Emoticon Analysis for Chinese Social Media and E-commerce: The AZEmo System

SHUO YU, HONGYI ZHU, and SHAN JIANG, University of Arizona, USA
YONG ZHANG and CHUNXIAO XING, Tsinghua University, China
HSINCHUN CHEN, University of Arizona, USA and Tsinghua University, China

This article presents a novel system, AZEmo, which extracts and classifies emoticons from the ever-growing critical Chinese social media and E-commerce. An emoticon is a meta-communicative pictorial representation of facial expressions, which helps to describe the sender's emotional state. To complement non-verbal communication, emoticons are frequently used in social media websites. However, limited research has been done to effectively analyze the affects of emoticons in a Chinese context. In this study, we developed an emoticon analysis system to extract emoticons from Chinese text and classify them into one of seven affect categories. The system is based on a kinesics model that divides emoticons into semantic areas (eyes, mouths, etc.), with improvements for adaptation in the Chinese context. Machine-learning methods were developed based on feature vector extraction of emoticons. Empirical tests were conducted to evaluate the effectiveness of the proposed system in extracting and classifying emoticons, based on corpora from a video sharing website and an E-commerce website. Results showed the effectiveness of the system in detecting and extracting emoticons from text and in interpreting the affects conveyed by emoticons.

CCS Concepts: • **Information systems** → **Sentiment analysis**;

Additional Key Words and Phrases: Affect analysis, Chinese Internet, emoticon, social media

ACM Reference format:

Shuo Yu, Hongyi Zhu, Shan Jiang, Yong Zhang, Chunxiao Xing, and Hsinchun Chen. 2019. Emoticon Analysis for Chinese Social Media and E-commerce: The AZEmo System. *ACM Trans. Manage. Inf. Syst.* 9, 4, Article 16 (February 2019), 22 pages.

<https://doi.org/10.1145/3309707>

1 INTRODUCTION

The emergence of Web 2.0 technologies has dramatically reframed the definitions of communication and interpersonal connection. Although Facebook, Twitter, and other social media websites have integrated multimedia components into their services (e.g., pictures, video), text-based messages have not lost their popularity in computer-mediated communication (CMC) due to their

This study was supported by China Elite-1000 Program, Tsinghua University, the National Natural Science Foundation of China (Grant No.: 71110107027), USA NSF SES-1314631 and DUE-1303362.

Authors' addresses: S. Yu, H. Zhu, S. Jiang, and H. Chen, University of Arizona, Department of Management Information Systems, 1130 E. Helen St., McClelland Hall 430, Tucson, Arizona 85721-0108, USA; emails: {shuoyu, zhuhy, jiangs}@email.arizona.edu, hchen@eller.arizona.edu; Y. Zhang, C. Xing, H. Chen, Tsinghua University, Information Science and Technology Building, Room 1-310, Beijing 100084, China; emails: {zhangyong05, xingcx}@tsinghua.edu.cn.

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

© 2019 Association for Computing Machinery.

2158-656X/2019/02-ART16 \$15.00

<https://doi.org/10.1145/3309707>

convenience and compatibility across platforms. However, since few visual and vocal cues can be conveyed through text, text-based online communication suffers from the loss of non-verbal symbolic information compared with natural face-to-face communication or video/audio calls. As a compensation, symbols that mimic human faces to express more nuanced opinions and emotions are often used along with text to aid in text-based communication. These symbols are called emoticons.

An emoticon is a meta-communicative pictorial representation of facial expression, which serves to convey information about the sender's emotional state. Emoticons have been widely used by social media users and contribute to the facilitation of text-based communication technologies [9] [34]. They enable people to express their subtle mood changes, including strengthening a message, clarifying or emphasizing feelings, softening negative tones, and expressing humor [11]. Three types of emoticons have been identified: Western text-based emoticons, Eastern text-based emoticons, and image-based emoticons (also known as Emoji). Recently, Chinese social media (e.g., Weibo, WeChat) and E-commerce sites (e.g., Alibaba, Tmall) have experienced significant growth and success in the past decade, surpassing 629 million active users and \$438 billion in revenues in 2014, taking over the U.S. market and ranking first in the world in online retail sales. We have observed an increasing use of emoticons in Chinese web forums and other social networking sites that feature gaming and entertainment, friending and social interactions, health and fitness topics, and E-commerce. Since these symbols can reveal people's feelings and emotional states hidden under the text, emoticon analysis in these platforms can improve the performance of sentiment analysis for Chinese social media, monitor people's emotional conditions from social networks (especially for those who should avoid violent mood swings), and help readers understand customer opinions with Chinese dialects (e.g., Cantonese) in online reviews, even though people may be unfamiliar with those dialects. For example, "Danced so far without being short of breath. ^(\≡▽≡)/~" shows happiness and excitement; "I can't find my jewelry QAQ" shows a feeling of upset. In these sentences, the symbols at the end can help us identify the emotional status of the individuals, although the words do not explicitly convey the emotion. As a result, a growing number of studies have started to examine the use of emoticons in online forums, blogs, and other social media [7, 27, 28, 51].

However, very few works have thoroughly investigated emoticons in a context of Chinese social media. In Chinese as well as other Eastern social media contexts, symbols used for emoticon construction have a greater variety compared to Western texts, with concentration on Eastern text-based emoticons. Eastern text-based emoticons often use characters from local languages, which can cover the entire Unicode set. The relative complexity and flexibility of the components of Eastern text-based emoticons have greatly stimulated user creativity, which has resulted in the existence of tens of thousands of distinct emoticons. Such emoticons have high variability and can hardly be represented by well-defined grammars, thus the need feature representation and machine-learning methods for affect analysis.

Although there is increasing popularity of image-based emoticons, their types provided by platforms are small, typically a few dozen. Users might find that the pre-set Emoji cannot exactly express the subtlety of their affects. In contrast, users have full control of Eastern text-based emoticons. They can easily add or remove components (hands, hair, objects, etc.) or even adjust the sizes of eyes and noses, simply by choosing proper characters. The flexibility of Eastern text-based emoticons helps reflect users' personal characteristics. Furthermore, text-based emoticons are still the only choice in particular circumstances where image-based emoticons are unavailable. For instance, recently a new form of viewer comment, called bullet-screen comments ("danmu" in Chinese), have become prevalent on Chinese video sharing and streaming sites (e.g., bilibili.tv, douyu.com, etc.). Bullet-screen comments appear with the video, floating from right to left on the

video. This way, viewers can comment on video plots in real time. With the popularity of Chinese video sharing and streaming sites, plain text-based emoticons are still preferred as image-based emoticons interfere with the video and cannot be displayed. These unique features of text-based emoticons make it irreplaceable by image-based emoticons. Our work takes the first step in this area to explore effective methods to extract and classify Eastern text-based emoticons in a Chinese context, using advanced feature vector extraction for machine-learning methods. Specifically, we focus on the application of emoticons in Chinese social media and E-commerce contexts, which are two of the main applications in everyday Internet reality.

