An Evaluation Scheme for the Quality of Reviews

An Evaluation Scheme for the Quality of Reviews

Xuefei Zhang
Applied Informatics Major, Graduate School of Science and Engineering,
Hosei University,
Supervisor: Akihiro Fujii

***Abstract*—In recent years, with the development of e-commerce, the research of product reviews has become more and more important. The high-quality reviews can provide sufficient information to help customers choose the products. In this paper, we built an evaluation scheme for the quality of reviews base on machine learning. The dataset was obtained by the crawler and labeled through manual annotation. We vectorize the reviews by TF-IDF for the SVM model as the basic algorithm. We used the sentiment lexicon analysis method to get a score as the sentiment feature to improve the SVM model. We compared SENTIWORDNET, AFINN, VADER three different sentiment lexicons and chose the best lexicon—VADER to get the sentiment score and add it to the SVM model. The Self-training which is a semi-supervised learning method combined with the unlabeled dataset was built based on the SVM model to solve the problem of insufficient dataset. We also built the LSTM neural network with word embedding of global vectors for word representation and improved it by virtual adversarial training method. We compared the neural network methods with SVM methods. The result shows that the virtual adversarial training model worked better in the evaluation scheme for the quality of reviews.**

Keywords: Quality of Reviews, Sentiment lexicon, Semi-supervised learning.

I. INTRODUCTION

E-commerce is known as electronic commerce or Internet commerce, usually refers to the providing service of buying and selling on the Internet, as well as the transfer of money and data for these transactions. Different from traditional physical stores, e-commerce is based on existing network technology. It has the characteristics of shopping convenient, effective search, seamless checkout experience. E-commerce has developed rapidly because of these unique characteristics around the world. The revenue of B2B e-commerce exceeded US$2 trillion in 2019. The United States is one of the mature online shopping markets and the global standard for e-commerce in the world. Many of the best e-commerce companies are in a leading position in other countries. Amazon has a clear lead in the e-commerce market, with more than 2 billion visits per month.

People can choose more and more goods with the expansion of the e-commerce market scale. However, too much product information will make customers confused. It is not easy to compare similar products and make a selection. People tend to rely on the reviews of people who have bought goods before in this situation. We can get the characteristics of this product from their reviews and the real user experience. However, the problem is that there are not all customer reviews are useful. In the review section, high-quality reviews with useful information are usually mixed with low-quality reviews with useless information. People can't immediately find out which review is valuable, what kind of reviews can help people get the real evaluation of products. The main purpose of this paper is to establish an evaluation scheme for the quality of reviews. This evaluation scheme can help people find high-quality reviews. We use the crawler tool to get online product reviews from Amazon and obtain a small number of labeled data sets through manual annotation. The methods used in this paper mainly include traditional machine learning methods and neural network methods to build models. We use TF-IDF to get the text information from the product reviews and combine SVM as our baseline. The sentiment lexicon is also used as the sentiment feature for the SVM model. Because the size of the dataset is too small, we combined the self-training of semi-supervised learning to improve classification accuracy. This paper also used the method of LSTM neural network to build the model, obtains the vector through GloVe. We further improved the model through the method of adversarial training based on the LSTM model.

II. RELATED WORK

The researchers have realized the importance of product reviews. The review content, emotional diversity, user behavior analysis and classification training models have been used by many researchers to detect fake reviews. The problem of false reviews was originally proposed by Jindal and Liu in the context of product reviews [1] [2]. The detection of reviews content focuses on identifying and classifying fake reviews based solely on opinions expressed by opinions. In [3], the author handled the problem of fake review detection as a binary classification problem, using two popular classification techniques: LS-SVM and Naive Bayes classifier. They classified the reviews. The difference between score-based detections refers to this by identifying scores based on the content of the reviews. The model compared them with a given score to find deviations and identify fake reviews. If the rating of the review is high, the review can also be classified as fake. The author in [4] calculated the rating of the review based on the content and found the difference from the rating given by the reviewer to incorporate the difference between the rating and the review. This research is different from fake review detection, we focus on the evaluation scheme for the quality of reviews.

III. METHODS

A. Data acquisition

In the process of building a machine learning model, we need a training dataset. It is a dataset for the training model to acquire learning ability. But there is no public dataset on the Internet about the quality evaluation of review, we need to collect data by ourselves.

