Online Dating Analysis for Chinese Gay Men using Artificial Intelligence

Runyao Yu
TU Munich
Munich, Germany
runyao.yu@tum.de

Ye Wang
ETH Zurich
Zurich, Switzerland
wangye@ethz.ch

Meng Zhao
ETH Zurich
Zurich, Switzerland
zhmeng@ethz.ch

***Abstract*—Gay dating apps like Grindr and SCRUFF are widely regarded as the primary platforms for gay men to engage in online dating. However, we discovered that on Zhihu, a Chinese question-and-answer website, tens of thousands of users have been looking for romantic partners. To investigate which motivations these gay men have the most, which topics they frequently discuss, and how they interact with each other on Zhihu, We conducted a mixed-methods study, including utilizing conventional Machine Learning (ML) methods and Natural Language Processing (NLP) to analyze their dating motivations, behaviors, and interaction quantitatively and qualitatively. We first identified a set of motivation and topic classes, with which we trained the Logistic Regression (LR) model, Long Short-Term Memory (LSTM) model, and Bidirectional Encoder Representations from Transformers (BERT) classifier. We then analyzed at scale which motivations and topics are most prevalent, as well as how interactions with various topics affect users' engagement on the site. We end by discussing the significance of our findings for future study and online dating for Chinese gay men.**

***Index Terms*—Social Media, Online Dating, Chinese Gay Men, Machine Learning (ML), Natural Language Processing (NLP)**

I. Introduction

Gay-oriented websites have become a part of gay men's daily lives since the late 1990s. Finding prospective dating partners online has grown in popularity as an alternative to cruising bars [1]. Gay dating apps have been the most popular avenues for socialization within the gay community since the early 2010s due to the advancement of mobile communication technologies and underlying infrastructure [3]. A similar story is narrated in China. Since Guangtong's inception in 1998, Chinese gay men have gathered on gay-oriented websites and gay dating apps [2]. However, in recent years, an increasing number of Chinese homosexual men have shifted from these cyberspaces, which are explicitly designed for LGBTQ users, to mainstream social media for online activities [4]. To be more specific, Chinese gay men are engaging in online dating activities on non-dating platforms such as Zhihu rather than gay dating apps.

Zhihu (zhihu.com) is a Chinese question-and-answer website where users can post questions, answers, and articles. On Zhihu, 100,000 users follow the "gay" topic, and the community offers knowledge, facts, understandings, and experiences about homosexuality. Dating activities on Zhihu always take place in the context of "gay" queries, such as "How do males find boyfriends on Zhihu?" or "What is it like to be gay and single?". Zhihu users refer to these inquiries as "fishing questions." Tens of thousands of people respond to these queries with their own self-presentations as "bait". If other users, referred to as "fishes," read these posts and deem the authors to be possible romantic partners, they will leave comments for further connection. However, due to the rapid development of modern Chinese, it has become extremely difficult for researchers to understand the motivations and topics of this group. Therefore, We attempt to apply artificial intelligence (AI) to assist researchers in better comprehending this group.

AI is constantly expanding and has been one of the most prominent research topics in recent decades [5], [6]. The goal of AI is to empower systems with intelligence capable of human-like learning and reasoning. It has numerous advantages and has been successfully implemented in a variety of industrial fields, including Speech Recognition, Computer Vision, Natural Language Processing, and so on. In order to evaluate the distribution of motivations and topics of Zhihu users, we applied different classifiers on the data collected from Zhihu. Logistic Regression (LR) [27], as a very powerful algorithm, can well implement classification problems. We also used LSTM and the cutting-edge classifier BERT as a comparison. As a variant of Recurrent Neural Network (RNN), LSTM can solve the problem of gradient vanishing to a large extent and is widely used in time-series problems [26], while BERT is a transformer-based machine learning technique developed by Google and is significantly efficient for dealing with NLP problems [25]. Both models have their own areas of expertise. In this paper, we will compare their performance on Chinese texts. To improve the performance of models, we designed the algorithm for data augmentation to generate similar data and receive better generalization. After classifying motivations and topics corresponding to the user's comments, we applied data science to establish a connection between self-presentations and comments in order to analyze the affection of self-presentations on comments. This work aims at addressing the following three research questions:

- What motivations do users bring to browse on Zhihu?
- What topics do users discuss frequently?
- How does the user's self-presentation affect the motivation and topic distribution in comments?

TABLE I
VALIDATION EXAMPLE USING TF-IDF AND N-GRAM

Topic	Translation	Keywords (Unigram)	Translation
外貌与身材	Appearance and Figure	身材/好看/气质	Figure/Good Looking/Temperament
地理位置	Location	郑州/西亚斯/河南	Zhengzhou/Xiyasi/Henan
教育与职业	Education and Carrier	学习/考研/茶艺师	Study/Post-graduate Examination/Tea Master
Topic	**Translation**	**Keywords (Bigram)**	**Translation**
外貌与身材	Appearance and Figure	这么 好看/神仙 颜值/好看 可惜	Very Good Looking/Great Facial Attractiveness/Good Looking Pity
地理位置	Location	也是 东北/ 河南 老乡/ 也是 河南	Also from Northeast/Henan Folk/Also from Henan
教育与职业	Education and Carrier	努力 学习/考研 加油/茶员 茶艺师	Study Hard/Cheer for Post-graduate Exam/Tea Clerk Tea Master

II. BACKGROUND AND RELATED WORK

A. Ideological Change for Chinese Gay Community

Homosexuality is morally abhorrent in traditional Chinese society, particularly in Chinese Confucian culture [9] [10]. Given that practically everyone in traditional Chinese society married, homosexuality was viewed as a habit rather than an identity in ancient times [8] [11]. Moral criticism was leveled because same-sexual behaviors impede the development of the ideal Confucian personality [7].

