FAdR: A System for Recognizing False Online Advertisements

Yi-jie Tang and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
tangyj@nlg.csie.ntu.edu.tw;hhchen@ntu.edu.tw

Abstract

More and more product information, including advertisements and user reviews, are presented to Internet users nowadays. Some of the information is false, misleading or overstated, which can cause seriousness and needs to be identified. Authorities, advertisers, website owners and consumers all have the needs to detect such statements. In this paper, we propose a **F**alse **Ad**vertisements **R**ecognition system called **FAdR** by using one-class and binary classification models. Illegal advertising lists made public by a government and product descriptions from a shopping website are obtained for training and testing. The results show that the binary SVM models can achieve the highest performance when unigrams with the weighting of log relative frequency ratios are used as features. Comparatively, the benefit of the one-class classification models is the adjustable rejection rate parameter, which can be changed to suit different applications. Verb phrases more likely to introduce overstated information are obtained by mining the datasets. These phrases help find problematic wordings in the advertising texts.

1 Introduction

As online commerce and advertising keep growing, more and more consumers depend on information on the Internet to make purchasing decisions. This kind of information includes online advertisements posted by businesses, and discussions or reviews generated by users. However, false statements can also be presented to consumers. For example, some companies hire people to post fake product reviews in an attempt to promote their own products or reduce competitors' reputations (Ott et al., 2011). It is referred to as deceptive opinion spamming and explored in recent researches (Ott et al., 2011; Mukherjee et al., 2012; Mukherjee et al., 2013; Fei et al., 2013).

False statements and exaggerated content can also be seen in online advertisements. These statements can also be regarded as opinion spams, while the authors, that is, the advertisers, can be more easily identified. Yeh (2014) reported the top two types of illegal advertisements on the web, TV and broadcast are food (62.61%) and cosmetic (24.26%). Of the dissemination media, the web is the major source of false advertisements. Most inappropriate food-related advertisements contain overstated health claims. The medical effects and cure claims may also appear in cosmetic advertising. As a result, advertising regulations are enforced in many countries to protect consumers from fraudulent and misleading information. False, overstated or misleading information and mentions of curative effects can be prohibited by the authorities (FTC, 2000; DOH, 2009; CFIA, 2010).

To regulate online advertising, the authorities need to review a large number of advertisements and determine their legality, which is cost- and time-consuming. Advertisers also need to know the legality of their advertisements to avoid violating advertising laws. This becomes especially important when every Internet user can be an advertiser if s/he posts messages related to any product announcement, promotion, or sales. Website owners that accept advertisements have to present appropriate advertisement contents to users and avoid legal issues. Even Internet users should also identify false advertisements in order not to be misled. Thus, the recognition of false, misleading or overstated information is an emerging task.

This paper presents a **F**alse **Ad**vertisements **R**ecognition system called **FAdR**, and take two

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 103–108,
Baltimore, Maryland USA, June 23-24, 2014. ©2014 Association for Computational Linguistics

major sources of illegal advertisements on the web, i.e., food and cosmetic advertising, as examples. Section 2 surveys the related work. Section 3 introduces the datasets used in the experiments. Section 4 presents classification models and shows their performance. Section 5 mines the overstated phrases. Section 6 demonstrates the uses of **FAdR** system with screenshot. Both sentence and document levels are considered.

2 Related Work

Gokhman et al. (2012) collected data from the Internet and explored methods to construct a gold standard corpus for "deception" studies. Ott et al. (2011) studied methods to detect "disruptive opinion spams." Unlike conventional advertising spams, these fake opinions look authentic and are used to mislead users. Mukherjee et al. (2013) used reviewer's behavioral footprints to detect spammer. As they pointed out, one of the largest problems to solve this issue is that there is no appropriate datasets for fake and non-fake reviews.

Previous online advertising research mostly focuses on bidding, matching or recommendation of advertisements on websites. Ghosh et al. (2009) studied bidding strategies for advertisement allocations. Huang et al. (2008) proposed an advertisement recommendation method by classifying instant messages into the Yahoo categories. Scaiano and Inkpen (2011) used Wikipedia for negative keyphrase generation to hide advertisements that users are not interested in. This paper, in contrast, focuses on identifying false statements in online advertisements with classification models.

