INCORPORATING INFORMATIVE SCORE FOR INSTANCE SELECTION IN SEMI-SUPERVISED SENTIMENT CLASSIFICATION

VIVIAN LEE LAY SHAN

UNIVERSITI SAINS MALAYSIA

2022

INCORPORATING INFORMATIVE SCORE FOR INSTANCE SELECTION IN SEMI-SUPERVISED SENTIMENT CLASSIFICATION

by

VIVIAN LEE LAY SHAN

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

May 2022

DECLARATION OF ORIGINALITY

I hereby declare that this research together will all of its contents is none other than that of my own work, with consideration of the exception of research-based information and relative materials that were adapted and extracted from other resources, which have evidently been quoted or stated respectively.

Signed,

.....

VIVIAN LEE LAY SHAN

20th October 2021

School of Computer Sciences

UNIVERSITI SAINS MALAYSIA

ACKNOWLEDGEMENT

First and foremost, I would like to express my gratitude to my supervisor, Dr. Gan Keng Hoon for her comments, remarks and guidance throughout the learning processing of producing this master degree thesis. She has guided and motivated me in making this research a success. Besides that, I also would like to extent my gratitude to School of Computer Sciences, Universiti Sains Malaysia for giving me the opportunity to gain experiences in conducting this research.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
NB	Na ïve Bayes
NLP	Natural Language Processing
POS	Parts-of-speech
RNN	Recurrent Neural Network
SVM	Support Vector Machine

PENGGUNAAN SKOR INFORMATIF DALAM PEMILIHAN TIKA UNTUK KLASIFIKASI SENTIMEN SEPARA TERSELIA

ABSTRAK

Klasifikasi sentimen merupakan suatu alat berguna untuk mengklasifikasikan ulasan yang mengandungi maklumat mengenai sentimen dan sikap terhadap produk atau perkhidmatan. Kajian sekarang sangat bergantung pada kaedah klasifikasi sentimen yang memerlukan label input sepenuhnya. Dalam tahun-tahun kebelakangan ini, kaedah separa penyeliaan disarankan secara positif kerana ia hanya memerlukan input berlabel sebahagian dan prestasi setanding dengan kaedah pilihan semasa. Pada masa yang sama, terdapat beberapa karya yang melaporkan prestasi model separa diselia merosot selepas menambahkan contoh yang tidak berlabel ke dalam latihan model. literatur semasa menunjukkan bahawa bukan semua input tidak berlabel sama-sama berguna, dengan itu mengenal pasti input tidak berlabel yang berinformasi akan bermanfaat melatihkan model separa penyeliaan. Untuk mencapainya, skor maklumat dicadangkan dan dimasukkan ke dalam klasifikasi sentimen separa penyeliaan. Eksperimen membandingkan ketepatan dan kehilangan kaedah diselia, kaedah separa penyeliaan tanpa skor maklumat dan kaedah separa penyeliaan dengan skor maklumat. Dengan bantuan skor informatif untuk mengenal pasti kejadian tidak berlabel maklumat, model separa penyeliaan dapat menunjukkan prestasi yang lebih baik berbanding dengan model separa penyeliaan yang tidak memasukkan skor maklumat ke dalam latihannya. Walaupun prestasi model semi-selia yang digabungkan skor informatif tidak dapat melampaui model yang diawasi, tetapi hasilnya masih dijanjikan kerana perbezaan dalam prestasi adalah halus dan jumlah contoh berlabel yang digunakan sangat berkurang.

INCORPORATING INFORMATIVE SCORE FOR INSTANCE SELECTION IN SEMI-SUPERVISED SENTIMENT CLASSIFICATION

ABSTRACT

Sentiment classification is a useful tool to classify reviews that contain a wealth of information about sentiments and attitudes towards a product or service. Existing studies are heavily relying on sentiment classification methods that require fully annotated input. However, there are limited labelled text available, making the acquirement process of the fully annotated input costly and labour intensive. In recent years, semi-supervised methods have been positively recommended as they require only partially labelled input and performed comparably to the current preferred methods. At the same time, there are some works reported the performance of semi-supervised model degraded after adding unlabelled instances into training. The contrast of the current literature shows that not all unlabelled instances are equally useful; thus identifying the informative unlabelled instances is beneficial in training a semi-supervised model. To achieve this, informative score is proposed and incorporated into semi-supervised sentiment classification. The experiment compared the accuracy and loss of supervised method, semi-supervised method without informative score and semi-supervised method with informative score. With the help of informative score to identify informative unlabelled instances, semi-supervised models can perform better compared to semi-supervised models that do not incorporate informative score into its training. Although performance of semi-supervised models incorporated with informative score are not able to surpass the supervised models, the results are still found promising as the differences in performance are subtle and the number of labelled instances used are greatly reduced.