Since the 1920s, Western society has introduced a pathologized understanding of homosexuality into China, which has become the mainstream understanding of homosexuality in Chinese society. Homosexuality was decriminalized in China until 1997, and it is no longer regarded a mental disease until 2001 [13]. Meanwhile, since the first HIV infection in mainland China in the 1980s, ordinary people have associated homosexuality with AIDS, resulting in widespread public dread and misunderstanding of homosexuality [7].

In such a heteronormative environment, Chinese gay men usually hide their homosexual identity in public [12] [14]. Despite the fact that Chinese homosexual men have more freedom to express themselves online [16], they confront social and cultural pressures as a result of a lack of basic political and civil rights [15]. A large number of Chinese gay citizens are forced to marry (heterosexual) people of the opposite sex, and the ensuing social problems and conflicts exacerbate the stigmatization of homosexuality in modern culture [7] [15]. In this context, how Chinese homosexuals find partners on the Internet has emerged as an interdisciplinary topic of human, technology, society, media, and sexuality studies.

B. Social Media for Chinese Gay Community

Social media is crucial for Chinese gay individuals for allowing them to gain information and knowledge about the Chinese gay community, engage with other members of the community, and receive support from others. Social media, particularly in China, where the homosexual population is vast yet dispersed, allows individuals to connect regardless of geographic boundaries [7]. Previous studies have shown that social media platforms such as Tumblr and Facebook provide a "queer utopia" enabling homosexual individuals to exchange social supports, resulting in a positive mental influence on them [17] [18].

Social media is also intended to help people make new friends, seek entertainment, and engage in online dating [21] [23]. However, it appears that Chinese gay users are unable to fully utilize the capabilities of social media [22]. Because of the fear of stigma, Chinese gay users manage their social media ecosystem and publish different information on various social media platforms [19]. Gudelunas [20] discovered that Facebook is not regarded as a useful application for dating among gay men in New York and Dallas, owing to the fact that utilizing conventional dating apps exposes their identify to friends, family, and workplace.

The new occurrence on Zhihu implies that mainstream social media may be beneficial for the Chinese gay minority to pursue romantic connections as well. Understanding how the Chinese gay community uses social media for common purposes, as well as their experiences, difficulties, and challenges in using social media for dating needs, will assist us in better understanding how mainstream social media supports online dating for Chinese gay users. This encourages us to find out the motivations of Chinese gay men when using social media, the topics they are keen to talk about on the Internet, and how they affect each other.

III. RESEARCH AND METHODOLOGY

A. User Profile

This subsection aims to clarify the user profile of Zhihu [28] [30] and compare it with Blued and Fanka, which are the two most popular gay dating applications in China. Both of them are location-based real-time dating applications, while Fanka also provides a matching mechanism as Tinder.

First, Zhihu users are mainly young people, i.e., 72% of Zhihu users are between 20 and 29 years old, while other gay dating applications attract users from a much wider age range. The proportion of Blued and Fanka users aged from 20 to 29 is 24.75% and 47.68%, respectively.

Second, the education level of Zhihu users is high, i.e., more than 80% of Zhihu users have a bachelor's degree or higher [30]. However, there are no statistics of the users' educational level of Blued and Fanka. We consider two other numbers as a reference: 9.3% of all Chinese Internet users have a bachelor's degree or higher [29], and 51% of Chinese LGBT participants of an online survey have a bachelor's degree or higher [31].

These two characters suggest that Zhihu users are mostly young and relatively well-educated people.

TABLE II
LOOK-UP CLASS TABLE WITH KEYWORDS

Motivation	Translation	Keywords	Translation
想要认识对方	Want to have Relationship	找男朋友/普通朋友/炮友	Find a Boyfriend/Normal Friend/Sex Partner
想要更多信息	Want to have more Information	询问情感状态/身高/体重/年龄/位置	Ask for Relationship State/Height/Weight/Age/Location
分享看法与经历	Share Opinion and Experience	分享对情感的看法/过去经历	Share Opinions towards Relationship/Past Experience
与答主简单互动	Interact with "Fisher"	发布表情包/无观点性陈述/乱码	Post Emoji/Random Words/Garbled Character
Topic	**Translation**	**Keywords**	**Translation**
恋爱关系	Relationship	爱情/同居生活/前任	Love/Cohabitation/Ex-boyfriend
社交与朋友	Social Contact and Friendship	联系方式/人脉/朋友	Contact Information/Social Network/Friendship
性	Sex	约炮/性爱细节/攻/受	Hook-up/Sex Detail/Top/Bottom
外貌与身材	Appearance and Figure	长相/肌肉/胖/瘦	Face/Muscle/Overweight/Skinny Shape
地理位置	Location	街道/小区/北京/上海/广州	Street/Community/Peking/Shanghai/Guangzhou
教育与职业	Education and Career	大学/学历/专业/就业/工资	University/Degree/Major/Occupation/Salary
个人信息	Personal Information	身高/体重/年龄/星座/生肖	Height/Weight/Age/Constellation/Chinese Zodiac
日常生活	Daily Life	兴趣爱好/游戏/食物/厨艺/宠物	Hobby/Game/Food/Cooking/Pet
无意义话题	Meaningless Texts	表情包/无观点性陈述/乱码	Emoji/Random Words/Garbled Character

