Advertising Legality Recognition

Yi-jie Tang, Cong-kai Lin, Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University
#1, Sec.4, Roosevelt Road, Taipei, 10617 Taiwan
tangyj@nlg.csie.ntu.edu.tw,r00922106@ntu.edu.tw,hhchen@ntu.edu.tw

ABSTRACT

As online marketing and advertising keep growing on the Internet, a large amount of advertisements are presented to consumers. How consumers, advertisers and the authorities identify false and overstated advertisements becomes a critical issue. In this paper, we address this problem, and propose various classification models to detect illegal advertisements. Illegal advertisement lists announced by the government and legal advertising data crawled from an online shopping website are used for training and testing the classification models. Naïve Bayes and SVM classifiers with various feature settings are explored on food and cosmetic datasets to demonstrate their feasibility. The experimental results show that log relative frequency ratio can be used as weights for unigram features to achieve the best accuracy. The accuracies of SVM classifiers on food and cosmetic datasets are 93.433% and 86.037%; the false alarm rates are 0.083 and 0.166; and the missing rates are 0.053 and 0.115, respectively. Log relative frequency ratio is further used to mine verb phrases consisting of a transitive verb and an object noun from the illegal datasets. The mined verb phrases, which form an illegal advertising statement list, can be used as a reference for both the advertisers and the authority.

廣告合法性偵測

隨著線上廣告和行銷活動的快速發展，每天都有大量的廣告內容透過網際網路呈現在使用者眼前。因此，不論對於消費者、廣告主、或是政府相關單位來說，如何辨識誇大不實或具誤導性質的廣告，都已成為一項重要的課題。本研究提出數種分類模型來進行廣告合法性偵測。為了取得具有合法性標記的語料，我們採用政府單位公布的違規廣告資料，並從購物網站擷取合法廣告內容，以作為訓練和測試資料。資料分為食品廣告資料集和化粧品廣告資料集，分別以 Naive Bayes 和 SVM 分類器搭配不同特徵進行合法性偵測。實驗結果顯示使用相對頻率比率對數 (log relative frequency ratio) 來代表單字組 (unigram) 的權重並作為特徵時，能達到最佳準確率；在此模型下，食品和化粧品資料集的 SVM 分類準確率分別達到 94.433% 與 86.037%，其錯認率 (false alarm rate) 分別為 0.083 與 0.166，誤失率 (missing rate) 分別為 0.053 與 0.115。相對頻率比率對數也用於對非法廣告資料集進行動詞組的探勘，這些動詞組皆由動詞與其受詞組成，所形成的非法廣告用詞表可讓廣告主和政府單位作為辨識廣告合法性的參考依據。

KEYWORDS : Ad classification, collocation mining, computational advertising, legality recognition
關鍵詞 : 廣告分類, 詞語搭配探勘, 計算式廣告, 合法性辨識

Proceedings of COLING 2012: Posters, pages 1219–1228,
COLING 2012, Mumbai, December 2012.

1 Introduction

As online advertising keeps growing on the Internet, this new form of marketing has started to be regulated by advertising law. Two forms of advertising regulation exist, namely statutory regulation and self-regulation, to protect consumers from fraudulent and misleading advertising (FTC, 2000; CFIA, 2010; DOH, 2009). Under the food and cosmetic advertising regulations of Taiwan, food-related and cosmetic-related advertisements cannot be false, overstated or misleading, and should not mention any curative effects. Advertising statements that violate the regulations are called *illegal statements*. Advertisements containing illegal statements are regarded as illegal advertisements. Because of a large amount of advertisements are presented to consumers, how to recognize advertising legality automatically becomes an important task.

Besides consumers, several parties who are involved in online advertising can benefit from the automatic illegal advertisement recognition (*IDR*). On the one hand, the authority has to examine advertisements to decide which can be presented to Internet users. That requires a lot of time. An advertising legality recognition system not only saves much human effort, but also reports the illegal advertisements in real time. On the other hand, advertisers need to avoid legal issues while maximizing the effectiveness of their advertising. Even weblog and auction website users may need to take care of legal issues. Texts and images from their websites and auctions may be related to products, and thus may also be regarded as online advertising by the authority. Websites that deliver marketing messages from other companies may want to show only truthful advertisements to their users and avoid providing illegal and misleading ones.