This article presents a novel system, AZEemo, which extracts emoticons from posts and comments in Chinese social media and classifies them into one of the seven pre-defined categories automatically. The system extracts semantic components (representations of eyes, mouths, and other parts of human body) from emoticons and evaluates these areas integrally. It utilizes a large lexicon of emotion symbols whose construction was based on existing online emoticon dictionaries and adapts it to the Chinese context via emoticon truncation. For the classification task, AZEemo improves the heuristic method based on Reference [28] and proposes a novel approach to involve machine-learning methods based on emoticon kinesics features. The performance of the system is evaluated with empirical tests on a large corpus of more than one million sentences from different Chinese social media sites, which is an extension of Reference [48].

The main novelty of this work is as follows. First, we adapt Japanese emoticon dictionaries to create our Chinese emoticon dictionaries via emoticon truncation, which has not been done in prior works. This adaptation sheds light on the methodology of the initial construction of Chinese emoticon dictionaries. Second, we revise the heuristic method proposed in Reference [28] to account for more emotionally skewed kinesics components. We believe that this method better reflects the composition and actual use of emoticons. Third, we are the first to propose a new feature set (kinesics features) for machine-learning-based emoticon affect analysis. We believe this extension to emoticon feature generation can enhance emoticon classification as well as general sentiment analysis for sentences and articles.

The remainder of the article is organized as follows. Section 2 provides the scope of emoticons in this study, presents a review of relevant literature, and identifies potential research gaps. Section 3 presents our proposed AZEemo framework. Section 4 describes our datasets (one lexicon dataset and two test datasets) and experiment settings and discusses the results. Finally, Section 5 concludes the study and identifies future directions.

2 RELATED WORKS

2.1 The Scope of Emoticons

The intent of emoticons is to mimic human facial expressions. Although the denotation of emoticons can consist of symbols mimicking actions or objects other than facial expressions, we limit the scope of emoticons in this work to the more frequently used facial expressions. The main components of emoticons are human face parts (eyes, mouths, or noses, etc.), although additional parts (tools, arms, etc.) sometimes attach to them. Table 1 exhibits several emoticons in each category with similar meanings that were identified during our initial emoticon exploration.

Several types of emoticons have emerged during their development: Western text-based, Eastern text-based, and image-based emoticons. The use of Western text-based emoticons was first documented by Scott Falman at Carnegie Mellon University in 1982 [6]. They were exhibited in a 90°-rotated form, such as “:-)” (happy face) or “:- (“ (sad face). They typically consisted of two to four characters, representing eyes (“:”), noses (“-”), and mouths (“)” or “(”), mainly composed from the ASCII character set, and were relatively few in number (about several hundred). Eastern

Table 1. Samples of Emoticons Expressing Happiness, Sadness, Anger, and Surprise

Western	Eastern	Emoji	Meaning
:~)	(^_^)		Happiness
:-(>:(>:O	(` ; ω ; `) (# ° ▯ °) (□□°)		Sadness Anger Surprise

text-based emoticons, also known as “Kaomoji” in Japanese, have been popularized by Japanese users since 1986 and have been adopted by Chinese users with local modifications in recent years [46]. They can be understood without tilting one’s head to the left. Some examples are “\(^_^)/” (happy face with arms), “(T_T)” (sad, crying face), or “(╯°□°)╯┴┴” (an angry person turning a table upside-down). They may consist of 3 to 20 characters, depending on the presentation complexity. Eyes (e.g., “^”, “T”), mouths (e.g., “_”, “;”), and face borders (e.g., “()”) are the most frequent components in Eastern text-based emoticons; additional components such as arms (e.g., “\”, “/”), objects (e.g., “_┴┴_”), and cheeks (e.g., “*” in “(*^_^*)”), and so on, can significantly modify the affect of the emoticon to be conveyed. Image-based emoticons, also known as “Emoji” in Japanese, are small pictures representing faces or other objects, which can be sent along with electronic text. In the context of Chinese social media, Eastern text-based emoticons and image-based emoticons are the most widely used emoticons by users [21]. However, the site-specific implementations of image-based emoticons reduce their generalizability. For instance, some websites automatically render particular character sequences into their pre-defined image-based emoticons (e.g., “:-)” into a smiley face image) or convert particular tags (e.g., [smiley face]) into images to be exhibited on the website [21]. Considering their variety, flexibility and generalizability, we focused on Eastern text-based emoticons on Chinese social media websites as our research objects.

To form the basis of our research, we provide a review of recent emoticon studies. In general, four streams of emoticon research have been the most prevalent; they are discussed and summarized below.

2.2 Emoticons and User Behaviors

This stream of emoticon research focuses on the motivation for using emoticons, the outcomes of using emoticons, and how emoticons affect individuals’ behaviors. Individuals use emoticons for various reasons, but some of the most common motivations are to express friendliness and enhance sentimental expressions. For example, individuals tend to use more emoticons in socio-emotional contexts compared to that in task-oriented contexts [11], enhancing the smoothness of online communication processes. Another study also shows that students tend to use emoticons in less-formal conversations [22]. Researchers have found that by using negative emoticons, a negativity effect can shift message interpretation toward the emotion direction conveyed by the emoticon [39]. Attention saving is another important psychological motivations in using emoticon [7]. Individuals prefer using emoticons to substitute long text components. Emoticon usage can also lead to positive outcomes of conversations. Specifically, emoticons can help clarify and enhance the meaning of text messages. For example, an inter-university examination on e-mail communication concludes that inclusion of emoticons in e-mail messages increases the understandability of message content [31]. Individual behaviors of emoticon usage have also been studied extensively. Gender is a key factor to differentiate between different emoticon usage patterns. For example, female

groups tend to use more emoticons compared with male groups, but both genders perform similarly in mixed-gender groups. The major semantic category of female emoticon usage is humor, in contrast to male group's teasing or sarcasm [42]. In addition, individuals prefer emoticons that can be handily typed and intuitively understood [7]. Furthermore, emoticons are social-cultural norms whose meanings could vary [26].

This research stream contributed to investigating the pragmatics of emoticons used as a supplement of verbal languages in online communication. Unfortunately, these studies only investigated limited subsets of emoticons and their valence. Various emoticons and the affects they conveyed need further research.

2.3 Emoticons and Sentiment Analysis

This stream of emoticon research focuses on using the valence of emoticons as a complementary tool to improve classification performance in sentiment analysis. Emoticons can be adopted to reduce topic dependency while using Naïve Bayes and Support Vector Machines (SVM) classifiers for sentiment analysis instead of topic sensitive lexicons [30]. Furthermore, emoticons like Emoji can be used as an automatic annotation tool with algorithms like SVM and Conditional Random Fields (CRF) to tag blog sentences into four sentiment categories: happy, joyful, angry, and sad. Manual annotation can be reduced to save time and labor [45].

Analyzing sentiments for short texts such as Twitter is challenging, since fewer available features can be used compared with long articles. It has been a common practice to integrate the polarity of Western emoticons as features for supervised or unsupervised sentiment analysis in social media. This integration has improved sentiment analysis results in the Twitter context and helped [19] identify several categories of tweets that may cause erroneous classification results. To achieve better results, contributions of text and Western emoticons on sentence sentiment have been weighed with a probability model [20]. Machine-learning algorithms such as Naïve Bayes, SVM, and Multilevel Perceptron can utilize Western emoticons as features to improve their performance in sentiment analysis [27].

