Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2018 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

6-26-2018

Multiple-Domain Sentiment Classification for Cantonese Using a Combined Approach

E.W.T. Ngai
The Hong Kong Polytechnic University, eric.ngai@polyu.edu.hk

M.C.M. Lee
The Hong Kong Polytechnic University, Maggie.lee@polyu.edu.hk

Y.S. Choi
The Hong Kong Polytechnic University, star.choi@polyu.edu.hk

P.Y.F. Chai
The Hong Kong Polytechnic University, Paul.chai@polyu.edu.hk

Multiple-Domain Sentiment Classification For Cantonese Using a Combined Approach

Research-in-Progress

E. W.T. Ngai
Department of Management and Marketing
The Hong Kong Polytechnic University
Kowloon, Hong Kong, PR China
eric.ngai@polyu.edu.hk

M. C. M. Lee
Department of Management and Marketing
The Hong Kong Polytechnic University
Kowloon, Hong Kong, PR China
maggie.lee@polyu.edu.hk

S. Y. S. Choi
Department of Management and Marketing
The Hong Kong Polytechnic University
Kowloon, Hong Kong, PR China
star.choi@polyu.edu.hk

P. Y. F. Chai
Department of Management and Marketing
The Hong Kong Polytechnic University
Kowloon, Hong Kong, PR China
paul.chai@polyu.edu.hk

Abstract

In this study, we proposed a combined approach, which amalgamates machine learning and lexicon-based approaches for multiple-domain sentiment classification that supports Cantonese-based social media analysis. Our study contributes to the existing literature not only by investigating the effectiveness of the proposed combined approach for supporting social media analysis in the Cantonese context but also by verifying that the proposed method outperforms the baseline approaches, which are commonly used in the literature. We demonstrated that social media network-based classifiers can be general classifiers that support multiple-domain sentiment classification.

Keywords: Sentiment classification, sentiment analysis, Cantonese-based social media analysis

Introduction

Manufacturers and service providers are extremely interested in learning the polarity of comments and opinions written by people in social media because they enable companies to analyze the weaknesses of their products, promptly respond to negative comments, manage their brands, and improve their products and services. In particular, such entities note the comments issued by key opinion leaders because such statements influence the purchase decisions of other consumers. The text in social media is noisier, more informal, less grammatical, and typically shorter than edited text (Baldwin et al., 2013). Therefore, pre-processing and analysis should be designed for this type of short text.

Cantonese is an important dialect in several regions of Southern China (including Hong Kong, Macau, and Guangdong) and in overseas Chinese communities in other countries (such as Canada and the United States). Along with the pronunciation, grammatical, and lexical differences between Cantonese and Mandarin, a thousand extra characters invented specifically for Cantonese (Cheung & Bauer, 2002) makes written Cantonese be unintelligible to Mandarin speakers. Changes in the attitudes of the media and the people of Hong Kong toward written Cantonese have occurred over the recent decades, so the social role of written Cantonese may extensively expand in the future (Snow, 2008). Owing to the growth of Web 2.0, local online users frequently represent their ideas, opinions, and views in written Cantonese in social media. Research on sentiment classification has been conducted recently in English, Chinese, and Japanese. However, limited works have been performed on Cantonese sentiment classification for reviews in social media despite the global influence of Cantonese and its usage by 70 million people. We assume that the role of written Cantonese will become increasingly important, and analyzing written Cantonese in social media is a significant endeavor.

The crucial role of written Cantonese in social media and the limitations of existing sentiment analysis algorithms motivated us to identify an improved method for supporting Cantonese sentiment classification that can provide acceptable accuracy in different domains without causing costly training and maintenance. We assume that people discuss numerous issues on social media platforms; thus, the content of social media posts is general and may involve different topics. Therefore, we hypothesize that machine learning-based classifiers trained with the social media dataset perform better than the classifier trained with a specific domain dataset when performing the cross-domain analysis. We can utilize the classifier that is trained by social media data for classification in multiple domains with acceptable accuracy. No similar research has been performed on such classifier.

