.

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

[5] Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (2010), Association for Computational Linguistics, pp. 36–44.

[6] Benamara, F., Chardon, B., Mathieu, Y. Y., Popescu, V., et al. Towards context-based subjectivity analysis. In *IJCNLP* (2011), pp. 1180–1188.

[7] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., and Reynar, J. Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era* (2008), vol. 14, pp. 339–348.

[8] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022.

[9] Blitzer, J., Dredze, M., Pereira, F., et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL* (2007), vol. 7, pp. 440–447.

[10] Bobicev, V., and Sokolova, M. An effective and robust method for short text classification. In *AAAI* (2008), pp. 1444–1445.

[11] Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research 32*, suppl 1 (2004), D267–D270.

[12] Cambria, E. Affective computing and sentiment analysis. *IEEE Intelligent Systems 31*, 2 (2016), 102–107.

[13] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. Micro-wnop: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics* (2007), 200–210.

[14] Chee, B. W., Berlin, R., and Schatz, B. Predicting adverse drug events from personal health messages. In *AMIA Annu Symp Proc* (2011), vol. 2011, pp. 217–26.

[15] Chen, M., Jin, X., and Shen, D. Short text classification improved by learning multi-granularity topics. In *IJCAI* (2011), Citeseer, pp. 1776–1781.

[16] Dai, Z., Sun, A., and Liu, X.-Y. Crest: Cluster-based representation enrichment for short text classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2013), Springer, pp. 256–267.

[17] Dave, K., Lawrence, S., and Pennock, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web* (2003), ACM, pp. 519–528.

[18] Davidov, D., Tsur, O., and Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters* (2010), Association for Computational Linguistics, pp. 241–249.

[19] Denecke, K., and Deng, Y. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine 64*, 1 (2015), 17–27.

[20] Deng, L., Choi, Y., and Wiebe, J. Benefactive/malefactive event and writer attitude annotation. In *ACL (2)* (2013), pp. 120–125.

[21] Deng, Y., Stoehr, M., and Denecke, K. Retrieving attitudes: Sentiment analysis from clinical narratives. In *MedIR@ SIGIR* (2014), pp. 12–15.

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

[23] Du, W., Tan, S., Cheng, X., and Yun, X. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), ACM, pp. 111–120.

[24] Gamon, M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics* (2004), Association for Computational Linguistics, p. 841.

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

[26] Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)* (2006), pp. 1–4.

[27] Hassan, A., Qazvinian, V., and Radev, D. What's with the attitude?: identifying sentences with attitude in online discussions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (2010), Association for Computational Linguistics, pp. 1245–1255.

[28] Hatzivassiloglou, V., and Wiebe, J. M. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1* (2000), Association for Computational Linguistics, pp. 299–305.

[29] Hofmann, T. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (1999), Morgan Kaufmann Publishers Inc., pp. 289–296.

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

[31] Huang, M., Ye, B., Wang, Y., Chen, H., Cheng, J., and Zhu, X. New word detection for sentiment analysis. In *ACL (1)* (2014), pp. 531–541.

[32] Ikeda, D., Takamura, H., Ratinov, L.-A., and Okumura, M. Learning to shift the polarity of words for sentiment classification. In *IJCNLP* (2008), pp. 296–303.

[33] Jia, L., Yu, C., and Meng, W. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 1827–1830.

[34] Jin, W., Ho, H. H., and Srihari, R. K. A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), Citeseer, pp. 465–472.

[35] Jindal, P., and Roth, D. Extraction of events and temporal expressions from clinical narratives. *Journal of biomedical informatics 46* (2013), S13–S19.

[36] Kennedy, A., and Inkpen, D. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence 22*, 2 (2006), 110–125.

[37] Kim, S.-M., and Hovy, E. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions* (2006), Association for Computational Linguistics, pp. 483–490.

[38] Kiritchenko, S., Zhu, X., and Mohammad, S. M. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research 50* (2014), 723–762.

[39] Kobayashi, N., Inui, K., and Matsumoto, Y. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL* (2007), vol. 7, Citeseer, pp. 1065–1074.

