The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 68-80
© The Association for Computational Linguistics and Chinese Language Processing

Exploring Lavender Tongue from Social Media Texts

吳小涵　Hsiao-Han Wu
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
xiaohanwu.hanna@gmail.com

謝舒凱　Shu-Kai Hsieh
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

摘要

在性別與自然語言處理的脈絡下，大多數研究僅專注於生理性別的討論，對於性別文本的分類，更僅建立於異性戀男女的文本上。針對此一現象，本研究為中文性別與自然語言處理領域中，第一個由性取向的觀點出發，討論同性文本偵測的研究。首先本研究由網路論壇-PTT 收集同性戀文本並討論同性戀男女的語言學特徵。其次，藉由觀察到的語言學現象，利用 5 折交叉驗證支持向量機器(Support Vector Machine)與樸素貝葉斯分類器(Naive Bayes)模型，以機器訓練的方式，利用不同的語言學特徵組來偵測男同性戀與女同性戀的網路文本。機器訓練結果顯示，在同性文本的預測上，由於本研究使用了傳統性別與自然語言處理研究未考量到的同性戀特有詞彙特徵，而在同性文本偵測上達到了較佳的正確率。

Abstract

Under the issue of gender and Natural Language Processing (NLP), most papers aim at gender-norm language that spoken by biologically males and females with opposite-sex desires. However, from the point of view of sexual orientation, this study presents the first work in the task of Chinese homosexual identification. Firstly, we collect homosexual texts from social media, and secondly examine linguistic behavior found in gay and lesbian texts. In addition, we also provide sets of linguistic features to automatically predict homosexual language with the adoption of 5-fold cross-validation Support Vector Machine (SVM) and Naive Bayes (NB) models. Training procedure in the study resulted in promising f-score around 70% with the use

of particular lexicon-based feature set.

關鍵詞：同性文本偵測，薰衣草語言學，中文自然語言處理，支持向量機器，樸素貝葉斯分類器

Keywords: homosexual identification, lavender linguistics, Chinese NLP, Support Vector Machine, Naive Bayes

1. Introduction

Lavender Linguistics has been emerging as a linguistic sub-field which analyzes language used by gay, lesbian, bisexual, transgender, and queer (LGBTQ) speakers [1]. It is suggested that there is still considerable room for linguistic research based on fine-grained sexual orientation [2]. Previous studies of gender and NLP mainly focused on dichotomous genders in biological sense without considering the gender complexity of human beings in real world.

When it comes to gender, a general but complicated term with various dimensions involving both biology and psychology, anthropologists have divided it into three major classes:

1) *Sex* refers to physical or biological differences between males and females.

2) Opposite to physical characteristics, *gender* is characterized by self-identity, namely, whether one see himself/herself as male or female.

3) *Sexuality* is about one's sexual attraction and orientation. People who have opposite-sex desires are regarded as heterosexuals. Conversely, people who have same-sex desires are, therefore, regarded as homosexuals.

In the field of gender and NLP (abbreviated as GenderNLP), gender is usually considered with the norm that subjects are biologically males and females with heterosexual desires. However, based on the perspective towards *sexuality*, the present paper discusses lavender speakers and NLP (abbreviated as LavenderNLP) with the hypothesis that previous study on gender identification cannot correctly identify gender in a more complex dimension and that GenderNLP has failed to consider the complexity of *sexuality*.

Since GenderNLP only aims at biological gender, LavenderNLP, as a subclass of GenderNLP, targets not only at biological gender but also at psychological gender. Therefore, referring to Table 1, subjects in the current study are regarded as homosexual males (gays) and females (lesbians) who have same-sex desires regardless of whether he or she self-identifies as male or female; in other words, only *sex* and *sexuality* are taken into account in the definition of gay and lesbian in the study.

While studies on GenderNLP abound, there are still gaps in LavenderNLP to be explored. Accordingly, this study intends to explore lexicon-based cues of lavender speakers and applies all the investigated linguistic behavior to automatically predict homosexual texts from Chinese social media with the use of Support Vector Machine (SVM) and Naive Bayes (NB) models under the 5-fold cross-validation test.