Some Emoji and Eastern emoticons have been integrated into lexicons while dealing with non-English, Twitter-like corpora [49]. People express their emotions with Eastern emoticons systematically, and Emoji have been widely used in all kinds of situations including sarcasm contexts [50]. This integration has helped better classify tweet sentiments and monitor sentiment variance in Chinese social media [16, 51]. It has also contributed to a 92% sentiment analysis accuracy in Indonesian Twitter data [32]. These studies have contributed to the usability of emoticons as complementary tool to improve performance of sentiment analysis in different language contexts.

Research in this stream confirmed the positive influence of involving emoticons on sentiment analysis. However, most analyses of Chinese social media (e.g., Weibo) limited their scope by looking only at Emoji. Also, the size of their selected emoticon set was limited, and the structure of Eastern and Western emoticons was not thoroughly analyzed.

2.4 Emoticon Generation and Recommendation

This stream of studies focuses on emoticon generation and recommendation in human-computer interaction systems and input methods. Eastern emoticons were studied in this stream, since they have more variety and complexity compared to Western counterparts. The authors of Reference [23] implemented a neural network-based algorithm to learn how semantic areas of emoticons (mouths, eyes, hands, etc.) were related with the sentence. Although some emoticons generated after the learning process were less meaningful, the idea of exploiting semantic areas was promising. Researchers also paid attention on recommending specific emoticons based on the emotion categories of the text. With their novel definition of emotion categories, they were able to

recommend proper emoticons according to input text and outperform other emoticon recommend systems, such as iOS input methods in Japanese context [34, 38].

These studies typically conducted sentiment analysis on input text and recommended emoticons based on analysis results. Although we do not focus on emoticon recommendation, we are inspired by the idea of exploiting semantic areas of emoticons in a structured manner.

2.5 Emoticon Extraction and Classification

This stream of research aims to identify and extract emoticons from text and classify them into detailed affect categories (e.g., happiness, sadness, surprise, etc.). Two major approaches are identified to extract emoticons from corpora: lexicon matching and statistical and machine-learning-based extraction. Most prior studies applied lexicon matching to extract emoticons [1, 14, 15]. The emoticon lexicons contain a few pre-selected emoticons and their corresponding sentiment categories [10, 37, 44]. However, the lexicon matching method can only extract a limited portion of Eastern emoticons due to the flexibility of emoticon structures.

Several statistical and machine-learning-based methods have been developed and applied to these tasks. For example, the authors of Reference [35] proposed kernel methods for Eastern emoticon extraction and classification. They used POS type tags and word positions in sentences as features and extracted emoticons with SVM. Then they implemented the Dynamic Time Alignment Kernel (DTAK) and String Subsequence Kernel (SSK) in their SVM classifier for classification. Their kernels mainly focused on word length, character sequences, and substring similarity but did not investigate semantic areas of emoticons. Their research was one of the earliest to implement machine-learning methods on extraction and classification tasks.

Yamada et al. [43] used a BBS corpus and classified Eastern emoticons based on statistics of n-grams. They took a traditional text mining perspective, treating characters in emoticons as words in sentences. They experimented and concluded that the unigram method outperformed the bigram and trigram methods. These methods led to acceptable classification accuracies but generally overlooked the variety of meanings one character could have in different semantic areas.

Ptaszynski et al. [28] applied a kinesics based model on Eastern emoticon extraction and classification. They gathered a large emoticon lexicon from the Internet, semantically divided the emoticons into parts and then classified them into different affect types. They utilized a Japanese blog corpus to test their heuristic-based kinesics model and outperformed other methods such as n-grams. Their heuristic method adapted to the creativity of users' using new characters and constructing new emoticons. However, their evaluation criteria were disputable in the sense that their system ranked five most probable affect categories (1 of 10) for each emoticon with accuracy scores 5/5 to 1/5, for Rank 1 to Rank 5, respectively, which was not a common practice in calculating accuracy and may lead to an inflated result.

Bedrick et al. [3] proposed a context-independent emoticon extraction method using Probabilistic Context-Free Grammars to extract Eastern emoticons on Twitter with various kinds of languages. Based on emoticons' symmetrical visual characteristics, they grouped characters with symmetrical similarities and used them in pairs to identify emoticons. This research involved computer vision techniques and contributed a novel idea to this stream.

Yokoi et al. [47] proposed a regular expression based method to extract emoticons that contain symbols describing eyes on Japanese Twitter. Based on the symmetric structures of most Eastern emoticons, they adopted Damerau-Levenshtein Distance to determine emoticon boundaries by comparing the similarity between the strings on both sides of eye symbols. This method might find difficulty in handling asymmetric emoticons, which is common in real use.

Most of the above research was conducted in a Japanese context; similar studies do not appear to exist for the Chinese context, in which character and symbols used by Internet users may

Fig. 1. AZEemo framework for extracting and classifying emoticons in Chinese text.

differ. There appeared to be, therefore, no system adaptable to the Chinese context, with confirmed validity for Chinese social media.

2.6 Research Gaps and Questions

Based on the literature review, several research gaps can be identified. Most prior studies focused on the emotion valence (i.e., positive or negative) carried by emoticons [16, 27, 30, 45], while emoticons actually contain richer semantics that express the affects of users, such as happiness, sorrow, and anger. Several studies focused on emoticon affect extraction and refined classification [28, 35, 43], but most were conducted in a Japanese context, which may not be directly applicable to Chinese social media. For the emoticon classification task, we noticed that past work used substring similarity-based kernel methods [35], statistical n-grams [43], and dictionary-based heuristic methods [28]. However, to the best of our knowledge, few works have attempted to extract feature vectors from emoticons for machine-learning-based affect classification.

In this study, we developed an emoticon analysis system, AZEemo, to extract and classify text-based emoticons from text collected from Chinese social media. In terms of the affect classification task, we designed an improved heuristic method as well as an exploratory machine-learning method based on feature vector extraction. Specifically, we asked the following research questions:

- Can AZEemo effectively extract emoticons from text on Chinese social media?
- Can AZEemo accurately classify emoticons into different affect categories?

The following specific questions were asked for the emoticon classification task:

- Can our improved heuristic method outperform the existing heuristics method and other benchmark systems?
- Can our machine-learning methods with kinesics feature vectors outperform similar machine-learning methods without kinesics feature vectors, and as compared to other benchmark systems?
- Can our machine-learning methods with kinesics feature vectors outperform our improved heuristic method?

3 RESEARCH DESIGN

Figure 1 illustrates our proposed dictionary-based emoticon extraction and classification system, the Arizona Emoticon Analysis System (AZEemo). The major components including kinesics representation, affect mapping, emoticon extraction, and emoticon classification will be discussed in detail in the following subsections.