CHAPTER 1

INTRODUCTION

1.1 Introduction

The increasing development of web technology leads to the growth of different areas of evaluation. The original web had static pages and users were not allowed to manipulate its contents. Nevertheless, interactions on the web page are made possible by web technology advancements. Users are able to leave their comments and feedbacks, whereas the owners of the web pages are able to utilize users' ideas in improving future performances and adapting the products and services with their target group in an appropriate manner. Pang and Lee (2008) reported that there is 73% to 87% purchase decisions among the online reviews readers are greatly affected by the reviews of various services such as restaurants and hotels. However, manual analysis of a great number of opinions is very difficult, time consuming, and in some cases impossible. Therefore, sentiment analysis has been introduced to discover the knowledge through expressed comments in an effective way that can never be achieved with manual analysis. Some of the pioneer works (Dave, Lawrence, & Pennock, 2003; Pang & Lee, 2008) had successfully applied sentiment analysis to examine and analyse opinions within text. Sentiment classification, which is often mistaken for the same as sentiment analysis, is actually one of the tasks within the sentiment analysis process; its purpose is to classify the sentiment of a user's opinion towards a target that is expressing positive or negative polarity. Sentiment classification is similar to sentiment analysis, and can be divided into three extraction levels: document level, sentence level, and aspect level. Document level sentiment classification aims to classify the overall sentiment towards the target in a document. This task assumes each document expresses opinions on a single entity

from a single opinion holder; for example, product reviews (Liu, 2012). Sentence level sentiment classification is to determine the sentiment polarity of the sentence, whether it is positive or negative. The definition holds true when the sentence expresses a single opinion from a single opinion holder. This assumption is only appropriate for simple sentences expressing a single opinion; for example, "The grilled chicken chop served by this restaurant tastes good". In aspect level sentiment classification, its task is to identify separate sentiment degrees for different entities mentioned within a text. For instance, a review on a restaurant may consist of different aspects such as food and price. Besides, the sentiment expressed towards these different aspects can be different, e.g., "The grilled chicken chop served by this restaurant tastes good but it is pricy". Techniques used in solving sentiment classification are generally divided into two categories, lexicon-based approach and machine learning-based approach. Lexicon-based approaches rely on sentiment lexicon, which is a collection of known and precompiled sentiment terms, such as SentiWordNet. It can be further divided into corpus-based approach and dictionary-based approach. Machine learning-based approaches can be further catogorized into supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning relies on a fully annotated training dataset and gives good performances on predictions. On the other hand, unsupervised learning needs no labelled training datasets but suffers in model accuracy. Semi-supervised learning which lies in between supervised learning and unsupervised learning, requires a combination of a small part of labelled training data points and a large part of unlabelled training data points. Current sentiment classification works are more in using deep learning models instead of the conventional machine learning models, such as Naïve Bayes (NB) and Support Vector Machine (SVM). Deep learning is a subset of machine learning concerned with

algorithms inspired by the structure and function of the brain, known as artificial neural networks. Compared to traditional neural networks, deep learning is a very large neural network that can be trained using huge amounts of data and is scalable.

1.2 Motivation

A customer review is a textual review of a product or service made by a customer who had an experience with the product or service. The purpose served by these reviews is to share one's opinions towards the product or service and feed back one's opinion to the seller. Reviews contain a wealth of information about sentiments and attitudes towards a product or service. These reviews reflect the point of view and experiences of the reviewer, that are possibly helpful for potential buyers making their purchasing decisions and businesses to better understand how the customers feel about their products.

Sentiment classification, one of the important tasks in the process of sentiment analysis, classifies the reviews into the correct classes, either positive or negative (Cambria, Poria, Gelbukh, & Thelwall, 2017). There are two groups of methods in performing sentiment classification, lexicon-based methods and machine learning-based methods. Most of these methods require annotated input, that is costly and labour intensive, for model training. Currently, manual annotation is the most common way of acquiring high quality annotated input, the work is still manageable if the data size is small (up to 3000). On the other hand, training with deep learning usually requires the data size to be at least 10, 000 which is a huge number and not practical to annotate manually.

Semi-supervised methods which fall under the machine learning group, serve as a possible solution for reducing the needs for annotated input. They use a combination of a small part of labelled input and a large part of unlabelled input for model training. Compared to supervised machine learning methods, labour force and cost in acquiring labelled input are reduced as only a small number of labelled input is needed for training a semi-supervised model. Moreover, the accuracy of semi-supervised models is comparable to a supervised model. But at the same time, there are works reported that the performance of semi-supervised model degraded after adding unlabelled instances with predicted label into training (Iosifidis & Ntoutsi, 2017; Levatić, Ceci, Kocev, & Džeroski, 2017; W. Zhang, Tang, & Yoshida, 2015).

