ORIGINAL ARTICLE

Leveraging writing systems changes for deep learning based Chinese affective analysis

Rong Xiang[1] · Qin Lu[1] · Ying Jiao[1] · Yufei Zheng[1] · Wenhao Ying[1] · Yunfei Long[2,3,4]

Abstract
Affective analysis of social media text is in great demand. Online text written in Chinese communities often contains mixed scripts including major text written in Chinese, an ideograph-based writing system, and minor text using Latin letters, an alphabet-based writing system. This phenomenon is referred to as writing systems changes (WSCs). Past studies have shown that WSCs often reflect unfiltered immediate affections. However, the use of WSCs poses more challenges in Natural Language Processing tasks because WSCs can break the syntax of the major text. In this work, we present our work to use WSCs as an effective feature in a hybrid deep learning model with attention network. The WSCs scripts are first identified by their encoding range. Then, the document representation of the text is learned through a Long Short-Term Memory model and the minor text is learned by a separate Convolution Neural Network model. To further highlight the WSCs components, an attention mechanism is adopted to re-weight the feature vector before the classification layer. Experiments show that the proposed hybrid deep learning method which better incorporates WSCs features can further improve performance compared to the state-of-the-art classification models. The experimental result indicates that WSCs can serve as effective information in affective analysis of the social media text.

Keywords Affective analysis · Writing system changes · Deep learning network

✉ Yunfei Long
Yunfei.Long@nottingham.ac.uk

Rong Xiang
xiangrong0302@gmail.com

Qin Lu
csluqin@comp.polyu.edu.hk

Ying Jiao
naomi.jiao@connect.polyu.hk

Yufei Zheng
yufei.zheng@connect.polyu.hk

Wenhao Ying
yingwh123@sina.com

1 Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

2 Horizon Digital Economy Research, University of Nottingham, Nottingham, UK

3 University of Nottingham Innovation Park, Triumph Road, Nottingham, UK

4 Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China

1 Introduction

In social media, text is becoming increasingly important due to its effectiveness in disseminating information in highly individualized and opinionated context. Affective analysis has been studied using different Natural Language Processing (NLP) methods from a variety of linguistic perspectives such as semantic, syntactic, and cognitive properties [1–4]. In certain parts of the world, such as Mainland China and Hong Kong, social media text is often written in mixed scripts.

Below are three examples of text written in mixed scripts.

E1: 这个年每天都吃好饱！初三来点小朋友的最爱 麦当劳和pizza！！
(We are so full in every meal during this Spring Festival! We will take kids to their favorite place, MacDonald and pizza!)

E2: 我们会因为金希澈开了微博而看到许多无下限的nc发言。
(We will see a lot of stupid comments with no lower bound once Jin Xicheu opens her Weibo.)(*'nc' is short for Pinyin 'naocan', means stupid or retard.*)

E3: 真是速度啊，收到快递了。｡:P! *(It is so fast, I got the parcel already, happy!)*

From the above examples we can see that the major text is in Chinese, an ideograph-based writing system. The minor text can be written in English (as shown in E1), Pinyin[1] (phonetic denotation for Chinese) (as shown in E2 in short form), or other new Internet notations with Roman characters using some Latin-based writing system as well as other symbolic expressions, e.g. emoji symbols as shown in E3. This phenomenon of using mixed scripts in different writing systems is known as Writing Systems Changes (WSCs).

Previous work on lexicon based affective analysis primarily relies on syntactic information or semantic orientation to improve affective classification tasks since the positive/negative values of a lexicon contribute to the orientation of affections encoded in sentences [5, 6]. Syntactic and semantic knowledge is often used to transform raw data into feature vectors [7]. In social media text, however, WSCs can break the syntax of the major text and the switched minor text also lacks sufficient syntactic and semantic cues [8]. This makes syntax and semantics based methods difficult to work. Moreover, neologism in Internet forums increases the difficulty for both syntactic and semantic analysis. In particular, newly coined phrases tend to contain different types of symbols. Despite the additional challenges in affective analysis for social media datasets, this type of datasets is rich in shifts of writing systems orthography. The alternation between different writing systems is relatively common in real-time platforms like micro-blog in China. This feature offers reliable clues for affective analysis.

1.1 Definition of WSCs

The term Writing System Change (WSC) refers to the switching between two or more writing systems in a context where phonology, syntax, and semantics are coherent [9]. A narrower definition, often referred to as code-switching, is the use of more than one linguistic variety in a manner consistent with the syntax and phonology of each variety[2]. Using WSCs typically has socio-linguistic motivations such as identity and social position and it is also a way of organizing speech in spontaneous interaction [10, 11]. WSCs are also considered a strategy to create social interaction [12, 13]. Online social media forums for videos, news, and films often have intense and spontaneous social interactions. The use of WSCs is generally considered a common phenomenon. In some bilingual societies, it is customary to use WSCs. In more conservative communities such as Mainland

China, WSCs are even more used to express emotions that are easier to express in a different writing system to avoid the Internet inspection.

This study adopts a broad definition of WSCs, including switching between two languages, using punctuation markers to generate stickers such as a smiley face, and changing writing systems within the same language. In online platforms in China, the alternation between writing systems is quite common than in oral conversations. The cognitive and socio-linguistic motivations make WSCs a potentially effective predictor of affective analysis.

1.2 Types of WSCs in Chinese

The customary use of different writing systems or languages symbols is rooted in pragmatic and socio-linguistic motivations [11, 13]. The use of WSCs is considered a case of economy principle in language [14] which is pursued by people in various activities due to the innate indolence. It aims at the maximum effect with the least input. For instance, in Chinese social media, 'Good luck' becomes more popular than the corresponding Chinese version of 祝你好运 because inputting the English version takes much shorter time in expressing the same affection.

Studies in social psychology [15, 16] also show that WSCs are an effective and commonly used strategy to express affection or mark affective change especially in the society where social environment is more conservative [17]. Words of profanity, swearing, and cursing, which may not seem to be socially acceptable in its native form in Chinese communities, can come up in the text in their disguised form using different writing system counterparts in a different language. For example, a new-born swear word 'zz' is often used in place of the Chinese version of "moron". This is because 'zz', which is the acronym of the Pinyin 'zhi zhang (moron)', is less eye catching, and thus looks less disrespectful and relatively more acceptable in social media. With the rapid growth of globalization, Chinese youngsters also like to use English acronyms such as 'wtf' (what the fuck) 'stfu' (shut the fucking up), etc. Naturally, these negative comments using profanity appear frequently as well. Swearing words also occur with anger, passion, or some strong affection; yet they may be used in a protected way through WSCs since they are taboos [18]. People also choose WSCs to express idiosyncrasies using either English or some other languages for minor text because some popular words in other writing systems may not have appropriate short translations. Thus, writing words in their native form can make the comment distinctive. For example, the phrase 'hard core' '硬核' became very popular in Chinese communities to describe a dedicated person or a movement.

