The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 192-201

Extracting Semantic Representations of Sexual Biases from Word Vectors

Ying-Yu Chen
National Taiwan University
r06142009@ntu.edu.tw

Shu-Kai Hsieh
National Taiwan University
shukai@gmail.com

摘要

在台灣的標誌性論壇——PTT 論壇中，帶有性別偏見的線上仇恨言論問題發展已久，詞向量等計算語言學的應用也常帶有性別偏見。本研究利用詞向量表徵（Word2vec），從 PTT「母豬教」鄉民的言談來分析中文厭女言論的詞向量的語意分佈，發現母豬教徒相關言論的確帶有性別偏見與歧視。本研究採取質性加上量化分析，作為第一篇利用自然語言處理技術分析 PTT 論壇上的仇女歧視言論的研究。

Abstract

Sexually biased cyberhate speech has become a fast-growing problem on PTT (a representative online forum in Taiwan). The applications of computational linguistics like word embeddings would also carry similar biases. This paper analyzed the distribution of word representations of netizens from *mu zhu jiao* (a cult that often produces misogynistic cyberhate speech). Word vector representations (word2vec) was utilized for scrutinizing semantic representations of texts found on PTT. The findings from the distributed semantic representation of *mu zhu jiao* implied a sexual bias against them. This paper serves as the first study which investigates the distribution of word representations of the abusive language on PTT forum with an NLP method by taking advantage of both quantitative and qualitative methods.

1 Introduction

During the past few years, the tides of cyberhate speech have surged over online forums and social media. PTT forum, serving as the representative of online Chinese forum, has been to obtain evidence of abusive cyberhate speech on the basis of characterizing gender-biased language, mainly as

disparagingly underestimating the status of the female. Previous research focusing on cyberhate speech of misogyny (Citron, 2011) and negative sentiment of the related word representations (Yu, 2016) made up the big picture of the gender-related cyberhate speech. The emphasis on Distributional Hypotheses (Sahlgren, 2008) also depicts the importance of exploring distributional differences of word representations under the contexts related to a sexually biased speech based on a quantitative computational method. Instead of unsystematic regulations of stopping the propagation of cyberhate speech, there were some methods applying Natural Language Processing (NLP) models for efficiently detecting abusive language online with regard to, but not limited to, cyberbullying and gender-biased speech (Schmidt and Wiegand, 2017;Davidson et al., 2017;Burnap and Williams, 2016). On the ground of previous studies, the relationships and connections between cyberhate speech and gender-biased language on PTT could be reported in both quantitative and qualitative ways. The aim of the present paper is threefold: (1) to analyze the large text sources and compare the distributed representations of words represented by a word vector encoding semantic similarity;(2) to capture linguistic properties from the word vectors;(3) to specify the polarity fluctuation with regard to the keywords from the word vector.

2 Previous Studies

2.1 Misogyny and Cyberhate Speech

Misogyny complex, which Gilmore (2010) suggested best describing as a multi-dimensional phenomenon characterizing the woman hating emotion males direct toward females. This pervasive emotion over the world was specified by Gilmore as a result of the combination of psychogenic in origin and the influence caused by the environment. Specifically, on one hand, the inside conflict between men's abundant need for women and the equally intense fear of that dependence (or fear of losing that dependence) disappointed males themselves, thus leading to their implicit emotional dissatisfaction. Under this circumstance, the female had the possibility to become a convenient and better object to be bullied as a counterbalance to make up the defect. Although society appeared to make a fair amount of progress against gender discrimination, misogyny has, by all means, moved online due to the easy opportunities of speaking on condition of anonymity on the Internet, called "cyberhate attack (Citron, 2011)." Cyberhate attack contained not only speech, but other actions like doctored photographs, privacy invasions, and technical sabotage, while "cyber misogyny" referred to cyberhate attack such as abusive language specifically aiming at women.

As the representative online BBS (Bulletin Board System) forum in the Chinese community, PTT is the place where a wealth of cyberhate speech of misogyny takes place. Yu (2016) claimed that the key concept of misogyny on PTT appeared to follow two patterns: One was emphasizing on the beautiful figure and the virtues that female supposed to have, like performing both emotionally and physically weaker than male in concordance to meet the role expectation from Asian society. The other was restricting the completeness of female as an individual by devaluing female as only an object of sex. This objectification was to a large degree subjective to the critics who tended to take advantage of their ideology to label and define female by using words with specific semantic representations under the context. Such word representations with the aforementioned features were retrieved and visually demonstrated by Yu (2016) as a semantic network, which signaled the distributional word representations based on misogynistic cyberhate speech on PTT. This distribution to an extent specified the words appeared on PTT, such as *gan (fuck), qian (money), chou nu (misogyny), sui bian (easy especially referring to sex), gong zhu (princess syndrome), and po ma (bitch)*. On the basis of the literature review, this present study would be investigating the semantic networks from the abusive speech on PTT forum in the follow-up phases of the study.