Table 1. Definition of Gay and Lesbian in the present study

Sex	**Gender**	**Sexuality**	**Defined as**
Male	Male	Male	Gay
Male	Female	Male	Gay
Female	Male	Female	Lesbian
Female	Female	Female	Lesbian

2. Related Work

If males and females do have their own in-group language, gays and lesbians will also have their own language which is incomprehensible to outsiders [3]. Also, it had been noted that there is a relationship between language and *sexuality* [4]. Although studies rarely discuss *sexualities*, there is no doubt language can be classified via types of *sexuality*. People who have opposite- or same-sex desires will have different language behaviors. Since the present study discusses texting strategies of homosexual population, this section reviews previous works on how homosexual males and females produce language differently.

2.1 Homosexual Male Language

Compared to lesbian language, linguistic behavior of gay males has been studied extensively.

It has been claimed that gay people tend to use specialized lexicon, or argot, containing words not normally used in mainstream society [5][6][7]. However, not only argot but also gay language is in general characterized by the use of innuendo, categorizations, and strategic evasions such as omitting or changing gendered pronouns [4].

In the past, the word 'gay' was (and still) associated with negative thoughts, which is believed to be the main reason gay men shifted toward a more heterosexual masculine image [8] with their needs to distinct themselves from appearing obviously gay [9]. The appearance of masculine items [9] or the replacement of masculine pronouns with feminine pronouns [10] in gay men's language is considered strategies for homosexual males to behaves more heterosexually.

2.2 Homosexual Female Language

While linguistic features of gay language are believed to be more conspicuous, it is claimed that there are no linguistic features unique to lesbian text [11]. However, since lesbians can identify each other in a variety of settings but find it difficult to explain how the interaction mechanism works [12], four linguistic styles that may help lesbians identify each other are further proposed [2]: (a) stereotyped women’s language (hypercorrect grammar, tag questions); (b) stereotyped nonstandard varieties of working class urban male language (cursing, contracted forms); (c) stereotyped gay male language (specific words) and (d) stereotyped lesbian language (flat intonation, cursing) [4]. In other words, the mix of linguistic styles is the main reason why lesbian-specific language is less prominent than gay language.

3. Exploration of Gendered Features

In order to prove that previous studies on GenderNLP ignored homosexual language and language behavior should be categorized not only based on *sex* but also on *sexuality*, the present paper takes both heterosexual and homosexual linguistic features into consideration in the forthcoming tests. Since most of the studies on GenderNLP use both SVM and NB models to predict author's gender [13][14][15][16], this study will also adopt the same models under the 5-fold cross-validation test in predicting homosexual texts from Chinese social media.

This paper uses and translates all the gender-norm linguistic features from English to Chinese based on Huang, Li and Lin's study which detected author's gender with a number of linguistic cues [16]. However, features such as articles, capitalization, long/short words, abbreviation etc. which are absent in Chinese and statistical measures which do not fit our data are omitted. Also, Chinese-specific enumeration comma (、) is further added in the gender-norm feature list in our tests. It is worth noting that each type of punctuation will have two different forms due to Chinese text having no preference between using both full- and half-width punctuations on online social media.

Based on linguistic studies which discuss language features on homosexual texts [2][4][9][17] introduced in related work, eight convincing count-based homosexual-specific features are selected: (a) masculine words: words generally associated with masculine image; (b) feminine words: words generally associated with feminine image; (c) gay argot: a set of specialized lexicons used by gay community; (d) lesbian argot: a set of specialized lexicons used by lesbian community; (e) masculine pronouns: pronouns refer to male referent; (f) feminine pronouns: pronouns refer to female referent; (g) first person pronouns: pronouns refer to speaker or a community includes speaker as well as (h) swear words: a set of lexicons that is considered impolite or rude in mainstream society.