TABLE III
CLASS STATISTICS OF 500 SELF-PRESENTATIONS

Motivation	Translation	Pos.	Neg.
想要找男朋友	Want to find a Boyfriend	345	155
想要找炮友	Want to find a Sex Partner	28	472
想要找朋友	Want to find a Friend	98	402
分享看法与经历	Share Opinion and Experience	408	92
与以上无关	Not Relevant	38	462
Topic	**Translation**	**Pos.**	**Neg.**
恋爱关系	Relationship	347	153
社交与朋友	Social Contact and Friendship	116	384
性	Sex	231	269
外貌与身材	Appearance and Figure	197	303
地理位置	Location	406	94
教育与职业	Education and Career	357	143
个人信息	Personal Information	376	124
日常生活	Daily Life	370	130
无意义话题	Meaningless Texts	26	474

TABLE IV
CLASS STATISTICS OF 1500 COMMENTS

Motivation	Translation	Pos.	Neg.
想要认识对方	Want to have Relationship	200	1300
想要更多信息	Want to have more Information	209	1291
分享看法与经历	Share Opinion and Experience	209	1291
与以上无关	Not Relevant	920	580
Topic	**Translation**	**Pos.**	**Neg.**
恋爱关系	Relationship	277	1223
社交与朋友	Social Contact and Friendship	104	1396
性	Sex	120	1380
外貌与身材	Appearance and Figure	149	1351
地理位置	Location	278	1222
教育与职业	Education and Carrier	178	1322
个人信息	Personal Information	131	1369
日常生活	Daily Life	122	1378
无意义话题	Meaningless Texts	483	1017

B. Data Annotation and Validation

Initially, we developed a python crawler to collect data from Zhihu. We obtained 5891 self-presentations, each of which has between 0 and 1026 comments. The keywords were then determined by manually reading the randomly selected 500 self-presentations and 1500 comments. We created the class table based on the keywords and progressively adjusted it by performing validation of labeling experiments on other randomly selected 100 self-presentations and 300 comments. The validation experiments were conducted by three researchers, individually. The error rate for self-presentations is 27% and for comments is 18%, respectively.

One reason for such high error rates is that some texts are relatively long, and thus, the sub-classes covered by them are not unique. For instance, "I want to find a boyfriend! I'm currently studying for a master's degree in Shanghai", this comment's motivation is "Want to have Relationship", and the topics include "Relationship", "Location" and "Education and Carrier". If the three researchers only use one certain topic from the table to label, there will be a lot of disagreements. The fact that a text can have multiple labels inspired us to utilize multiple binary classifiers rather than multi-class classifiers.

After re-labeling these 100 self-presentations and 300 comments with the allowance of multiple labels, the disagreement rate reduced to 4% and 2%, respectively. Then, we used TF-IDF and N-Gram to verify the correctness of the class table and keywords. TF-IDF is a common method for weighing words in NLP tasks. It provides a value to a word based on its relevance in a text multiplied by its importance across all papers. N-Gram is a contiguous sequence of N items from a text. An N-Gram of size 1 is referred to as a "Unigram" and of size 2 is a "Bigram". TF-IDF and N-Gram are frequently combined to understand the context of texts. We utilized TF-IDF to acquire high-frequency terms and then applied N-Gram to obtain the corresponding keywords under each class, as shown in table I, and compare them with our class table. As a result, our class table, as shown in table II, share similar keywords to table I. Thus, the class table is ultimately validated as a vital reference for data annotation. We randomly re-sampled 500 self-presentations and 1500 comments and labeled these texts separately using the validated class table. After the first round of labeling, We marked the texts that had been labeled

TABLE V
BACK-TRANSLATION EXAMPLE

Index	Sentence
C1	我是重庆人，喜欢吃火锅。在一段感情里比较被动，比较享受被照顾的感觉。
E1	I am from Chongqing, I like hot pot. I am passive in relationship and enjoy the feeling of being taken care of.
C2	我出生于重庆，爱吃火锅。我的性格比较被动，喜欢被人照顾的感觉。