WSCs in this study are not limited to switching between different languages. Generally speaking, our study includes

---

[1] https://en.wikipedia.org/wiki/Pinyin.

[2] https://en.wikipedia.org/wiki/Code-switching.

the more liberal sense of writing systems changes, which can be between different writing systems of the same language. For Chinese, this means switching between the Chinese characters (logo-graphic systems) and alphabetic systems, such as Pinyin or acronyms written through Latin alphabet. Users in the social media are quite creative in employing such WSCs for euphemism and for other rhetoric effects. Abbreviated Pinyin alphabet sequences are often used for profanity, such as swearing and curse words. For instance, frequently used WSCs terms include 'tm', an abbreviated of Pinyin 他妈 as a common profanity by cursing one's mother, and *nc*, abbreviated from 脑残 to mean ('brain-damaged' or 'moron'). Similar to the use of alphabetical writing for profanity, another typical type of WSCs is also due to euphemism mainly to avoid directly confronting social norms or expectations. A very interesting example involves interspersing of character and Pinyin text with opposite meanings. For example, in the text 丧*(gan)*心*(de)*病*(piao)*狂*(liang)*, the Chinese part 丧心病狂 is an idiom and is extremely negative, but the interspersed Pinyin *gan de piao liang* actually stands for "well done (干得漂亮). Thus, even though the Chinese text was against certain action, the writer was in fact supporting the action she/he commented on. Another common scenario is to use Pinyin to replace sociologically or politically sensitive terms, partly to avoid getting attention or the risk of being targeted. For example, the Chinese term for 'government' (政府 with Pinyin 'zheng fu") may occur in the form of '*zf*' to avoid internet surveillance. It is important to note that no regular rules can be applied for these substitutes using WSCs, as one of the main purposes is to escape from detection.

Furthermore, there are other diversely types of WSCs in Chinese social media such as expressing named entities using full English names, abbreviations, or Pinyin abbreviations. These types of WSCs are generally not collected for affective analysis. For example, 'CBD' is quite often used in contemporary commercial conversations and it is a WSC use for efficiency. In online shopping comments and catering comments, some translations are used to indicate the product and service, barely relating to emotion expressions.

1.3 Our approach

This work studies WSCs related textual features at the orthography level to explore their effectiveness as affective indicators in social media and review text. In this work, we propose a Hybrid Neural Network with Attention Network (HAN-WSC), a novel deep learning based method to incorporate textual features associated with WSCs via an attention mechanism. More specifically, the proposed HAN-WSC first identifies all WSCs points. Representation of the major text is learned through a Long-Short Term Memory (LSTM) model whereas the presentation of the minority text

is learned by a separate Convolution Neural Network (CNN). Affection expressed in both major and minor text is further highlighted through an attention mechanism before affective classification. In HAN-WSC, the whole text, which is generally coherent both syntactically and semantically, is learned through an LSTM network at the sentence level. The minor text, containing both Chinese Pinyin and other types of WSCs, is extracted out from the main text first and then processed by a CNN network to learn their representation vector. The attention mechanism is achieved by projecting the major text representation into attention vectors aggregated with the representation of informative tokens from WSCs context.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 describes the HAN-WSC model. Section 4 gives performance evaluation. Section 5 concludes this paper with future directions.

2 Related work

Text in casual genres may adopt different combination of writing systems. In Chinese speaking communities, semantics of the Chinese writing system is encoded as ideograph character sequences [19] and phonology-based Pinyin system. Modern Chinese also adopts the Pinyin system as a supplementary phonetic system to denote pronunciations for natives as a mandatory part of first language learning as well others as second language learning. Pinyin[3] is a phonology-based system, similar to those Latin-based languages such as English. Pinyin system also provides the most effective method for inputting Chinese characters into computer systems. In social media, WSCs can serve as an emphasis which effectively indicates the delivery of a particular type of affection [15, 16]. The orthography linked to WSCs in Chinese text can be motivated by socio-linguistic factors.

WSCs have been recognized to be relevant to affections [20, 21]. These statistical studies show that WSCs frequently occur in social media. As a typical type of WSCs, code-switching documents have received considerable attention in the NLP community. Several studies have focused on WSCs identification and analysis, including mining translations in WSCs documents [22], predicting WSCs points [23], identifying WSCs text [24], language modeling [25], and Part-of-Speech tagging [26]. In affective analysis, WSCs in text are less studied. Li et al. [27] proposed a machine translation based approach to predict affection in WSCs text with various external resources.

Affective analysis, either aims to identify sentiment as binary classification problem or affection as multiple labeled emotion identification, is approached based on

---

contextual information because they offer assessment of the affective value of a phrase for automatic classification [28]. Semantic orientation has been employed to estimate the positive or negative orientation of a phrase based on its association with positive or negative evaluations [28, 29].

Early tasks in affective analysis are based on lexical rules. Hatzivassiloglou et al. [30] proposed an affective analysis task explicitly based on adjectives for English with available linguistic resources. The proposed linguistic rules based on 21 million words of English. Rule based methods are simple but lack generalization ability. Later studies in affective analysis are based on linear classifier with feature engineering. The Support Vector Machine (SVM) classifier has achieved great success in text classification [31, 32]. SVM, if used with effective feature engineering, was considered as a commonly used affective classification method before the drastic performance improvement by deep learning methods. In recent years, deep learning based methods have greatly improved the performance of affective analysis. Commonly used models include Convolutional Neural Network (CNN) [33], Recursive Neural Network (ReNN) [34], and Recurrent Neural Networks (RNN) [35]. RNN naturally benefits affective classification because of its ability to capture sequential information when processing text. However, standard RNN suffers from the so-called gradient vanishing problem [36] where gradients may grow or decay exponentially over long sequences. To address this problem, Long-Short Term Memory (LSTM) model is introduced by adding a gated mechanism to keep long term memory [37]. Each LSTM layer is generally followed by mean pooling and then fed into the next layer. Experiments in datasets which contain long sentences and long documents demonstrate that the LSTM model outperforms the traditional RNN [38–40]. Attention mechanisms are also proposed to highlight the difference in contribution of different words [41]. Attention layer can be built from local context, or external knowledge from cognitive science [42, 43]. Wang et al. [44] proposed a Bilingual Attention Network (BAN) model to aggregate the monolingual and bilingual informative words to form vectors from document representation and integrated the attention vectors in affective prediction. However, previous work suffered from two main problems. Firstly, WSCs are defined as switching of text between two languages. However, WSCs in Chinese communities essentially are mainly in the change of writing systems. In other words, the WSCs can co-occur with or without switching to a different language, such as in the switching between Chinese characters and Chinese Pinyin. The characteristics for such switching are different from code switching between languages. Secondly, dispersion of alphabetic writing text is quite unique in the Chinese social media text which requires new methods to handle.