Table 2. Examples for types of homosexual-specific feature

Homosexual-specific features	**Example**	**Numbers of word in each lexicon**
Masculine word	*badao* 霸道 'domineering'; *wangzi* 王子 'prince'	122
Feminine word	*wenrou* 溫柔 'soft'; *gongzou* 公主 'princess'	148
Gay argot	*linghao* 零號 'bottom'; *yihao* 一號 'top'	99
Lesbian argot	*oulei* 歐蕾 'old lady'; *lala* 拉拉 'lesbian'	28
Masculine pronoun	*ni* 你 'you'; *ta* 他 'he'	4
Feminine pronoun	*ni* 妳 'you'; *ta* 她 'she'	4

First person pronoun	*wo* 我 ‘I’; *women* 我們 ‘we’	3
Swear word	*gaisi* 該死 ‘damn it’; *qu ni de* 去你的 ‘fuck off’	166

With the extraction of 24 sex-oriented gender-norm features from Huang, Li and Lin's study [16] together with 8 sexuality-oriented homosexual-specific features, a total of 32 gendered linguistic cues are included in our training procedures.

4. Training the Classifier

4.1 Framework

The LavenderNLP framework has five major components to automatically detect homosexual language from unstructured data from social media.

1. **Raw data**: In this study, experiments are conducted with a dataset containing 1433 homosexual male, 1481 homosexual female, 1476 heterosexual male and 1475 heterosexual female texts collected from the *gay*, *lesbian*, *mentalk* and *womentalk* boards on PTT[1]. Besides, in order for the collected data to be unbiased and informative, only long posts in specific topic associated to emotion venting are considered.
2. **Preprocessing**: Since stop words and punctuations are also regarded as important linguistic cues for various linguistic styles, only redundant information like web links and forum rules which appear in texts are removed during data preprocessing. Furthermore, in word segmentation, we apply *jieba* library with an additional user-defined dictionary containing all the words list in our selected heterosexual and homosexual feature sets.
3. **Annotation**: After data cleaning, the 32 types of gender-norm and homosexual-specific features are annotated automatically post by post. Considering that the annotated values may range from 0 to more than 5000, each value is normalized to a z-score so that all the computed results are treated equally across different features.
4. **Feature selection:** To test the hypothesis that previous GenderNLP studies are unable to perform expected results in detecting homosexual texts with gender-norm features and that homosexual languages do have their own unique linguistic styles, tests with three

[1] As the most popular online bulletin board and social media in Taiwan, PTT has more than one hundred and fifty thousand registrations. Due to its accessibility, PTT has been widely used in academic studies related to Chinese social media.

different feature sets are conducted: (a) gender-norm features; (b) homosexual-specific features; and (c) both gender-norm and homosexual-specific features.

5. **Classifier:** In recent years, studies on GenderNLP in identifying author's gender generally make use of both SVM and NB models [13][14][15][16]. Accordingly, this study follows the same route and reports the resulted F-scores average over the 5-fold cross-validation test. The main purpose of the current study is to automatically detect unstructured homosexual texts from Chinese social media with a number of investigated gender-specific linguistic features. Among the collected data, only homosexual male and homosexual female texts will be taken into account in the training procedure. Heterosexual male and heterosexual female texts, on the other hand, are used to evaluate how languages are produced differently by speakers with different types of *sex* and *sexuality*.

4.2 Feature Evaluation and Result

With the collected data from Chinese social media, tests with different feature sets and different machine learning models introduced in framework are conducted. This subsection discusses how homosexual-specific language are expressed and which feature set and model yield the best result in recognizing unstructured homosexual data from the viewpoints of linguistics and NLP, respectively.

Linguistically, there are tendencies for homosexuals and heterosexuals to use homosexual-specific features differently in text-making. Figure 1 demonstrates how such linguistic features are distributed in texts. The number in each bin denotes the average count of features per post.