Furthermore, we implemented our proposed ensemble-based methods to further improve the classification result. Theoretically, if the base classifiers are highly accurate and diversified, then the ensemble classifier can outperform the base classifiers (Polikar, 2006). This condition is exactly the reason we consult the opinions of different experts when we have problems to reduce the risk of following the advice of a single expert whose experience significantly differs from that of other experts. In addition, many investigations confirmed that combining different classifiers produces better results than base classifiers. However, most extant combined approaches in sentiment analysis literature focus on English. To the best of our knowledge, no study is conducted for Cantonese sentiment analysis using the combined approach.

Hence, in this work, we will fill the aforementioned gaps and address the following research questions:

- Do machine learning-based classifiers trained with a social media dataset outperform a classifier trained with a specific domain dataset and provide an acceptable accuracy when performing the multiple-domain analysis?
- Is a Cantonese lexicon necessary for analyzing the sentiment from a Cantonese content?
- Do ensemble-based methods improve the accuracy of sentiment classification in Cantonese content that is written for social media?
- Which combination rule is optimal for the ensemble-based methods for Cantonese sentiment analysis?

To resolve the abovementioned questions, we will first examine if a classifier that is trained with a social media dataset can alleviate error given a cross-domain analysis and if the dictionaries we developed can handle content in Cantonese. Second, we will combine the lexicon- and machine learning-based approaches. We will investigate if the combination approach is more effective for capturing the sentiment from reviews and posts written in Cantonese at multiple domains and ascertain which rules are optimal. In this paper, although we cannot answer all the research questions completely, we will present the proposed approach and provide some preliminary results.

Proposed Approach

In this study, we propose the combination of three member classifiers, that is, Support Vector Machines (SVM) with Naïve Bayes (NB) features (NBSVM)-based, convolutional neutral network (CNN)-based, and lexicon-based classifiers. The three member classifiers were selected to form the ensemble classifier considering their high accuracy and popularity in the literature. Ensemble classifier outperforms the base classifiers only when the base classifiers are highly accurate and diversified (Polikar, 2006).

We reviewed the findings from the literature and the result in the classifier training stage to assess the accuracy of the member classifiers. The results reported in existing studies indicate that all three classification methods achieve high accuracy (Wang & Manning, 2012; Deriu et al., 2016; Zhang et al., 2012). Meanwhile, the outcomes of the NBSVM- and CNN-based classifiers on the validation training dataset in this study involved scores that are greater than 72. We also used the lexicon-based classifier to classify the labels for the validation data to estimate its performance and obtained a result over 65. Therefore, the three classifiers are robust.

In the proposed approach, we aim to combine the results of the member classifiers by using several combined approaches to determine the final polarity. The flowchart of our approach in this study is shown below.

Input messages, such as segmentation based on some dictionaries, are pre-processed and then fed into the next step to develop the matrix, in which each word in the message is converted to a vector. The matrix is entered into a trained NBSVM-based and a trained CNN-based classifier to obtain two sets of sentiment analysis results. The features of the input messages, including lexicons, negations, and modifiers, are extracted for calculating the sentiment score in accordance with several semantic rules and then passed onto the lexicon-based classifier. The results of the three classifiers are combined by the combination rules. The following paragraphs introduce the detail of each step and the member classifiers.

