Quantum-Inspired Density Matrix Encoder for Sexual Harassment Personal Stories Classification

Peng Yan[1, 2], Linjing Li[1, 3], Weiyun Chen[4], Daniel Zeng[1, 3]
[1] The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Shenzhen Artificial Intelligence and Data Science Institute (Longhua), Shenzhen, China
[4] School of Management, Huazhong University of Science and Technology, Wuhan, China
{yanpeng2017, linjing.li, dajun.zeng}@ia.ac.cn, chenweiyun@hust.edu.cn

***Abstract*—Nowadays, more and more sexual harassment personal stories have been shared on social media. To better monitor and analyze the extent of sexual harassment based on these social media data, we need to automatically categorize different forms of sexual harassment personal stories. Existing methods apply convolutional neural network (CNN) with different convolution window sizes to this text classification task. However, the previous CNN models do not provide an effective way to synthesize window size-related local representations, but simply concatenate all local representations together. To address this problem, we propose a new density matrix encoder, inspired by quantum mechanics, to encode local representations as particles in quantum state and generate a global representation as quantum mixed system for each story. Experiment on SafeCity dataset shows that our model outperforms CNN baseline and achieves better performance than the state-of-the-art model when considering both accuracy and speed, demonstrating the effectiveness of the proposed density matrix encoder.**

***Keywords*—Density Matrix, Quantum Mechanics, Text Classification, Sexual Harassment**

I. Introduction

Nowadays, with the development of social media, more and more people are willing to share their lives and experiences on social media. Victims of sexual harassment also share their sexual harassment personal stories on different online platforms. For instance, the global #MeToo movement[1] demonstrates that the widespread prevalence of sexual assault and harassment has caused a great impact on public security. So it is of great significance to monitor and analyze these shared harassment data on social media for combating sexual assault and harassment. The classification of different types of sexual harassment personal stories (e.g., commenting, ogling/staring, and touching/groping) can help authorities and the public improve awareness and increase understanding of the extent of sexual harassment. However, it is impossible to manually sort and comprehend large-scale stories shared on social media in a timely manner. Thus, we need the power of artificial intelligence, especially natural language processing (NLP) technology, to automatically classify vast amounts of sexual harassment personal stories, which is an important task in security-related social media analytics.

Existing methods apply traditional convolutional neural network with different convolution window sizes (e.g., [3, 4, 5]) to this text classification task[1]. Each convolutional filter with the given window size n generates a local representation to encode n-grams related semantic information in text. However, traditional convolutional neural network models do not provide an effective way to synthesize these local representations, but simply use concatenation of local representations as global representation for each story, which may be sub-optimal compared to jointly optimizing them.

To deal with this weakness, we propose a new density matrix encoder, inspired by quantum mechanics, to encode a mixture of local representations as particles in quantum state and obtain a global representation as quantum mixed system for each sexual harassment personal story. Such a density matrix encoder for better semantic representation fusion can be integrated into neural network architecture. Experiment on SafeCity dataset shows that our model outperforms several baseline models, including a traditional CNN baseline. Futhermore, our model achieves comparable accuracy of state-of-the-art model and has faster training speed advantage, which demonstrates the effectiveness of the proposed density matrix encoder.

II. Related Work

Recently, researches on harassment and abuse information on social media have got more and more academic and industrial attention, especially for sexual harassment. Karlekar et al. [1] assembles the SafeCity Dataset of sexual harassment personal stories, discussed further in Section IV.A. Also, they apply multiple classifiers (e.g., logistic regression, Gaussian Naive Bayes, convolutional neural network (CNN), recurrent neural network (RNN), etc) to sexual harassment personal stories classification task. And our proposed model is improved from their convolutional neural network model by adding a density matrix encoder on the top of convolution encoder.

Other related works includes building a quality annotated corpus and an offensive words lexicon for different types of harassment content [2] and utilizing an annotated corpus from Twitter to analyze the language for different types of harassment, including sexual harassment [3].

[1] https://metoomvmt.org/

978-1-7281-2504-6/19/$31.00 ©2019 IEEE

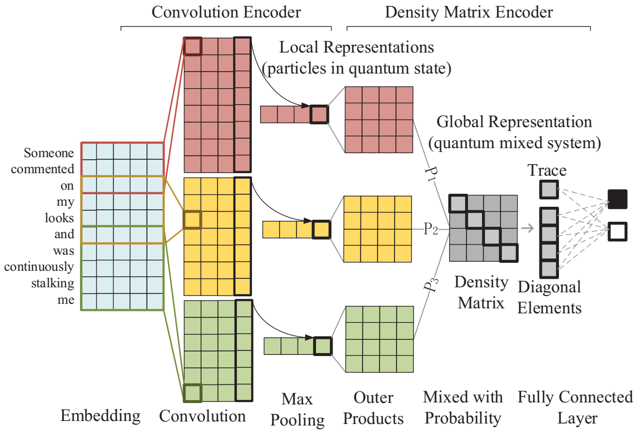

Fig. 1: The overall architecture of the proposed model

III. Model

A. Overall Architecture

We use one-versus-rest method to convert sexual harassment personal stories multi-label classification to several single-label binary classification tasks for different harassment forms. The binary classification labels are [commenting, non-commenting], [ogling, non-ogling], and [groping, non-groping].