3 Datasets

We use the illegal advertising lists and statements made public by the Taipei City Government[1] as the illegal advertising datasets. The contents of the government data are split into sentences by colon, period, question mark and exclamation mark. Two types of datasets are built for illegal food and cosmetic advertising, named FOOD_ILLEGAL and COS_ILLEGAL, respectively. Some illegal sentences in the illegal food advertising dataset are shown below:

(1) 減少代謝廢物的堆積。
Reduces waste produced by metabolism process.

(2) 減少失眠及疼痛。
Stops insomnia and pain.

(3) 治療高血壓。
Cures hypertension.

In the government website, the authority does not regularly announce legal advertising data. We adopt one-class classifiers with only illegal data for this scenario, as shown in Section 4.1. To experiment on binary classifiers, we collect product descriptions from a shopping website[2] and verify their legality manually to construct the legal advertising datasets. The legal food and cosmetic adverting datasets are named FOOD_LEGAL and COS_LEGAL, respectively. The numbers of the sentences in FOOD_LEGAL, FOOD_ILLEGAL, COS_LEGAL, and COS_ILLEGAL are 5,059, 7,033, 10,520, and 11,381, respectively.

4 Classification Models

One-class Naïve Bayes and Bagging classifiers, and binary classifiers based on Naïve Bayes and SVM models are implemented.

4.1 One-Class Classifiers

We adopt the OneClassClassifier module (Hempstalk et al., 2008) in the WEKA machine learning tool to train one-class classifiers with illegal statements only. The OneClassClassifier module provides a rejection rate parameter for adjusting the threshold between target and non-target instances. The target class, which corresponds to the illegal class in this study, is the single class used to train the classifier. Higher rejection rate means that more legal statements will be preferred, but illegal statements may be still incorrectly classified into legal ones. Naïve Bayes and Bagging classifiers are chosen because they achieve best performance among the algorithms we have explored in this experiment.

Each instance in the dataset, i.e., a sentence, 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 vocabulary is selected from the illegal advertising datasets. To properly filter out common words, we count top 1,000 frequent words in the Sinica Balanced Corpus of Modern Chinese[3] and remove them from the vocabulary. The remaining top 1,000 words are used for vector representation.

Total 532 illegal statements provided by the Department of Health form the training set. An

[1] http://www.health.gov.tw/Default.aspx?tabid=295

[2] http://www.7net.com.tw

[3] http://app.sinica.edu.tw/kiwi/mkiwi/

illegal and a legal advertising dataset make up the test set. The former consists of 317 illegal sentences from Taipei City Government's lists, and the latter contains 203 legal statement examples from the Department of Health.

Table 1 shows the accuracies of Naïve Bayes and Bagging classifiers in the food dataset. The rejection rates from 0.7 to 0.8 are preferable for most applications, because they result in higher accuracy for legal statement classification while not significantly reducing the performance of illegal statement detection. Using the 0.7 rejection rate produces high performance for the illegal class while 0.8 rejection rate does better for the legal class. The actual choice of rejection rate depends on the demands of users. For an advertiser, it is important to avoid all possible problematic statements. Thus, a lower rejection rate will be more suitable. If the system is used by the authorities, a rejection rate higher than 0.7 may be preferable because they don't misjudge too many legal advertisements.

Rejection rate		0.4	0.5	0.6	0.7	0.8	0.9
Naïve Bayes	Illegal	85.33%	82.39%	79.01%	74.49%	68.17%	59.14%
	Legal	31.07%	39.81%	53.40%	63.11%	72.82%	86.41%
Bagging	Illegal	92.78%	88.49%	84.65%	74.94%	69.07%	0.23%
	Legal	3.88%	17.48%	27.18%	65.72%	82.52%	99.77%

Table 1: Accuracies of Classifiers in Different Rejection Rates.