Table 2. Samples of Emoticon Truncation

Emoticons before Truncation	Emoticons after Truncation
ㄣ(>▽<)oきやはははっ!	ㄣ(>▽<)o
ムカ...(-_-メ)	(-_-メ)
きゃハハ!! ㄣ(@^▽^@)ノ☆ハハン!!	ㄣ(@^▽^@)ノ☆

3.1 Lexicon Development

The first step in our system was to develop a lexicon of emoticon symbols that could be used to suggest the meanings of emoticons appeared in Chinese texts. However, since we were not aware of any comprehensive online resources for large-scale Chinese emoticon dictionaries, we utilized several online Japanese emoticon dictionaries to construct the base lexicon, due to the fact that Eastern text-based emoticons were first developed in Japan and that Chinese and Japanese share some cultural roots (using Chinese characters, etc.). Meanwhile, we recognized that Chinese and Japanese are distinct languages and adapted Japanese emoticons to fit into Chinese contexts. The adaptation process is explained in the next paragraph. The Japanese emoticon dictionaries included Facemark Party, Kaomoji-café, Kaomoji Paradise, Kaomoji Station, Kaomojisyo, Kaomoji-toshokan, and Kaomojiya,¹ which had been used by prior emoticon research [28]. Each emoticon dictionary site had its taxonomy of affect categories for collected emoticons. Example categories from one dictionary site include greetings, farewell, talking, suffix, response, surprise, shyness, happiness, sadness, fear, laugh, fondness, support, action, life, and so on. Each tag contained a list of emoticons, and each emoticon belonged only to a single tag. Since obvious emotional inclinations cannot be inferred from some tags including greetings and farewells, we only collected emoticons that were labeled by tags suggesting emotional meanings. Only tags with names suggesting emotional contents, such as happiness, sadness, and fear, were selected, and emoticons from those tags were collected. In total, 11,988 emoticons were collected from the list of websites, of which 10,560 were unique.

We observed that a large proportion of the collected emoticons containing Japanese characters that are communicative rather than meta-communicative. That is, those characters are Japanese language helping verbally explain the semantics of the corresponding emoticon, instead of presenting human facial expressions. Those “verbal” Japanese characters usually appear consecutively on the leftmost, rightmost, or both in an emoticon. As the Japanese language is rarely seen in Chinese online social media, “verbal” Japanese characters are hardly adaptable to our Chinese context. Therefore, before using kinesics to represent the collected emoticons, we truncated continuous consecutive preceding and following Japanese alphabets characters (including some punctuation marks) from raw emoticons, but not touching those characters in the middle of them. Table 2 shows several examples of this truncation process. “きやはははっ!” means “hahahaha,” “ムカ...” means “enraged,” and “きゃハハ!!” and “ハハハハ!!” mean “hahahaha” and “bangbang,” respectively. Most of the truncated parts are imitative words, which are not part of emoticons by definition. Although “メ” is

¹Face-mark Party: <http://www.facemark.jp/facemark.htm>.

Kaomoji-café: <http://kaomojicafe.jp/>.

Kaomoji Paradise: <http://kaopara.net/>.

Kaomoji Station: <http://kaosute.net/jisyo/kanjou.shtml>.

Kaomojisyo: <http://matsucon.net/material/dic/>.

Kaomoji-toshokan: <http://www.kaomoji.com/kao/text/>.

Kaomojiya: <http://kaomojiya.com/>.

Table 3. Sample Emoticon Segmentation Using the Proposed Kinesics Model

Emoticon	S_1	B_1	S_2	E_L	M	E_R	S_3	B_2	S_4
$\overset{\wedge}{-}\overset{\wedge}{-}$	null	null	null	$\wedge$	$-$	$\wedge$	null	null	null
$(-_{-}\lambda)$	null	(	null	$-$	$-$	$-$	λ	)	null
$\text{灬}(>\nabla<)\circ$	灬	(	null	$>$	∇	$<$	null	)	$\circ$
$\text{灬}(@\wedge\nabla\wedge@)\text{丕}\star$	灬	(	@	$\wedge$	∇	$\wedge$	@	)	$\text{丕}\star$

a Japanese character, we did not remove it from “ $(-_{-}\lambda)$,” as it appears in the middle of the emoticon. Semantically it represents veins standing out from one’s face and thus should not be removed.

3.2 Kinesics Representation

In our research, the theory of kinesics, originally developed by Birdwhistell [4, 5], is applied to determine which emoticon is represented by which body movement or facial expression. Since emoticons are treated as non-verbal cues in online communication, i.e., emoticons convey emotions by mimicking human faces, it is reasonable to research emoticons in a framework of kinesics. It is common that in an emoticon, some parts (character n-grams) represent human eyes, some parts represent mouths, and so forth. To segment emoticons into different semantic parts that represent different functional components, we adopted the kinesics model in Reference [28] and represented each emoticon as the following nine components:

$$\{S_1\}\{B_1\}\{S_2\}\{E_L\}\{M\}\{E_R\}\{S_3\}\{B_2\}\{S_4\},$$

where $\{B_1\}$ and $\{B_2\}$ are the borders of the emoticon (mostly parentheses), $\{E_L\}$ and $\{E_R\}$ are left and right eyes, $\{M\}$ is the mouth, and $\{S_1\}\{S_2\}\{S_3\}\{S_4\}$ are the remaining components representing other parts that can vary depending on context (e.g., sweat, arms, tools, breath, etc.). Table 3 presents some examples how emoticons are segmented into the nine components.

Some components in the model can be empty (shown as “null” in Table 3), except for the $\{E_LME_R\}$ triplet. This eye-mouth-eye triplet is the essential component of any emoticon, and thus the triplet was extracted as an integral part from the collected emoticons. For better generalizability, we also allow partial forms of $\{E_LM\}$, $\{ME_R\}$, or $\{E_LE_R\}$, while still referring to them as triplets.

To segment the emoticons, we collected into kinesics components, we applied the following procedures to the emoticons. First, face borders were defined, and triplets were extracted as a whole. From the triplets, we separated left eyes $\{E_L\}$, mouths $\{M\}$, and right eyes $\{E_R\}$. Typically, the eyes were represented in identical strings (some were in reverse order); if not, then we extracted the leftmost and the rightmost character in the triplet as eyes, and the remaining parts as mouths. We also recombined eyes and mouths to form $\{E_LM\}$, $\{ME_R\}$, or $\{E_LE_R\}$, since we observed that there were variations that omitted a single eye or the mouth, e.g., “ $(\wedge\wedge)$ ” indicating a smiling face, although they only accounted for a small proportion. Consequently, we located and extracted additional components $\{S_1\}$ to $\{S_4\}$. $\{S_1\}$ is the part between the leftmost and the left border $\{B_1\}$, $\{S_2\}$ between the left border and the triplet, $\{S_3\}$ between the triplet and the right border $\{B_2\}$, and $\{S_4\}$ between the right border and the rightmost.

Table 4 shows the number of distinct character sequences extracted for each component from our test-bed of 3,518 triplets. From this table, we can see that $\{E_L\}$, $\{E_R\}$, and $\{M\}$ were frequently reused in different emoticons, and thus had smaller numbers of distinct character sequences, whereas the remaining components $\{S_1\}\{S_2\}\{S_3\}\{S_4\}$ had higher numbers of distinct character sequences, especially for $\{S_4\}$. This suggested that people tended to describe their emotions more specifically by adding different characters after the $\{E_LME_R\}$ triplet.