Attention based neural networks are proposed to highlight the difference of words in contribution to semantic expressions [41]. Attention mechanism is introduced because not all words contribute equally to the meaning of a sentence: some are more informative, and others are more functional [45]. In document classification, both sentence level attention and document level attention are proposed. In sentence level attention layer, an attention mechanism identifies informative words that are important in each sentence. Those informative words are aggregated as attention weights to form sentence embedding representation. This method is generally called local context based attention method. Similarly, some informative sentences can also be highlighted to indicate their importance in a document. It is also proven that eye-tracking data can be used as weights of words in attention mechanism, showing the weighted model could further improve attention-based neural networks [42, 43].

3 Hybrid neural model with attention network

In this paper, we propose a Hybrid Neural Network with Attention Network. (HAN-WSC) to incorporate the implicit information expressed by WSCs text. HAN-WSC is a deep learning based method to combine the use of both LSTM and CNN with attention mechanism to better capture different textual features associated with WSCs.

The whole text, mainly written in the Chinese ideographic writing system, provides descriptive information. Therefore, it is reasonable to use LSTM as the learning model as the semantic information from text is rather coherent and complete. For the minor text written in alphabet or other writing systems, which occurs more as isolated instances, an additional layer can be provided such that their features can be captured. CNN is more suited to extract information from minor text written in other writing systems.

3.1 Task definition

Let D be a collection of documents for affective classification. Each document d_i is an instance in D, $d_i \in D$. In multi-class based sentiment analysis, the sentiment to be labeled to each document in D is often a numerical value to indicate both polarity and strength. In multi-label based emotion analysis, the goal is to predict whether a certain type of emotion is expressed for each d_i. The set of emotion labels in multi-label emotion analysis usually contains a set of emotion labels. The most popular ones include {*Happiness, Sadness, Anger, Fear, Surprise*}. In deep learning based models, each document d_i is first tokenized. The embedding vector of each token is denoted as $\overrightarrow{w_i}$. In our work, every WSC token in d_i is also identified and the embedding vector of each WSC is denoted as $\overrightarrow{w_j^s}$.
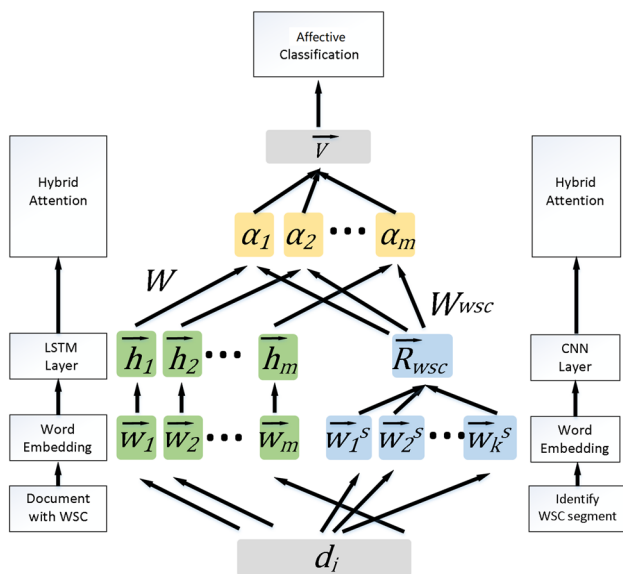
Fig. 1 Hybrid deep learning model with Attention Network framework

3.2 WSC identification

We use the term WSCs segments to refer to the minor WSCs text pieces. Note that in computer systems, Chinese characters and other scripts such as Romanized Pinyin or English text are coded in different code ranges. Thus, WSCs segments can be easily marked in a pre-processing step by separating text through their internal code ranges (Unicode). According to Unicode standard, any token with all characters encoded between [0x4E00, 0x9FA5] shall be identified as Chinese characters. Along with punctuation set, their union is then regarded as the major text. All remaining tokens are referred to as minor text.

3.3 The hybrid neural network structure

To make better use of WSCs scripts, our system explicitly assembles WSCs information separately from the original text learning, and then an attention layer is also utilized to WSCs segments. Figure 1 shows the framework of HAN-WSC. This model contains four components after pre-processing, (1) LSTM for both Chinese and WSCs text, (2) CNN for WSCs text, (3) combined attention layer, and (4) output layer for classification. More specifically, the complete text is learned though the LSTM model on the left side, marked in green in Figure 1, to generate the representation of a document including the WSCs segments. This is because documents with included WSCs are generally syntactically coherent and intact despite the fact that few WSCs segments may break the semantics of the main writing system. The CNN model on the right, marked in blue, is used to learn the representation of WSCs

segments extracted from the sentence because they often occur discontinuously without syntactic structure. It should be noted that CNN models do not take n-grams as a sequence. Rather, it is learned as a bag-of-words without consideration of sequences which is quite reasonable for WSCs. The outputs of both models are then integrated into one unified attention layer before classification is carried out.

Using deep learning methods, token representation in $d_i = \vec{w}_1, \ldots, \vec{w}_m$, is learned using two networks. d_i is fed into an LSTM to generate the hidden vector $\vec{h}_1, \vec{h}_2 \ldots \vec{h}_m$ from d_i.

In Chinese social media, WSCs segments are generally dispersed sporadically. To distinguish the WSCs units, each WSCs token is also extracted to form a designated vector $\vec{w}_j^s$ ($\vec{w}_j^s \subset d_i, j = 1 \ldots k$). These WSCs vectors are then fed into another CNN to separately learn their representations. So, for d_i with k WSCs segments, the convolution is calculated using a sliding window of size $2n + 1$:

$$\overrightarrow{conv_p} = \sum_{j=p-n}^{p+n} \vec{w}_j^s, \tag{1}$$

and

$$\overrightarrow{R_{WSC}} = \frac{\sum_{p=1}^{k} \overrightarrow{conv_p}}{k}. \tag{2}$$

The WSCs feature vector $\overrightarrow{R_{WSC}}$ is generated by average pooling. Attention model was used in affective analysis by Yang et al. [41] to show different contributions of different tokens semantically. For a token w_p, to include both the information learned from LSTM and CNN, the consolidated representation, $\vec{u}_p$, includes both $\vec{h}_p$, and $\overrightarrow{R_{WSC}}$ into a perceptron defined below:

$$\vec{u}_p = \tan h(W\vec{h}_p + W_{WSC}\overrightarrow{R_{WSC}} + b). \tag{3}$$

In order to evaluate the significance of each token $\vec{w}_p$, a coefficient vector $\vec{U}$ is introduced as an informative representation of the words in a network memory. The representation of a token $\vec{u}_p$ and the corresponding token-level context vector $\vec{U}$ is integrated with dot product to obtain a normalized attention weight:

$$\alpha_p = \frac{\exp(\vec{U} \cdot \vec{u}_p)}{\sum_p \exp(\vec{U} \cdot \vec{u}_p)}. \tag{4}$$

The updated document representation $\vec{v}$ can be generated as a weighted sum of the token vectors given below:

$$\vec{v} = \sum_p (\alpha_p \vec{h}_p), \tag{5}$$

where $\vec{v}$ contains both document information and WSCs representation with attention weights in the final Softmax function, producing the output vector. Lastly, an Argmax classifier is used to predict the class label C_i of $i-th$ instance.