The Web crawler is a kind of computer program, which can browse and find all the information on the target website page by itself. Then the program stores the target data to the local computer. ASIN is the abbreviation of Amazon standard identification number. We used the ASIN code to find the target products. The specific implementation principle and process as shown in Fig.1.

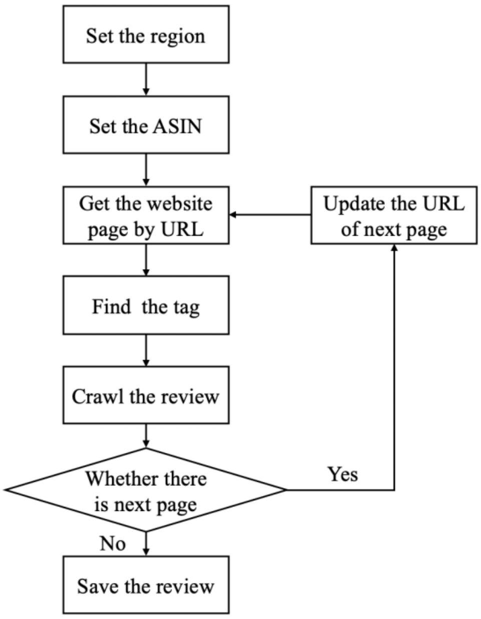

Set the region
Set the ASIN
Get the website page by URL
Update the URL of next page
Find the tag
Crawl the review
Whether there is next page
Yes
No
Save the review

Fig 1. The crawling steps

Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML for scraping. We implement the crawling steps by Beautiful Soup to get the reviews. Our target review dataset is the game product review, and we labeled the dataset by manual annotation. We classify the review as a high-quality review if the review gives us the specific information of the game such as the picture and music, otherwise it is a low-quality review. The same category of reviews was downloaded for semi-supervised learning. We also got the public review dataset with sentiment label for comparing the sentiment lexicons.

B. Data Processing

Processing natural language text and extract useful information from the given word, a sentence using machine learning and deep learning techniques requires the string or text needs to be converted into a set of a vector. Text vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities and semantics.

- Term Frequency-Inverse Document Frequency (TF-IDF), it is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. Although the algorithm is simple, it can efficiently implement functions.
- GloVe, coined from Global Vectors, is a model for distributed word representation [5]. The model is an unsupervised learning algorithm for obtaining vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. Is this research, we used pre-trained word vectors to obtain a single word vector. We matched it with the words of the reviews to get the word embedding and applied it in neural networks.

In the process of building a general machine learning model, we want to use sentiment scores as text features to further improve the model. A common method of sentiment analysis is the sentiment lexicon. The algorithm of lexicon will perform semantic analysis based on the words and phrases that appear in the entire sentence or text and get a sentiment score. We can know the sentiment trend of the sentence or the whole text through the score. If we want to use this method for sentiment analysis, we need the lexicon of sentiment polarity words. Each sentiment polarity word in the lexicon needs to have a score that can express the degree of sentiment. The sentiment score was calculated by the lexicon. There are three mainstream sentiment analysis lexicons:

- SENTIWORDNET is the result of automatically annotating all WORDNET synsets according to their degrees of positivity, negativity, and neutrality [6]. We made program to calculate the sentiment score by SENTIWORDNET lexicon text file.

- AFINN can be used extensively for sentiment analysis. The current version of the lexicon is AFINN-en-165. txt and it contains over 3,300+ words with a polarity score associated with each word [7]. We can obtain sentiment scores by AFINN framework.
- VADER (Valence Aware Dictionary and sEntiment Reasoner) uses a combination of A sentiment lexicon is a list of lexical features which are generally labelled according to their semantic orientation as either positive or negative [8]. VADER framework can be used to get the sentiment scores.

KNN (k-nearest neighbors) is classified by measuring the distance between different eigenvalues. In pattern recognition, the KNN is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space. In KNN, the distance between objects as the non-similarity index between objects is calculated by Euclidean distance or Manhattan distance. We used KNN and SVM to compare the three different lexicons with the public sentiment dataset with and choose the best lexicon as the features。

C. Model Building

The method of the quality evaluation model is machine learning. Machine learning is the general description of many algorithms. The purpose of these algorithms is trying to find the hidden rules from a large number of historical data and use them for prediction or classification.

support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The authors of paper [9] compared SVM, KNN and naive Bayesian algorithms. The experimental data includes four categories of documents: environment, sports, politics and art, and the result shows that SVM classification method is better than other classification methods. Thus, SVM with the TF-IDF was used as the basic model. We also added the sentiment score of the lexicon to the SVM model and compared the result.