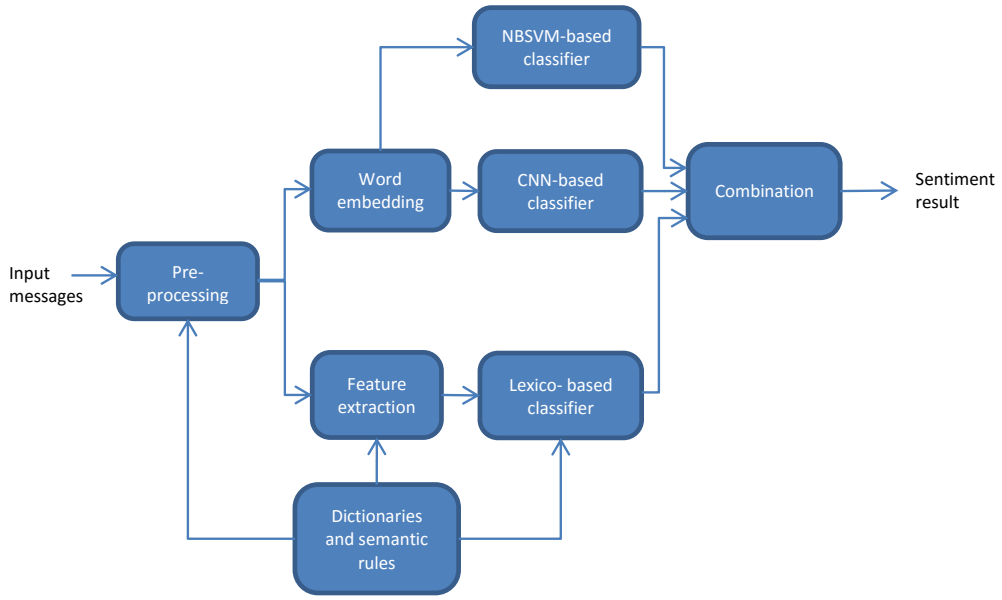

Figure 1. Flowchart of the proposed approach.

Pre-processing

First, messages longer than 100 words are trimmed to 100 words because messages are limited to 7–100 Chinese words. Unlike in English, no space is found between words in the Cantonese-written sentence. Thus, we must perform segmentation on the input messages. Features, such as usernames, URLs, and hashtags, are also removed from the dataset.

Word embedding

In natural language processing, we must represent words in a manner that allows a computer to "understand" natural language and use them as an input for any model. Thus, we encode each word token into the vector that represents a point in a certain "word" space (Chaubard et al., 2016); these word representations are called word embeddings or word vectors.

In this study, word embeddings are initialized using word2vec (Mikolov et al., 2013), which is a common tool for computing vector representations of words, and trained by a corpus of 1.6 M articles from the Chinese Wikipedia. The meaning of words and relationships between words are encoded spatially by using word2vec. A word vector is produced by applying a skip-gram model with a window size of 5. Words with a total frequency that is lower than 5 were ignored. The dimensionality of the vector representation is 300.

Feature extraction

Different features, which are words or terms that determine the sentiment of the message, must be extracted from the input message to calculate the sentiment polarity score. In our approach, extracted features include lexicons, negation, modifier, and neutralization. Table 1 describes the required features in detail.

Table 1. Features extracted for the lexicon-based classifier.

Features	Description
Lexicons	Lexicons are the opinion words extracted from the sentence. A polarity score that ranges from −2 to 2 (except 0) is assigned to each word depending on the sentiment strength. “好靚” (beautiful) and “舒服” (comfortable) are examples of positive opinion words, whereas “醜樣” (ugly) and “艱難” (difficult) are examples of negative opinion words.
Negation	A negation word or phrase typically reverses the polarity of the sentence. It includes “不” or “唔” (not) and “但係” (but). For example, “呢個牌子唔好用” (this brand is not good).
Modifier	A modifier strengthens or weakens the polarity score of the opinion word. Different degrees of modification ranging from −2 to 2 (except 0) are assigned to the modifiers. “十分” (very), “強烈” (strongly), and “輕微” (slightly) are examples of modifiers.
Neutralization	A zero score is assigned for the whole sentence when the conjunction for conditional sentences, such as “假如” (if), or when words that express doubt, uncertainty, or ambiguity, including “可能” (maybe) and “不怎麼” (not really), appears.

Dictionaries and semantic rules

A lexicon-based classifier relies on dictionaries and semantic rules to calculate sentiment polarity scores. A list of seed opinion words is available and is used for locating other opinion words (synonyms and antonyms) to build the basic sentiment word dictionary for Chinese. Words, such as “廣州話俗語詞典” and “實用廣州話分類詞典”, in several Cantonese dictionaries are added in the basic Chinese dictionaries to enrich the dictionaries with additional Cantonese words or phrases. Furthermore, certain online dictionaries for local Hong Kong-Cantonese (e.g., “香港常用及地道式廣東話”, “粵典”, “開放詞典” and “香港網絡大典”) are also added into the dictionary database to develop the dictionary for local Cantonese particularly for Hong Kong-Cantonese. Next, we expanded the dictionaries by incorporating trendy Cantonese words used in social media on the basis of words discovered in an expansive corpus in Hong Kong (e.g., from online news, magazines, and discussion forums).