3.4 Objective functions

For each instance, let y_i denote the ground truth class, let p_i denote the predict value and T indicates the total number of instance. Affective analysis commonly includes either sentiment analysis or emotion analysis. Considering the characteristics of evaluated dataset, which shall be elaborated in 4.1, different loss function is used.

For the multi-class based sentiment analysis task, the labels are numerical and thus, root-mean-square error (RMSE) is used to measure the distance between y_i and p_i as shown below by the loss function L, where y_i and p_i are the predicted value and true value, respectively.

$$L = \sqrt{\frac{\sum (y_i - p_i) * (y_i - p_i)}{T}} \tag{6}$$

For the multi-label based emotion analysis task, we use cross entropy for each emotion type C.

$$L_C = - \sum_{l(y_i) \in C} y_i ln(p_i) \tag{7}$$

Note that for each class C, the emotion label of y_i, denoted by $l(y_i)$, must be in C.

4 Performance evaluation

Two Chinese social media datasets are used for evaluation. The first dataset is collected from the food critics website OpenRice by this project for sentiment classification. The second set is a publicly available Chinese micro-blog. A number of variants of HAN-WSC are implemented and evaluated to show the contributions of different languages resources. A few crucial parameters in HAN-WSC are investigated, aiming to obtain a better understanding of this model. Lastly, visualization and case studies are presented for intuitional and quantitative study. This can also set direction for further improvement of HAN-WSC.

4.1 Datasets

Recently, identifying people's attitude towards food comments has become a famous research topic in sentiment analysis. We provide a new dataset collected from Openrice[4].

---

[4] https://www.openrice.com.

Table 1 Dataset information of openrice

Class	Proportion (%)
1	0.4
2	2.6
3	16.8
4	59.5
5	20.7

The instances in this dataset are mainly written in Cantonese Chinese. In this dataset, WSCs segments, which are mainly written in English, are widely in comments. For example, it is of less necessity to its Chinese translation for some food names such as Fettuccine. Some phrases like 'very family taste' and 'lovelp' can be directly considered as the evidence of affective analysis. There are 14,608 instances and corresponding sentiment labels. These labels are rated from 1-star to 5-star. Thus, this task is regarded as multi-class problem. The instances are mostly long paragraphs. On average, the sentence number is 13.2 and each sentence contains 74.5 characters. Detailed information is indicated in Table 1. Using the stratified stochastic sampling, 90% of them are used as training set and the remaining ones are regarded as testing set. Based on the tokenized instances, the proportions of English, pinyin and other types of WSCs take 5.6%, 1.2% and 16.8%, respectively. In Openrice, there are abundant WSCs of other types, including French, Japanese, emoji, symbolic expression, etc. This is not surprising as Openrice is a website used by food critics, and French and Japanese tokens are often directly used in Hong Kong.

A publicly available and widely used dataset containing WSCs for emotion analysis is collected from Chinese micro-blog [20]. Every instance is written in Mandarin Chinese with at least one WSCs segment. The 8,728 instances in the collection are evenly divided into the training set and the testing set. Providing with segmentation in pre-processing, each instance is a micro-blog massage with short sentences. The average length of an instance is 46.8 tokens which can be a character or phrase. The longest document contains 119 tokens whereas the shortest document only contains 4 tokens. Each sentence is annotated with the class of whether it contains one or more emotion types including *happiness, sadness, anger, fear, surprise*. Therefore, this dataset can be regarded as a multi-label problem. Separate annotations are also given to indicate whether the Chinese, WSCs or both scripts contribute to the emotion. More details can be found in Table 2. Based on tokenized instances, English, pinyin and other types of WSCs take proportions of 4.3%, 1.6% and 5.3%, respectively.

4.2 Baseline systems and performance measures

A set of experiments is conducted to evaluate the performance of affective prediction. The following gives the list

Table 2 Dataset information of microblog

Emotion	Percentage in corpus (%)
Happy	30.0
Sad	17.8
Anger	10.2
Fear	10.8
Surprise	11.3

Table 3 Comparison with baselines in Openrice

	Acc	RMSE
SVM	0.587	NA
CNN	0.643	0.401
LSTM	0.654	0.362
BAN	<u>0.662</u>	<u>0.329</u>
HAN-WSC	**0.672**	**0.308**

Best result in accuracy is marked bold; second best is underlined

of baseline models to be compared to our proposed HAN-WSC algorithm.

- *SVM* is the basic model that uses features of all the Chinese and English words. We use the mean of token vectors to generate the document representation.
- *CNN* uses a convolution layer to capture feature of adjacent tokens. Then affection is classified with a perceptron.
- *LSTM* uses mixed WSCs text as the input to train a basic LSTM model. This serves as a neural network baseline without separate processes for WSCs.
- *BAN* uses LSTM with attention mechanism to capture informative words from both monolingual and bilingual context [44]. BAN is the current state-of-the-art algorithm.
- *HAN-WSC* is our proposed model, which feeds supplemental minor WSCs texts to an attention layer.

Since the first dataset Openrice is annotated with sentiments, the performance of sentiment analysis is measured by accuracy and RMSE. To calculate accuracy, we use the following notations: TP = True positive; FP = False positive; TN = True negative; FN = False negative. For RMSE, y = ground true label; p = predicted label; T = total number of instances in the testing set test instance number. Accuracy and RMSE are then computed using the following formulas.

$$accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{8}$$

$$RMSE = \sqrt{\frac{\sum (y - p) * (y - p)}{T}} \tag{9}$$

For the multi-labeled Chinese blog dataset, F1 score is used as the performance measurement. Since the proportion of the five emotion types are imbalanced, both the average F1-score and weighted F1-score are provided[5]. In Chinese Microblog dataset, the relative weights W_i for the five classes are 26%, 16%, 9%, 9%, and 11%, respectively.

$$F_{1avg} = \frac{\sum F_{1i}}{5} \tag{10}$$

---
5 https://en.wikipedia.org/wiki/F1_score.

$$F_{1wgt} = \frac{\sum_i F_{1i} * W_i}{\sum W_i} \tag{11}$$

4.3 Affective analysis

The experiments on affective analysis are set up to evaluate the performance of both sentiment analysis on Openrice and emotion analysis on Chinese micro-blog. SVM, CNN, LSTM, and BAN are used as baselines. BAN implemented by Wang [44] is implemented and tuned as the main comparison.