Semi-supervised learning is a method between supervised learning and unsupervised learning. We know that the class labels of the samples are known in supervised learning. But there is the problem is that the cost of manually labeling samples is very high, which leads to the scarcity of labeled samples. On the other hand, unlabeled samples are easily collected. The semi-supervised learning models use a large number of unlabeled samples and a small number of labeled samples to train the classifier. This method can solve the problem of insufficient labeled samples. Self-training is a kind of semi supervised learning. Self-training first trains a basic model through a supervised algorithm, and then uses this model to predict unlabeled data to obtain prediction results. The model analyzed whether the new prediction data could be added to the labeled dataset which we already have by the possibility of the result. The specific steps are as shown in Fig 2.

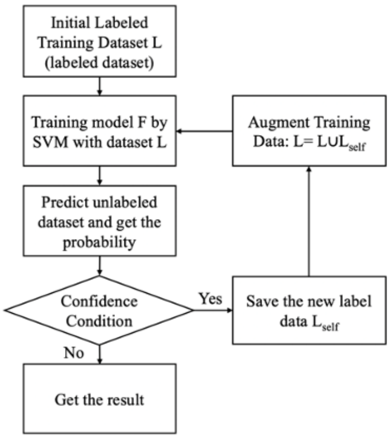

Initial Labeled Training Dataset L (labeled dataset)
Training model F by SVM with dataset L
Augment Training Data: L= L∪L$_{self}$
Predict unlabeled dataset and get the probability
Confidence Condition
Yes
No
Save the new label data L$_{self}$
Get the result

Fig 2. Self-training

Recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. In the paper [10], the author compared different neural models in the classification task of semantic relations between pairs of nominals. The result shows that the RNN model is more appropriate for the relation classification task. We used Long Short-Term Memory (LSTM) in our paper, which is a kind of special RNN could solve the problem of gradient vanishing problem and gradient exploding problem in the process of long sequence training. The structure as shown as Fig 3.

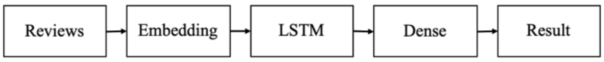


Fig 3. LSTM model

Adversarial training is the process of training a model that can effectively identify the original sample and the adversarial sample model [11]. In image classification, a perturbation which is difficult to distinguish for human eye can greatly change the judgment of model on the image category. Adversarial training usually requires labeled samples to provide supervised loss because the perturbation is designed to improve the model. The virtual adversarial training (VAT) extends the adversarial training to the semi-supervised field by adding regularization to the model so that the output distribution of the sample is the same as the output distribution with the perturbation [12]. We used VAT method to improve the LSTM model as shown as Fig 4.

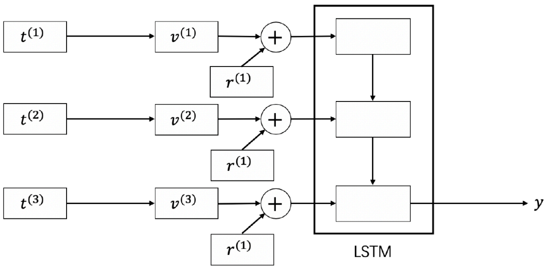


Fig 4. The virtual adversarial training

The r is the perturbation. The all word embedding vectors $v^{(1)}, v^{(1)}, \dots, v^{(t)}$ as s, and the model conditional probability of y given s as $p(y|s;\theta)$ where θ are model parameters. The adversarial perturbation r_{adv} on s as:

$$r_{v\text{-}adv} = \epsilon g / \|g\|_2$$

$$\text{where } g = \nabla_{s+d} KL[p(y|s;\hat{\theta}) || p(y|\, s+d;\hat{\theta})] \quad (1)$$

The d is a TD-dimensional small random vector. The loss function of virtual adversarial loss is:

$$\frac{1}{N^{'}} \sum_{n^{'}=1}^{N^{'}} KL\left[p(y|s_{n^{'}};\hat{\theta}) || p(y|\, s_{n^{'}} + r_{v-adv,n^{'}};\theta)\right] \quad (2)$$

The $N^{'}$ include both labeled and unlabeled examples. We can improve the model by this virtual adversarial training method.