Suggested solution by some works is to add in only informative unlabelled instances that have positive impacts on model performances (He, Huang, Tsechpenakis, Metaxas, & Neidle, 2005; Tian, Yu, Xue, & Sebe, 2004). Understanding what are informative unlabelled instances in the semi-supervised training context helps to reduce the time taken to train a good model and avoids wasting time in finding the unlabelled instances that can positively impact the model performance. However, there was less discussion on the what are informative unlabelled instances in semi-supervised training.

1.3 Problem statement

Along with the success of semi-supervised learning in many other application domains such as computer vision, semi-supervised methods is gaining enormous attention in sentiment classification community. The methods receiving positive encouragement but there are also studies pointed out that although semi-supervised sentiment classification are able to work with limited amount of labelled text, however

observed a slow decline in accuracy of model after augmenting pseudo-labelled into each training iteration (Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, & Javier González-Castaño, 2016; Hu, Tang, Gao, & Liu, 2013; Iosifidis & Ntoutsi, 2017). Suggested solution from studies is to include only informative unlabelled data that have a positive impact in the training of semi-supervised sentiment classification (He et al., 2005; Tian et al., 2004; Zhou, Kantarcioglu, & Thuraisingham, 2012). However, current existing methods have substantially focused on selecting confidently predicted instances only. On the other hand, proposed methods that work on selecting informative predicted instances require high computational resources (Han, Liu, & Jin, 2019; Zhou, Kantarcioglu, & Thuraisingham, 2012). This necessitates the need for a simple scoring formula that can represent the informativeness of the unlabelled data.

Semi-supervised sentiment classification methods require only a small portion of a labelled dataset for model training. But the ratio of labelled and unlabelled data is rarely reported or suggested in studies. Without a guide, it is difficult to estimate when to stop collecting labelled text and this makes it difficult for resource management. Therefore, a suggestion on the optimal ratio of labelled and unlabelled data is required to minimize the time and effort in acquiring labelled text for semi-supervised training.

Having an explanation of the problems, the research questions of concern are:

1) What are informative unlabelled instances in the context of semi-supervised sentiment classification?

2) What is the optimal ratio of labelled and unlabelled data for semi-supervised sentiment classification?

3) What are the optimal parameters for review informative score and instance selection strategy proposed?

1.4 Research objectives

The fundamental goal of the research is to harness the advantage of semi-supervised methods in sentiment text classification without using a large set of annotated training data. To achieve this, the research comprises the following objectives.

1) To incorporate review informative score into instance selection process of semi-supervised sentiment classification

2) To determine optimal ratio of labelled and unlabelled data for semi-supervised sentiment classification

3) To determine the optimal parameters for review informative score and instance selection strategy proposed

1.5 Research contribution

In this thesis, the main contribution is the proposed methodology for semi-supervised sentiment classification with review informative score. The proposed review informative score enables evaluation of unlabelled reviews, checking its informativeness, and brings positive impacts to semi-supervised model performances. Moreover, the proposed instance selection strategy is able to select the confidently predicted and informative predicted instances. This allows the community to move

towards creating powerful semi-supervised or even unsupervised sentiment classification models with satisfactory performances.

Our research also suggested an optimal ratio of labelled and unlabelled data in semi-supervised model training, making it possible for semi-supervised model users to roughly estimate the number of labelled text they should be collecting. Besides, the proposed methodology also automates the data annotation using only a small amount of labelled data and the results are comparably good as supervised models. This in turn allows to reduce dependency on supervised models which require fully annotated inputs that are costly to acquire.

1.6 Significance

The research will provide new insights into semi-supervised sentiment classification, especially on selecting informative samples. Through this research, the community will start to understand informative unlabelled instances in the context of semi-supervised sentiment classification. This helps the community to form the definition of informative unlabelled instances and eventually shifts the heavy dependence on supervised learning, enabling the researchers to produce powerful unsupervised sentiment classification models.

Besides, there are potential benefits for business companies and consumers. It is crucial for business companies to take advantage of sentiment classification since its impact on business decision-making is undeniable. Therefore, research targeting sentiment classification will provide them opportunities to gain insights and observe market trends, keep track of companies image and reputation, evaluate consumers' feedback, and propose better products. Consumers also benefit from sentiment

classification applications such as product comparison tools to aid them in making wise buying decisions.