The result on Openrice is given in Table 3. Our proposed HAN-WSC has the best performance compared to all the baseline models including the state-of-the-art BAN. Considering that 4-star rating comments account for 59.5%, the best performance by HAN-WSC only reaches 0.672. The relative low accuracy by all five methods shows prediction based on the Openrice dataset is very challenging. In fact, the performance of SVM is even worse than the proportion of the 4-star group, showing that token level approach is not effective. Obviously, the mean of token embedding fails to provide useful information for long paragraphs. CNN gives about 5% boost compared to the 4-star ratings. The convolution of adjacent tokens is more informative in affective analysis. Among the three deep learning algorithms, the performance of BAN using LSTM with attention mechanism is better than LSTM, showing the effectiveness of attention mechanism. Since our proposed HAN-WSC is also based on LSTM with attention mechanism, the additional gain in performance is attributed to learning features of WSCs in a separate CNN.

In the task of emotion classification using the Chinese Microblog dataset, we follow the 50–50% ratio for splitting training and testing for fair comparison with Wang's work [44]. From Table 4, we can see that the performance of SVM ranks the lowest since it lacks phrase level analytic capability. Although using token embedding tricks can improve the performance of vector-based modelling, each token in SVM is only considered independently. Unlike sequence based deep learning models, insufficient information can

Table 4 Comparison with baselines in Chinese Blog

	Hap	Sad	Anger	Fear	Surprise	Avg. F1	Wgt. F1
SVM	0.693	0.560	0.640	0.549	0.593	0.607	0.623
CNN	0.675	0.618	0.671	0.596	0.603	0.633	0.641
LSTM	0.717	0.642	0.704	0.606	<u>0.628</u>	0.659	0.671
BAN	<u>0.724</u>	<u>0.649</u>	<u>0.712</u>	**0.627**	<u>0.628</u>	<u>0.668</u>	<u>0.678</u>
HAN-WSC	**0.729**	**0.658**	**0.729**	<u>0.625</u>	**0.641**	**0.676**	**0.688**

Best result is marked bold; second best is underlined (*Avg.* short for average, *Wgt.* short for weighted

Table 5 Performance using single writing system

Network	Text Set	Hap	Sad	Anger	Fear	Surprise	Avg. F1	Wgt. F1
LSTM	WSCs	0.631	0.546	0.682	0.589	0.529	0.595	0.598
LSTM	CN	0.695	0.632	0.671	0.612	0.615	0.645	0.656
LSTM	CN+WSCs	0.717	0.642	0.704	0.606	0.628	0.659	0.671
BAN	WSCs	0.631	0.551	0.681	0.589	0.529	0.596	0.599
BAN	CN	0.698	0.626	0.669	0.613	<u>0.631</u>	0.647	0.658
BAN	CN+WSCs	<u>0.724</u>	<u>0.649</u>	<u>0.712</u>	**0.627**	0.628	<u>0.668</u>	<u>0.678</u>
HAN-WSC	CN + WSCs; WSCs	**0.729**	**0.658**	**0.729**	<u>0.625</u>	**0.641**	**0.676**	**0.688**

Best result is marked bold; second best is underlined

be learned in SVM. The improved performance by CNN in both measures shows that introducing phrase level features by a convolution layer can improve the overall classification performance. However, the F1 scores of CNN are noticeably smaller than LSTM, indicating that the gated memory mechanism is effective when learning information in text which is sequentially coded. The 3.0% gain in the micro F1 shows that the order of tokens should not be neglected in emotion analysis. The attention mechanism used in BAN makes a 0.7% improvement to LSTM in micro F1. Our proposed HAN-WSC shows a comprehensive improvement compared to BAN. Since we model WSCs as individual information, they are learned by a separate CNN network. An additional CNN shall not introduce too much computational complexity and yet the result shows that the attention-based LSTM model by BAN can be further improved by about 1.0% on micro F1 by integrating WSCs representation.

4.4 Writing system investigation

When handling text with mixed writing systems, previous tasks translate the text of the minor writing system. After pre-processing, the syntax of the sentences can then be reconstructed. This method may work for traditional mixed language with code switch. However, they would not work in social media text as many of the WSCs are not proper tokens of any language. They can be short hands and transformed representations. The implicit intention and emotion of using such WSCs is not common in traditional text with code switches.

To further investigate the impact of WSCs in social media, Another set of experiments is conducted to observe the effect of WSCs. We divide text in the micro blog dataset into three categories:
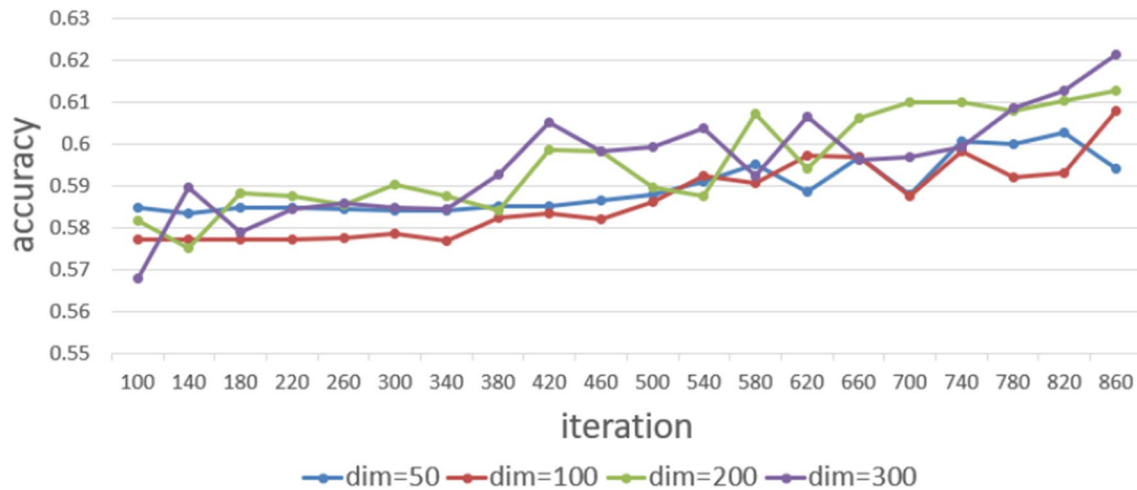
- *CN* refers to all the Chinese text with all the WSCs removed;
- *WSCs* refers to the WSC tokens that can either be in English, Pinyin or other types of WSCs;
- *CN+WSCs* refers to the complete text including both Chinese and WSCs.