2.2 Word Vectors

"You shall know a word by the company it keeps (Firth, 1957)." Distributional hypothesis depicts the idea of the correlation between distributional similarity and meaning similarity so as to make a prediction of semantics by utilizing distributional properties of linguistic entities (Sahlgren, 2008). Exploring the distributional properties of the misogynistic cyberhate speech and the differences between the different distributions are needed. Now, this could be implemented via a neural network inspired vector semantic model called Word2vec (Mikolov et al., 2013;Mikolov et al., 2013) which has been widely used for gathering high-quality vector representations of words from large amounts of unstructured data by using the advantageous Skip-gram model. Previous work (Heuer, 2016) utilized this simple but robust model to provide a view of text source and to compare the words with semantically similar or contrasting relationships, like mapping the country *Finland* to its national sport *hockey*. Schmidt and Wiegand (2017) also gave an overview of hate speech detection, which indicated distributed word representations based on neural networks was an effective way yielding a good classification performance from working the features of word generalization out. Such vector representations were used as classification features as it had the

preference that semantically similar words may end up with similar vectors. With the assumption that distributional models are models of word meaning, this study took advantage of such technique to scrutinize the large text sources for revealing the linguistic properties from the distributional models.

3 Methodology

3.1 Data Collection

Over the decades, PTT has derived various subcultures, among which a new force suddenly arose as *mu zhu jiao*, a cult whose followers believe in the existence of *mu zhu* and worship the popes obov (also known as a000000000) and sumade. On the online forum, everyone can easily share and exchange opinion with each other on more or less condition of anonymity, resulting in the gender-biased speech contributed from the popes of *mu zhu jiao*. With the purpose of getting evidence of the characteristics and nature of the abusive speech on PTT, this study favored making a comparison of the speech between general *xiang min* and the popes of *mu zhu jiao*, trying to draw the differences between them. Hence, the data population of this study consisted of two datasets: Dataset I is made up of the posts from 180 PTT netizens (also known as *xiang min*) randomly chosen from the online PTT user list, while dataset II involved posts from two remarkable popes of *xiang min* of *mu zhu jiao.*

The big picture of the procedure was to collect, organize, and integrate the data. The data were first gathered through crawling the posts from the aforementioned *xiang min* on PTT, after which a popular tool "jieba" was used for segmentalizing text into tokens. In the meantime, a list of 1218 stopwords (e.g., emojis and interjections) was ruled out to avoid getting too much uninformative data. Apart from the list of stopwords, the user dictionary[1] which is comprised of a bunch of frequently used terms and multi-word expressions (MWEs) on PTT was taken into consideration. This user dictionary is consistent with the MWEs and catchphrases employed by pttpedia with the further trimming to 5294 tokens. With the help of the user dictionary, a number of essential terms and fixed constructions were maintained in the data for further analysis. Table 1 provides the overall information of two datasets, including the number of articles, word tokens, word types, and

[1] The user dictionary is retrieved with permission from Liao (https://liao961120.github.io/PTTscrapy/).

vocabulary richness (VR). VR indicated that the words xiang min from dataset II used were slightly richer than that from xiang min from mu zhu jiao.

Table 1: The number of articles, token, types, and vocabulary richness (VR) of two datasets

Population	sumade	obov	a000000000	Dataset I	Dataset II
Articles	1230	573	340	**2143**	**2855**
Tokens	226589	56017	35694	**329127**	**333575**
Types	45339	16508	12501	**63039**	**72648**
VR (%)	0.2001	0.2947	0.3502	**0.1915**	**0.2178**

3.2 Data Analysis

This study employed both quantitative and qualitative approaches, comprising of text analysis and questionnaire of the authentic data from PTT. The quantitative data analysis was performed by using the powerful package word2vec from NLP architecture. Word2vec can be seen as a fashionable and efficient model utilizing large data to compute vector representations of words (Mikolov et al., 2013). With this hypothesis, it was carried out so as to map the representations of words into a vector space to gain word similarity and relatedness. This paper drew a comparison of word representation between the two groups: (1) two well-known *xiang min* of high reputation from *mu zhu jiao*, obov (also known as a000000000) and sumade;(2) 180 randomly chosen *xiang min*. After respectively training the dataset I and dataset II from 329127 and 333575 tokens, two comparing models of representation of vector space were created.