Table 4. Number of Distinct Character Sequences Extracted for Each Emoticon Component

	Triplet	E_L	M	E_R	S_1	S_2	S_3	S_4
Number of distinct character sequences	3,518	343	623	339	1,704	1,444	1,405	4,109

Table 5. Number and Percentage of Emoticons in Seven-Type Categorization

Affect Category	# of Emoticons	Percentage (%)
Happiness	3,879	32.4
Sadness	1,832	15.3
Fear	899	7.5
Anger	1,360	11.3
Disgust	286	2.4
Surprise	1,437	12.0
Fondness	2,295	19.1
Total	11,988	100.0

3.3 Affect Mapping

As described before, we observed that each of the seven emoticon dictionary sites has its own taxonomy on categorizing emoticons. A unified categorization schema is necessary to suggest precise semantics for those emoticons. One of the major emotional categorization schemas in psychology was proposed by Ekman [12], who suggested a six-type classification of affect categories: happiness, sadness, fear, disgust, anger, and surprise. In a Chinese context, the authors of Reference [8] categorized Chinese emotional words into 20 subtypes, including joy, ease, respect, praise, trust, fondness, anger, sorrow, disappointment, guiltiness, longing, panic, dread, shame, boredom, hatred, censure, envy, suspicion, and surprise, and further aggregated them into seven main categories, which are happiness, fondness, anger, sadness, fear, disgust, and surprise, based on a literature review on Chinese affect categorization. Compared to Ekman's categorization, a new category, "fondness," has been introduced into this schema. It is considered to be more comprehensive and balanced to describe Chinese emotions. To unify the *ad hoc* taxonomies applied by dictionary sites and map them onto our seven-type categorization, five experts specified in both Chinese and Japanese languages (one native in Chinese and Japanese, the other four native in Chinese and holding college majors in Japanese) examined the categories applied by the dictionary sites and found the most appropriate mapping. For instance, emoticon "(^_^)" was found under a category of "smiling," and the experts decided to treat "smiling" as a sub-category of "happiness." In this way, all the emoticons with an original category "smiling" became part of "happiness." The mapping and conflict resolving process is as follows. First, each expert maps the lexicon taxonomy labels to the seven categories individually. The five experts then meet and compare their mappings and mark out inconsistent ones. With the help of Japanese language dictionaries, the five experts discuss about the meanings of labels to figure out whether they can reach an agreement on the conflicts. For those with no agreements the conflicts are resolved by a majority vote. All conflicts were resolved up to this point. Table 5 shows the resulting categorization of the collected emoticon lexicon.

Furthermore, we separately mapped the $\{E_L M E_R\}$ triplets and $\{S_1\}\{S_2\}\{S_3\}\{S_4\}$ components into affect categories, because the same triplet could be used to express different emotions. Table 6 shows several examples of how the same triplet could be assigned to different affect categories when other kinesics components were attached. We then counted the number of times each triplet

Table 6. Examples of Emoticon Triplets Associated with Different Affect Categories

Triplet	Happiness	Sadness	Fear	Disgust	Anger	Surprise	Fondness
° ▯ °	(° ▯ ° °)	(11 ° ▯ °)	(* ° ▯ °) ♂	↘ (° ▯ ° ;)	(# ° ▯ °)	Σ (° ▯ °)	* ° ▯ ° ♀
˘ 艸 ˘	(★ ˘ 艸 ˘)	-	-	-	-	(○ ˘ 艸 ˘ ·)	-
• ω •	((p • ω • q))	(˘ • ω • ˘)	-	-	(/ • ω • \)	Σ (• ω • ;)	(* • ω • *)

Table 7. Examples of Affect Category Counts and Frequencies for Kinesics Components

Symbol	Position	Type	Happiness	Sadness	Fear	Disgust	Anger	Surprise	Fondness
˘ ▯ ˘	Triplet	Count	260	2	32	2	24	8	145
		Freq.	0.55	0.004	0.07	0.004	0.05	0.02	0.31
↘	S ₁	Count	157	21	14	13	54	38	51
		Freq.	0.45	0.06	0.04	0.04	0.16	0.11	0.15
˘ ▯ ˘	S ₁	Count	6	9	6	9	13	184	4
		Freq.	0.03	0.04	0.03	0.04	0.06	0.80	0.02

and component were assigned to each affect category, and used them to model the probability of each triplet being used to express certain affects. Suppose a triplet “° ▯ °” appeared in 7 happiness emoticons, 3 sadness emoticons and 10 surprise emoticons in our database. The triplet would be assigned “Happiness (7), Sadness (3), Surprise (10)” as its affect category count, along with “Happiness (0.35), Sadness (0.15), Surprise (0.5)” as its affect category frequency. Some specific character sequences appeared in different emoticons more than 100 times. Several sample components along with their affect category counts and frequencies are exhibited in Table 7. To account for Chinese emoticons containing triplets that did not exist in the constructed lexicon, we further separately modeled the possible affects of each individual part of a triplet $\{E_L\}$, $\{M\}$, and $\{E_R\}$ using a similar process. These separately modeled triplets or single parts were used to infer the semantic meanings of emoticons in the Chinese context.

3.4 Emoticon Extraction

To extract emoticons from Chinese texts, we leveraged the constructed emoticon lexicon with triplets and other semantic components in the following four steps:

- Step 1** Search the emoticon lexicon to find if the text contains any integral emoticon. If yes, then output the emoticon and end the procedure; if no, then go to Step 2.
- Step 2** Search the triplet lexicon to find if the text contains any triplet. If yes, then go to Step 4; if no, then go to Step 3.
- Step 3** Search the component lexicons $\{E_L M\}$, $\{M E_R\}$ and $\{E_L E_R\}$ to find if a partial triplet exists. If no, then return null and end the procedure; if yes, then, depending on the component that is missing, search one of the lexicons $\{E_L\}$, $\{M\}$ or $\{E_R\}$ to see if an integral triplet can be constituted in the text. If yes, then use the integrated triplet and go to Step 4; if no, then use the partial triplet and go to Step 4.
- Step 4** Use the regular expression “ $(.*)\{B_1\}(.)\{\text{triplet}\}(.)\{B_2\}(.)$ ” to extract the remaining components $\{S_1\}$ to $\{S_4\}$, where $\{B_1\}$ and $\{B_2\}$ are typically parentheses. Output the emoticon.

The detailed algorithm is shown in Algorithm 1.

ALGORITHM 1: The AZEmo Emoticon Extraction Algorithm

Input: Message/Sentence
Output: Extracted emoticon/partial emoticon

if *input contains an entry in emoticon dictionary* **then**
 return *that entry*;

else
 find a triplet from the input based on triplet lexicon;
 if *a triplet is found* **then**
 go to RegEx module;
 else
 if *a partial triplet is found* **then**
 go to RegEx module;
 else
 return *Null*;
 end
 end
Regex Module:
 $(B_1, B_2) = \text{localize_borders}(\text{input});$
 $\text{regex} = "(*)" + B_1 + "(*)" + \text{triplet} + "(*)" + B_2 + "(*)";$
 $(S_1, S_2, S_3, S_4) = \text{regex.match}(\text{input});$
 return *matching output*;

end

3.5 Emoticon Classification

The emoticon classification was conducted in an improved heuristic method and a machine-learning method based on feature vector extraction, which is this article's principal contribution to the emoticon research field. The classification procedure was based on the previous extraction procedure, where kinesics components $\{E_LME_R\}$, $\{S_1\}$, $\{S_2\}$, $\{S_3\}$, $\{S_4\}$ were separately located in the extracted emoticon.