1.7 Research scope

Although there is a range of text resources contributing to online reviews including blogs, forums, comments and tweets, the scope of this work is defined around reviews posted on e-commerce platforms written in English. This work focuses only on document level sentiment classification and binary classification.

1.8 Thesis organization

The remaining chapters in this thesis are organized as follows:

Chapter 2 presents the existing literature with analysis and comparison. Chapter 3 elaborates on the methodology with details intended to solve research questions stated. Chapter 4 shows the experiment settings and discusses the results of the experiment. Chapter 5 concludes the thesis stating the summary, contributions and potential directions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This review focuses on the problem of minimizing labour forces and costs in acquiring a small set of high-quality annotated text and training a model with good performance. The study first discusses the methods used in sentiment classification. The review phase presents a survey of existing techniques, instance selection strategy and their research purposes for semi-supervised sentiment classification. The limitations of current practices highlighting research gaps are outlined at the end.

2.2 Sentiment classification

Text data is usually in unstructured form; however rich information such as opinions, sentiments, attitudes and emotions towards entities are hidden inside. To retrieve this meaningful information, sentiment analysis is required. In the process of sentiment analysis, sentiment classification is an important subtask. Sentiment classification is defined as the identification of sentimental orientation of a piece of text towards a given topic, whether it is positive or negative (Cambria et al., 2017). The given definition is identical to sentiment analysis. Due to the limited definition of "sentiment analysis", it has long been mistaken as a simple classification task (Cambria et al., 2017). Thus, these two terms are often used interchangeably in existing research work.

Sentiment analysis has been investigated mainly at three levels: document level, sentence level and aspect level. Document level sentiment classification classifies whether a whole opinion document expresses a positive or negative

sentiment. An assumption is made that each document expresses opinions on a single entity; for example a product review (L. Zhang, Wang, & Liu, 2018). Thus, it is not applicable to documents evaluating multiple entities. In sentence level sentiment classification, the sentence is first identified as to whether an opinion exists, then assigns the sentiment polarity to the sentence. This level of classification is strongly related to subjectivity classification, one of the subtasks in the sentiment analysis process. Aspect level performs finer-grained analysis which focuses on a target and its sentiment in text. This level of analysis turns unstructured text into a structured form of opinions about entities and their aspects.

Techniques of sentiment classification can be divided into two categories, linguistic-based methods and machine learning-based methods (Hemmatian & Sohrabi, 2019; Ravi & Ravi, 2015; Rodrigues, Camilo-Junior, & Rosa, 2018). The first category, lexicon-based methods are used to calculate the orientation of text according to sentiment words in text which have their own polarities. Lexicon-based methods can be further divided into two groups, corpus-based methods and dictionary-based methods. In dictionary-based method, sentiment classification is performed with the help of online dictionaries such as SentiWordNet and WordNet. On the other hand, corpus-based sentiment classification relies on statistical analysis of the contents of a collection of text instead of a predefined dictionary. The other category, machine learning-based methods can be further divided into three groups; supervised learning, semi-supervised learning and unsupervised learning. The main difference between these methods is the amount of labelled data needed for model training. Supervised learning requires fully labelled training data, whereas unsupervised learning does not require any. Semi-supervised learning is the combination of supervised learning and unsupervised learning, and requires a small part of labelled training data and a large

part of unlabelled data to train a model. Figure 2.1 below shows the taxonomy of sentiment classification techniques.