Computational advertising has attracted much attention in recent years. How to "best match" between a given user in a given context and a suitable advertisement is one of the major issues. Gabrilovich, Josifovski and Pang (2008, 2009) gave tutorials on this trend in ACL 2008 and IJCAI 2009. Previous Internet advertising focuses on bidding (selecting) advertisements and placing them in the best (right) positions. Ghosh et al. (2009) proposed bidding strategies for the allocations of advertisements. Edelman, Ostrovsky and Schwarz (2007) investigated generalized second-price (GSP) auction for online advertising. Huang, Lin, and Chen (2008) classified instant messaging dialogues into the Yahoo categories, and applied the method to advertisement recommendation. Cheng and Cantú-Paz (2010) proposed a framework to predict the probability that individual users click on ads. Scaiano and Inkpen (2011) used Wikipedia as an annotated corpus to find negative key phrases to avoid displaying advertisements to non-target audience.

Unlike advertisement bidding, matching and recommendation in computational advertising, this paper deals with illegal advertisement recognition. Illegal advertising is similar to ad spam[1] in financial gain, but the former exploits false, overstated or misleading statements to defraud customers, and the latter creates artificial ad traffic, inflates click/impression, and so on, to defraud online advertising systems like AdWords. To the best of our knowledge, advertising legality recognition is a pilot study in this research direction. Food, cosmetic, and medicine are three major sources of illegal advertising. Since advertisements that make health claims are highly regulated in many countries, we focus on food-related and cosmetic-related advertising in this paper. We introduce NLP techniques to extract critical features for illegal statement detection. Section 2 introduces the experimental datasets. Section 3 presents legality recognition methods. Section 4 proposes an approach to illegal verb phrase mining. The last concludes the remarks.

[1] http://support.google.com/adwordspolicy/bin/answer.py?hl=en&answer=50424

2 Datasets

The first step for the advertising legality recognition research is to obtain advertisements with appropriate labels. Since advertising legality can only be determined by the authority, we need to obtain official announcements regarding illegal advertisements. We collect the illegal food and cosmetic advertisement lists made public by the Taipei City Government[2] from July 2009 to November 2011. Each item in the list contains a product name and the corresponding problematic advertising statements. Figure 1 shows a food advertisement consisting of a product name (the 1st line) and illegal food advertising statements (the 2nd-4th lines). The legal parts are removed and denoted by "…". English translation is listed after Chinese food advertisement for reference.

活百O2高溶氧水
可潤腸通便，改善腸胃道的血行，清除宿便，預防痔瘡及治療高血壓、低血壓、肥胖症…活化細胞…改善腦細胞的血液體環境，血液黏稠度…增加唾液之分泌，血液的循環和 血紅球隊氧合(活血)…減少代謝廢物的堆積…失眠及疼痛…消除宿醉…

HOPPER High Oxygen Water
Can remove intestinal obstruction, improve blood flow of the stomach and intestines, prevent hemorrhoid, and cure hypertension, hypotension, and obesity … Activates cells … Improves blood conditions of brain cells and blood concentration … Increases saliva and promotes blood circulation … Reduces waste produced by metabolism process … Stops insomnia and pain … Stops hangover

FIGURE 1– An Example of an illegal food advertisement

The above example shows the fact that the government prohibits the use of statements related to curative effects and improvement of physical conditions. Most illegal statements listed by the government are verb phrases consisting of a verb and an object noun. According to the observed patterns, we propose methods to expand these terms and find similar phrases in the datasets, as described in Section 3.2 and Section 4. This can improve the recognition tasks and help the authorities and advertisers to find problematic expressions.

Since the government web site does not announce the legal advertisements, we need to collect legal advertising data from other sources. An online shopping website in Taiwan[3] is used to collect legal food and cosmetic advertising items. We assume most of these advertisements comply with advertising regulations, and these data are examined by human to make sure that unsuitable data are removed. Food and cosmetic product descriptions are used to build two legal advertising datasets: FOOD and COS, respectively. To obtain a balanced dataset, each dataset is collected from all related categories listed on the website.