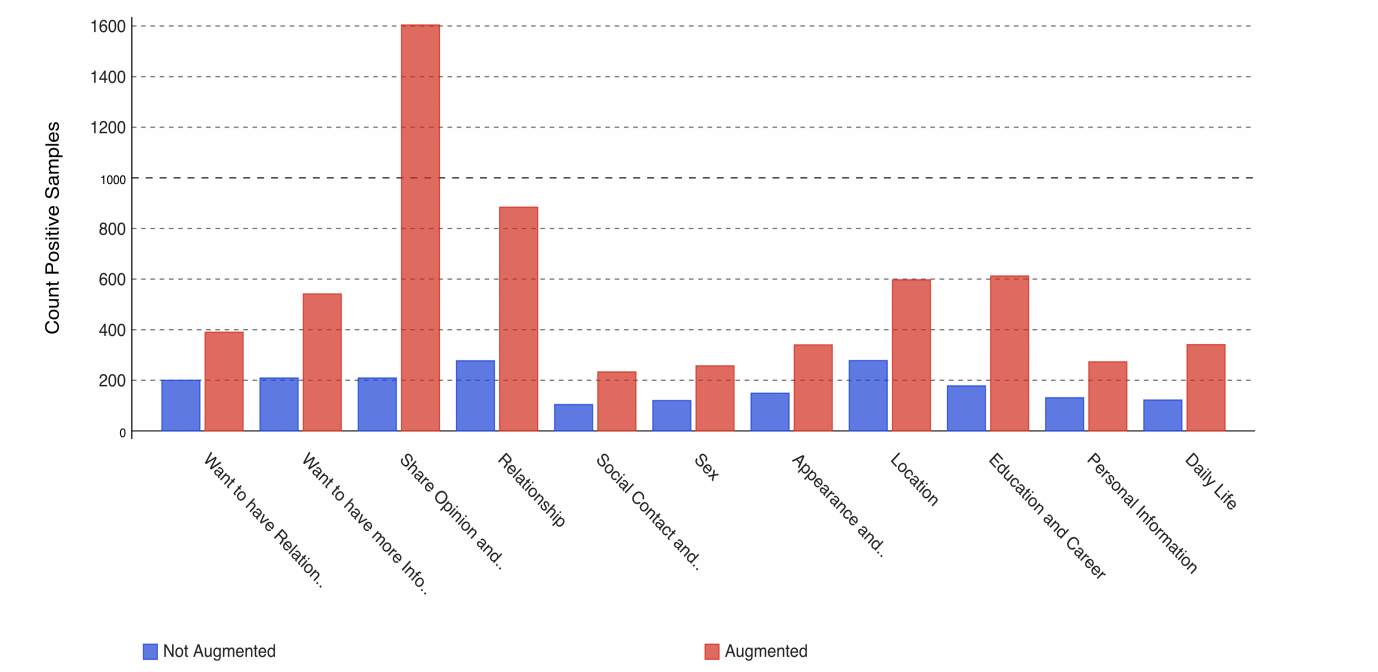

Fig. 1. Distribution Comparison

differently by two or three authors. We repeated the procedure until no more different labels appeared.

We observed the distribution of 500 self-presentations and 1500 comments, as given in table III and IV. we evaluated statistics for each subclass and discovered that the difference between the number of positive and negative samples was excessive (positive label means that the text is related to the sub-class, and a negative sample means that it is not involved). Due to the data imbalance, texts classification becomes much more challenging.

C. Data Augmentation

Data augmentation [32] is a method of data generation based on visual or semantic invariance, and it is the most straightforward and concise way to increase model performance. When dealing with imbalanced or small data sets, the data augmentation strategy can help the model generalize and perform better.

In text classification tasks, a large amount of labeled data is frequently required to obtain adequate results for a classifier model. However, in many circumstances, such as product reviews, the volume of label data is minimal and the acquisition cost is considerable [33].

The two most prevalent data augmentation approaches in NLP tasks are Easy Data Augmentation (EDA) and Back-Translation [34].

The first-place solution for Kaggle's "Toxic Comment Classification Challenge" employed Back-Translation. The winner deployed machine translations to supplement both the train and test data sets, using translations from French, German, and Spanish that were then translated back into English. Back-Translation is also commonly applied in machine translation [35] to ensure that the translation is accurate [36]. It has been demonstrated that the Back-Translation approach outperformed EDA in general [41]. Therefore, we utilized Back-Translation to augment the data in this paper.

Back-translating the source text typically involves two translations, as shown in table V, with the original sentence C1 being translated into other languages (such as English) as E1 and then back-translated into the original language as C2. The augmentation corpus is constituted of C1 and C2 after back-translation of target language texts. Owing to variation in translation software and grammar, Back-Translation differs from the source text in several ways, from words to grammatical structure, which can be comprehended as raising the number of datasets. Back-Translation can generate a variety of paraphrases while maintaining the semantics of