3.5.1 Heuristic Method: The Classification Algorithm. For the first method, we calculated the probability of each component being used to express certain affects from the lexicon using affect category frequencies and inferred the type of the most probable affect that is expressed by the emoticon. This procedure was similar to the frequency subsection in the output calculation section of Reference [28]. The difference was that the procedure described in that work summed up the affect category frequencies over all the components to infer the affect category, whereas we looked at the component that dominates the affect expressed by the emoticon. This modification was due to our observation that, it was common that an identical triplet appeared in different emoticons expressing different affects, and what counted for the final affect were the additional components. The procedure is described as follows:

- Step 1** Based on the previous extraction procedure, the triplet and additional components $\{S_1\}$ to $\{S_4\}$ could be identified.
- Step 2** Read the affect category frequencies for the triplet and $\{S_1\}$ to $\{S_4\}$. Each component has seven values corresponding to seven affect categories. Thus, there are altogether 35 frequency values for the five components.
- Step 3** Compare the 35 values and find the largest one. Fetch the affect category for the largest value and output it as the affect category for the entire emoticon.

ALGORITHM 2: The AZEemo Heuristic Emoticon Classification Algorithm**Input:** Emoticon(triplet, S1, S2, S3, S4)**Output:** Affect Class for the Emoticon(happiness[0], sadness[0], fear[0], disgust[0], anger[0], surprise[0], fondness[0]) =
affect_category_frequency_for_triplets(triplet);**for** $i = 1$ to 4 **do**(happiness[i], sadness[i], fear[i], disgust[i], anger[i], surprise[i], fondness[i]) =
affect_category_frequency_for_si(s[i]);**end**largest_frequency = max(happiness[0~4], sadness[0~4], fear[0~4], disgust[0~4], anger[0~4], surprise[0~4],
fondness[0~4]);**for** $i = 1$ to 4 **do****if** largest_frequency == happiness[i] **then**
 return "happiness";**else****if** largest_frequency == sadness[i] **then**
 return "sadness";**else**

.....(omitted)

if largest_frequency == fondness[i] **then**
 return "fondness";**end****end****end****end**

The detailed algorithm is also shown in Algorithm 2.

3.5.2 Machine-learning Method: Emoticon Feature Vector Representation. Machine-learning methods were seldom applied onto emoticon classification in past literature. Reference [35] was the first to utilize kernel methods based on substring similarity for emoticon classification. However, few studies have extracted feature vectors from emoticons for emoticon classification. Since the kinesics model has provided rich information on the components of emoticons, it is potentially appropriate to involve the kinesics model in feature vector extraction to better describe the emotional nature of emoticons.

To the best of our knowledge, extracting comprehensive and relevant feature vectors from emoticons is a novel method in emoticon classification with few preceding studies. As an exploratory work on feature extraction for emoticons, we analyzed the status and nature of emoticons and developed a set of features on our own, including 4 structural features, 7 categorical features, and 35 kinesics features. Table 8 summarizes the complete list of the features involved, which are explained in detail in the subsequent subsection.

Extracting structural, syntactic and semantic features is widely adopted in text mining and sentiment analysis research. However, since the nature of emoticons is different from language sentences, some features (e.g., syntactic features) may be not applicable to emoticons. Hence, we developed and reorganized the features extracted from the emoticons, and categorized them as three types: Structural Features, Categorical Features, and kinesics Features.

The *Structural Features* take the structural information of emoticons into account. This type includes the total length of the emoticon, the proportion of ASCII characters, the number of different types of characters, and the number of the most frequently recurring character in the emoticon.

Table 8. List of Features Extracted from Emoticons

Type	Feature	Explanation
Structural Features	length	Total length of emoticons
	proportionASCII	Proportion of ASCII characters appeared in emoticon
	charTypes	Number of different types of characters in emoticons
	maxRecurringCharCount	Number of character that occurs the most frequently in emoticons
Categorical Features	freq_{h,s,f,a,d,u,l}	Sum of category frequencies for all characters used in emoticons
Kinesics Features	triplet_{h,s,f,a,d,u,l}	Affect category frequencies on seven categories for triplet appeared in emoticons
	s1_{h,s,f,a,d,u,l}	Affect category frequencies on the seven categories for specific additional component S_1 appeared in emoticons
	s2_{h,s,f,a,d,u,l}	Affect category frequencies on the seven categories for specific additional component S_b appeared in emoticons
	s3_{h,s,f,a,d,u,l}	Affect category frequencies on the seven categories for specific additional component S_3 appeared in emoticons
	s4_{h,s,f,a,d,u,l}	Affect category frequencies on the seven categories for specific additional component S_4 appeared in emoticons

Note: h: happiness, s: sadness, f: fear, a: anger, d: disgust, u: surprise, l: fondness

These factors can be related to the affect conveyed by emoticons, based on an observation that emoticons with fewer characters, higher proportion of ASCII characters, and fewer types of characters are considered structurally simpler, potentially related to certain affect categories, whereas emoticons with more recurring characters are considered to be conveying more intense affects.

The *Categorical Features* are drawn from emoticons by summing up the affect category frequencies for each of the characters in the emoticon. This feature was first developed in Reference [43]. We divide the occurrence of a character in a specific affect category by the total occurrence of the character across all the emoticons to get the frequency of this character expressing that specific affect.

The *Kinesics Features* are drawn from emoticons by calculating the affect category frequencies for triplets and additional components. The affect of an emoticon is constituted on its kinesics components, which suggests that the affect category frequencies of the kinesics components should explain a large proportion for the affect of the emoticon. If two kinesics components have similar affect category frequencies across the seven affect categories, then they are considered to have similar functionalities for constructing emoticons. Triplets $\{E_LME_R\}$ and additional components $\{S_1\}$ to $\{S_4\}$ are taken into consideration. Each of the five components has seven kinesics features, corresponding to the frequency for each affect category. For those emoticons not containing certain components, the features are left null. For instance, emoticon “(^_^)” contains only a triplet without additional components. Thus, it will have null feature values for $\{S_1\}$ through $\{S_4\}$. Altogether, 35 feature values (some null) are represented in the feature vector.

After the extraction of feature vectors from the emoticons, we apply two state-of-the-art machine-learning algorithms, Naïve Bayes and Random Forest, via Weka [13]. The classification results for both heuristics and machine-learning methods are presented in the Section 4.

4 RESEARCH TEST-BEDS AND EVALUATIONS

4.1 Review on Chinese Social Media, E-commerce Sites, and Research

Chinese social media and E-commerce sites have experienced significant growth in recent years. In 2014, China's E-commerce market value has reached \$1,980 billion with a 21.3% growth rate compared with 2013. Online retail revenue took up \$438 billion, accounting for 22% of the market value. Research agencies such as iResearch are expecting continuous growth of this market. According to Cicero Group's report on Chinese social media, 60% of Internet users in China access at least one social media platform every day. The largest Chinese microblog site, Weibo, has 175.7 million Monthly Active Users (MAU). QQ and WeChat, two of the most used IM software by the Chinese, have 800 million and 500 million MAU, respectively. Internet users are generating huge volumes of data that can be used in all sorts of research. With the growing significance of Chinese social media and E-commerce, an increasing number of studies involved Chinese test-beds (e.g., Weibo data). For instance, Nie et al. [24] presented a scheme to learn descriptive tags for each question in Chinese and English question answering forums. Quan et al. [29] used corpora crawled from two Chinese news portals to evaluate their models for social emotion detection. Bao et al. [2] collected Chinese news from Sina to construct a social emotion miner using latent Dirichlet allocation (LDA). Su et al. [33] used Chinese online reviews for opinion mining research. Jiang et al. [18] collected Renren (a Chinese social network site) and Tencent Weibo (a Chinese microblog site) data for developing a recommendation system with social contextual information. Ou et al. [25] used Tencent Weibo for constructing a test-bed for a heterogeneous hashing topic, and Jiang et al. [17] used Tencent Weibo as well to model synchronized behaviors in large networks using graph mining. Tang et al. [36], Wang et al. [40], and Wang et al. [41] collected Weibo data for sentiment analysis. Zhao et al. [51] developed an emoticon-based sentiment analysis system for Chinese tweets, which showed high relevance to our research.