[40] Lee, J., Scott, D. J., Villarroel, M., Clifford, G. D., Saeed, M., and Mark, R. G. Open-access mimic-ii database for intensive care research. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2011), IEEE, pp. 8315–8318.

[41] Li, J., Zhou, G., Wang, H., and Zhu, Q. Learning the scope of negation via shallow semantic parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics* (2010), Association for Computational Linguistics, pp. 671–679.

[42] Li, S., Lee, S. Y. M., Chen, Y., Huang, C.-R., and Zhou, G. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics* (2010), Association for Computational Linguistics, pp. 635–643.

[43] Lin, C., and He, Y. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 375–384.

[44] Liu, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*, 1 (2012), 1–167.

[45] Madhoushi, Z., Hamdan, A. R., and Zainudin, S. Sentiment analysis techniques in recent works. In *Science and Information Conference (SAI), 2015* (2015), IEEE, pp. 288–291.

[46] McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association For Computational Linguistics* (2007), vol. 45, Citeseer, p. 432.

[47] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web* (2007), ACM, pp. 171–180.

[48] Mukherjee, A., and Liu, B. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (2012), Association for Computational Linguistics, pp. 339–348.

[49] Na, J.-C., Kyaing, W. Y. M., Khoo, C. S., Foo, S., Chang, Y.-K., and Theng, Y.-L. Sentiment classification of drug reviews using a rule-based linguistic approach. In *International Conference on Asian Digital Libraries* (2012), Springer, pp. 189–198.

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

[51] Polanyi, L., and Zaenen, A. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*. Springer, 2006, pp. 1–10.

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

[53] Quan, X., Liu, G., Lu, Z., Ni, X., and Wenyin, L. Short text similarity based on probabilistic topics. *Knowledge and information systems 25*, 3 (2010), 473–491.

[54] Riloff, E., Patwardhan, S., and Wiebe, J. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (2006), Association for Computational Linguistics, pp. 440–448.

[55] Riloff, E., and Wiebe, J. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (2003), Association for Computational Linguistics, pp. 105–112.

[56] Saif, H., He, Y., Fernandez, M., and Alani, H. Adapting sentiment lexicons using contextual semantics for sentiment analysis of twitter. In *European Semantic Web Conference* (2014), Springer, pp. 54–63.

[57] Somasundaran, S., and Wiebe, J. Recognizing stances in online debates. 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 1-Volume 1* (2009), Association for Computational Linguistics, pp. 226–234.

[58] Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. Short text classification: A survey. *Journal of Multimedia 9*, 5 (2014), 635–643.

[59] Stone, P. J., Dunphy, D. C., and Smith, M. S. The general inquirer: A computer approach to content analysis.

[60] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. Lexicon-based methods for sentiment analysis. *Computational linguistics 37*, 2 (2011), 267–307.

[61] Täckström, O., and McDonald, R. Discovering fine-grained sentiment with latent variable structured prediction models. In *European Conference on Information Retrieval* (2011), Springer, pp. 368–374.

[62] Täckström, O., and McDonald, R. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (2011), Association for Computational Linguistics, pp. 569–574.

[63] Titov, I., and McDonald, R. T. A joint model of text and aspect ratings for sentiment summarization. In *ACL* (2008), vol. 8, Citeseer, pp. 308–316.

[64] Turney, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (2002), Association for Computational Linguistics, pp. 417–424.

[65] Wiebe, J. Learning subjective adjectives from corpora. In *AAAI/IAAI* (2000), pp. 735–740.

[66] Wiebe, J., and Riloff, E. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics* (2005), Springer, pp. 486–497.

[67] Wilson, T., Wiebe, J., and Hwa, R. Just how mad are you? finding strong and weak opinion clauses. In *aaai* (2004), vol. 4, pp. 761–769.

[68] Wu, Y., Denny, J. C., Rosenbloom, S., Miller, R. A., Giuse, D. A., and Xu, H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA* (2012).

[69] Zhuang, L., Jing, F., and Zhu, X.-Y. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (2006), ACM, pp. 43–50.