The overall architecture of our model is shown in Fig. 1. First, a convolution encoder reads word embeddings and computes local representations for each sexual harassment personal story by a convolution neural network with max pooling. Then, a density matrix encoder computes a density matrix based on local representations and produce a global representation by concatenating the trace and diagonal elements of density matrix. Finally, the produced global representation is passed through a fully connected layer to produce probabilities over the output classes. The first two complex modules, convolution encoder and density matrix encoder, will be explained in detail in the following sections.

B. Convolution Encoder

We use a convolutional neural network to encode local representations for each sexual harassment personal story. A similar architecture has been proved effective in sentiment analysis tasks [4]. For each input story s consisting of a sequence of N words $(w_1, w_2, \cdots, w_N)$, each word is converted to a d-dimensional vector by an embedding matrix layer W_{emb}. Thus, a story s can be represented as a matrix $W_s \in R^{N \times d}$. In Fig. 1, we give an example of 10 words with 5-dimensional word embedding.

Then the matrix W_s is fed through multiple convolution filters of varying window sizes with relu activation. After that, results of multiple convolution filters are applied max pooling to produce multiple m-dimensional local representations, where m is the number of hidden units for each window size. As illustrated in Fig. 1, we use a list of filter window sizes [3, 4, 5] with 4 hidden units for each window size and thus can get three 4-dimensional local representations.

In actual implementation, we use 300-dimensional word embedding and a list of filter window sizes [3, 4, 5] with 128 hidden units for each window size.

C. Density Matrix Encoder

Recently, density matrix in quantum mechanics has been applied in information retrieval and natural language processing tasks [5] [6] and been proved effective in modeling text data as a quantum mixed state. For example, Zhang et al. proposes a question answering model by learning to compare the density matrices of a question and an answer [6]. Inspired by quantum mechanics, we propose density matrix encoder to encode a mixture of convolution-encoder-generated local representations as particles in quantum state to produce a global representation as quantum mixed system. We will use Diracs notation in quantum mathematics to describe how this density matrix encoder works. The Diracs notation denotes a unit vector $\vec{u}$ as a ket $|u\rangle$ and its transpose $\vec{u}^T$ as a bra $\langle u|$. And an outer product (also called dyad) of $|u\rangle$ is $|u\rangle\langle u|$.

First, we get multiple local representations from convolution encoder, represented by m-dimensional vectors $\vec{l_1}, \vec{l_2} \cdots \vec{l_c}$, where c is the number of filter window sizes. In actual implementation, c equals three as three different window sizes [3, 4, 5] are used. Such a local representation vector $\vec{l_i}$ can be regarded as a quantum state for a story s. We normalize it to get a unit state vector $|l_i\rangle$

$$| \, l_i\rangle = \frac{\vec{l_i}}{\| \, \vec{l_i} \, \|_2}. \tag{1}$$

Then, according to the definition of density matrix in quantum probability, we compute outer products of the state vector $| \, l_i\rangle, \forall i \in \{1, 2, \cdots, c\}$ and generate density matrix ρ as a mixture of outer products with probability P_i, as illustrated in Fig. 1

$$\rho = \sum_{i \in \{1,2,\cdots,c\}} P_i \, | \, l_i\rangle\langle l_i \, |, \tag{2}$$

where $P_i, i \in \{1, 2, \cdots, c\}$ are trainable parameters with unit normalization: $\sum_{i \in \{1,2,\cdots,c\}} P_i = 1$. The parameters $P_i, i \in \{1, 2, \cdots, c\}$, which can be explained as weights for n-grams related semantic information, are set trainable to get better semantic representation fusion. Density matrix is a generalization of the conventional finite probability distributions. Inspired by Gleasons Theorem [7] in quantum measurement probability, we compute a global representation g as a quantum mixed state by concatenating the trace and diagonal elements of the density matrix ρ

$$g = [trace(\rho); \rho_{diag}], \tag{3}$$

where ρ_{diag} is the diagonal elements vector of density matrix ρ. The global representation g is the final output of the density matrix encoder, which will be fed to the binary classifier, a fully connected layer with a dropout of 0.80 applied, to produce probabilities over the output classes.