4.2 Research Test-beds

To test whether our proposed system was effective in extracting and classifying emoticons in a Chinese context, we gathered two large sets of test-bed corpora from the Internet. The first test-bed was a health related corpus collected from two video sharing websites. The second test-bed was in an E-commerce related context gathered from two online E-commerce websites. These two test-beds were used in emoticon extraction and classification tasks. Information about the test-beds is discussed below.

4.2.1 Test-bed: Health and Fitness Social Media. The health dataset was gathered from real-time video comments and replies from two large, well-known Chinese video sites, Bilibili (www.bilibili.tv) and AcFun (www.acfun.tv), which are popular video sharing sites among Chinese young people. These two websites have around 8.8 million registered users and on average 158 million user access per week. Users can upload user-generated video clips onto these websites to share their creativity, attitudes, and feelings with other users. We specifically focused on topics relating to food and sports. young people interested in these topics often share their thoughts and feelings while watching these videos, thus allowing us to obtain a corpus with relatively rich emoticons. In total, 1,003,244 comments were collected and the statistics are summarized in Table 9.

Within the health and fitness test-bed, we sampled 985 comments from those with emoticons and 1,015 comments from those without emoticons (2,000 in total). This is for the evaluation purpose as a relatively balanced dataset is preferred by machine-learning algorithms. For those comments with emoticons, one of the seven-type affect categories was assigned to them: happiness, sadness, fear, disgust, anger, surprise, or fondness. The annotations were done by two individual

Table 9. Statistics for Health and Fitness Test-beds

Number of comments	Acfun.tv	Bilibili.tv
Food related	463,434	215,144
Sports related	253,780	70,866

Table 10. Annotation Statistics for the Two Test-beds

Number of annotations	Happiness	Sadness	Fear	Disgust	Anger	Surprise	Fondness
Health and Fitness	383	391	10	7	79	50	65
E-commerce and ACG	307	131	30	26	16	36	54

annotators who are familiar with Chinese social media and not related to the research. Each annotator annotated the 2,000 comments separately. They read the whole comments and judged the affect of the emoticons in the comments. The Cohen's Kappa coefficient was 0.957. The disputation resolving process was as follows. First, each annotator individually judges the affect category for emoticons appearing in the comments. The two annotators then meet and compare their judgments and mark out inconsistent ones. They discuss about the affects conveyed by emoticons in the comments to figure out whether they can reach an agreement on the conflicts. For those with no agreements, the annotation administrator is involved in the discussion. After another round of discussion, the conflicts are resolved by a majority vote.

4.2.2 Test-bed: E-commerce and ACG. We gathered our E-commerce related test-bed from Taobao (taobao.com) and Tmall (tmall.com), two of the biggest online C2C and B2C E-commerce websites in China. These sites had 307 million active buyers in the fourth quarter, Year 2014. According to statistics provided by Alexa, Taobao.com and Tmall.com ranked as the second and sixth among all the websites in China and as the top two among all the E-commerce websites. These statistics indicated a large user base from which we could expect to gather rich emoticon corpus. We focused on Animations, Comics, and Games (ACG) related shops, since sellers and customers of these shops are mainly young people who are more inclined to using emoticons. We used product reviews as the target texts to build the corpus, because, in product reviews, customers often express their objective judgment as well as subjective emotions and sellers respond accordingly. From the top-10 shops with the highest credit ratings, we collected 1,200 reviews. Following the same annotation process, we sampled 600 reviews from those containing emoticons and the remaining 600 from those not containing emoticons. Another two individual annotators finished the annotation. The Cohen's Kappa coefficient was 0.963. The disputation resolving process was identical as that of the Health and Fitness test-bed.

Annotation statistics for the two datasets can be found in Table 10. We can observe different distributions of emoticon affects across the two datasets. Happiness is the majority in the E-commerce and ACG dataset, whereas happiness and sadness are the two dominant affect categories in the health and fitness dataset. This might reveal the different patterns of users' usage of emoticons on the two platforms.

Although it is possible that a single comment contains multiple emoticons, such cases are relatively rare. As the focus of this research is on the affect classification on single emoticons but not on studying the effect of multiple emoticons appearing in one sentence, we did not include those comments with multiple emoticons in our datasets. In other words, we focused on sentences with only one emoticon.

Table 11. Results for Emoticon Extraction

Dataset	Precision	Recall	F-measure
Health and Fitness	99.4%	90.9%	95.0%
E-commerce and ACG	95.3%	83.8%	89.2%

4.3 Evaluation of the Emoticon Extraction Task

As the baseline for our study, we evaluated the performance of AZEmo on emoticon extraction by comparing the results of AZEmo with manual annotations on the test dataset. Results of precisions, recalls and F-measures are summarized in Table 11.

The overall performance of AZEmo on the emoticon extraction task was promising, with F-measure between 89.2% and 95.1%. The high precision rate indicates that emoticons often consist of very unique characters that are elsewhere unused. Therefore, if any entry in our lexicon, including complete emoticons or parts of the emoticons (e.g., triplets), appeared in the given Chinese sentence, then the probability of those characters being used as emoticons would be higher than those used for any other purposes. The remaining small portion indicates, however, that there indeed exist some entries in the lexicon not fully served for emoticons in the Chinese context. These were instead used as plain numbers or characters with other specific meanings. For example, “00” and “==” could both represent two eyes, or, respectively, could also be interpreted as two consecutive zeroes in numbers, or could be used as an equal sign in programming.

The relatively lower recall indicated that AZEmo was weaker in recognizing different characters with high visual similarity. For example, “V” and “^O^” could not be properly recognized and extracted, although the lexicon contained numerous entries containing “V” and “^O^”. Also, the system could not recognize those characters not included in the lexicon, and those characters in most cases were unique to the Chinese encoding system, such as “丿” and “㇏”. Occasionally we could even find in the emoticon some characters outside the range of CJK (Chinese, Japanese, and Korean) encoding system, such as “ཀ” from the Tibetan alphabet, in the emoticon. To improve the AZEmo system, these encoding problems must be addressed properly in the future.

4.4 Evaluation of the Emoticon Classification Task

To evaluate the effectiveness of the proposed system in classifying extracted emoticons, we compared our system with six benchmark classification systems used in prior research:

Benchmarks 1, 2, and 3. Classification based on the unigram, bigram and trigram models, implemented in Reference [43]. In this benchmark, the number of times each character was assigned to each affect category was counted without considering its position in the kinesics model, i.e., “o” in “^o^” and “o_o” were treated equally. These benchmarks considered the affect category frequencies of each n-gram in the emoticon and inferred the affect category for the whole emoticon. These benchmarks were further denoted as “Uni,” “Bi,” and “Tri,” respectively.