Classifiers

Three member classifiers, namely, CNN-, NBSVM-, and lexicon-based classifiers, are used to form an ensemble classifier. We experimented with the three standard classifiers (algorithms) that have been proven effective in previous text categorization studies.

CNN-based classifier

A CNN is a special kind of NN with a distinctive architecture. The CNN architecture typically consists of three layers, namely, a convolutional, a pool, and a fully connected layer. These layers are stacked to form a basic CNN architecture (O'Shea & Nash, 2015). In the present study, we simplified the approach that was proposed by Deriu et al. (2016). We only used a two-layer CNN to obtain the

classification results instead of combining the results of two two-layer CNNs. We developed the CNNs using the parameters as shown in Table 2.

Table 2. The summary the parameters used in our CNN system.

Parameters	Values
Number of convolutional filers	m1 = m2 = 200
Filter window size h	h1 = [3,4,5], h2 = 2
Size of first max-pooling interval	Width = 2, striding =2
Activation function α	relu

NBSVM-based classifier

SVMs were first introduced by Cortes and Vapnik in 1995 (Cortes & Vapnik, 1995). SVMs are the learning methods used for binary classification, and their basic idea is to identify a hyperplane that separates the d-dimensional data perfectly into its two classes (Boswell, 2002). In the present study, we used the NBSVM, which consistently performs well on snippets and extended documents in the topic of sentiment and subjectivity classification (Wang & Manning, 2012). NB is another common robust classification technology, whereas NBSVM is a simple model variant in which the SVM is built over NB log-count ratios as feature values (Wang & Manning, 2012). In the present study, we implemented the NBSVM using parameters, which are the same as the values mentioned in Wang and Manning (2012): $\alpha = 1$, $C = 1$, and $\beta = 0.25$.

Lexicon-based classifier

The lexicon-based classifier in the present study determines the sentiment based on the values of the sentiment polarity scores of the messages. The following passages describe the pseudo code for calculating the sentiment polarity scores for each message. Pre- and post-factors mean the negation or modifier before or after the sentiment word. A maximum of two post- or pre-factors, such as "不會太高興" (will not be very happy), are considered. The sentence score was calculated by using this algorithm. Moreover, the classifier determines the polarity based on the value of the score.

Classifier combination

In the present study, we implemented and compared three non-trainable fixed combination approaches, which are commonly used for ensemble systems. We combined the results of the member classifiers by majority voting, weighted voting, and a hybrid (sequential) approach.

Majority voting

Majority voting is the most popular voting method (Zhou, 2012) for classification. We are given a set of individual classifiers, and our task is to combine the classifiers to predict the class label from a set of possible class labels. The ensemble selects the class that receives more than half of the votes (and must be the highest number of votes) from the individual classifiers. In the present study, we considered the strategy of neutral default, in which if no prediction is created, then the classification result of that classifier is set to neutral as a default.

Weighted voting

Majority voting and weighted voting for classification are widely used. Zhou (2012) defined weighted voting as a weight which is assigned to each member classifier in voting. We obtain a set of the weights, which are theoretically in proportion to the performance of the individual classifier, such that the classifier with better performance will be assigned a greater weight and vice versa. The weights are normalized and larger than 0 and their sum are equal to 1.

Hybrid approach

The hybrid approach was based on the idea of hybrid classification proposed by König and Brill (2006). A three-stage classifier is constructed in the present study. The first classifier categorizes the input message as positive or negative; otherwise, the result of the second classifier is verified. If the second classifier cannot categorize its polarity, then the last classifier outcome is considered the final sentiment classification result. The sequence of the classifiers will be determined on the basis of their performance in the training and development stage. The first classifier should be the most accurate among the three classifiers.