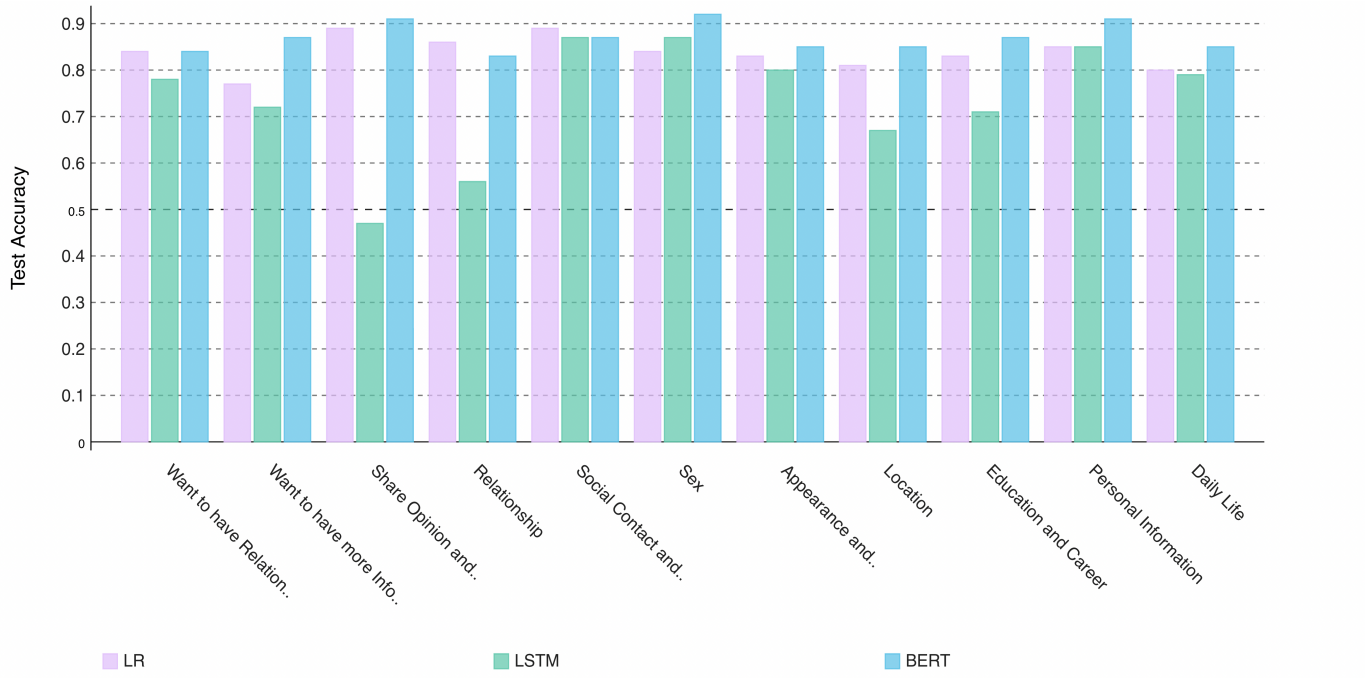

Fig. 2. Test Accuracy Comparison

the original sentences, including the synonym replacement [38] [39] [40], syntactic structure substitution, and deletion of irrelevant words [37]. As shown in table V, take a Chinese computer review as an example of Back-Translation using Baidu Translate.

We designed an algorithm to remotely connect to Baidu Translate through Baidu API. Baidu Translate takes an average of 18 seconds to deliver a response, with 1-7 back-translated texts. The amount of back-translated data varies depending on the text's complexity. We employed data augmentation on comments in order to use machine learning to better accomplish classification task. The disparity between the number of positive and negative samples is substantially reduced after the original data is augmented. The comparison is shown in Figure 1.

D. Predicting Motivations and Topics

The average length of self-presentations is 1810, and over 10.26% of texts are greater than 4096 characters. The BERT classifier has a constraint in that the maximum text length should never exceed 512 characters. As a result, the model is inapplicable in this circumstance. Although there are various models that can handle long texts, such as Longformer, its maximum allowable length remains 4096. Even if we utilize Longformer, a big portion of the texts will be cut off, resulting in poor performance. Furthermore, for self-presentations, 500 texts that we have already labeled are sufficient to represent user's behavior. Thus, forecasting thousands of self-presentations and observing the distribution makes little difference.

However, we need to investigate how self-presentations influence the distribution of comments for the subsequent analysis. Therefore, it is essential to know the motivations and topics associated with each comment. There are a total of 46241 comments under 500 self-presentations. As a solution, we considered training our existing 1500 labeled comments with machine learning algorithms to forecast the motivations and topics relating to the remaining 44741 comments.

To automatically predict the motivations and topics of each comment, we divided this classification task into several binary classification tasks. Namely, each task is to predict whether it is about a certain sub-class. We abandoned the class "Not Relevant" and "Meaningless Texts", respectively. These two classes stand for non-sense characters or random words and thus are not necessary to be trained or predicted. Consequently, we have a total of 11 classification tasks, and furthermore to predict the remaining 44741 comments. We evaluated three machine learning models commonly used in text classification tasks to predict these sub-classes unilaterally.

Logistic Regression (LR) [27]: Logistic Regression is one of the simplest machine learning algorithms, and it is straightforward to implement while providing excellent training efficiency. Additionally, training a model with this algorithm does not necessitate a large amount of computational power. We developed the generic logistic regression model using Word2Vec representation [43] and trained it on the GPU "TITAN Xp".

Long Short-Term Memory (LSTM) [26]: LSTM networks are a type of RNN that can learn order dependence through some variant of gradient descent. However, the gradient in the Deep Neural Network (DNN) is unstable and tends to

TABLE VI
MACRO TEST ACCURACY COMPARISON

Model	Macro Test Accuracy
Logistic Regression Classifier	83.7%
LSTM Classifier	73.5%
BERT Classifier	86.9%

TABLE VII
CHI-SQUARE TEST

Type	P-Value
Motivation-Motivation	2.37e-5
Motivation-Topic	8.08e-3
Topic-Motivation	6.68e-7
Topic-Topic	5.05e-5

either increase or decrease exponentially. This is known as the gradient vanishing or exploding problem. The concept of cell states is introduced in LSTM, which allows the gradient to freely flow backward through time, helping to make it more resistant to the vanishing gradient problem. The upper limit of LSTM is relatively high. Nevertheless, one of its limitations is that the amount of data required for training is typically greater than that required by other commonly used algorithms. After searching for hyperparameters, we decided to use a single-layer LSTM model with 128 units implemented in TensorFlow. We trained all 11 sub-classes on the CPU.

Bidirectional Encoder Representations from Transformers (BERT) [42]: BERT is a pre-trained language representation model that has outperformed baseline methods on NLP tasks, achieving state-of-the-art results. BERT has inspired many recent NLP architectures and language models, such as XLNet, ERNIE2.0, RoBERTa, Google's, TransformerXL, OpenAI's GPT-2, etc. As a very powerful model, BERT is expected to perform well even on a small data set. We implemented the BERT algorithm in PyTorch and ran it on the GPU "TITAN Xp".

