# CIEA: A Corpus for Chinese Implicit Emotion Analysis

Dawei Li, Jin Wang and Xuejie Zhang
*School of Information Science and Engineerin*
*Yunnan University*
*Kunming, P.R. China*
*Contact: xjzhang@ynu.edu.cn*

***Abstract*—The traditional cultural euphemism of the Han nationality has profound ideological roots. China has always advocated Confucianism, which has led to the implicit expression of Chinese people's emotions. There are almost no obvious emotional words in spoken language, which poses a challenge to Chinese sentiment analysis. It is very interesting to exploit a corpus that does not contain emotional words, but instead uses detailed description in text to determine the category of the emotional expressed. In this study, we propose a corpus for Chinese implicit sentiment analysis. To do this, we have crawled millions of microblogs. After data cleaning and processing, we obtained the corpus. Based on this corpus, we introduced conventional models and neural networks for implicit sentiment analysis, and achieve promising results. A comparative experiment with a well-known corpus showed the importance of implicit emotions to emotional classification. This not only shows the usefulness of the proposed corpus for implicit sentiment analysis research, but also provides a baseline for further research on this topic.**

***Keywords*-Chinese corpus; implicit emotion; masking keywords; baseline;**

## I. INTRODUCTION

Nowadays, individuals or organizations can share and send information on social networks anytime and anywhere. As an example, on the popular Chinese microblogging website Sina Weibo, users can share information, which mainly includes the things in their daily lives and comments on hot news on the site. These messages can be positive, neutral, negative, or one of six basic emotions (e.g., anger, happiness, fear, sadness, disgust and surprise) [1] in these microblog contents. However, Chinese people often express their emotions implicitly and do not directly express happiness or sadness. This leads to no obvious emotional words in the text, and it is necessary to judge the expressed emotions through context. When there are no obvious emotional words in the text, the analysis of emotions will be challenging.

Whether they are used in a conventional model or a neural network, the models built on previous corpora rely on emotional vocabulary (or corresponding representations) and do not focus on the causes of emotions or events. In previous corpus, emotion words appears in the text, which causes the model to pay more attention to emotional words. The model will not predict the correct emotions when there are no obvious emotional words in the text. When emotions are inferred, the type of emotion is often related to the context described in the text [2]. For this purpose, it is necessary to set up an emotional classification system without emotional words, and the emotion depends on the context depicted.

Table I
THE EXAMPLE OF CORPUS.

| ID | Text | Label |
|---|---|---|
| 1 | 提交完论文的我[#关键词#]得不是一点点。<br>(I am so [#keyword#] that I submitted the paper.) | 开心<br>(happy) |
| 2 | 我好[#关键词#]！爸爸突然买了条边牧，说好的金毛和拉布拉多没有了。<br>(I am so [#keyword#] that dad bought a border collie. We negotiated to buy a golden lab.) | 生气<br>(angry) |
| 3 | 有点[#关键词#]，今晚我要开着灯睡觉<br>(I am so [#keyword#] that i have to sleep with a lights tonight.) | 害怕<br>(fear) |
| 4 | 刚喝完酒回来寝室就遭贼了，[#关键词#]，你们有什么办法对付这种事吗？<br>(I am so [#keyword#] that I was stolen after drinking the wine and returning to the bedroom. Do you have any way to deal with this kind of thing?) | 伤心<br>(sad) |
| 5 | 招商银行这是疯了么，一天发那么多条短信给我，已经有点[#关键词#]的感觉了！<br>(This is crazy for China Merchants Bank. It is [#keyword#] to send me so many text messages a day.) | 厌恶<br>(disgusting) |
| 6 | 一张老照片虽然不值钱，但它却保留了你人生中最难忘的一幕。当你若干年后，再看到后说不定会[#关键词#]地说：“咦，这就是我吗？”<br>(An old photo is worthless, but it retains the most memorable scene of your life. When you see it for a few years, you may be [#keyword#] to say: ”Hey, is this me?”.) | 惊讶<br>(surprised) |