4.2 Binary Classifiers

We use FOOD_LEGAL and FOOD_ILLEGAL datasets, and COS_LEGAL and COS_ILLEGAL datasets to build binary classifiers for food and cosmetic advertising classification, respectively. Naïve Bayes classifiers and SVM classifiers implemented with libSVM (Chang & Lin, 2011) are adopted. Ten-fold cross validation is used for the training and testing tasks. Total 1,000 highly frequent words are selected in the same way as in Section 4.1 to form a word-based unigram feature set.

Two weighting schemes are considered. In the binary weighting, each sentence 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. In the weighting of log relative frequency ratio, we follow the idea of collocation mining (Damerau, 1993). 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. It has been successfully applied to mine sentiment words from microblog and to model reader/writer emotion transition (Tang and Chen, 2011, 2012).

The log relative frequency ratio (*logRF*) is defined formally as follows. Given two datasets A and B, the log relative frequency ratio for each $w^i \in A \cup B$ is computed with the following formula.

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

$logRF_{AB}(w^i)$ 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, respectively, and $|A|$ and $|B|$ are total words in A and in B, respectively. *logRF* values are used to estimate the distribution of the words in datasets A and B. If w^i has higher relative frequency in A than in B, then $logRF_{AB}(w^i)>0$, and vice versa. In our experiments, *logRF* is used to present each unigram's distribution in the legal and illegal datasets, replacing the binary value for a unigram feature.

Tables 2 and 3 show the results of the classification models with different combinations of feature sets. When *logRF* is combined with Unigram, the accuracy is significantly improved in both the food and cosmetic datasets. We can also see that the performance of all FOOD models are higher than equivalent COS models. Possible reasons may be that the effects of cosmetics are related to body appearance, and inappropriate cure claims are also related to body improvement and appearance changes. There can be some overlaps between the words used in legal and illegal cosmetic advertising.

Classification Models →	Naïve Bayes		SVM	
Illegal vs. Legal → Features ↓	Illegal	Legal	Illegal	Legal
Unigram	92.59%	85.06%	89.46%	88.00%
Unigram + logRF	**94.32%**	**86.37%**	**94.70%**	**91.68%**

Table 2: Classification Accuracies for FOOD Datasets.

Classification Models →	Naïve Bayes		SVM	
Illegal vs. Legal → Features ↓	Illegal	Legal	Illegal	Legal
Unigram	86.48%	77.63%	82.47%	82.36%
Unigram + logRF	**88.20%**	**83.06%**	**88.46%**	**83.41%**

Table 3: Classification Accuracies for COS Datasets.

5 Overstated Phrase Mining

Since the authority focuses on health claims in advertising, almost all illegal statements announced by the government include an action related to health improvement and a name that refers to diseases or body conditions. Thus, we can observe that most of the illegal statements recognized and forbidden by the authority contain a health-related verb phrase consisting of a transitive verb and an object. These illegal advertising verb phrases can be mined from the datasets for the government's and advertisers' reference. We can also use these verb phrases to help the users of our system understand possible reasons why the sentences in advertisements are labeled as illegal.

We propose a mining method based on log relative frequency ratio, which is described in Section 4.2. We compute $logRF_{AB}(w^i)$ to obtain the words that are most likely to be used in illegal advertising. We identify transitive verbs and nouns in the word list based on POS tagging results generated by the CKIP parser[4], and then use them to examine if a verb phrase is presented in a sentence. Total 979 verb phrases are mined from the FOOD datasets, and 2,302 from the COS dataset. Table 4 shows some examples.

Dataset	Illegal advertising verb phrases	
	Transitive verb	Object noun
FOOD	增強 (improve)	體質 (physical condition)
	抑制 (inactivate)	細菌 (bacteria)
	分解 (decompose)	膽固醇 (cholesterol)
COS	淨化 (purify)	體質 (body)
	舒緩 (ease)	疼痛 (pain)
	治療 (cure)	面皰 (acne vulgaris)

Table 4: Example illegal verb phrases mined from the FOOD and COS datasets.