IV. RESULTS

We used Beautiful Soup to get the reviews of game. These reviews are categorized by manual annotation. The high-quality reviews are represented by "1", which provides valid information for customers. The low-quality reviews are represented by the number "0", which cannot provide enough information. There are 300 reviews in this dataset. And we also use an amazon reviews dataset of the video game published by Amazon which includes 2317780 reviews without the label as the unlabeled dataset.

We also used the public Amazon review dataset with emotion labels and compared three different lexicons with SVM and KNN algorithms as shown as Table 1.

Table 1. The Lexicons Comparison

Accuracy	Algorithm	
	SVM	KNN
VADER	0.72	0.70
AFINN	0.70	0.64
SENTIWORDNET	0.69	0.65

The result shows that the sentiment score of the VADER lexicon is better than the AFINN and SENTIWORDNET lexicons. Therefore, we used VADER as the sentiment score feature of the review.

The SVM algorithm with TF-IDF features was trained as the basic evaluation model. Learning curve is a widely used diagnostic tool in machine learning for algorithms. Reviewing learning curves of models during training can be used to diagnose problems with learning, such as underfitting or overfitting. The x-axis of learning curve is the number of samples, and the y-axis is the accuracy of the model. It shows the accuracy of the training model and the test model to compared results. The learning curve of SVM model as shown in the Fig 5.

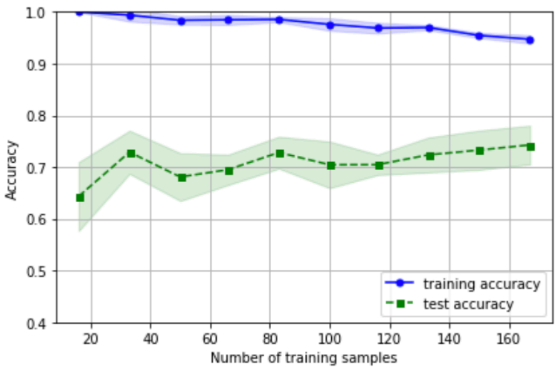


Fig 5. The SVM learning curve with TF-IDF.

As can be seen from the figure, the accuracy of the model increases with the number of training samples. But the model is still underfitting. In order to improve the model, we got the sentiment score as the feature based on the VADER lexicon to retrain the SVM model and applied the self-training method to the model.

Table 2. The Self-training Accuracy

	The Classification Accuracy			
Original Training Data Size	20	50	100	210
SVM_TF-IDF	0.64	0.68	0.74	0.75
SVM_TF-IDF_ Sentimentscore	0.64	0.68	0.75	0.77
Self_Training_ SVM_TF-IDF_ Sentimentscore (K=0.7)	0.64 new labele d 0	0.7 new labeled 10933 5	0.76 new labeled 168475	0.78 new labeled 170600

From the Table 2 we can see that the model with TF-IDF and sentient score features has better results than the model with only used TF-IDF feature.

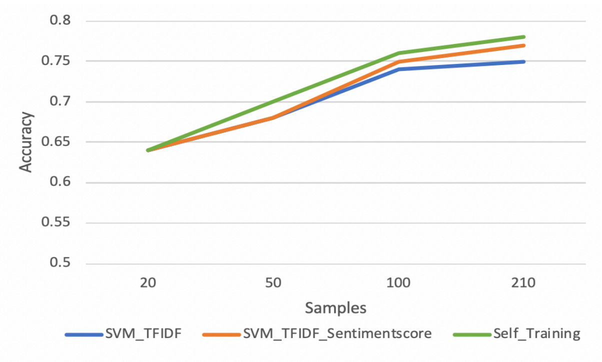
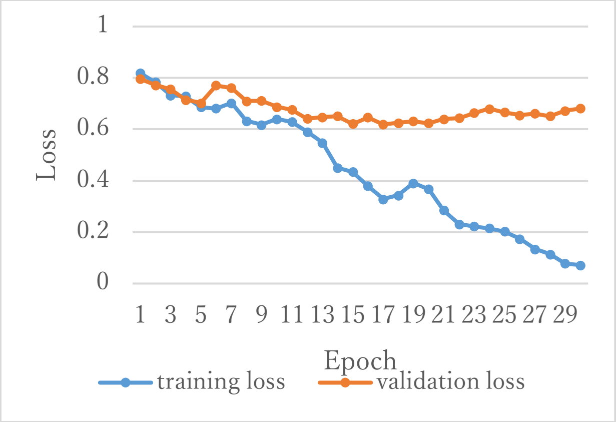