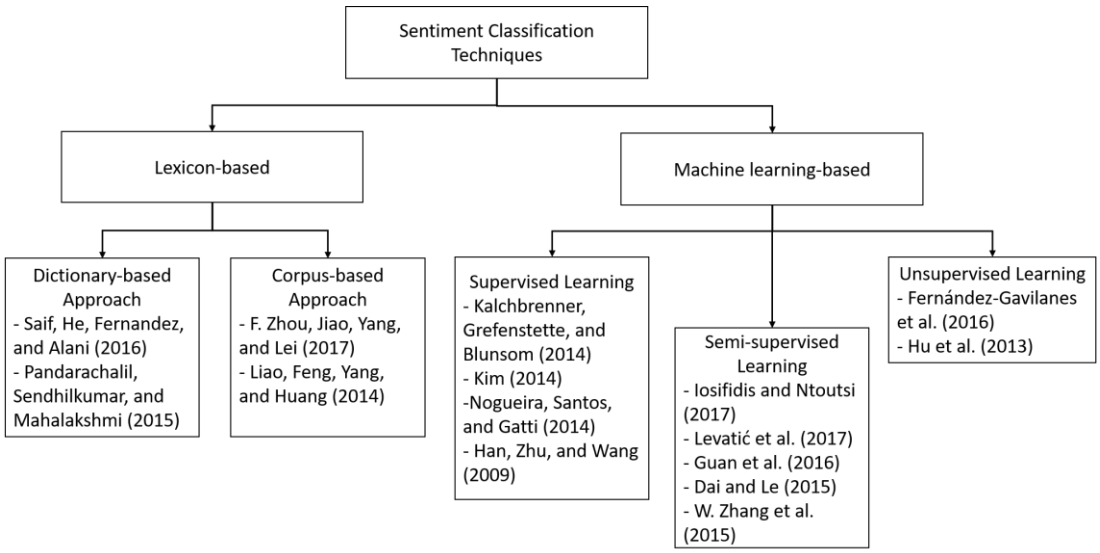

Figure 2.1 Taxonomy of sentiment classification techniques

2.2.1 Lexicon-based approach

Dictionary-based methods require manual identification of sentiment words; their synonyms and antonyms are then lookup in online dictionaries. The process stops when there are no new words can be detected. Pandarachalil, Sendhilkumar, and Mahalakshmi (2015) used three public sentiment lexicons, SentiWordNet, SenticNet, and SentislangNet, in determining the polarity of tweet sentiments. SentiWordNet is a domain-independent public lexical source for sentiment analysis that is derived from WordNet. But it contains only unigrams, thus providing sentiment scores only at the syntactic level. For example, SentiWordNet fails to provide sentiment scores for terms such as "not good" and "sunny day". The SenticNet is another public lexical source that contains not only unigrams but other ngrams as well, a total of 14,244 concepts. Whereas SentislangNet is a sentiment lexicon for slangs created by the authors, they make use of SenticNet and SentiWordNet for evaluating sentiment scores of slang

scraped from the Web. Saif, He, Fernandez, and Alani (2016) proposed a lexicon-based approach called SentiCircle for sentiment analysis on Twitter. Differ from typical lexicon-based approaches which determines sentiment polarities of words regardless of context, their approach builds a dynamic representation of words that captures the contextual semantics in tweets and tune their pre-assigned strength and polarity in sentiment lexicons.

In corpus-based methods, fully annotated corpora are needed to perform statistical analysis on the contents of a collection of text. F. Zhou, Jiao, Yang, and Lei (2017) proposed to incorporate customer preference information into feature models using sentiment analysis, a hybrid combination of affective lexicons and a rough-set technique. In this work, they focus only on Kindle Fire HD Tablets. However, a general sentiment lexicon based on ANEW and WordNet is built instead of a domain-dependent sentiment lexicon. A word may be positive in one domain and negative in another domain; for example, "long battery life" and "long queue". The "long" in first phrase holds positive sentiment as a device is preferably runs longer time before it needs to be recharged, while "long" in second phrase suggests negative sentiment as people dislike waiting in a long queue which might require them to spend a long time in queue. Except for English, the other languages lack of linguistic sources. Liao, Feng, Yang, and Huang (2014) presented a new approach named PDSP which incorporates domain lexicon with groups of features using syntax and semantics. They worked on Chinese microblogs and outperformed other baselines for opinion target extraction.

2.2.2 Machine learning-based approach

In recent years, machine learning-based methods have been preferable in sentiment classification. Han, Zhu, and Wang (2009) tried to tackle the domain adaptation problem in sentiment classification by applying stacked denoising auto-encoders to learn text reviews representation in an unsupervised fashion. Their proposed method showed positive results as they succeeded in performing domain adaptation on a large dataset composite of twenty-two domains. Kim (2014) examined convolutional neural networks (CNN) for sentence classification and found their work outperformed Kalchbrenner, Grefenstette, and Blunsom (2014). Nogueira, Santos, and Gatti (2014) proposed a modified CNN that exploits different levels of information, from character- to sentence-level, in order to address the limited contextual information in short text sentiment analysis. However, these supervised methods mentioned require fully annotated input that takes up a lot of time and effort in labelling the data. The other two methods, semi-supervised learning and unsupervised learning, require smaller sets of training data or none, but have their own disadvantages too.