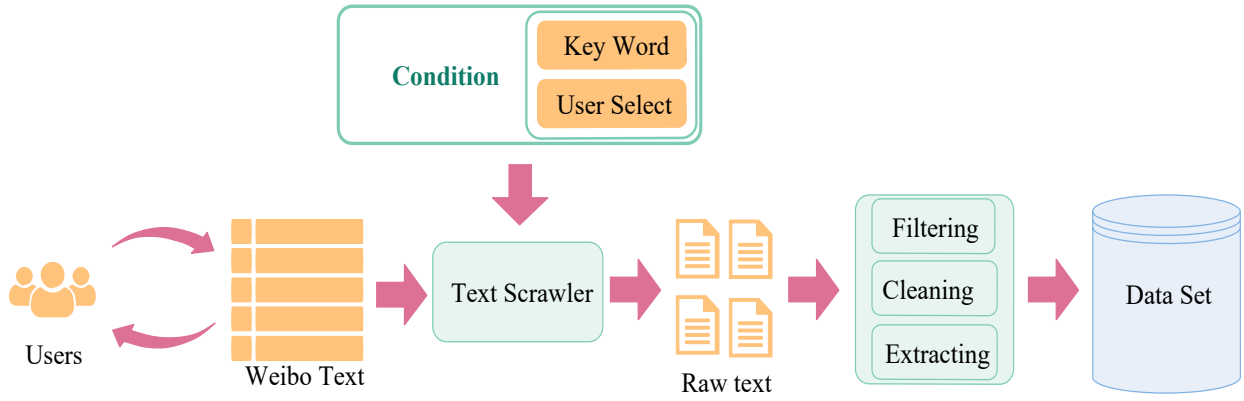


Figure 1. Diagram of the process for creating the corpus.

With the corpus, we aim to solve the problems mentioned above. We chose to build a microblog corpus, because Chinese microblogs are convenient to use, and easy to access. The data on microblogs show diversity and richness. The corpus contains a large amount of self-labeling annotation of the microblog data. We extract the microblogs (including emojis) with a certain number of words and clean the data. Then, we mask the "emotional word" and use context information to predict the masked word. The words masked are as follows:生气,厌恶,害怕,开心,伤心,惊讶 (angry, disgusting, fear, happy, sad, and surprised).

As shown in **Table I**, based on context analysis of Example 1, when an important thing is done, people will be happy. In Example 2, one thing was determined after by consultation, but not done according to the plan. This will make people angry or sad. In the example, the matter of selecting a dog is already discussed, but different dog was bought than that discussed. Therefore, the emotional word of the case should be angry. In example 3, the hidden emotional word should be fear which can be inferred from "i have to sleep with a lights tonight". There are no obvious emotional words in the example, but the masked emotions can be inferred by analyzing context information.

We used several well-known affective analysis corpora in the experiment. Delete the emotional words in the well-known corpus with obvious emotional trends, and build a comparative corpus. Using conventional models and deep learning models to model the corpus. The experimental results show that the performance of model is significantly reduced after deleting the emotional words with obvious emotional trends. This shows that emotional words have an impact on whether the model can predict the classification correctly. At the same time, the corpus we proposed did not appear to drop significantly.

We introduce a Chinese implicit emotional corpus. To our knowledge, it is the first such corpus (masking emotion words) to date. We split the corpus into a standard training set, experimental sets, and test sets to a facilitate baseline for related methods. We use a neural network to evaluate the corpus and achieve promising results, which can be used as a baseline for the task. Finally, we compare the different corpora to verify the impact of emotional words on the prediction results.

The rest of this paper is organized as follow. Section 2 introduces existing corpora in different fields. Section 3 describes the process of building the Chinese implicit emotional corpus and its properties. Section 4 analyzes the quality of the corpus through experiments and compares it with famous corpora. We conclude this work in Section 5.

## II. Related Work

There are many works on exploiting corpora to promote the development of NLP. Previous sentiment analysis involved multiple text types, including product [3], movie reviews [4], fairy tales [5], news stories, social commentary and commentary articles [6], blogs [7] [8] and microblogs [9]. Some well-known emotional resources are as follows.

The subjunctive mood in fairy tales is one of the first corpora to annotate the expression of emotions. Annotated a corpus of approximately 185 children stories including Grimms', H.C. Andersen's and B.Potter's stories. The goal is to classify the emotional affinity of sentences in the narrative domain of children's fairy tales, for subsequent usage in appropriate expressive rendering of text-to-speech synthesis [5]. Then emotions were annotated according to the headlines of news. The aim is to predict emotions from news headlines. [10]. Movie reviews (MR) [11], Large Movie Review Dataset (IMDB) [12] and the Stanford sentiment tree-bank (SST) [13] were used. Use movie reviews to do a polar analysis of emotions, and achieve success. Twitter is one of the most popular sites, and in a variety of topics, it contains various emotions. Therefore. the researchers' enthusiasm for evaluating Twitter is very high. This has been researched by creating a corpus from Twitter posts using emotion-word hashtags [14], tracking the public sentiment of 2012 US presidential election on the polarization of the candidates [15], and modeling the Twitter text to determine whether the author is in favor of, against, or neutral to something [16]. These work on twitter corpus is to collect twitter prediction first. Then manually check the self-annotated text to ensure the correctness of the self-annotated text. These studies play