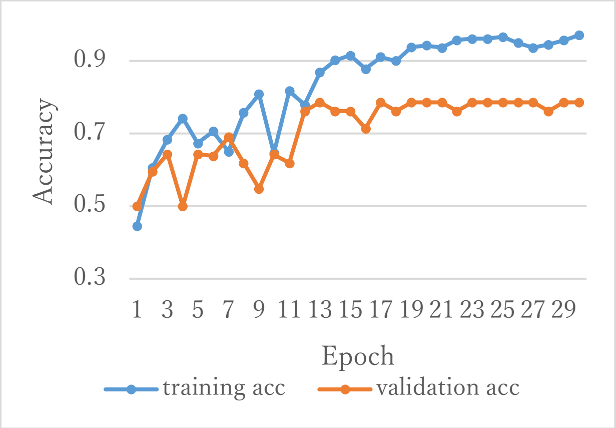
Fig 6. The Accuracy Comparison

As shown in the Table 2 and Fig 6, after self-training, we found that there are more reviews added to the training set. We could get good results when the K is 0.7. The self-training got higher accuracy than SVM with the increase of the training dataset by combining labeled and unlabeled data.

According to the network structure of LSMT, we can change the training rounds of the model by adjusting epochs. The result as shown in the Fig 7.



Loss value

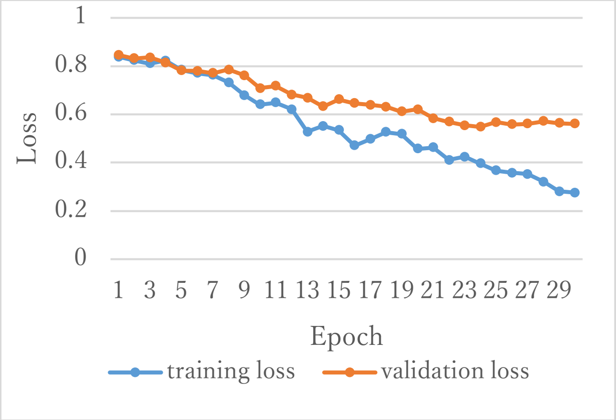


Accuracy

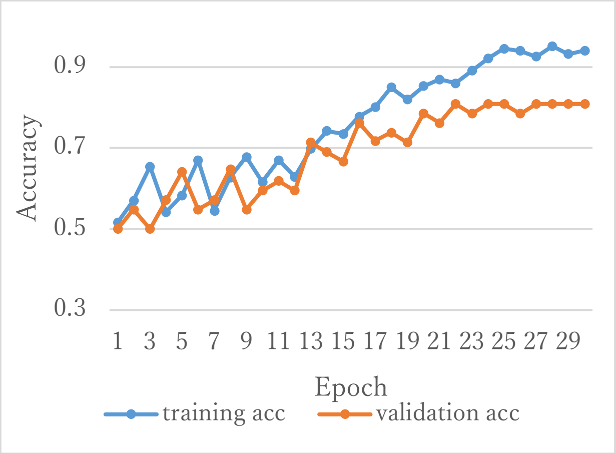
Fig 7. The loss value and accuracy of LSTM

As can be seen from the figure, we find that when the epoch is 20, the model accuracy of the validation set tends to be stable. The loss value increased a little with the epochs. We think that the model is overfitting.

We built the VAT model based on LSTM and set the epochs is 30, as shown in the Fig 8.



Loss value



Accuracy

Fig 8. The loss value and accuracy of VAT

From the figure we can see that the model tends to be stable and converges when epoch is 23. The results of all models as shown in the Table 3.

Table 3. The Result Comparison

	SVM	Self-training	LSTM	VAT
Accuracy	0.75	0.78	0.78	0.80

From the table, we can see that the accuracy of SVM is the lowest, and the accuracy can be further improved by combining self-training. The result of LSTM is similar to the self-training. The VAT method could further improve the LSTM model and worked better than other models.