Experiment

Data are collected from several leading and popular social media platforms and review sites in China and Hong Kong for training and testing. Dianping is specific to shopping, while Hong Kong Movies involves film reviews in Hong Kong. Unlike the review websites, Sina Weibo and Facebook users post short messages about a variety of topics and not on a specific issue. In order to ensure that the majority of posts are written in Cantonese, the data collected is from the posts which are written by the users who registered their regions as Hong Kong or the posts which are related to Hong Kong products and services. The datasets are polarity balanced. Table 3 summarizes the number of messages that we collected from the sources.

Table 3. The total amount of messages collected from each source.

Source	Domain	Dataset Size
Sina Weibo	Social media	~3,200,000
Facebook	Social media	~100,000
Dianping	Shopping	~155,000
Hong Kong Movie	Movies	~77,000

Classifier training

All the training datasets involve data that were collected from Facebook and Sina Weibo, which represent the writing style in Hong Kong and Mainland China, respectively. The NBSVM- and CNN-based classifiers involve three training stages, namely, distant supervised, supervised, and validation. Table 4 displays the data counts in the datasets in different stages of training. We used Facebook and Sina Weibo messages with emoticons as the training data to train the NBSVM- and CNN-based classifiers through distant supervised training (Go et al., 2009). Then, we performed supervised training with a limited number of hand-labeled data from Facebook and Sina Weibo. The supervised training data entail posts related to the 25 most controversial organizations and people in China and Hong Kong in recent years. After the training, the trained classifier was validated with another hand-labeled dataset, which is equal to approximately 10% of the amount of the training data, to ensure that the classifier is not overfitting.

Table 4. Stage of training and dataset size for the social media based classifiers.

Stage of Training	Sources	Dataset Size
Distant supervised	Facebook (HK)	89,790
	Weibo (China)	3,138,906
Supervised	Facebook (HK)	19,824
	Weibo (China)	16,656
Validation	Facebook (HK)	1,941
	Weibo (China)	2,079

Preliminary Results

We attempted to combine the results of the three base classifiers with three fixed combination approaches and compared the accuracy of the sentiment classification to evaluate the combined approaches and identify the optimal combination rules in this study. Moreover, we compared the results with the baseline approaches (the three base classifiers) to assess if the combined approaches are superior.

For majority voting, the message was classified as neutral in the event of a tie. The weights, which are proportional to $\log\frac{p_i}{1-p_i}$ are: w_{cnn} = 0.43, w_{NBsvm} = 0.34, and $w_{lexicon}$ = 0.23, for weighted voting in accordance with the performance of the individual classifiers, p_i, in the validation stage. The sequence of classifiers for the hybrid approach in accordance with their performances should be as follows: CNN → NBSVM → lexicon.

We have calculated the F_1 scores of the three combination approaches after applying them on the Weibo validation dataset, and compare them with the F_1 scores of the baseline methods at the validation stage. Also, we have applied the three combination approaches and the baseline methods on a shopping (Dianping) dataset of size 15,492 (which has not been used in the training). The labeling of the dataset is based on the star rating (scale of 1 – 5 stars) given by users to the product or services. In Table 5, comparing the combination rules reveals that the weighted voting rule provides optimal classification result in different domains.

Table 5. Performance results in F1 score of different approaches in two domains.

Classification Technology	Social media		Shopping
	Weibo	Facebook	Dianping
Weighted voting	77.6	72.0	72.4
Majority voting	77.3	72.0	71.9
Hybrid approach	76.8	70.3	71.7
NBSVM	72.5	68.1	58.9
CNN	77.1	70.4	72.0
Lexicon (with Cantonese dictionaries)	65.6	62.8	68.5
Lexicon without Cantonese dictionaries	43.8	42.8	59.6

Table 5 comparing the combination rules reveals that the weighted voting rule provides optimal classification result in different domains. Such outcome is due to we considered the accuracy of the individual classifier as a factor for determining the weights. Differences of 12 in Weibo, 9.2 in Facebook, and 10.5 in Dianping are noted between the best (weighted voting) and worst member classifiers. However, the scores of the weighted and majority voting rules were very close. This result is due to three classifiers are too few to allow us to demonstrate the importance of the weights. The results also show that the social media-based classifiers can also provide a high accuracy in other domain. Besides, comparing to the lexicon approach without adding the Cantonese dictionaries, the lexicon approach shows a significant improvement (+15% to 50%), so we believe that the Cantonese dictionaries play an important role in improving the accuracy of the lexicon approach.