All data are separated into sentences according to punctuations, including period, question mark, and exclamation. Only sentences with more than 3 characters are collected. Any expressions containing only product names are filtered out because product names cannot be used to determine its legality. All sentences are in Traditional Chinese. We perform Chinese word segmentation and part-of-speech tagging using the CKIP segmentation and POS tagging system.[4]

[2] http://www.health.gov.tw/Portals/0/%E8%97%A5%E7%89%A9%E9%A3%9F%E5%93%81%E8%99%95/10010food.pdf
[3] http://www.7net.com.tw
[4] http://ckipsvr.iis.sinica.edu.tw/

Thus, we have four datasets for legal food advertising, illegal food advertising, legal cosmetic advertising, and illegal cosmetic advertising. For clarity, they are named as FOOD_LEGAL, FOOD_ILLEGAL, COS_LEGAL, and COS_ILLEGAL. The numbers of instances in the four datasets are 5,059, 7,033, 10,520, and 11,381, respectively.

3 Advertising Legality Recognition

Advertising legality statement recognition aims at determining if an advertising statement is legal or illegal, so that it can be regarded as a binary classification problem. In the development processes, Naïve Bayes classifiers and SVM classifiers implemented with libSVM (Chang & Lin, 2001) are adopted. All training and test processes are based on 10-fold cross validation and every training model was tuned with the optimized parameters to achieve the best performance. Accuracy is adopted as an evaluation metric. Table 1 shows the experimental results. Two classification models (Naïve Bayes and SVM) with different feature settings are explored on food (FOOD) and cosmetic (COS) datasets. The following sections describe how various features are extracted for legality classification.

Classification Models →	Naïve Bayes		SVM	
Features ↓ Materials →	FOOD	COS	FOOD	COS
Unigram	89.148%	81.357%	88.851%	82.416%
Unigram + CILIN	88.950%	81.311%	89.728%	82.759%
Unigram + DOH	89.182%	81.553%	89.554%	83.658%
Unigram + CILIN + DOH	89.000%	81.439%	89.727%	83.325%
Unigram + logRF	**90.695%**	**85.179%**	**93.433%**	**86.037%**

TABLE 1 – Accuracies of advertising legality recognition models

3.1 Feature Set 1: Unigrams

Unigrams are considered as a fundamental feature set. We select the top 1,000 most frequent words from the legal and the illegal advertising datasets as features. Only content words including verbs, nouns and adjectives are included in order to remove the words that may not be relevant. Every sentence separated by punctuations forms an instance of the datasets, and each instance is represented by a word vector (w_1, w_2, …, w_{1000}), where w_i is a binary value indicating whether a word occurs in the sentence or not. The 3rd row of Table 1 shows the accuracies of Naïve Bayes classifiers and SVM classifiers on FOOD and COS datasets are (89.148%, 81.357%) and (88.851%, 82.416%), respectively. Bigram features are also tested, but the performance is lower than that of unigrams, so the results of bigram features are not included in this paper.

3.2 Feature Set 2: Health Related Terms

Advertising regulations are announced along with illegal advertising statement examples for advertisers' reference. Table 2 shows some illegal examples for food related regulations. The 1st type listed in the 1st column denotes mention of any curative effects and the 2nd type denotes false, overstated or misleading cases. Several subtypes along with the corresponding examples are listed in the 2nd and the 3rd columns, respectively.

Advertisers should not mention any curative effects on food and cosmetic advertisements under advertising regulations. We expand the words related to curative effects by a thesaurus to increase the coverage of the feature sets. These statements are used as auxiliary features, and are

Type	Sub-Type	Example
1	宣稱預防、改善、減輕、診斷或治療疾病或特定生理情形 (Claim of prevention, improvement, reduction, diagnosis or cure of diseases or physical conditions)	減輕過敏性皮膚病 (reduce allergic skin disease)
	宣稱減輕或降低導致疾病有關之體內成分 (Claim of elimination of substances that cause diseases)	解肝毒, 降肝脂 (remove poison and fat in liver)
	宣稱產品對疾病及疾病症候群或症狀有效 (Claim of effectiveness to diseases and symptoms)	消除心律不整 (cure arrhythmia)
	涉及中藥材之效能者 (Related to effects of Chinese medicine)	補腎 (improve health condition of kidney)
	引用或摘錄出版品、典籍或以他人名義並述及醫藥效能 (Reference to publications, books or statements by others with medical effects)	「本草綱目」記載：黑豆可止痛 (according to the book "Bencao Gangmu," black beans can ease pain)
2	涉及生理功能者 (Related to physiological functions)	分解有害物質 (decompose toxicants)
	涉及五官臟器者 (Related to organs)	增加血管彈性 (increase elasticity of blood vessel)
	涉及改變身體外觀者 (Related to change of appearance of human body)	防止老化 (prevent aging)
	引用本署衛署食字號或相當意義詞句者 (Reference to DOH permission numbers or related expressions)	通過衛生署配方審查 (pass formula review by DOH)

TABLE 2 – Illegal advertising statement examples announced by the government

combined with unigram features. Two kinds of auxiliary features shown as follows are used.