V. DISCUSSION

The quality evaluation scheme of reviews was built in this paper. We get the dataset product review from Amazon by Beautiful Soup. The dataset was classified into high-quality and low-quality reviews manually. We used SVM with TF-IDF as the basic model. The sentiment scores and self-training could improve the SVM model. LSTM model was also used in this research and can be improved by VAT method. We found that semi-supervised learning methods could improve the supervised learning methods and solve the problem of insufficient dataset in this paper. There are some other methods that can be used for the scheme, we need to try and compare more algorithms. We can also consider adding more features to improve the model.

VI. CONCLUSION

In this work, we propose an evaluation scheme for the quality of reviews to help people choose high-quality product reviews with valid information. We used the machine learning method to build the quality evaluation model of the review. Since Amazon has a large market size in e-commerce, we choose Amazon product reviews as our data and get the product review from Amazon through the crawler. The dataset was classified into high-quality and low-quality reviews manually. We tokenized the reviews and removed the stopwords. The TF-IDF was used to get the vectors from the review, and we trained the SVM as the basic model. The sentiment feature of the review can be used as a feature in the model. The mainstream methods of sentiment analysis are machine learning and sentiment lexicons. We choose the sentiment lexicon method to obtain the sentiment features and compared SENTIWORDNET, AFINN, VADER three different sentiment lexicons. The result shows that VADER works better than other lexicons. We added the sentiment features to the SVM model. The model with sentiment features and TF-IDF is better than the SVM only used TF-IDF as the feature. And we built the self-training model based on the SVM model. This method is one of the semi-supervised learning methods by combining the unlabeled data. For LSTM neural network, we choose GloVe to get word embedding. And the VAT method was used to improve the LSTM model by adding perturbations to the word embedding. We compared the four algorithm models. The result shows that VAT model performs best in this paper. This method can be used to build the evaluation scheme for the quality of reviews to help people get high-quality product reviews.

ACKNOWLEDGEMENT

I would first like to sincerely thank Professor Akihiro Fujii for his great patience and consistent encouragement. He gave me a lot of help， provided me with the necessary materials, valuable suggestions and inspiration of new ideas. Then, I am extremely grateful for Professor Kazuo Yana and Professor Jinjia Zhou for their help in life. Finally, I would like to express my gratitude to my parents for their unfailing support my years of study and the process of writing this thesis.

REFERENCES

[1] N. Jindal and B. Liu, “Analyzing and Detecting Review Spam,” *Seventh IEEE International Conference on Data Mining*, pp. 547-552, 2007.

[2] N. Jindal, B. Liu, “Opinion spam and analysis,” *International Conference on Web Search and Data Mining ACM*, pp. 219—230, 2008.

[3] R. Patel and P. Thakkar, “Opinion Spam Detection Using Feature Selection,” *2014 International Conference on Computational Intelligence and Communication Networks*, pp. 560-564, 2014.

[4] S. P. Algur and J. G. Biradar, “Review spamicity based on rank and content of the review,” *2015 International Conference on Applied and Theoretical Computing and Communication Technology*, pp.140-145,2015.

[5] Pennington, J., Socher, R., & Manning, C. D, “Glove: Global vectors for word representation.” *Proceedings of the 2014 conference on empirical methods in natural language processing*, pp. 1532-1543. 2014.

[6] Baccianella, S., Esuli, A., & Sebastiani, F, “Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining,” *International Conference on Language Resources and Evaluation*, pp. 2200-2204, 2010.

[7] Nielsen, Finn Årup, “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.” *arXiv preprint*, arXiv:1103.2903, 2011.

[8] Hutto, Clayton J., and Eric Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text.” *Eighth international AAAI conference on weblogs and social media*. 2014.

[9] Liu, Z., Lv, X., Liu, K., & Shi, S, “Study on SVM compared with the other text classification methods,” *Second international workshop on education technology and computer science*, Vol. 1, pp. 219-222, IEEE, 2010.

[10] Zhang, D., & Wang, D, “Relation classification: CNN or RNN.” *In Natural Language Understanding and Intelligent Applications*, pp. 665-675, 2016.

[11] [18] Miyato, Takeru , A. M. Dai , and I. Goodfellow . “Adversarial Training Methods for Semi-Supervised Text Classification.” *International Conference on Learning Representations*. 2017.

[12] Miyato, Takeru, et al. “Distributional Smoothing with Virtual Adversarial Training,” *Computer ence* ,2015.