Conclusion and Future Work

This study preliminarily demonstrates that social media-based classifiers can function as a general classifier that supports multiple-domain sentiment classification. Such general classifier can handle

cross-domain classification without costly training and maintenance. In addition, the Cantonese dictionaries we developed in this study were confirmed to be effective. Furthermore, we investigated several combined approaches and verified that the combined approach enhances the accuracy of Cantonese sentiment classification and outperforms the extant approaches that are commonly used in the literature. We assume that our empirical results are likely to lead a new trend of Cantonese sentiment analysis for Cantonese reviews.

As the next step, we will test the proposed method in more testing datasets and in other different domains, including reviews for restaurants, cosmetics, vehicles, or cell phones. Besides, we will implement and test the feedback mechanism for continuous improvement by feeding the discrepancy. The discrepancy among classifiers will be reviewed manually for future improvements. Any discrepancy among the three classifiers will be reviewed and examined manually. The dictionaries and semantic rules may be modified or the NBSVM- and CNN-based classifiers may be re-trained based on the discrepancy to increase accuracy.

Acknowledgements

This research was supported in part by the Innovative Technology Fund provided by the Innovation Technology Commission (grant number ZR1Y) and The Hong Kong Polytechnic University under a research grant (grant number ZVK9). The authors are grateful for the constructive comments of the two referees on an earlier version of this paper.

References

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. 2013. "How noisy social media text, how diffrnt social media sources?," In *Proceedings of IJCNLP*. pp. 356-364.

Chaubard, F., Mundra, R. and Socher, R. 2016. *Lecture Notes: Part I of CS224d: Deep Learning for Natural Language Processing*, Stanford University, URL: http://cs224d.stanford.edu/lecture_notes/notes1.pdf.

Cheung, K., and Bauer, R. 2002. "The representation of Cantonese with Chinese characters," *Journal of Chinese Linguistics, Monograph Series number 18*. Berkeley: University of California.

Cortes, C., and Vapnik, V. 1995. "Support-vector networks," *Machine learning* (20:3), pp. 273-297.

Boswell, D. 2002. *Introduction to support vector machines*, University of carlifornia, San Diego.

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., and Jaggi, M. 2016. "SwissCheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision," In *Proceedings of SemEval*, pp. 1124-1128.

Go, A., Bhayani, R., and Huang, L. 2009. "Twitter sentiment classification using distant supervision," *CS224N Project Report*, Stanford, (1:12).

König, A. C., and Brill, E. 2006. "Reducing the human overhead in text categorization," In *Proceedings of the 12th ACM SIGKDD conference on knowledge discovery and data mining Philadelphia*, Pennsylvania, USA, pp. 598–603.

Mikolov, T., Le Q. V., and Sutskever, I. 2013. "Exploiting Similarities among Languages for Machine Translation," *arXiv*.

O'Shea, K., and Nash, R. 2015. "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*.

Polikar, R. 2006. "Ensemble based systems in decision making," *IEEE Circuits and systems magazine* (6:3), pp. 21-45.

Snow, D. 2008. "Cantonese as written standard?" *Journal of Asian Pacific Communication* (18:2), pp. 190-208.

Wang, S., and Manning, C. D. 2012. "Baselines and bigrams: Simple, good sentiment and topic classification," In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics, pp. 90-94.

Zhang, W., Xu, H., and Wan, W. 2012. "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," *Expert Systems with Applications* (39:11), pp. 10283-10291.

Zhou, Z. H. 2012. Ensemble methods: foundations and algorithms. CRC press.