(1) All verbs related to curative effects from a Chinese thesaurus *Tongyicicilin* (Mei et al., 1984): This feature set is called CILIN in Table 1.
(2) Illegal statement examples listed by Department of Health (DOH) of Taiwan: This feature set is called DOH in Table 1.

The 4th-6th rows of Table 1 show the accuracies of using the above feature sets. Thesaurus expansion (Unigram + CILIN) has some positive effects in SVM classifiers. Comparing with pure unigram feature sets, integrating features selected from illegal advertising statement examples of DOH is also useful (refer to Unigram + DOH). However, the accuracy is not further improved, when all the three kinds of features are combined (refer to Unigram + CILIN + DOH). A possible reason is that the number of terms in the CILIN feature set is high, and a thesaurus always tries to collect as many terms as possible. Thus, many uncommon words are included as incorrect expansion. The DOH feature set includes lists that are edited by professionals in the

government, so it captures illegal advertising statements that are in actual use. However, the coverage is an issue. Section 4 discusses how to expand this list.

3.3 Feature Set 3: Log Relative Frequency Ratio

Relative frequency ratio between two datasets has been shown to be useful to discover collocations that are characteristic of a dataset when compared to the other dataset (Damerau, 1993). It is also used to model emotion transition between writers and readers (Tang and Chen, 2012). We extend this idea to select the critical features that capture the legality transition. The log relative frequency ratio *lr* of words in two datasets *A* and *B* are defined as follows. For each $w^i \in A \cup B$, we compute

$$lr_{AB}(w^i) = \log \frac{\frac{f_A(w^i)}{|A|}}{\frac{f_B(w^i)}{|B|}}$$

where $lr_{AB}(w^i)$ is a log ratio of relative frequencies of word w^i in *A* and *B*, $f_A(w^i)$ and $f_B(w^i)$ are frequencies of w^i in *A* and in *B*, and $|A|$ and $|B|$ are total words in *A* and in *B*, respectively. The log relative frequency ratios are used to estimate the distribution of the words in datasets *A* and *B*.

The interpretations of $lr_{AB}(w^i)$ are shown as follows.

(1) If w^i has higher relative frequency in *A* than in *B*, then $lr_{AB}(w^i)>0$. Those words of positive ratio form a set *A*-*B*.
(2) If w^i has higher relative frequency in *B* than in *A*, then $lr_{AB}(w^i)<0$. Those words of negative ratio form a set *B*-*A*.
(3) If w^i has similar relative frequency in both sets, then $lr_{AB}(w^i)\cong 0$.

In our experiments for food advertising, *A*=FOOD_LEGAL and *B*=FOOD_ILLEGAL. As for the experiments for cosmetic advertising, *A*=COS_LEGAL and *B*=COS_ILLEGAL. We employ the log relative frequency ratio as a weight of each unigram in a dataset. Each sentence in the datasets is represented by a vector $(w_1, w_2, \ldots, w_n)$, where w_i is the weight of i^{th} word from the unigram feature set. The 7th row of Table 1 lists the accuracy of the log relative frequency ratio feature set for the FOOD and COS advertising legality classification. The performance of both Naïve Bayes and SVM classifiers with Unigram + logRF feature settings are higher than those with the unigram and the auxiliary feature settings on both FOOD and COS datasets. The differences of accuracies between Unigram + logRF and all the other feature settings for both datasets are statistically significant ($p<0.01$).