E. Finding Self-Presentation Affection in Comments

After using machine learning tools to predict all the remaining 44741 comments, we applied data science to establish a connection between self-presentations and comments.

We first investigated the distribution of motivations and topics of self-presentations, and then tracked down the index of the set of self-presentations tied to a particular sub-class. We used Chi-square test to verify the influence and utilized the index to find the motivations and topics that corresponded to all of the comments in these self-presentation sets in order to explore how self-presentations affect the distribution of their comments.

IV. RESULT AND DISCUSSION

We applied 80% of the augmented data as the training set and trained them on three models separately for 11 sub-classes. We used 20% of the augmented data as the test set. Figure 2 depicts the accuracy on the test set for three models.

It can be seen that, in addition to the higher accuracy of LSTM on the sub-class "Sex" than that of Logistic Regression, the accuracy of LSTM on all other sub-classes is lower or equal than that of the other two models. As shown in the table VI, the average accuracy of LSTM is 10.2% and 13.4% lower than the other two models, respectively. The main reason for this issue is that the LSTM model requires a much larger training set than the other two models. The training set in the order of thousands is obviously not enough for LSTM. In contrast, Logistic Regression has a simpler model and requires fewer training sets than LSTM. Thus, it is more expressive than LSTM model in this case. BERT, as we expected, can achieve excellent accuracy even on a small training set. As compensation, the BERT classifier spent the most computational time on the GPU.

Then, we decided to use BERT, the best classifier from the trained model, to predict all comments corresponding to 500 self-presentations. The distribution of 46241 comments can be seen in Figure 3. From the overall distribution point of view, the motivation of Chinese gay men to comment on Zhihu is more inclined to express their personal views towards relationships and share past experiences (13.6%) than looking for a partner (7.7%) or requesting more information (8.0%). The topics frequently discussed are "Relationship" (17.0%), "Location" (10.5%), "Education and Career" (9.7%), and "Personal Information" (9.2%). In the public's perception, the topics that gays are keen to discuss might be sex or appearance. Such statistical findings differ from the public's stereotype of gay people. One reason for this statistical result is that Zhihu users are generally better educated. In comparison to seeking simple and fast pleasure, Zhihu users likely appear to care about critical factors that influence relationship such as geographic location, personal information, and social background.

In order to understand how self-presentations influence the comments, we applied Chi-square test to investigate the affection. We first counted the number of positive sub-classes of comments corresponding to each sub-class of self-presentations. Next, we divided the amounts into 4 parts: motivations of self-presentations vs motivations of comments; motivations of self-presentation vs topics of comments; topics of self-presentations vs motivations of comments; topics of self-presentations vs topics of comments. Afterwards, we conducted Chi-square test on 4 parts, individually. We set the alpha as 0.05, the P-values of 4 tests can be found in table VII. They all indicated that we should reject the hypothesis and therefore, the influence of self-presentations on comments cannot be ignored.

To investigate the influence quantitatively, We began by observing the distribution of motivations and topics of comments and defining it as the baseline, as shown in Figure **??**. Then, for each sub-class of self-presentations, we computed the index and used it to calculate the proportion of each sub-class of comments. Next, we calculated the increment and decrement based on the baseline and marked the increment as green and decrement as red, if the absolute value is higher than 1.0%,

TABLE VIII
SELF-PRESENTATION'S AFFECTION ON MOTIVATIONS OF COMMENTS

Self-Presentations\Comments	Want to have Relationship	Want to have more Information	Share Opinion and Experience
Want to find a Boyfriend	+0.96%	+0.29%	-1.26%
Want to find a Sex Partner	+0.45%	-2.29%	+1.83%
Want to find a Friend	-2.56%	+1.47%	+1.08%
Share Opinion and Experience	-0.39%	-0.16%	+0.55%
Relationship	-0.68%	-0.14%	0.83%
Social Contact and Friendship	-1.84%	-0.67%	+2.52%
Sex	+1.68%	+0.59%	-2.27%
Appearance and Figure	+0.67%	-0.17%	-0.49%
Location	+0.54%	+0.24%	0.79%
Education and Carrier	-0.07%	+0.06%	+0.00%
Personal Information	+0.21%	+0.26%	-0.47%
Daily Life	-0.39%	+0.18%	+0.21%