Table II
THE BASIC PROPERTIES OF THE CORPUS. TOKEN REPRESENTS THE AVERAGE NUMBER OF WORDS FOR EACH MICROBLOG.

| Label \ Token | Train | Dev | Test |
|---|---|---|---|
| 伤心(sadness) | 65.9 | 64.39 | 63.79 |
| 开心(happiness) | 59.03 | 60.68 | 59.64 |
| 厌恶(disgust) | 64.54 | 63.63 | 65.41 |
| 惊讶(surprise) | 75.88 | 77.5 | 75.3 |
| 生气(anger) | 65.97 | 64.72 | 64.74 |
| 害怕(fear) | 66.59 | 64.94 | 67.54 |

an important role in promoting the emotional analysis of online texts.

Wassa-2018 Task uses a tweet with a distinct emotional expression and then hides the emotional words [2]. Then use the tweet text which was masked emotional word to predict the masked emotional word belongs to. Our work is related to this work, using a large number of self-labeled Chinese Microblogs which emotional word was masked to predict the implicit emotion.

## III. CHINESE IMPLICIT EMOTION RESOURCE CONSTRUCTION

We chose to use web crawler technology to obtain the corpus of Sina-Microblogs for its diversity and richness. We use six keywords:" 生气", "厌恶", " 害怕", "开心", "伤心" and " 惊讶" as the condition to obtain a large amount self-labeling data. Anyone can create an account (individual, media or organization) on Sina-Microblogs and the content and form of each microblog are different. In order to guarantee the quality of the content, we only crawl personal original microblog data. In this way, the data is closer to daily life and more realistic. The process of the data collection is shown in **Figure 1** and summarized as follows:

(1)Use the key words: "生气", "厌恶", " 害怕", "开心", "伤心" and " 惊讶" as a condition to crawl personal original microblog text to ensure that the text contains clear emotional words.

(2)Extract the data which the length of sentence is between 15 and 200 as the original data. Delete negative data such as abusive content, advertising and pornography.

(3)Discard sentences which contains synonyms and antonyms related to keywords. Ensure only one type of word with a clear emotional tendency in each sentences. Then delete the repeated sentences.

(4)Ten volunteers were invited to manually repeat the steps 3.

(5)Mask emotional words (label) in the data. Replace the emotional word with "[# 关键词#]".

After data cleaning and processing, the corpus we proposed does not contain emotional words with obvious emotional tendencies. The corpus consists of 285000 pieces of microblogs data, including 156000 training posts, 39000 developments posts, and 90000 test posts. The three subcollection are equally divided into six categories. In this way, the distribution of data is balanced. The basic corpus information is shown in the **Table II**. The table is analyzed to compare the average number of microblogs words in the same category for different subsets. It can be seen that the difference in the average number of microblogs for each class is very small. There are some differences in the average number of microblogs in different categories. This does not affect the training and prediction of the model.

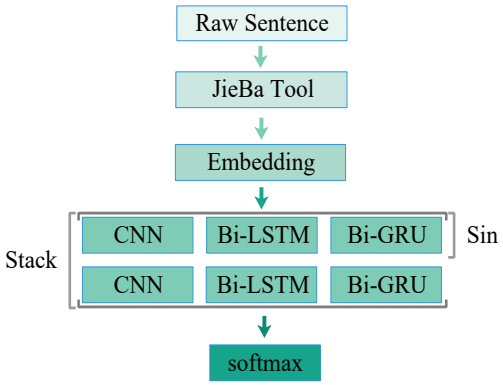


Figure 2. Stacked and non-stacked neural network models. Sin means that the neural network has only one layer, and stack means that there are two layers of neural networks.