6 System Architecture

The **FAdR** system is composed of preprocessing (Pre-Processor), recognition (Recognizer), and explanation (Explainer) modules. Figure 1 shows the overall system architecture.

6.1 Pre-processing Module

Our classification models are sentence-based, so the main purpose of the Pre-processor in the system is detecting sentence boundaries. Four types of punctuations, including period, colon, exclamation, and question mark, are used to segment a document into sentences. Line breaks are also regarded as a sentence boundary marker because many advertisements in Chinese put sentences in separate lines and do not include any punctuation. Sentences with less than three characters or more than 80 characters are ignored.

Word segmentation is performed by using the CKIP segmenter, which is an online service and can be accessed through the TCP socket. Segmented data will be represented by the corresponding feature sets based on classification model and converted to a format that the Recognizer can read as input.

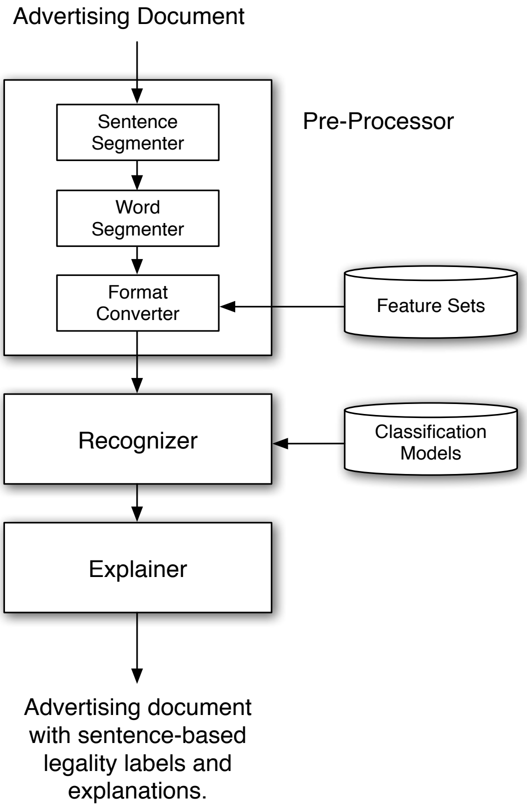


Figure 1. System architecture of **FAdR**

6.2 Recognition Module

All processed sentences are sent from the Pre-Processor to the Recognizer for legality identification.

Since our training tasks are done in WEKA, we can use the model files generated by WEKA for implementing the Recognizer. The Recognizer loads the pre-trained SVM models for food and cosmetic advertising classification, and then uses them for labeling the incoming sentences.

For the One-Class models, the model files are pre-generated by training with different rejection rates from 0.4 to 0.9. When the user adjusts the threshold, the Recognizer chooses the corresponding model to perform illegal sentences identification.

[4] http://ckipsvr.iis.sinica.edu.tw

6.3 Explanation Module

To give users more information on the possible reasons why the advertising contents are considered illegal, the Explainer uses the illegal verb phrase list, which is discussed in Section 5, to extract the problematic words from the input sentences. If the verb and the object noun in a verb phrase from the list both occur in an illegal sentence, then the verb phrase will be shown besides the recognition results in the user interface.