3.4 Discussion

We further examine the individual accuracies of illegal advertising detection and legal advertising detection. Tables 3 and 4 show the experimental results of Naïve Bayes and SVM classifiers with different feature settings on food and cosmetic datasets, respectively. We can summarize some conclusions from these two tables. Firstly, Unigram+CILIN does not improve the accuracy of Unigram. The *Cilin* thesaurus contains many words that are not commonly used. Besides, its purpose is to help people find similar and related words conveniently. Thus, its organization of lexical terms may not be suitable for our classification tasks. Secondly, the accuracies of illegal advertising detection with both classifiers on both datasets are better than

Classification Models → Features ↓ Illegal vs. Legal →	Naïve Bayes		SVM	
	Illegal	Legal	Illegal	Legal
Unigram	92.592%	85.058%	89.463%	88.000%
Unigram + CILIN	93.367%	83.851%	90.330%	88.889%
Unigram + DOH	92.705%	84.994%	89.875%	89.106%
Unigram + CILIN + DOH	93.421%	83.902%	90.159%	89.126%
Unigram + logRF	**94.317%**	**86.371%**	**94.696%**	**91.677%**

TABLE 3 – Individual accuracies of illegal and legal advertising recognition on food dataset

Classification Models → Features ↓ Illegal vs. Legal →	Naïve Bayes		SVM	
	Illegal	Legal	Illegal	Legal
Unigram	86.479%	77.632%	82.470%	82.357%
Unigram + CILIN	86.812%	77.374%	83.287%	82.186%
Unigram + DOH	86.944%	77.658%	83.375%	83.964%
Unigram + CILIN + DOH	87.075%	77.431%	83.384%	83.260%
Unigram + logRF	**88.197%**	**83.060%**	**88.463%**	**83.413%**

TABLE 4 – Individual accuracies of illegal and legal advertising recognition on cosmetic dataset

those of legal advertising detection with the same classifiers on the same datasets. The accuracy difference between illegal and legal advertising recognition with SVM classifier is comparatively smaller than that with Naïve Bayes classifier. Note that the ratio of legal instances versus illegal instances in the food dataset is 41.84:58.16, and the ratio in the cosmetic dataset is 48.03:51.97. Thirdly, in the first four feature settings, Naïve Bayes classifiers perform illegal advertising detection better than SVM classifiers in both datasets. In contrast, SVM classifiers achieve better legal advertising detection than Naïve Bayes classifiers. Fourthly, when log relative frequency ratio is introduced, i.e., the Unigram+logRF feature setting, SVM classifier achieves the best performance in both illegal and legal advertising recognition on both datasets. The false alarm rates, a ratio of legal statements mis-recognized as illegal ones among all the legal statements, in food and cosmetic datasets are 0.083 and 0.166, respectively. The missing rates, a ratio of illegal statements mis-recognized as legal ones among all the illegal statements, in food and cosmetic datasets are 0.053 and 0.115, respectively. That illustrates the feasibility of log relative frequency ratio and SVM classifier.

4 Illegal Verb Phrase Mining

Effective identification of illegal advertising is a challenge for the authority and advertisers. Table 2 shows that almost all illegal advertising statements listed by DOH are verb phrases consisting of a transitive verb and an object noun. Thus, the usage of these verb phrases is a key criterion. To realize how illegal advertising uses verb phrases, we mine illegal advertising verb phrases from the illegal food and cosmetic datasets. The results can be used to extend the official list of illegal statements to improve illegal advertising recognition processes by the authority, and to help advertisers prepare legal advertisements.

The first step of mining illegal advertising verb phrases is to obtain the words that present more frequently in the illegal datasets. We adopt the same formula of log relative frequency ratio mentioned in Section 3.3. If $lr_{AB}(w^i)$ is a negative value, then w^i is more frequently used in illegal advertising. In our experiments, only the words with a log relative frequency lower than -0.1 and

with appropriate POS tags will be selected. The verb must be a transitive verb or nominalize verb, and the noun must be a common noun.

Then, we examine each sentence in the datasets to determine whether it contains a verb phrase consisting of a verb and a noun from our word list or not. Since we do not use a parser in the current stage, and an object noun does not necessarily immediately follow its verb, we identify a VP by the following criteria.

(1) The verb should occur before the noun.
(2) The distance between the verb and the noun should not exceed 3 words.

The noun should be the head of the noun phrase where it presents. That is, the noun should be the last word in the noun phrase. In Chinese, the head of a noun phrase is preceded by its adjectives and noun modifiers in most cases.

There are 979 and 2,302 verb phrases mined from the FOOD and the COS datasets, respectively. Some examples of these phrases are listed in Table 5. Log relative frequency ratio can be used with a POS tagger to mine illegal verb phrases consisting of a transitive verb and an object noun. We can observe that most verbs in the verb phrase lists are related to curative effects, and the objects are related to the human body, nutrients and diet. Similar structure and properties can been seen in the sample illegal expressions provided by the government. Thus we can conclude that log relative frequency ratio is an effective method to mine illegal expression lists.