Table III
THE RESULT OF BASELINE SYSTEM. SIN MEANS SINGLE LAYER. THE NUMBER REPRESENTS MACAVG-$F_1(\%)$.

| Model | MacAvg-$F_1(\%)$ | | | |
|---|---|---|---|---|
| | Test | | Dev | |
| | Sin | Stack | Sin | Stack |
| Logistic | 59.12 | - | 58.5 | - |
| Random Forest | 52.51 | - | 51.4 | - |
| CNN | 55.73 | 59.65 | 55.39 | 59.14 |
| Bi-LSTM | 72.63 | 74.47 | 72.42 | 74.29 |
| Bi-GRU | 73.06 | 75.06 | 72.47 | 74.99 |

## IV. EXPERIMENT

In this section, we provided several baseline systems for evaluating our corpus. We set several baseline systems to test the basic performance of our corpus and provide meaningful comparisons for different systems and corpora.

### A. Evaluation Script

In order to facilitate the evaluation of classification, we provide an evaluation script that measures the precision, recall, and F1 score for each emotional classification. We also added the $F_1(\%)$ score of micro-average and macro-average. Because the amount of data in each category in our corpus is almost average, we use the macro-average $F_1(\%)$ scores as the evaluation criteria.

### B. Baseline System

In this study, we provide five baseline systems, including two conventional ones and three neural network

Table IV
THE RESULT OF CONTRAST EXPERIMENT. **MAX** REPRESENTS THE MAXIMUM NUMBER OF SENTENCES CONTAINING A SENTIMENT WORD; **AVERAGE** REPRESENTS THE AVERAGE NUMBER OF SENTIMENT WORDS IN EACH SENTENCE; **RATE** REPRESENTS THE PROPORTION OF SENTENCES IN THE CORPUS CONTAINING EMOTIONAL WORDS; **W/ EMOTION** REPRESENTS THE TEST SET ACCURACY RATE CONTAINING EMOTIONAL WORDS, AND **W/O EMOTION** REPRESENTS THE TEST SET ACCURACY RATE WITHOUT EMOTIONAL WORDS.

| Corpus | Category | Max | Average | *Rate(%))* | *Acc(%)* | |
|---|---|---|---|---|---|---|
| | | | | | w/ Emotion | w/o Emotion |
| SST-2 | 2 | 20 | 1.71 | 36.68 | 83.47 | 73.09 |
| IMDB | 2 | 209 | 34.3 | 99.98 | 89.3 | 80.69 |
| MR | 2 | 22 | 2.06 | 41.25 | 77.92 | 67.84 |
| CIEA-b | 2 | 51 | 1.65 | 68.8 | 87.48 | 84.51 |

models. In the data processing phase, we only processed the text, and no special processing was used for emotions. We used the Jieba [17] word segmentation tool to segment Chinese sentence and retain the stop words. The goal is to enable a baseline system to accurately reflect the quality of the corpus. For the neural networks, Tencent AI Lab Embedding Corpus [18] is used as input to the word embedding layer, and provide a baseline for stacked and non-stacked models. Architecture shown in Figure 2.

- Conventional model: We used logistic regression [19] and the random forest [20] model and initialize a random forest classifier with 100 trees. Both models used the bag-of-words as features.
- Neural networks: We used the deep learnings models of the convolutional neural network (CNN) [21], and bidirectional long short-term Memory (LSTM) and gated recurrent unit (GRU) base on LSTM [22] and GRU [23]. Stacking and non-stacking were used in the neural network model. Stacked consists of two layers of deep learning model components, while non-stacked consists of one layer of deep learning components. We initialized the batch size to 1024 and the dropout to 0.25. For CNN, we initialized the kernel size to 3 and filter size to 60, and used the Soft-Max function to output predictive probability values. For LSTM or GRU, we initialized hidden units to 120 and the recurrent dropout to 0.25. Finally, the Soft-Max function was used to output Predictive probability value.

We used the conventional models and the neural network to evaluate the corpus. The results are showed in **Table III**.The performance of the neural network model is better than conventional model, while the stacked sequence model has a significant improvement over the non-stacked model. The results of the stacked Bidirectional GRU network is the best. For neural networks, the results of each model of the development set and the test set differ between 0.07% and 0.56%. The predicted results of development set are close to the test set prediction results. There is no significant difference in text characteristics between the Two data sets. It's means that the corpus we propose has good generalization capabilities, and follow the principle of independently identically distribution.