6.4 User Interface

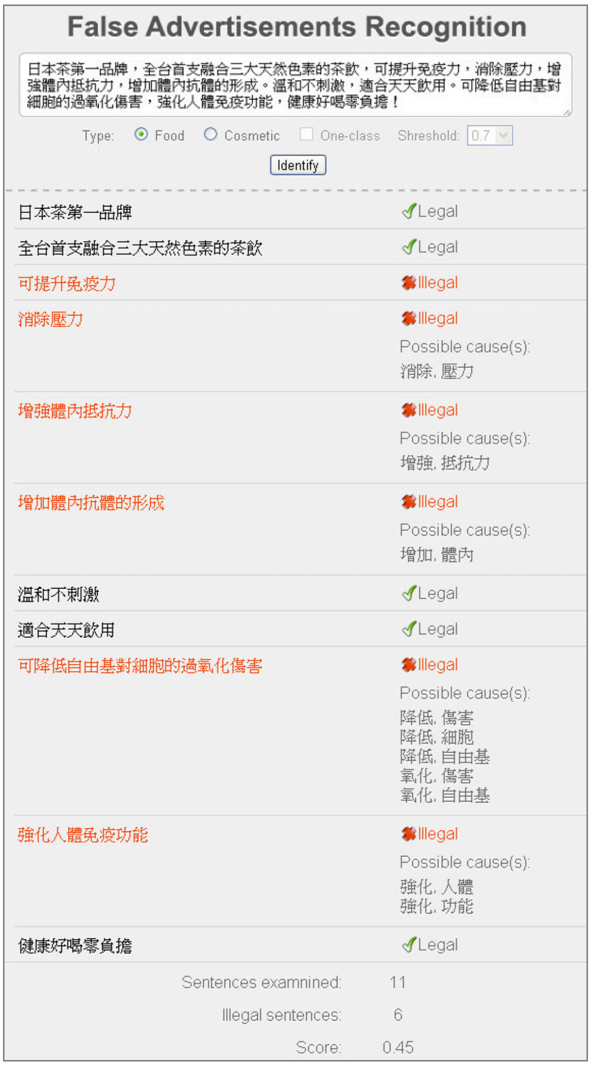
Users can copy and paste the advertising contents to be recognized to the text field, or upload a document to the system. It usually takes less than 10 seconds on our server to process a document with 200 characters, so the system is suitable to quickly process a large amount of data.

If the users choose to use the one-class models, they can adjust the threshold value to fit different needs and receive useful results. Lowering the value can find as many problematic sentences as possible, but more legal sentences can also be misjudged. Increasing the value can avoid wrongly labeling legal sentences as illegal, but more illegal sentences can be missed.

Figure 2 shows a system screenshot. The recognition results of a food advertisement with 11 sentences are demonstrated. Sentences labelled as illegal are highlighted in red. Verb phrases possibly causing illegality are listed in grey colour for illegal sentences. The number of all sentences, the number of illegal sentences, and the final score are shown at the bottom. The correct score of an advertisement is defined as the number of correct sentences divided by total sentences in this advertisement. The sample advertisement used in Figure 2 and its English translation are shown as follows.

<A food advertisement>
日本茶第一品牌，全台首支融合三大天然色素的茶飲，可提升免疫力，消除壓力，增強體內抵抗力，增加體內抗體的形成。溫和不刺激，適合天天飲用。可降低自由基對細胞的過氧化傷害，強化人體免疫功能，健康好喝零負擔！

(The leading brand for Japanese tea. The first tea product combining three kinds of natural colourings in Taiwan. Can improve immunity. Can relieve stress. Can strengthen resistance to disease. Can increase antibodies in your body. It is mild and not irritative. Good for daily use. Can prevent body cells from being harmed by free radicals. Can strengthen immunity. It is healthy and tasty, and brings no body burden.)



False Advertisements Recognition

日本茶第一品牌，全台首支融合三大天然色素的茶飲，可提升免疫力，消除壓力，增強體內抵抗力，增加體內抗體的形成。溫和不刺激，適合天天飲用。可降低自由基對細胞的過氧化傷害，強化人體免疫功能，健康好喝零負擔！

Type: ● Food ○ Cosmetic □ One-class Shreshold: 0.7

日本茶第一品牌	✓Legal
全台首支融合三大天然色素的茶飲	✓Legal
可提升免疫力	✗Illegal
消除壓力	✗Illegal Possible cause(s): 消除, 壓力
增強體內抵抗力	✗Illegal Possible cause(s): 增強, 抵抗力
增加體內抗體的形成	✗Illegal Possible cause(s): 增加, 體內
溫和不刺激	✓Legal
適合天天飲用	✓Legal
可降低自由基對細胞的過氧化傷害	✗Illegal Possible cause(s): 降低, 傷害 降低, 細胞 降低, 自由基 氧化, 傷害 氧化, 自由基
強化人體免疫功能	✗Illegal Possible cause(s): 強化, 人體 強化, 功能
健康好喝零負擔	✓Legal