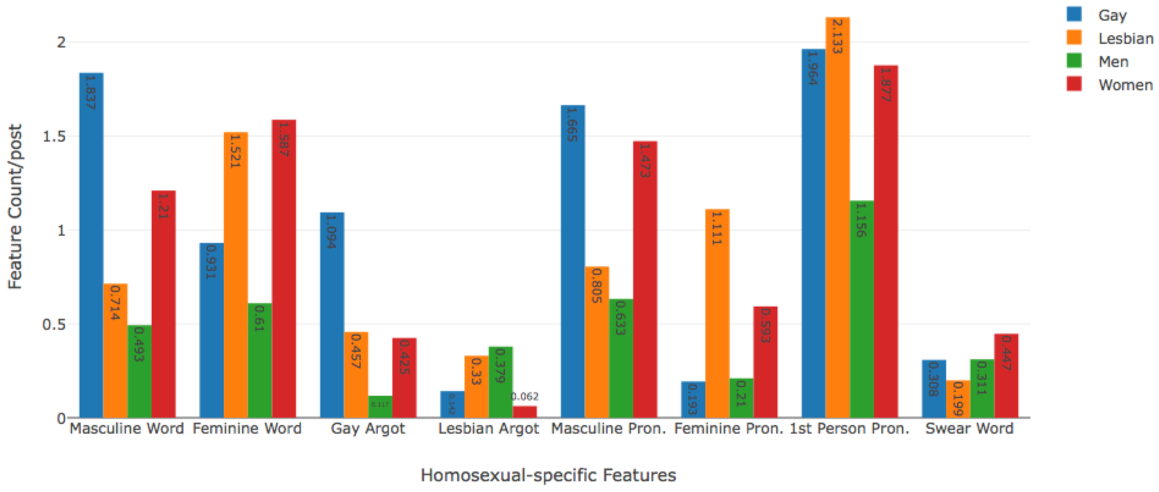

Figure1. Homosexual-specific Features in Gay, Lesbian, Men and Women Texts

With regard to the findings, several observations can be drawn below:

1. Gendered term: The saliency of masculine word counts in gay texts verifies the assumption that gay men tend to emphasize their masculinity with words associated with the stereotyped masculine male. Example (1a-e) are the five most frequently used masculine words in gay males' texts. Conversely, heterosexual males have no such needs to emphasize their masculinity through language behavior. For lesbians and heterosexual females, the use of feminine terms is similar.

   (1) *a. nanyou* 男友 'boyfriend'

   b. *nansheng* 男生 'male'

   c. *nanren* 男人 'man'

   d. *nanhai* 男孩 'boy'

   e. *nanpengyou* 男朋友 'boyfriend'

2. Homosexual argot: In terms of homosexual argot, it is clear that gay males have a strong preference in constructing texts with gay argots, see example (2-3). It is also interesting to note that the use of lesbian argot in heterosexual male language is relatively more than the use of lesbian argot in lesbian language. In addition, while gay-specific language is avoided by heterosexual males, lesbian-specific language is also avoided by heterosexual females in text-making.

   (2) 出櫃是一種同時具備堅持、面對和承受的行為。

chugui shi yizhong tongshi jubei jianchi miandui han chengshou de xingwei
'Coming out is a behavior of insisting, facing and bearing.'

(3) 從猴吃成熊，從熊瘦成猴。有人健身，有人節食。

cong hou chicheng xiong cong xiong shoucheng hou youren jianshen youren jieshi
'Some people work out in order to shape into bear from the monkey; some people go on diet in order to shape into monkey from the bear.'

(*xiong* 熊 'bear' means a hairy, hefty gay male; *hou* 猴 'monkey' means a skinny gay male)

3. Pronoun: Self-awareness is reflected by the use of self-referring statements which can lead to increased self-esteem and positive affect [18]. Compare to heterosexual male language, the wide use of first person pronoun in gay, lesbian and heterosexual female language indicates their refusal to be viewed negatively and to be accepted by society, especially, for lesbians.
4. Taboo: While the occurrence of swear words in both gay and heterosexual male's texts are about the same, it is quite different between lesbian and heterosexual female's texts. Obviously, heterosexual females swear more than lesbians on online social media, which conflicts with previous studies which claimed that lesbians are characterized by the use of cursing, taboo words, and progressive forms [2][4]. Example (4a-e) are the five most frequently used taboo words in heterosexual females texts.

(4) a. *gan* 幹 'fuck'

b. *kao* 靠 'damn'

c. *qiang* 嗆 'diss'

d. *biantai* 變態 'pervert'

e. *pishi* 屁事 'crap / (none of) one's business'

5. Others: Besides homosexual-specific features, there is an interesting finding opposite to the idea that homosexual language is marked by exaggerations [4][17]. Generally, exaggerations are expressed by means of punctuations such as single or multiple exclamation or multiple question marks; nevertheless, use of punctuations as such was not found in our annotated data. This may indicate that Chinese homosexuals are likely to hide emotions on social media and protect themselves from others.