### C. Corpus Comparison

In order to verify the impact of emotional words on the classification results, we designed the experiment. In the experiment we used the text of the movie review, which is the IMDB, MR and Stanford sentiment tree-bank-2 (SST-2) corpus. Simultaneously, extract data from our corpus labeled "开心" (happy) and "伤心" (sad). This binary classification corpus (CIEA-b) without strong emotional words is compared with the three corpora above. We use opinion lexicon [24] to process English text and [1]HowNet to process the binary corpus from our corpus. We filter the sentiment dictionary to obtain a secondary sentiment dictionary with obvious or strong emotions. We mask the emotional words which represent strong emotions in the text and generate a corpus that dose not contain emotional words. For the corpus with emotional words and the corpus of hidden emotional words, a stacked bidirectional GRU deep learning model is used to do a comparative experiment. In the case that the corpus is different and the rest of the conditions are identical. The model parameters corresponding to different datasets in the experiment are optimal. The result is shown in Table **Table IV**.

As indicated in **Table IV**, it is clear that the results of the corpus (SST, IMDB, MR) that doesn't contain emotional words compared to the original corpus have declined a lot. The corpus of the three binary classifications decreased from 8.61% to 10.38%. This result shows that the emotional word with obvious emotional tendency have a great influence on the performance of the model. Comparing the binary datasets extracted from the corpus we propose, the accuracy rate dropped by 2.97% in the case of masking a large number of emotional words. It proves that our corpus contains very few obvious emotional trend words.

## V. CONCLUSION

In this study, we propose a new Chinese implicit sentiment corpus for implicit sentiment analysis. The corpus includes a large number of self-labeling training set data, validation set data and test set data. In the baseline experiment, the baseline of the stacked model reached 75.06% (macro-average $F_1$ scores). In baseline

[1] 知网:http://www.keenage.com/

system, we just used the word segmentation tool and no adjustment parameters. This shows that the quality of the corpus we proposed is promising. We used comparative experiments to prove the effectiveness and necessity of our proposed corpus. At the same time, the corpus we proposed is the first Chinese implicit sentiment analysis corpus. It effectively fits the implicit expression of Chinese expression and predicts implicit emotions. The new corpus can be used for research purposes. We hope that the corpus we proposed will help researchers better study sentiment analysis of Chinese texts, and encourage more researchers to perform experiments on this data.

## VI. Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No.61966038, No.61702443 and No.61762091.

## References

[1] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] R. Klinger, O. De Clercq, S. M. Mohammad, and A. Balahur, "IEST: WASSA-2018 Implicit Emotions Shared Task," vol. 1, 2018. [Online]. Available: http://arxiv.org/abs/1809.01083

[3] P. D. Turney, "Thumbs up or thumbs down?" *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, vol. 12, no. 3, p. 417, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1073083.1073153

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Antike und Abendland*, vol. 57, no. July, pp. 151–168, 2002. [Online]. Available: http://arxiv.org/abs/cs/0205070

[5] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, no. October, pp. 579–586, 2005.

[6] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[7] G. A. van Kan, Y. Rolland, M. Houles, S. Gillette-Guyonnet, M. Soto, and B. Vellas, "The assessment of frailty in older adults," *Clinics in Geriatric Medicine*, vol. 26, no. 2, pp. 275–286, 2010.

[8] S. Aman and S. Szpakowicz, "Identifying Expressions of Emotion in Text," *Text, Speech and Dialogue*, pp. 196–205, 2007. [Online]. Available: http://link.springer.com/10.1007/978-3-540-74628-7_27

[9] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS ONE*, vol. 6, no. 12, 2011.

[10] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective Text," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007, pp. 70–74.

[11] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *Association for Computational Linguistics*, vol. Proceeding, no. 1, pp. 115—-124, 2005. [Online]. Available: http://aclweb.org/anthology/P05-1015

[12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 2011)*, pp. 142–150, 2011. [Online]. Available: http://www.aclweb.org/anthology/P11-1015%7D

[13] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank ," *PLoS ONE*, vol. 8, no. 9, 2013.

[14] S. Mohammad, "#Emotional Tweets," {**SEM 2012*}*: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* {*(SemEval 2012)*}, pp. 246–255, 2012. [Online]. Available: http://www.aclweb.org/anthology/S12-1033

[15] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Information Processing and Management*, vol. 51, no. 4, pp. 480–499, 2015.

[16] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and Sentiment in Tweets," vol. 0, no. 0, 2016. [Online]. Available: http://arxiv.org/abs/1605.01655

[17] J. Sun, "'jieba' chinese word segmentation tool," 2012.

[18] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 175–180. [Online]. Available: https://www.aclweb.org/anthology/N18-2028

[19] R. E. Wright, "Logistic regression." 1995.

[20] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[21] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1746–1751, 2014.

[22] S. J. Hochreiter Sepp, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078

[24] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, no. November, p. 168, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1014052.1014073