Semi-supervised learning performs comparably to supervised learning and requires a smaller set of labelled data, but there are mixed reviews on the performance of semi-supervised models. Guan et al. (2016) introduced a semi-supervised framework that incorporated rating information in training a sentence sentiment classifier and the results showed that the proposed network worked better than the statistical approaches and supervised deep learning network. Dai and Le (2015) reported that it is possible to train a semi-supervised network to achieve good performance with some simple pre-training steps and careful tuning of hyper-

parameters. Besides, their work also found out that the usage of unlabelled data from related tasks may improve the generalization of the subsequent supervised model. In contrary, Iosifidis and Ntoutsi (2017) examined self-learning and co-learning to annotate a huge collection of tweets. In their experiments, they found out both methods produced unbalanced outputs, models were augmented with more positive examples and observed a slow decline in models' accuracy after each iteration. Moreover, this situation is also reported by other works (Levatić et al., 2017; W. Zhang et al., 2015). They also observed an unbalanced output was produced, and the accuracy of the model increased until a few iterations and then degraded afterwards.

Although unsupervised learning methods require no labelled data, they have lower accuracy compared to other learning methods. Hu et al. (2013) incorporated emotional signals, i.e., emotion indication and emotion correlation, into an unsupervised learning framework for sentiment analysis. They tested the framework on two Twitter datasets; the model accuracy falls on 0.6 to 0.7 averagely. On the other hand, Fernández-Gavilanes et al. (2016) presented an unsupervised dependency parsing-based classification method to predict sentiment in online text. Accuracy of the model surpassed Hu et al. (2013), yet still lacks its competitiveness compared to supervised and semi-supervised models.

2.3 Related works on semi-supervised learning in sentiment classification

Recently, semi-supervised learning techniques have been vastly applied to improve efficiency of sentiment classification. Semi-supervised sentiment classification studies mainly aim to reduce the burden for acquiring labelled datasets while attaining high classification accuracy. Table 2.1 shows the summary of related works described in this section.

2.3.1 Semi-supervised approaches

One of the approaches is to utilize unsupervised learning in deriving new knowledge from unlabelled instances or to use publicly available knowledge bases to improve semi-supervised model performance. Da Silva, Coletta, Hruschka, and Hruschka Jr. (2016) integrated unsupervised knowledge derived from unlabelled instances into model training. Khan, Qamar, and Bashir (2016) performed semi-supervised sentiment analysis using revised sentiment scores based on SentiWordNet. Whereas S. Lee and Kim (2017) extended the number of labelled data using an unsupervised joint sentiment/topic model and filter confidently predicted instances for better model performance.

Another approach is to make use of different views of the data. This approach usually pairs together with co-training, one of the popular semi-supervised methods, which involves two or more classifiers for sentiment classification. Chen, Feng, Sun, & Liu (2019) describe how they use word embedding and character-based embedding of forum posts to improve the accuracy of sentiment classification. The two views of embedding ensure the deep neural network can extract different features from texts. Besides, the proposed double-check strategy is applied to select samples with the same pseudo-labels from both classifiers. Their experimental results demonstrate the effectiveness of their proposed methods for training models on limited labelled data. Note that training multiple classifiers requires more resources than training only one. In addition, graph-based models and generative models are also commonly used semi-supervised methods for solving challenges in sentiment analysis. Bijari, Zare, Kebriaei, and Veisi (2019) employed graph-based text representation for sentiment

analysis whereas Duan, Luo, and Zeng (2020) utilized a generative model to perform sentiment feature extraction for classification.

2.3.2 Instance selection strategy

Semi-supervised methods are undeniably powerful and competent. But due to its own model predictions are used in the training process iteratively; thus it is hard to guarantee the introduction of unlabelled data will not degrade performance. In extreme cases, the classifier will classify all the unlabelled data into one of the classes which is undesirable. Hence, there are studies suggested to include only informative pseudo-labelled instances that have positive impacts on classifiers.

Han, Liu and Jin (2019) proposed dynamic thresholds for different iterations to maximize the number of accurately labelled data selected. The proposed dynamic threshold is based on two factors, the quality and the quantity of pseudo-labelled training data. Evaluation of predicted label quality is related to specific prediction models. SVM classifier was adopted in this work; thus the quality of predicted labels is defined as the distance between the pseudo-labelled instance and the hyperplane. For example, pseudo-labelled instances A and B fall on the side of the positive hyperplane. The label quality of sample B is higher than sample A as B is much further away from the hyperplane. They proposed to set a high threshold in former iterations and decrease the threshold when the iteration number increases. A higher threshold for the first few iterations is proposed as the quantity of initial labelled training data is small. High quality pseudo-labelled data is preferred to prevent deterioration of classifier performance at later iterations. At later iterations, the threshold is lower to guarantee enough labelled data.

In Lee and Kim (2017), they select a set of class-balanced and confidently predicted instances to be included into the next iteration of classifier training. The predicted instances are first ranked according to confidence, then a set of top N instances is selected. The selected set is further divided into two disjoint sets based on sentiment polarities of each pseudo-labelled instance. The set with more instances is taken as the major class, whereas the other set is taken as the minor class. Instances from both classes are sampled based on the number of instances in the minor class to form a class-balanced set. This class-balanced and confidently predicted instances are included into the next training iteration and subtracted from unlabelled dataset.