When it comes to homosexual text recognition in LavenderNLP, the averaged 5-fold cross-validation f-score performances of SVM and NB models with different linguistic feature sets are shown in Table 3.

Table 3. Five-fold Cross-validation SVM and NB Averaged F-score Performances of Homosexual Texts Recognition with Sets of Linguistic Features

Gendered Feature Set	**SVM**	**NB**
Gender-norm Feature Set	57.11	33.18
Homosexual-specific Feature Set	69.57	66.49
Both Feature Set	74.54	58.75

Among performances with types of feature set, it is clear that in both SVM and NB models, the gender-norm feature set yields the lowest f-score. The low accuracy of gender-norm feature set verifies the hypothesis that previous research on GenderNLP ignores the homosexual group and implies that gender-norm linguistic features are not able to recognize homosexual texts as expected. Taking the resulting f-scores of gender-norm feature set as our baseline, Table 4 demonstrates the effectiveness of homosexual-specific features in identifying gay and lesbian texts.

Table 4. Effectiveness of Different Feature Sets in Identifying Homosexual Texts

Model	**Baseline**	**Best Result**	**Feature Set Taken**	**Improvement**
SVM	57.11	74.54	Both Feature Sets	+ 17.43
NB	33.18	66.49	Homosexual-specific Feature Set	+ 33.31

Based on the best results present in Table 4, one can see that the f-score in NB model is doubled with the use of homosexual-specific feature set alone. As for SVM, the f-score reaches up to 74.54% with both gender-norm and homosexual-specific feature sets. As shown in Figure 2, although the best result of SVM was produced by the use of both gender-norm and homosexual-specific feature sets, the homosexual-specific feature set still contributes more than gender-norm features to the resulted accuracy since it increases the accuracy 12.46% from the baseline while the gender-norm feature set, only 4.97%.

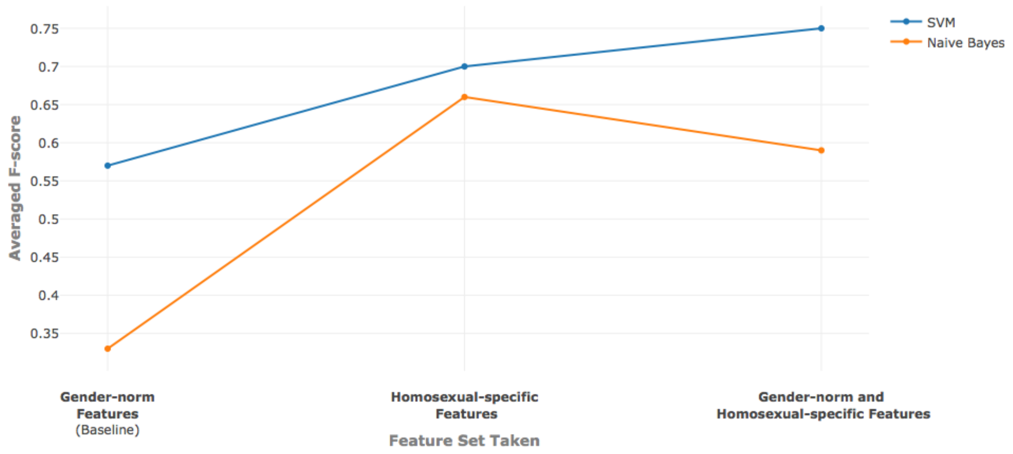

Figure2. SVM and NB F-score Comparison with Different Linguistic Feature Sets

The results reveal that gendered language can not only be divided into biological genders but also ones' sexual orientations. Apart from people with opposite-sex desires, gays and lesbians also have their own unique language styles. While heterosexual males and females are likely to produce languages with gender-norm features in order to meet social expectations, gays and lesbians' utterances are full of particular lexical items that have to do with their culture uniqueness and self-awareness.