Dataset	Illegal advertising verb phrases
FOOD	增強體質 (improve physical condition) 抑制細菌 (inactivate bacteria) 對抗年齡 (fight against aging) 分解膽固醇 (decompose cholesterol)
COS	淨化體質 (purify human body) 舒緩疼痛 (ease pain) 供給氧氣 (provide oxygen) 治療面皰 (cure acne vulgaris)

TABLE 5 – Examples of illegal advertising verb phrases mined from the FOOD and COS datasets.

Conclusion

This paper addresses the importance of legality recognition in Internet advertising. We use Naïve Bayes and SVM classifiers to perform the recognition tasks. The experimental results show that log relative frequency ratio can be used as weights for unigrams to improve performance of advertising legality recognition, and achieve the best accuracy in our experiments. We also use log relative frequency ratio to mine verb phrases consisting of a transitive verb and an object noun from illegal advertising statements. We find that this is an effective way to obtain a list of verb phrases that are related to problematic advertisements.

The recognition models proposed in this paper can be employed to build an automated illegal advertising recognition system in order to identify a huge number of advertisements automatically. The illegal verb phrase lists can also be used in a computer assisted system to help both the authority speed up the illegal advertising identification processes, and the advertisers to prepare suitable advertisements. As future work, we will extend the methodology to other types of advertising legality recognition task such as medicine domain.

Acknowledgments

This research was partially supported by Excellent Research Projects of National Taiwan University under contract 101R890858.

References

CFIA (2010). Advertising Requirements. Canadian Food Inspection Agency. Available at http://www.inspection.gc.ca/english/fssa/labeti/advpube.shtml.

Chang, C. and Lin, C. (2001). LIBSVM: a Library for Support Vector Machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cheng, H and Cantú-Paz, E. (2010). Personalized click prediction in sponsored search. In *Third ACM International Conference on Web Search and Data Mining* (*WSDM 2010*), pages 351-359, New York, USA.

Damerau, Fred J. (1993). Generating and Evaluating Domain-Oriented Multi-Word Terms from Text, *Information Processing and Management*, 29:433-477.

DOH (2009). Legal and Illegal Advertising Statements for Cosmetic Regulations. Department of Health of Taiwan, Available at http://www.doh.gov.tw/ufile/doc/0980305527.pdf.

Edelman, B., Ostrovsky, M., Schwarz, M. (2007). Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars Worth of Keywords, *American Economic Review*, American Economic Association, 97(1):242-259.

FTC (2000). Advertising and Marketing on the Internet: Rules of the Road, Bureau of Consumer Protection. Federal Trade Commission, September 2000, Available at http://business.ftc.gov/sites/default/files/pdf/bus28-advertising-and-marketing-internet-rules-road.pdf.

Gabrilovich, E., Josifovski, V. and Pang, B. (2008). Introduction to Computational Advertising. Tutorial Abstracts of ACL-08: HLT, page 1.

Gabrilovich, E., Josifovski, V. and Pang, B. (2009). Introduction to Computational Advertising. IJCAI 2009 Tutorial, http://research.yahoo.com/tutorials/ijcai09_compadv/

Ghosh, A., McAfee, P., Papineni, K., and Vassilvitskii, S. (2009). Bidding for Representative Allocations for Display Advertising. CoRR, abs/0910-0880, 2009.

Huang, H.C., Lin, M.S. and Chen H.H. (2008). Analysis of intention in dialogues using category trees and its application to advertisement recommendation. In *the Third International Joint Conference on Natural Language Processing* (*IJCNLP 2008*), pages 625-630, Hyderabad, India.

Mei, J., Zhu, Y., Gao, Y. and Yin, H. (1982). *Tóngyìcícílín*. Shanghai Dictionary Press.

Scaiano, M. and Inkpen, D. (2011). Finding negative key phrases for internet advertising campaigns using wikipedia. In *Recent Advances in Natural Language Processing* (*RANLP 2011*), pages 648–653, Hissar, Bulgaria.

Tang, Y.J. and Chen, H.H. (2012). Mining sentiment words from microblogs for predicting writer-reader emotion transition. In *the 8th International Conference on Language Resources and Evaluation* (*LREC 2012*), pages 1226-1229, Istanbul, Turkey.