For conducting the analysis, previous researchers typically compared the words with semantically similar or contrasting relationships (Heuer, 2016). Following this, the study was proceeded by generating a bunch of co-occurred words from the model to represent the learned distributional relationships. In the two models, the core word *mu zhu* was queried to predict 20 words with strong context dependency and their corresponding degrees of cosine similarity in the vector space in an effort to detect the distributed difference of word representation between *xiang min* from *mu zhu jiao* and others. Subsequently, the above distributions would be scrutinized to obtain the authentic context and compare the different semantic implication of the keywords in two comparing models. The corresponding concordances of the above 20 words with strong context dependency with *mu zhu* were analyzed as in section 4.1. Furthermore, in the interest of

investigating how the polarity of the related language varied from people who happened to have the related knowledge of misogynistic hate speech on PTT, this study employed a questionnaire to investigate the polarity of the keywords between two groups of people: thirty senior *xiang min* who have logged in PTT more than 3000 times, and thirty people who are not *xiang min*. The keywords were manually selected under human determination in order to filter out the 25 words with the highest correlation with the context of misogynistic subculture on PTT. The result would be discussed in section 4.2.

4 Result and Discussion

4.1 Co-occurred Word and Concordance

The study was proceeded by generating a bunch of co-occurred words from the model to represent the learned distributional relationships. The outcome of the distributed representation from comparing word vectors suggested two preliminary significant differences. First, the words similarly distributed with *mu zhu* shown in the word vector of *mu zhu jiao* indicated a significant semantic consistency with gender-biased language, while the demonstration of the contrasting representation was comparatively neutral without connecting with specific contexts. Second, some famous catchphrases in terms of the abusive language on PTT were presented as similarly appeared word in the word vectors from *mu zhu jiao*. However, the evidence seemed not to imply any connection between the represented words and the implicit implication in word vector from general *xiang min.* That is, the representation of the distributed words from general *xiang min* appeared not to link to any specific context.

Mu zhu, in the configuration of the speech from *mu zhu jiao*, had no doubt serving as the most essential core, assembling its multi-valency on bullying female. In order to compare different semantic implication under the context of two comparing representation, the corresponding concordances of *mu zhu* from two word vectors were extracted. From the concordances, a preliminary difference of semantic implication of *mu zhu* between the two groups of *xiang min* appeared to be detected. In every situation, general *xiang min* talked about *mu zhu* theoretically under the context of breeding pigs, such as the way a boar started to breed a sexually receptive gilt. By contrast, *xiang min* from *mu zhu jiao* strongly criticized *mu zhu* a lot, comparing *mu zhu* to women who have specific characteristics, like double standards, preferring Cross Culture Romance (CCR), worshiping money, and other features. It is noteworthy that some features were judged so hard under this context while

they have no bad essence at all. Take CCR for example, *xiang min* from *mu zhu jiao* considered Taiwanese female having a relationship with a foreigner (especially Caucasian) as a heinous behavior. However, a Taiwanese male would be priding himself if he is dating a Caucasian female. To sum up, the evidence suggested that the speech from *mu zhu jiao* to some extent testified itself as an abusive language on the basis of showing the represented offensive words alluded to gender discrimination. After all, a couple of inferences of *mu zhu* were drawn as a network trying to specify the related semantic implications (Figure 1).

Figure 1: The network of the related semantic implications of *mu zhu* on PTT forum

On the other hand, in contrast to *mu zhu*, this paper also reported the concordances of the counterpart *gong zhu (male pig)* from two word vectors to see if the literally contrasting words *gong zhu (male pig)* are being used as much as *mu zhu (female pig)* in common usage. It turned out that the concordances of *gong zhu* from the discourse from general *xiang min* were exactly the same the concordances of mu zhu, which meant there is little additional use among them. In comparison, *gong zhu* from the discourse from general *xiang min* appeared twice in the overall texts, primarily arguing the issue regarding the existence of "gong zhu jiao (a cult from gong zhu)". In brief, the concordances of the core word *mu zhu* explained its multi-valency from the perspective of language usage on PTT. By contrast, the concordance of *gong zhu* helped to specify that the metaphorically abusive expression of the pig was originated from *mu zhu* and there was not enough evidence to show the gender equality with female in usage. Besides, compared to the word vector of general *xiang min*, the

concordances of both *mu zhu* and *gong zhu* in the word vector of *mu zhu jiao* provided solid evidence of *mu zhu*'s strong context dependency.