TABLE IX
SELF-PRESENTATION'S AFFECTION ON TOPICS OF COMMENTS

Self-Presentations\Comments	Relationship	Social Contact and Friendship	Sex	Appearance and Figure
Want to find a Boyfriend	-0.66%	+0.25%	+0.02%	-0.21%
Want to find a Sex Partner	+0.49%	-1.56%	+1.76%	+0.05%
Want to find a Friend	+0.50%	-0.11%	-0.10%	-0.41%
Share Opinion and Experience	0.00%	-0.07%	+0.03%	-0.04%
Relationship	+0.07%	+0.12%	+0.33%	-0.16%
Social Contact and Friendship	+0.79%	-0.41%	-0.47%	+0.00%
Sex	-0.73%	+1.00%	-0.02%	-0.35%
Appearance and Figure	+0.07%	+0.20%	+0.03%	-0.10%
Location	-0.10%	+0.14%	-0.07%	-6.00%
Education and Carrier	-0.27%	+0.07%	-0.07%	-0.30%
Personal Information	-0.18%	+0.20%	+0.11%	-0.24%
Daily Life	-0.07%	-0.09%	+0.05%	-0.11%
Self-Presentations\Comments	Location	Education and Career	Personal Information	Daily Life
Want to find a Boyfriend	+0.55%	+0.01%	+0.07%	-0.13%
Want to find a Sex Partner	+0.06%	-0.79%	+0.26%	-0.26%
Want to find a Friend	-0.27%	+0.61%	-0.63%	+0.40%
Share Opinion and Experience	-0.01%	-0.01%	+0.07%	+0.03%
Relationship	+0.14%	-0.42%	+0.09%	-0.18%
Social Contact and Friendship	-0.22%	+0.03%	-0.35%	+0.62%
Sex	+0.66%	-0.08%	-0.08%	-0.41%
Appearance and Figure	-0.10%	-0.38%	+0.30%	-0.01%
Location	+0.11%	-0.16%	+0.13%	-0.05%
Education and Carrier	+0.20%	+0.45%	-0.01%	-0.06%
Personal Information	+0.22%	+0.03%	-0.01%	-0.13%
Daily Life	+0.09%	-0.02%	+0.08%	+0.06%

as shown in table VIII and table IX.

According to the table, if self-presentation's motivation is to find a boyfriend, the number of comments driven to share decreases by 1.26%, while the number of comments motivated to have relationship increases slightly by 0.96%. This implies that when Zhihu users encounter someone looking for a boyfriend on the Internet, they generally keep their own remarks regarding emotional perspectives and prior experiences to a minimum. In reality, when most individuals strive to form new connections, they tend to conceal their genuine selves.

What's noteworthy is that when the motive for self-presentation is to find a sex partner, less people want to know or understand these "fishers" (-2.29%), and people are less inclined to form friendships with these "fishers" (-1.56%). Simultaneously, the amount of comments discussing and sharing their previous sexual experiences has substantially grown (+1.83%). The majority of highly educated Zhihu users are more concerned with efficiency. Those who just want to have sex seek to capture the other people's attention through more direct communication while minimizing their own time expense as much as possible.

In contrast to the preceding, when the aim of "fisher" is to discover a friend with whom to form a deep relationship, the person who replies will want to better understand each other (+1.47%) by expressing their own perspectives on emotions and life, as well as discussing their prior experiences to allowing the "fisher" to better understand themselves (+1.08%). Undoubtedly, the number of people who respond with the desire to have relationship will drop as a result (-2.56%).

Sex has always been a sensitive topic. It is worth noting that when self-presentation discusses the topic of sex rather than finding a sex partner, people become willing to establish

connection with the "fisher" (+1.68%), and more of these people want to become friends with those who are willing to discuss sex on Zhihu (+1.00%). What's remarkable is that, as a result of the system for banning sensitive terms, most individuals will become cautious, resulting in 2.27% less comments regarding their experience. Instead, people choose to interact with "fisher" via other social applications, such as Wechat and QQ.

Another intriguing finding is that when the topic of self-presentation is tied to a geographic place, people are less likely to address appearance and body shape in their responses (-6.00%). Geographic location is crucially important to Zhihu users while seeking for a spouse. We also discovered that certain Zhihu users are completely opposed to "long-distance relationships." For the purpose of geographical location, Zhihu users will even compromise or decrease the requirements for partner's appearance.

V. Conclusion

In this paper, we searched for keywords, applied TF-IDF to validate their correctness, and then manually labeled the data. To address the issue of data imbalance, we employed the data augmentation method i.e. Back-Translation. We used the three most prevalent machine learning algorithms to train the models. We weighed the benefits and drawbacks of each model and selected the best-performing model to create text-based predictions. We utilized the Chi-Square Test technique to qualitatively confirm the effect of self-presentations on comments, and then we applied data science to establish the connection between self-presentations and comments. Finally, we evaluated the influence of self-presentations on comments quantitatively.

We presented a comprehensive set of research methodologies on motivation and topics for Chinese question-answer social platforms in this study. We now have a better knowledge of Chinese gays with higher education as a result of our research, which will enable software developers design more appropriate regulations to assist Chinese homosexuals have a better social experience.

In the future work, we plan to use computer vision technology to process user avatars and chat content-related pictures, supplemented by NLP technology to conduct a more comprehensive analysis of user's behavior.

References

[1] Bettina Heinz, Li Gu, Ako Inuzuka, and Roger Zender. 2002. Under the rainbow flag: Webbing global gay identities. International Journal of Sexuality and Gender Studies 7, 2 (2002), 107-124.

[2] Hui Jiang et al. 2011. ICCGL: cultural communication via the internet and GLBT community building in China. (2011).

[3] Shangwei Wu and Janelle Ward. 2018. The mediation of gay men's lives: A review on gay dating apps studies. Sociology Compass 12, 2 (2018), e12560.

[4] Tianyang Zhou. 2018. Jack'd, Douban Group, and Feizan. com: The impact of cyberqueer techno-practice on the Chinese gay male experience. In Exploring Erotic Encounters. Brill Rodopi, 27-43.

[5] C.M. Bishop,Pattern Recognition and Machine Learning. Berlin, Germany: Springer-Verlag, 2006.