Zhou et al. (2012) proposed a self-training algorithm that decreases the disagreement region of hypotheses based on three properties of informative unlabelled data. Informative unlabelled data improve classifier performance if they provide additional information on the true decision boundary, retain the overall data distribution and introduce bounded noise. They transform the problem of selecting a set of informative unlabelled data to the problem of rejecting the same set of unlabelled data with its labels inverted. This is due to when the inclusion of an unlabelled dataset has a marginal impact on classification of labelled data, it is difficult to conclude whether the classifier is approaching the true decision boundary. On the other hand, when the addition of unlabelled data introduced a much greater misclassification of the labelled data, the unlabelled data with current assigned labels can be confidently recognized as harmful, because they are incorrectly labelled, or they have been sampled disproportionately. Furthermore, not all unlabelled data is helpful in improving the classifier's performance despite being correctly labelled. Thus, only the subset of unlabelled data that deteriorates the performance of a new classifier trained

using the training set including the same set of unlabelled data with inverted labels, will be added into the training set.

In active learning settings, a special case of semi-supervised learning is in which a learning algorithm can interactively query a user to label new data points with desired outputs. Chosen data points are usually the ones that the model is uncertain about and are believed to have huge impacts on model performances. In Sharma & Bilgic (2017), they proposed an evidence-based framework that can uncover the reasons for why a model is uncertain on a given instance. Two reasons for uncertainty in a model are discussed. A model can be uncertain about an instance because it has strong but conflicting evidence for both classes, introduced as conflicting-evidence instances. On the other hand, a model can be uncertain about an instance because it does not have enough evidence for either classes or known as insufficient-evidence instances. Their work found out the conflicting-evidence instances significantly improved the learning efficiency of a model, whereas the insufficient-evidence instances provided least value to a model.

Table 2.1 Summary of related works for semi-supervised learning in sentiment classification

Study	**Research Work**	**Method**	**Instance Selection Strategy**	**Target**
Duan et al. (2020)	Semi-supervised generative model for sentiment classification	Decision tree, SVM, NB		Sentiment feature extraction from stock messages with semi-supervised generative emotion model using categorized words
Gengo & Verri (2020)	Semi-supervised sentiment analysis of Portuguese Tweets with Random Walk	Random walk in Feature-Sample Networks (graph-based)		Improve model performances on classifying short text with unknown class prior probabilities
Ji, Yan, Ying, Chen, & Su (2020)	Semi-supervised sentiment analysis for Chinese stock texts	Multi-Layer Perceptron, LSTM, GRU		Increase sentiment analysis efficiency on Chinese financial text data
Bijari et al. (2019)	Graph-based sentiment classification	CNN		Improve performance in sentiment classification task leveraging graph-based representation and deep learning
Chen et al. (2019)	Co-training semi-supervised for sentiment classification	CNN		Improve classification accuracy leveraging word embedding and character-based embedding

Table 2.1 Continued

Study	Research Work	Method	Instance Selection Strategy	Target
Y. Han, Liu, & Jin (2019)	Semi-supervised sentiment analysis based on dynamic threshold and multi-classifiers	SVM	Dynamic threshold based on quality and quantity of auto-labelled training data	Improve accuracy of semi-supervised sentiment analysis by extending labelled data with dynamic threshold and multi-classifiers
Xu & Tan (2019)	Semi-supervised target-oriented aspect-based sentiment classification	Recurrent Neural Network (RNN)		Target-oriented aspect-based sentiment analysis with deep generative model using context variable and sentiment variable
Fu, Sun, Wu, Cui, & Huang (2018)	Weakly supervised generative model for sentiment analysis	Gibs sampling		Improve topic identification and sentiment recognition by incorporating word embeddings and HowNet lexicon for sentiment classification
S. Lee & Kim (2017)	Using sentiment labelling for extending labelled data for semi-supervised sentiment classification	Unsupervised joint sentiment/topic model combine with semi-supervised training	Confidence-based and class-balanced instances	Extend number of labelled data and improve self-training model performance by filtering confidently predicted instances

Table 2.1 Continued

Study	Research Work	Method	Instance Selection Strategy	Target
Sharma & Bilgic (2017)	Evidence-based uncertainty sampling for active learning	NB	Evidence-based uncertainty	Improve performance of learning by uncover the reasons for a model's uncertainty
Khan et al. (2016)	Semi-supervised sentiment analysis using revised sentiment score based on SentiWordNet	SVM combined with lexicon-based approach		Build a Sentiment Knowledge Base using Information Gain and Cosine Similarity to improve sentiment analysis performance
Y. Zhou et al. (2012)	Self-training with selection-by-rejection	Logistic regression and SVM	Rejection strategy	Improves performance of self-training by decreasing the disagreement region of hypotheses