4.2 Keywords Polarity and Knowledge-based Features

Abundant semantic implications of the represented words had the tendency to be sentimentally negative. However, the sexually biased hate speech on PTT could be completely understood largely based on related knowledge of misogynistic expressions on PTT because it often employed context-dependent metaphors. In other words, people who didn't engage in PTT would probably have difficulty understanding the implicit meanings from the misogynistic subculture. In order to know how the polarity of the related speech varied from people happened to have the related knowledge, in this subsection, two groups of people were invited to evaluate the polarity of the "keywords" of the abusive speech on PTT: thirty senior *xiang min* who have logged in PTT more than 3000 times and thirty people who are not *xiang min*. 25 keywords were the terms manually selected under human determination to specify their high context-dependency from the speech from m*u zhu jiao*. They might contain words, catchphrases, fixed constructions which often employed metaphoric expressions so as to implicitly identify the intrinsic biased nature. In terms of this questionnaire, the higher the polarity score is, the positive the keyword is, while the lower the polarity score is, the negative the keyword is. The polarity tendency is calculated from 1 (the lowest) to 5 (the highest).

Afterward, the average polarity scores of the 25 keywords from two groups were calculated, rounding to one decimal place. It was not surprising that, 72 % (18 out of 25 terms) of the scores from non *xiang min*, suggested to be lower (from average 0.1 point to 1.2 points) than those from senior *xiang min*. Also, the average polarity score from senior *xiang min* (average 2.16) is 0.24 points lower than which of non *xiang min* (average 2.4). The finding indicated that the perception of particular ideology could be strongly influenced by language use. In this case, it would be more difficult for people without the knowledge of the subculture to perceive the implicit information than people with that knowledge. This situation, in some circumstances, proved the knowledge-based nature of the abusive speech on PTT.

5 Conclusion

This study examined the word representations of biased speech on PTT forum with the application of NLP tools to detect the abusive language online by comparing the distributed representations of

words represented by a word vector encoding semantic similarity and capturing significant linguistic properties from the word vectors. Also, this study took advantage of questionnaires to see the fluctuation of the polarity of the word representation between the party causing injury and others. The findings showed that the distributed semantic representations of *mu zhu jiao* and those of general *xiang min* differed significantly in terms of several perspectives: First, the distributed representations of words represented by a word vector encoding semantic similarity. Second, the keyword concordance and the encoding semantic implications. Third, the related semantic network of the keyword *mu zhu*. Last, the polarity implementation of the related words of the keyword *mu zhu.* The overall results indicated that the distribution of the surrounding words of *mu zhu* tended to display sexually biased semantic implications, making the word vector of *mu zhu jiao* a blatantly sexist.

This study served as the very first paper analyzing the abusive language on PTT with an NLP method by taking advantage of both quantitative and qualitative methods. However, it should be noted that this study has been primarily concerned with posts originating from the popes of *mu zhu jiao* because of the difficulty of defining who really are its followers, and has only addressed some features to detect sexually biased cyberhate speech on PTT instead of developing a well-established computational model to automatically detect the abusive language. Despite its preliminary nature, the study may offer some insight into the gender bias amplification of the distribution of word representations on PTT. To further capture the subtle and highly context-dependent abusive speech in an efficient and scalable computational method, future research may use a text processor to capture the characteristics of biased language while explicitly seen as neutral speech. Hopefully, the embedding model could be further modified to remove gender stereotypes and biases from training data.

References

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357).

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526.

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 11.

Citron, D. K. (2011). Misogynistic Cyber Hate Speech.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Gilmore, D. D. (2010). *Misogyny: The male malady*. University of Pennsylvania Press.

Grus, J. (2015). *Data science from scratch: first principles with python*. O'Reilly Media, Inc.

Heuer, H. (2016). Text comparison using word vector representations and dimensionality reduction. *arXiv preprint arXiv:1607.00534*.

Jane, E. A. (2016). Online misogyny and feminist degilantism. Continuum: Journal of Media & Cultural Studies.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pagano, R. R. (2012). *Understanding statistics in the behavioral sciences*. Cengage Learning.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33-53.

Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10).

余貞誼. (2016). 我說妳是妳就是：從 PTT 母豬教的仇女行動談網路性霸凌的性別階層. *婦研縱橫*, (105), 22-29. [Yu, Zhen-Yi. (2016). Wo shuo ni shi ni jiu shi: Cong PTT muzhujiao de chounu xingdong tan wanglu xing baling de xingbie jieceng. Fu yan zong heng