Table 5 shows the performance of LSTM, BAN, and HAN-WSC by using different types of data in the dataset. The data used as input to the models in Table 5 is noted in the parenthesis. Since HAN-WSC has two inputs, one to LSTM and the other to CNN, input in CNN follows the semicolon in the parenthesis. It is shown that Chinese text carries more emotional information than WSCs text. The use of both Chinese text and WCSs has the best performance, showing that WSCs is also contributing to the information delivery in sentences. However, using the complete text with WSCs without distinction does not highlight the importance of WSCs for emotion analysis. That is why the F1 score of most emotion types in HAN-WSC is considerably better than BAN (CN + WSCs).

It is interesting to note that BAN (CN) using only Chinese text has a comparable result to BAN (CN+WSCs) with only a slight performance loss of 0.3% in the emotion of *surprise* even though WSCs are not used. This means

Table 6 Performance by multiple writing systems; best result in accuracy is marked bold; second best is underlined

Network	Text Set	Set	Sad	Anger	Fear	Surprise	Avg. F1	Wgt. F1
HAN-WSC	CN; WSCs	0.629	0.613	0.682	0.588	0.531	0.606	0.613
HAN-WSC	CN+WSCs; CN	<u>0.720</u>	<u>0.646</u>	<u>0.698</u>	<u>0.624</u>	<u>0.616</u>	<u>0.661</u>	<u>0.673</u>
HAN-WSC	CN+WSCs; WSCs	**0.729**	**0.658**	**0.729**	**0.625**	**0.641**	**0.676**	**0.688**

Fig. 2 Dimension comparison in the conditions: learning rate = 0.001; dropout keep rate = 0.9; convolutional window size = 3

that BAN is not making good use of WSCs contained in the text.

Table 6 shows a more detailed performance analysis of HAN-WSC with different data as input to the hybrid model. The first two experiments show that the input pair (CN, WSCs) is better suited for our model than that of (WSCs, CN). This is because CN basically maintained the syntactic and semantic sequence which is better suited for LSTM. WSCs, on the other hand, are in sporadic use, and their information is better learned using CNN. In the last two experiments, LSTM with complete text can catch more emotional information. Using only Chinese text for attention information, HAN-WSC (CN+WSCs; CN), will make the result 1.5% worse than that of HAN-WSC (CN+WSCs; WSCs). This gap could be caused by integrating too much information from Chinese characters which contain some irrelevant tokens.

4.5 Parameter tuning

In this section, we use the Openrice dataset to show how the parameters of HAN-WSC are tuned. Three main parameters include token embedding dimension, dropout rate and window size of CNN. HAN-WSC is trained using different random seeds a number of times. To show fair comparison of different settings, these experiments are conducted with same training data batches size and content. The first 100 iterations are trained as the warm-up phrase.

In NLP, the choice of embedding dimension often depends on the scale of the problem under consideration. Since Openrice is a domain specific dataset on food reviews, its vocabulary size is usually limited, yet the lengths of paragraphs can be rather long. To find the appropriate dimension, the initial learning rate is set to 0.001, dropout keep rate to 0.9, and the convolutional window size to 3. A few typical dimensions including 50, 100, 200 and 300 are compared.

Figure 2 shows the accuracy of the testing result with iteration as the variable. The performances in dimensions 50 and 100 do not show significant difference. Although the model at 200 and 300 dimensions experience some fluctuations, their overall performance is obviously higher than smaller dimensions. This shows that Cantonese style writing is as comprehensive as Mandarin Chinese and thus, the suitable dimensions are also similar. Consequently, embedding dimensions should be between 200 and 300, and the latter has more potential to achieve higher performance since the capability of representation is expended with higher dimensional vector.

Dropout keep rate, an effective way for model regularization, is regarded as a key parameter for deep learning algorithms. By using this parameter, the right dropout rate can alleviate the over fitting problem in the learning model. To find the best dropout rate, we have the initial learning rate set to 0.001, token embedding dimension to 300, and convolutional window size to 3. The experiments covered dropout keep rate (DKR) 0.7, 0.8, 0.9 and 1.0.
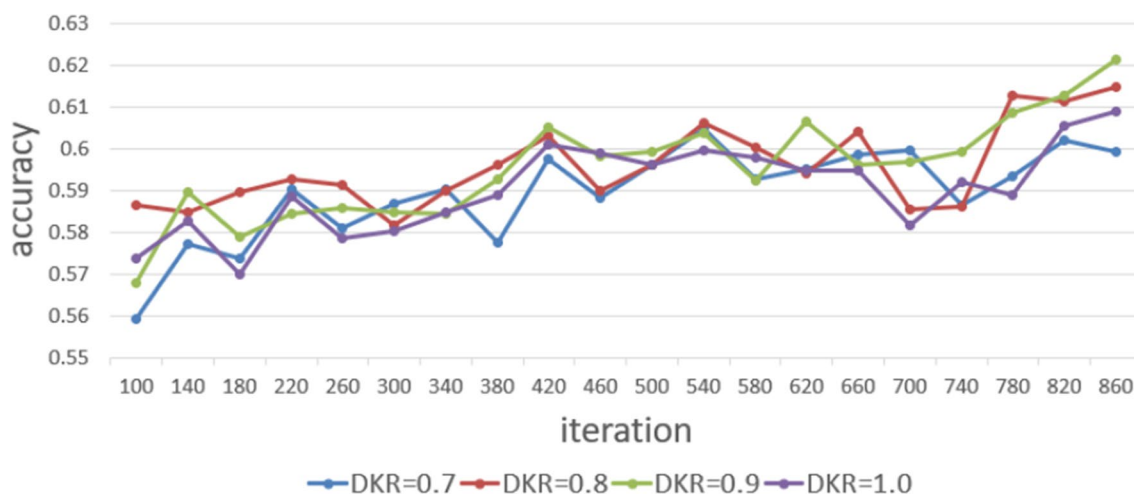
Fig. 3 Dropout keep rate comparison in the conditions: learning rate = 0.001; token embedding dimension = 300; convolutional window size = 3
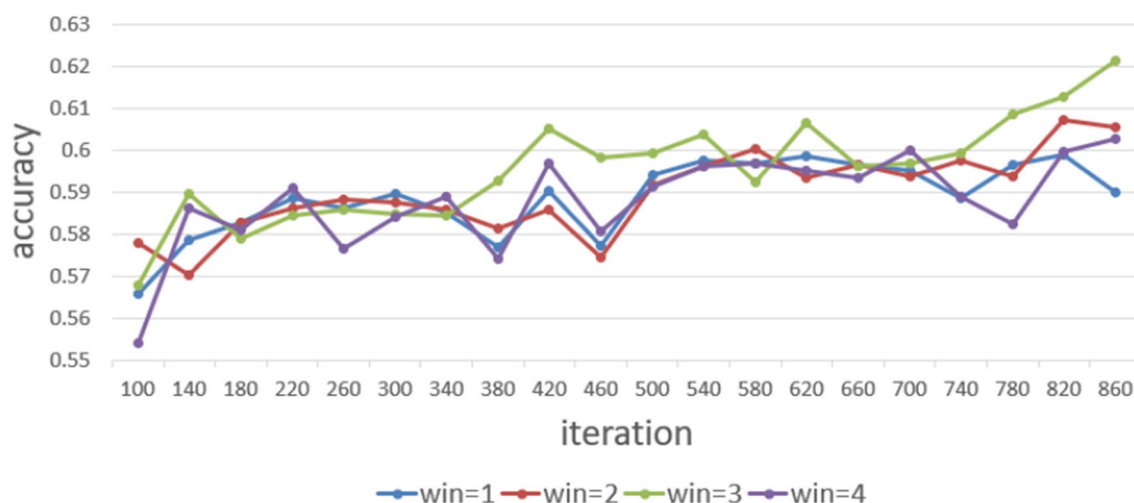