[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[7] Jin Cao and Xinlei Lu. 2014. A preliminary exploration of the gay movement in mainland China: Legacy, transition, opportunity, and the new media. Signs: Journal of Women in Culture and Society 39, 4 (2014), 840-848.

[8] Edmond J Coleman and Wah-Shan Chou. 2013. Tongzhi: Politics of same-sex eroticism in Chinese societies. Routledge.

[9] Nan Zhang and Jing Zhang. 2010. The influence of traditional ethical views on the phenomenon of homosexuality in Chinese and Western history. Journal of Heilongjiang College of Education 3 (2010).

[10] Xianglong Zhang. 2018. How Should Confucianism View the Legalization of Same-sex Marriage? INTERNATIONAL JOURNAL OF CHINESE AND COMPARATIVE PHILOSOPHY OF MEDICINE 16, 2 (2018), 53-72.

[11] Zaizhou Zhang. 2001. Aimei de lichen: zhongguo gudai tongxinglian shi/An Ambiguous Trajectory: History of Homosexuality in Pre-modern China.

[12] Petula Sik Ying Ho, Stevi Jackson, Siyang Cao, and Chi Kwok. 2018. Sex with Chinese characteristics: Sexuality research in/on 21st-century China. The Journal of Sex Research 55, 4-5 (2018), 486-521.

[13] T Mountford. 2010. China: the legal position and status of lesbian, gay, bisexual and transgender people in the People's Republic of China. International Gay and Lesbian Human Rights Commission.

[14] Tiantian Zheng. 2015. Tongzhi living: Men attracted to men in postsocialist China. U of Minnesota Press.

[15] Min Liu. 2013. Two gay men seeking two lesbians: An analysis of Xinghun (formality marriage) ads on China's Tianya.cn. Sexuality and Culture 17, 3 (2013), 494-511.

[16] Shuaishuai Wang. 2020. Chinese affective platform economies: dating, live streaming, and performative labor on Blued. Media, Culture and Society 42, 4 (2020), 502-520.

[17] Andre Cavalcante. 2019. Tumbling into queer utopias and vortexes: Experiences of LGBTQ social media users on Tumblr. Journal of Homosexuality 66, 12 (2019), 1715-1735.

[18] Alexander Cho. 2018. Default publicness: Queer youth of color, social media, and being outed by the machine. New Media and Society 20, 9 (2018), 3183-3200.

[19] Michael A DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook' Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1-23.

[20] David Gudelunas. 2012. There's an app for that: The uses and gratifications of online social networks for gay men. Sexuality and Culture 16, 4 (2012), 347-365.

[21] Tiffany A Pempek, Yevdokiya A Yermolayeva, and Sandra L Calvert. 2009. College students' social networking experiences on Facebook. Journal of applied developmental psychology 30, 3 (2009), 227-238.

[22] Ellen Simpson and Bryan Semaan. 2021. For You, or For 'You'? Everyday LGBTQ+ Encounters with TikTok. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (2021), 1-34.

[23] Lesa A Stern and Kim Taylor. 2007. Social networking on Facebook. Journal of the Communication, Speech and Theatre Association of North Dakota 20, 2007 (2007), 9-20.

[24] Sepp Hochreiter and Juergen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735-1780.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2018.

[26] L. Yang et al., "A LSTM Based Model for Personalized Context-Aware Citation Recommendation," in IEEE Access, vol. 6, pp. 59618-59627, 2018.

[27] T. Haifley, "Linear logistic regression: an introduction," IEEE International Integrated Reliability Workshop Final Report, 2002., 2002, pp. 184-187.

[28] Baidu Zhishu. 2021. Zhihu user demographic profile. http://zhishu.baidu.com/

[29] China Internet Network Information Center. 2021. The 47th China Statistical Report on Internet Development. (2021).

[30] I-Research. 2018. 2018 China Knowledge Marketing White Paper - Taking Zhihu as an Example. (2018).

[31] Statista Research Department. 2014. 1st China LGBT Community Survey.

[32] S. Wong, A. Gatt, V. Stamatescu, and M. D. Mcdonnell, "Understanding Data Augmentation for Classification: When to Warp?," in digital image computing techniques and applications, 2016, pp. 1-6.

[33] A. Anabytavor et al., "Do not have enough data? Deep learning to the rescue!," in national conference on artificial intelligence, 2020.
[34] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," arXiv: Computation and Language, 2019.
[35] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in meeting of the association for computational linguistics, 2016, vol. 1, pp. 86-96.
[36] M. Miyabe and T. Yoshino, "Evaluation of the Validity of Back-Translation as a Method of Assessing the Accuracy of Machine Translation," in international conference on culture and computing, 2015, pp. 145-150.
[37] Q. Xie, Z. Dai, E. Hovy, M. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," arXiv: Learning, 2019.
[38] X. Zhang, J. Zhao, and Y. Lecun, "Character-level convolutional networks for text classification," in neural information processing systems, 2015, pp. 649-657.
[39] W. Y. Wang and D. Yang, "That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using petpeeve Tweets," in empirical methods in natural language processing, 2015, pp. 2557-2563.
[40] S. Kobayashi, "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," arXiv: Computation and Language, 2018.
[41] Ma, J. and Li, L., "Data Augmentation For Chinese Text Classification Using Back-Translation", in Journal of Physics Conference Series, 2020, vol. 1651, no. 1. doi:10.1088/1742-6596/1651/1/012039.
[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019).
[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. (2013).