2.4 Gap analysis

Based on the literature, existing works is improving sentiment classification models using semi-supervised learning approaches. However, the review shows that there are limited studies focused on instance selection strategies to improve the performances of semi-supervised sentiment classification models. The focuses is commonly on unsupervised pre-trained networks and external knowledge bases. Besides, instance selection strategies proposed are predominantly to identify confidently predicted instances instead of informative instances.

Quality and the quantity of pseudo-labelled data are two key factors in the dynamic threshold proposed by Han, Liu and Jin (2019). Quality of pseudo-labelled instances is defined accordingly to the prediction model used. However, high and low thresholds mentioned in the research were not defined or suggested. The instance selection strategy proposed by Lee and Kim (2017) is simple and easily adapted to other works but the threshold value suggested is extensively tested on joint sentiment/topic model only. Adapting the proposed instance selection strategy requires the examination of an optimal confidence threshold value. These proposed methods identify confidently predicted instances only. Three properties of informative unlabelled data are introduced by Y. Zhou et al. (2012), they should be able to provide additional information on the true decision boundary, retain the overall data distribution, and introduce bounded noise. Nevertheless, the method proposed is not efficient because it involves training and testing of all unlabelled subsets, reverting the label back and forth. This requires high computational resources for model training. All this leads to the need for an efficient method to discover informative data points that contribute positively to model performance.

Besides, the ratio or number of labelled and unlabelled training instances used in experiments are rarely reported in studies, despite the fact that semi-supervised learning has been used extensively in sentiment classification. This leads to resources such as time and money being not utilized to its fullest as there is no guidance on the estimated number of labelled text required in the training. Therefore, a suggestion on the optimal ratio of labelled and unlabelled instances is required.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter presents the research methodology to be implemented, with the purpose of identifying informative unlabelled instances. Section 3.2 presents a snippet of data used for the ease of understanding proposed methodology which will be introduced in section 3.3. Details of the flow are explained in sections 3.3.1 and 3.3.2. The concept and formula of the proposed review informative score will be further explained in sections 3.3.2(a) to 3.3.2(d).

3.2 Data preview

An example of two datasets, product metadata and product review, are provided in Table 3.1 and Table 3.2 for the ease of following the flow of proposed methodology which will be explained in section 3.3. productID acts as the key attribute that links the relationship between product metadata and product review. A full description of the dataset used will be introduced in Chapter 4.

Table 3.1 An example of product metadata

Attributes	**Description**	**Value**
productID	ID of product	"0000069512"
productTitle	Name of product	"Refrigerator Storage Organizer"
productFeature	Features of product	["Great organizer for fridge", "Easy to carry", "Clear material"]
productDesc	Description of product	"This organizer is great for fridge. Cutout side handles for easy carry. Clear material enable easy see through what are stored inside the bin."

Table 3.2 An example of product review

Attributes	Description	Value
productID	ID of product	"0000069512"
helpful	Helpful votes of review	3
review	Text of review	"Love these bins to help me keep my fridge organized. These bins not only has helped me see what I have but makes me happy seeing how tidy it looks now too!"
overall	Rating of product	5.0
reviewTime	Time of review	"01 28, 2009"

3.3 Methodology description

Given a dataset R consists of n review entries, after pre-processing and matching with their respective product ID, the product metadata (product title, product features, and product description), review posted date and helpful votes (similar to upvotes and like) are used to calculate the review informative score S(I). The proposed review informative score S(I) consists of two parts: content score S(C) and popularity score S(P). S(C) is calculated using the product metadata and with the help of SentiWordNet whereas S(P) is calculated using the review posted date and its helpful votes. S(I) is the sum of both S(C) and S(P) multiplied by weights, where the sum of two weights is one. The dataset is then split into two sets, labelled set L and unlabelled set U, and then transformed into vectors. The labelled set L is used to train a model, then the model is used to predict labels for unlabelled set U. The predicted labels are either positive or negative. Along with the confidence given by the model, the pseudo-labelled set is ranked using the sum of both confidences and review informative score S(I) multiplied with weights, where the sum of two weights is one. Confidence is a value representing the confidence level of the model predicting the label correctly. The top ten percent of predicted instances are chosen for data augmentation and dropped from the unlabelled set U. The model is then retrained using the augmented data and