Fig. 4 Convolutional window comparison in the conditions: learning rate = 0.001; Token embedding dimension = 300; dropout keep rate = 0.9

Figure 3 shows that the system has the best performance when DKR is set to 0.9. The second best occurs in the group of 0.8 DKR. Although breaking some connections in deep network layers can avoid the over-fitting problem in some extent, the weight learned in training could be wrongly ignored by the same reason. The deterioration is more apparent in the setting of 0.7 DR, indicating that an over-simplified model is not a good approach neither. Based on above result, we can observe that choosing the right dropout rate can improve the generalization of a model. But, a good ratio should be determined cautiously.

CNN, used as a deep learning approach to extract N-gram features [46], requires a window size for taking the n-grams features. In this experiment, we measure the performance with window size from 1 to 4. Window size of 4 makes sure that commonly used 4-word WSCs scripts are included. The initial

learning rate is set to 0.001. Token embedding dimension is 300 and the dropout keep rate is 0.9.

Figure 4 shows that window size 1 has very poor performance. In this case, the model can be regarded as the hybrid of LSTM and a mean of token level representative vector and thus, noise can be introduced from average computation of all minor writing system tokens. Wider window sizes have better performance since the evidence becomes stronger considering phrase or multiple token co-occurence. For example, 'familp' and 'taste' are basically neutral in affective values. When they are combined in 'family taste', it becomes quite positive. However, wider windows do not always show improvement. The accuracy of window size 4 only reaches the same level as that of size 2. Both of them are significantly worse than 3-length window. One possible

Fig. 5 Word cloud for happiness; left is depicted by the complete text; right is depicted by the WSCs scripts

Fig. 6 Word cloud for anger; left is depicted by the complete text; right is depicted by the WSCs scripts

reason is that the dataset of Openrice contains more triple token phrases of WSC tokens so 3-length window can naturally match those features. For example, 3-token expressions like 'who tm (fucking) care', 'what the fuck', and 'stay with me' can be easily observed in the dataset. However, 4-token expressions in WSCs rarely occur in this dataset.

4.6 Visualization and case study

To make a general perspective of Chinese text and WSCs, Word Cloud graphs[6] are used as a visualization tool to intuitively identify the most frequent scripts in the Chinese Microblog dataset. Figures 5 and 6 show the word clouds for *happiness* and *anger*, respectively. In each figure, the result for the complete text is shown on the left whereas the WSCs only collection is depicted on the right.

From Figs. 5 and 6, a reasonable consistence of writing system expressions can be observed for both *happiness* and *anger*. The most frequent positive Chinese tokens are 太 ('high'), '爱' ('love') and '开心' ('happp') whereas the negative ones are '死' (*fuck off*) and '讨厌' (*hate*). 'fuck' and 'shit' are also often used to strongly reflect negative emotion. In general, English words are the majority among all kinds of WSCs. There are, however, other types of interesting WSCs

---

6 https://en.wikipedia.org/wiki/Tag_cloud.

Fig. 7 Case study 1 with attention heat map of BAN and HAN-WSC

Fig. 8 Case study 2 with attention heat map of BAN and HAN-WSC

tokens, e.g. '*ja*' (an onomatopoeic token to describe the complacent laugh) and 'lol' ('laughing out loud'). The WSCs 'qaq'(the emoji for tearing) and 'tmd'('ta ma de', a curse word like fuck) are commonly used words for negative expressions.

E5 Wuli super junior 永远是最棒的...爱你们。
Our favorite super junior is always the best. I love you.
Emotion: happiness (Fig. 7)

Two emotion analysis examples and attention heat maps are provided to demonstrate the differences between the state-of-the-art BAN and our HAN-WSC. In example E1, the predicting task is difficult for BAN since WSCs tokens cannot be explicitly used. In fact, 'wuli' is a Korean word yet spelled using the Mandarin Pinyin system in the Internet community to show their enthusiasm. Comparing attention weights (the lighter color indicates higher weight, vice versus), BAN puts more weights on Chinese words '棒','爱'. Since 'wuli' is neither an English word nor a Korean script, BAN does not have any knowledge to put attention to this script. On the other hand, this problem can be easily solved in HAN-WSC as it uses a separate learning framework for WSCs, granting more weight to 'wuli'.

E6 ccav 直播巨大的bug啊！李娜说成"澳网冠军、法网亚军"啊！sb了啊！
Ccav live has a huge bug! Li Na was described "Australian Open champion, the French Open runners" Ah! Idiot!
Emotion: anger (Fig. 8)

In example E2, BAN gives more attention to the exclamatory marks, which are often used in either positive or negative intense emotion. On the contrary, HAN-WSC gives the most significant attention weights to the two WSCs which

are 'bug' and 'sb', leaving the third WSCs script 'ccav' with a smaller weight. The token 'ccav' emphasized by BAN, in general, is not related to *anger*. On the other hand, HAN-WSC does not assign much attention to this WSC token during the training process. Moreover, 'bug', 'sb' identified by HAN-WSC efficiently catch the negative sense. 'sb' (short hand for *sha bi*, '傻逼'), the commonly used new-born WSC, is generally used to describe an idiot, and the hybrid model is more effective to handle these odd cases.