5. Conclusion

Though previous studies on GenderNLP deal with gender from the biological perspective only, the present paper takes the psychological viewpoint into account as well. With the examination of linguistic behavior of homosexuals, it has been proved that traditional GenderNLP models are unable to detect gender in more complex dimensions. Also, with the adoption of homosexual-specific features, our NLP models resulted in promising accuracy in detecting unstructured homosexual texts automatically.

LavenderNLP has one important application in homosexual e-commerce. That is, while several online businesses are able to automatically recognize potential customers from biological genders, homosexual market is a segment that has often been ignored in the marketing strategies of businesses [19] and only few marketing departments pay attention to homosexual customers or do not even know how to find their potential customers. With its rapid growth, the homosexual community has attracted a great deal of attention and LavenderNLP should be able to keep up with the changes caused by this aforementioned growth.

As the very first work on Chinese LavenderNLP, there are more points to be considered under the lavender issue. For example, speakers of lavender language contain not only gay and lesbian, but also no-sex, bi-sex and transgender groups that further studies should also examine such linguistic behavior in order to enhance the field of LavenderNLP.

Although research on Chinese LavenderNLP lags far behind GenderNLP and is still at a nascent phase, it is believed that the fast-growing homosexual community is a sign that this issue will be regarded as important in the near future.

Reference

[1] W. Leap, "Beyond the lavender lexicon," Amsterdam: Gordon & Breach, 1995.

[2] R. M. Queen, "I don't speak spritch": Locating les- bian language," Queerly phrased: Language, gender, and sexuality, pp. 233–256, 1997.

[3] E. W. Burgess, "The sociologic theory of psychosexual behavior," Psychosexual Developments in Health and Disease, pp. 227–243, 1949.

[4] D. Kulick, "Gay and lesbian language" Annual Review of Anthropology, vol. 29, no. 1, pp. 243–285, 2000.

[5] R. A. Farrell, "The argot of the homosexual subcul- ture," Anthropological Linguistics, pp. 97–109, 1972.

[6] J. P. Stanley, "Homosexual slang," American speech, vol. 45, no. 1/2, pp. 45–59, 1970.

[7] P. Baker, "What can I do with a naked corpus?" Public Discourses of Gay Men, pp. 1–37, 2005.

[8] C. M. Nash, "Review of: Public discourses of gay men," Gender and Language, vol. 3, pp. 279–282, 2010.

[9] P. Baker, "'no effeminates please': Discourses on gay men's personal adverts," Public Discourses of Gay Men, pp. 131–153, 2005.

[10] G. Legman, "The language of homosexuality: an American glossary," Sex variants: A study of homosexual patterns, vol. 2, pp. 1149–1179, 1941.

[11] B. Moonwomon, "Lesbian discourse, lesbian knowledge," Beyond the Lavender Lexicon, pp. 45–64, 1995.

[12] D. S. Painter, "Recognition among lesbians in straight settings," Gayspeak: gay male & lesbian communication, pp. 68–79, 1981.

[13] M. Vicente, F. Batista, and J. P. Carvalho, "Twitter gender classification using user unstructured information," in 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, pp. 1–7, 2015.

[14] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating gender on twitter,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1301–1309, 2011.

[15] C. Zhang and P. Zhang, “Predicting gender from blog posts,” University of Massachussetts Amherst, USA, 2010.

[16] F. Huang, C. Li, and L. Lin, “Identifying gender of microblog users based on message mining,” International Conference on Web-Age Information Management, pp. 488–493, 2014.

[17] E. P. Johnson, “Mother knows best: Black gay vernacular and transgressive domestic space,” Speaking in queer tongues: Globalization and gay language, pp. 251–278, 2004.

[18] D. Davis and T. C. Brock, “Use of first person pronouns as a function of increased objective self- awareness and performance feedback,” Journal of Experimental Social Psychology, vol. 11, no. 4, pp. 381– 388, 1975.

[19] DeLozier, Dr. M. Wayne, and Jason Rodrigue. "Marketing to the homosexual (gay) market: A profile and strategy implications." Journal of homosexuality 31.1-2 (1996): 203-212.