Sentences examnined: 11
Illegal sentences: 6
Score: 0.45

Figure 2: Screenshot for Illegal Sentence Recognition

7 Conclusion

Detecting false information on the Internet has become an important issue for users and organizations. In this paper, we present two types of classification methods to identify overstated sentences in online advertisements and build a false online advertisements recognition system **FAdR**. The recognition on both document and sentence levels is addressed in the demonstration.

In the binary models, using combinations of unigrams and the log relative frequency ratio as features can achieve highest performance. On the other hand, the one-class models can be used to build a system that is adjustable by users for different application domains.

The authorities or website owners can use a rejection rate of 0.7 or 0.8 to highlight most serious illegal advertisements. An advertisement with a score lower than 0.5 means it may critically violate the regulations, and need to be regarded as illegal advertising. Since not all advertisement posters are professional advertisers, they may need detailed information on the legality of every sentence. The illegal verb phrases found in a sentence provide clues to the advertiser. The system is also useful for consumers, as they can check if the advertisement contents can be trusted before making a purchase decision.

As future work, we will extend the methodology presented in this study to handle other types of advertisements and the materials in other languages. We will also investigate what linguistic patterns can be used to mine the overstated phrases in different languages.

Acknowledgments

This research was partially supported by National Taiwan University and Ministry of Science and Technology, Taiwan under 103R890858 and 102-2221-E-002-103-MY3.

References

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a Library for Support Vector Machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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

Fred J. Damerau. 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.

Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Exploiting Burstiness in Reviews for Review Spammer Detection. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (*ICWSM-2013*), 175-184.

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.

Stephanie Gokhman, Jeff Hancock, Poornima Prabhu, Myle Ott, and Claire Cardie. 2012. In Search of a Gold Standard in Studies of Deception. In *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*, 23–30.

Arpita Ghosh, Preston McAfee, Kishore Papineni, and Sergei Vassilvitskii. 2009. Bidding for Representative Allocations for Display Advertising. CoRR, abs/0910-0880, 2009.

Hung-Chi Huang, Ming-Shun Lin and Hsin-Hsi Chen. 2008. Analysis of Intention in Dialogues Using Category Trees and Its Application to Advertisement Recommendation. In *Proceedings of the Third International Joint Conference on Natural Language Processing* (*IJCNLP 2008*), 625-630.

Kathryn Hempstalk, Eibe Frank, and Ian H. Witten. 2008. One-Class Classification by Combining Density and Class Probability Estimation. In *Proceedings of the 12th European Conference on Principles and Practice of Knowledge Discovery in Databases and 19th European Conference on Machine Learning*, 505-519.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting Fake Reviewer Groups in Consumer Reviews. In *Proceedings of the International World Wide Web Conference* (*WWW 2012*), 191-200.

Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Spotting Opinion Spammers using Behavioral Footprints. In *Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD 2013*), 632-640.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 309–319.

M. Scaiano and D. Inkpen. 2011. Finding Negative Key Phrases for Internet Advertising Campaigns Using Wikipedia. In *Recent Advances in Natural Language Processing* (*RANLP 2011*), 648–653.

Yi-jie Tang and Hsin-Hsi Chen. 2011. Emotion Modeling from Writer/Reader Perspectives Using a Microblog Dataset. In *Proceedings of IJCNLP Workshop on Sentiment Analysis where AI Meets Psychology,* 11-19.

Yi-jie Tang and Hsin-Hsi Chen. 2012. Mining Sentiment Words from Microblogs for Predicting Writer-Reader Emotion Transition. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (*LREC 2012*), 1226-1229.

Ming-kung Yeh. 2014. *Weekly Food and Drug Safety*. No. 440, February, Food and Drug Administration, Taiwan. Available at http://www.fda.gov.tw/TC/PublishOther.aspx.