5 Conclusion and future work

This paper presents a hybrid deep learning model with attention network for affective analysis in the context of writing system changes. We argue that WSCs text is potentially informative and a proper learning model needs to be designed such that additional information can be captured in deep learning based models for emotion classification. Based on the hypothesis, our proposed hybrid neural network model offers a new way to integrate multiple types of writing systems into attention-based LSTM model. Along with WSCs, the text of the major writing system, which reveals the events, is regarded as informative resources by using LSTM. WSCs can be used to generate representation especially linked to emotional feature using a CNN model. Through performance evaluation, we also show that the LSTM model is more suited to major writing system, and CNN is more suited for WSCs. Experiments show that the proposed hybrid deep learning method which better incorporates WSCs features can further improve performance compared to the state-of-the-art classification models. It clearly indicates that WSCs can serve as an effective information in affective analysis of the social media text.

Future work will include two directions. One is to investigate the performance of our proposed HAN-WSC on more datasets as currently only one publicly accessible dataset is available for writing system changes focusing on Chinese text. The other direction is to explore the use and types of WSCs to express affections in other language communities.

References

1. Barbosa L, Feng J (2010) Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd international conference on computational linguistics: posters, Association for Computational Linguistics, pp 36–44
2. Balamurali A, Joshi A, Bhattacharyya P (2011) Harnessing wordnet senses for supervised sentiment classification. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 1081–1091
3. Liu B, Zhang L (2012) A survey of opinion mining and sentiment analysis. In: Mining text data, Springer, pp 415–463
4. Wilson T, Kozareva Z, Nakov P, Rosenthal S, Stoyanov V, Ritter A (2013) Sentiment analysis in twitter. In: Proceedings of the international workshop on semantic
5. Joshi NS, Itkat SA (2014) A survey on feature level sentiment analysis. Int J Comput Sci Inf Technol 5:5422–5425
6. Mishra A, Dey K, Bhattacharyya P (2017) Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), vol 1, pp 377–387
7. Kanter JM, Veeramachaneni K (2015) Deep feature synthesis: towards automating data science endeavors. In: Data science and advanced analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, IEEE, pp 1–10
8. Dos Santos CN, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. In: COLING, pp 69–78
9. Clyne M (2000) Constraints on code-switching: How universal are they. The bilingualism reader, pp 257–280
10. Pujolar J (2001) Gender, heteroglossia and power: a sociolinguistic study of youth culture, vol 4. Walter de Gruyter, Berlin
11. Cromdal J (2001) Overlap in bilingual play: some implications of code-switching for overlap resolution. Res Lang Soc Interact 34(4):421–451
12. Auer P (2013) Code-switching in conversation: language, interaction and identity. Routledge, Abingdon
13. Musk N (2012) "Performing bilingualism in wales," Pragmatics. Q Publ IPrA 22(4):651–669
14. Vicentini A (2003) "The economy principle in language," Notes and Observations from early modern english grammars. Mots Palabras Words 3:37–57
15. Bond MH, Lai T-M (1986) Embarrassment and code-switching into a second language. J Soc Psychol 126(2):179–186
16. Heredia RR, Altarriba J (2001) Bilingual language mixing: why do bilinguals code-switch? Curr Direct Psychol Sci 10(5):164–168
17. Wei JM (2003) Codeswitching in campaigning discourse: the case of taiwanese president chen shui-bian. Lang Linguist 4(1):139–165
18. Bergen B (2016) What the F: what swearing reveals about our language, our brains, and ourselves. Basic Books, New York
19. Huang C-R, Shi D (2016) A reference grammar of Chinese. Cambridge University Press, Cambridge
20. Lee S, Wang Z (2015) Emotion in code-switching texts: corpus construction and analysis. In: Proceedings of the Eighth SIGHAN workshop on chinese language processing, pp 91–99

21. Wang Z, Lee SYM, Li S, Zhou G (2017) Emotion analysis in code-switching text with joint factor graph model. IEEE/ACM Trans Audio Speech Lang Process 25(3):469–480

22. Adel H, Vu NT, Schultz T (2013) Combination of recurrent neural networks and factored language models for code-switching language modeling. In: Proceedings of the 51st annual meeting of the association for computational linguistics (Volume 2: Short Papers), vol 2, pp 206–211

23. Solorio T, Liu Y (2008) Learning to predict code-switching points. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 973–981

24. Lignos C, Marcus M (2013) Toward web-scale analysis of codes-witching. In: 87th Annual Meeting of the Linguistic Society of America

25. Li Y, Fung P (2012) Code-switch language model with inversion constraints for mixed language speech recognition. Proc COLING 2012:1671–1680

26. Jamatia A, Gambäck B, Das A (2015) Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In: Proceedings of the international conference recent advances in natural language processing, pp 239–248

27. Li S, Huang L, Wang R, Zhou G (2015) Sentence-level emotion classification with label and context dependence. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long Papers), vol 1, pp 1045–1053

28. Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pp 417–424

29. Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 174–181

30. McKeown K, Jordan D, Hatzivassiloglou V (1998) Generating patient-specific summaries of online literature. In: Proc. of Intelligent Text Summarization, AAAI Spring Symposium, Citeseer

31. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp 79–86

32. Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci 181(6):1138–1152

33. Socher R, Pennington J, Huang EH, Ng AY, Manning CD (2011) Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 151–161

34. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, Citeseer, p 1642

35. Irsoy O, Cardie C (2014) Opinion mining with deep recurrent neural networks. In: EMNLP, pp 720–728

36. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166

37. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

38. Tang D, Qin B, Liu T (2015) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing (2015), pp 1422–1432

39. Tang D, Qin B, Liu T (2015) Learning semantic representations of users and products for document level sentiment classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long Papers), (Beijing, China), Association for Computational Linguistics, July, pp 1014–1023

40. Long Y, Ma M, Lu Q, Xiang R, Huang C.-R (2018) Dual memory network model for biased product review classification. arXiv preprint arXiv:1809.05807

41. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489

42. Long Y, Qin L, Xiang R, Li M, Huang C.-R (2017) A cognition based attention model for sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 473–482

43. Long Y, Xiang R, Lu Q, Huang C-R, Li M (2019) Improving attention model based on cognition grounded data for sentiment analysis. In: IEEE transactions on affective computing

44. Wang Z, Zhang Y, Lee S, Li S, Zhou G (2016) A bilingual attention network for code-switched emotion prediction. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 1624–1634

45. Keller MHF (2016) Modeling human reading with neural attention. In: Proceedings of the conference on empirical methods in natural language processing, p 95

46. Wang J, Yu L.-C, Lai K. R, Zhang X (2016) Dimensional sentiment analysis using a regional cnn-lstm model. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 2: Short Papers), vol 2, pp 225–230