TABLE I: Experimental results (accuracy) on SafeCity dataset

Models	Commenting	Ogling	Groping
Non-neural models			
Logistic Regression	61.4	78.0	69.1
Gaussian Naive Bayes	46.8	74.7	66.0
SVM	65.5	79.0	70.3
Neural models			
CNN+Simply concatenation	80.9	82.2	86.0
RNN	81.0	82.2	86.2
CNN-RNN	81.6	**84.1**	**86.5**
Our model(CNN+Density matrix)	**82.2**	83.0	**86.5**

TABLE II: Results of training time experiment

Models	First Epoch(s)	Training Phase(s)
CNN+Simply concatenation	10	116
RNN	609	9665
CNN-RNN	899	13271
Our model(CNN+Density matrix)	10	119

IV. Experiment

A. Datasets

The proposed model is evaluated on a publicly-available dataset SafeCity[2], which is built based on stories shared on the online forum SafeCity[3] by Karlekar et al. [1]. SafeCity is a crowd-sourcing online forum for victims of sexual harassment sharing personal stories. Each story includes a description of the occurrence submitted and tagged forms of sexual harassment by forum users. We use standard splits of single-label classification on the top-3 dense categories of sexual harassment: commenting, ogling and groping. The whole dataset includes 9,892 stories with 7201 training samples, 990 development samples, and 1701 test samples.

B. Hyper-parameter Settings

We use a 300 dimensions GloVe word embedding[4] as initialization and keep word embedding trainable during training. For the convolution encoder, we use a list of filter window sizes [3, 4, 5] with 128 hidden units per filter window size, which is the same as CNN baseline's setting [1] for comparison. A dropout of 0.80 is applied in the final fully-connected layer to overcome overfitting problem. In the training phase, batch size is set to 64 and Adam optimizer is used with learning rate 0.0001.

C. Results and Analysis

We compare our model with multiple baselines, including logistic regression, Gaussian Naive Bayes, SVM, CNN+Simply concatenation that simply uses the concatenation of local representations as a global representation, RNN, and CNN-RNN model that consists of RNN on top of CNN. The results of baselines are borrowed from [1].

Table I shows classification accuracy of all models on three harassment categories tasks. As shown, our model outperforms all non-neural model baselines by a wide margin. As for neural baselines, our model also outperforms CNN+Simply concatenation and RNN model, and is comparable to the state-of-the-art CNN-RNN model in terms of accuracy. We conduct a training time experiment on SafeCity dataset commenting task for all neural models. Table II shows results of training time experiment. As shown in Table II, our model is training much (about a hundred times) faster than the models using RNN structure, which is not suitable for parallelization. Therefore, our model with quantum-inspired density matrix encoder achieves better performance than the state-of-the-art model when considering both accuracy and speed.

V. Conclusion

In this paper, we applied a quantum-inspired density matrix encoder to sexual harassment personal stories classification, which is an important task in security-related social media analytics. The proposed density matrix encoder fuses semantic representation better by modeling global text representation as quantum mixed system. Experimental results on the SafeCity dataset indicated that our model achieves comparable accuracy of the state-of-the-art model with significant speed advantage. In the future, we will continue to evaluate the effectiveness of the proposed density matrix encoder on other NLP tasks.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2016QY02D0305 and 2017YFC0820105, the National Natural Science Foundation of China under Grants 71702181, 71621002, as well as the Key Research Program of the Chinese Academy of Sciences under Grant ZDRW-XH-2017-3. Linjing li is the corresponding author.

References

[1] S. Karlekar and M. Bansal, "Safecity: Understanding diverse forms of sexual harassment personal stories," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 2805–2811.

[2] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth, "A quality type-aware annotated corpus and lexicon for harassment research," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '18. New York, NY, USA: ACM, 2018, pp. 33–36.

[3] M. Rezvan, S. Shekarpour, K. Thirunarayan, V. L. Shalin, and A. P. Sheth, "Analyzing and learning the language for different types of harassment," *CoRR*, vol. abs/1811.00644, 2018.

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

[5] A. Sordoni, J.-Y. Nie, and Y. Bengio, "Modeling term dependencies with quantum language models for ir," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, NY, USA: ACM, 2013, pp. 653–662.

[6] P. Zhang, J. Niu, Z. Su, B. Wang, L. Ma, and D. Song, "End-to-end quantum-like language models with application to question answering," in *AAAI*, 2018.

[7] A. M. GLEASON, "Measures on the closed subspaces of a hilbert space," *Journal of Mathematics and Mechanics*, vol. 6, no. 6, pp. 885–893, 1957.

[2] The SafeCity dataset is avaliable at https://github.com/swkarlekar/safecity

[3] http://safecity.in

[4] http://nlp.stanford.edu/data/glove.840B.300d.zip