Benchmark 4. Heuristic classification method proposed in Reference [28]. This method used the summed statistics of affect categories over all the components, rather than the statistics of the dominating component in the emoticon. It was further denoted as “OH” (Old Heuristics).

Benchmarks 5, 6, 7, and 8. Machine-learning classification methods without kinesics features. In these benchmarks, each feature vector contained 4 structural features and 7 categorical features. Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SVM),

Fig. 2. Accuracies for emoticon classification: (a) lexicon, (b) health and fitness, (c) e-commerce and ACG.

Table 12. Paired *t*-test Results between AZEmo Methods vs. Corresponding Benchmarks

Methods	Lexicon	Health and Fitness	E-commerce and ACG
IH vs. OH	12.0**	2.42*	4.56**
NB vs. NB/k	19.2**	6.05**	4.84**
RF vs. RF/k	48.7**	2.10*	1.75 ⁺
SVM vs. SVM/k	78.1**	11.8**	4.80**
LR vs. LR/k	75.9**	15.4**	5.05**

⁺p-value < 0.1; *p-value < 0.05; **p-value < 0.01.

and Logistic Regression (LR) were applied in this benchmark set. Other experimental settings remained the same with the methods used in AZEmo. These benchmarks were further denoted as “NB/k” (NB without kinesics), “RF/k” (RF without kinesics), “SVM/k” (SVM without kinesics), and “LR/k” (LR without kinesics), respectively.

Three methods were used in our AZEmo system: the improved heuristic method (IH), and NB, RF, SVM, and LR, with kinesics features, respectively. The classification procedure only dealt with the extracted emoticon itself, without considering the surrounding text. Three datasets were used to evaluate the classification task: the original lexicon dataset of 10,560 emoticons, the health and fitness dataset of 985 emoticons, and the E-commerce and ACG dataset of 600 emoticons. Each instance in the three datasets belonged to a confirmed affect category (tagged or annotated). Each system inferred the affect category compared to ground truth. Accuracy was calculated based on the proportion of correct inferences. Each dataset was stratified into 10 equal proportions for further *t*-tests on classification accuracies. Tenfold cross-validation tests were performed on the lexicon dataset to avoid over-fitting. Paired *t*-tests were conducted between AZEmo methods (IH, NB, RF, SVM, LR) and benchmarks (Uni, Bi, Tri, OH, NB/k, RF/k, SVM/k, LR/k).

Figure 2 shows the classification accuracies for AZEmo and benchmark systems. Accuracies for methods used by the AZEmo system are shown in black bars, in comparison with grey bars that denote benchmark systems. Table 12 shows the paired *t*-test results testing whether AZEmo methods outperform corresponding benchmark methods without improvement (i.e., IH vs. OH, NB vs. NB/k, RF vs. RF/k, SVM vs. SVM/k, LR vs. LR/k).

The classification accuracies of methods used by AZEemo significantly outperformed the benchmarks in the original lexicon (our methods: between 79.4% and 88.5%; benchmarks: between 45.8% and 69.5%) and health and fitness dataset (our methods: between 71.6% and 76.6%; benchmarks: between 40.1% and 68.6%). The relatively low accuracy for n-gram methods suggests that a pure n-gram algorithm is insufficient for describing the structure of emoticons, especially longer ones. For instance, the unigram “o” in “o_o” and “^o^” are in different positions (eye vs. mouth) and thus should express different affects. The first one is a surprised face, whereas the second one is a happy face. Without specifying the character into detailed kinesics components, the affect it conveyed could not be properly analyzed. The relatively low accuracy for the old heuristic method suggests that the analysis of dominating component in the emoticon can improve the performance of the system. For example, “Σ (° ▯ °)” was a surprised face, “(# ° ▯ °)” an angry face, and “↘ (° ▯ ° ;)” a disgusted face. The triplets in these emoticons were identical (“ ° ▯ ° ”), thus the additional components “Σ”, “↘”, “#”, “↘”, and “;” counted for their affect inclinations.

However, in the E-commerce and ACG dataset, our three methods did not significantly outperform the bigram and trigram methods (our methods: between 73.7% and 90.0%; bigram benchmark: 88.9%; trigram benchmark: 87.5%). A possible reason is that the E-commerce test dataset contained emoticons that were more commonly used and relatively shorter, which could be better classified by bigram and trigram. Although the extracted features can assist us better understand relatively complicated emoticons, a rule-based matching system (n-gram system) might be more straightforward and have better performance for a small set of common short emoticons. This can be a separate sub-system for our AZEemo system to improve its overall performance.

In all evaluations, IH, NB, RF, SVM, and LR significantly outperformed OH, NB/k, RF/k, SVM/k, LR/k, respectively (Table 12). This indicates that our improvement on the heuristics had a significant improvement on classification accuracy, and that kinesics features can improve classification accuracy for machine-learning methods. However, the machine-learning methods did not outperform the improved heuristics method (Machine learning: between 71.6% and 88.2%; heuristics: between 72.7% and 90.0%). This may be due to the incompleteness of the features extracted from emoticons. Improvement on the selection of feature vectors may potentially increase the classification accuracy.

5 CONCLUSION AND FUTURE DIRECTIONS

In this study, we developed a novel kinesics model with affect mapping to perform emoticon analysis for Chinese social media and E-commerce sites. We developed an emoticon analysis system, AZEemo, to extract and classify emoticons to better shed light on people’s feelings and emotional states as hidden in text. Health and fitness-related comments and E-commerce online reviews were used for empirical tests. With the adaptation from Japanese emoticon dictionaries (emoticon truncation) and the kinesics decomposition, we are able to extract emoticons from online text and analyze their affect inclinations in a Chinese context. This approach adds to the current sentiment analysis literature with a unique perspective on human emotion. It also develops a novel set of features (kinesics features) that can be used by future machine-learning algorithms, which provides a more delineated view of emoticons used in Chinese social media and E-commerce sites, a large and growing Internet community.

From a business point of view, emoticon analysis helps companies better understand their customers. For example, an analysis on online product reviews help online sellers generate a high-level report on their customers’ affect tendencies. Sellers can inspect those reviews with “angry” emoticons to identify potential problems. Another example can be on social networking platforms or online health communities. Platform providers can track users’ mood by analyzing the emoticons they are using to provide precise and personalized services, such as alerts on fierce emotional

fluctuations. Emoticon analysis can also help users improve their experience by enabling them getting aware of other users' opinion and mood. This work echoes the core of the MIS discipline, which has a concentration on how people create, use, and adopt IT artifacts in a business context.

Due to the lexicon nature of AZEmo, it was relatively weak in processing characters that did not exist in the lexicon and in recognizing visually similar characters. Emoji were also not taken into account due to their site-specific representations. Computer image recognition technology may potentially improve the system's effectiveness as well as functionality on Emoji recognition. We assumed that each emoticon only conveyed one category of affect in this research, which may actually have multiple dimensions of affects in real life, e.g., happy and sad simultaneously. A proper research framework of affect representation could be designed to resolve this problem. For the machine-learning methods in the classification task, more refined feature vectors could be drawn from emoticons to improve classification accuracies after a more comprehensive understanding of the nature of emoticons. AZEmo could also be a supplement to existing valence-based sentiment analysis techniques, providing a more comprehensive and reliable view for underlying emoticons expressed by social media users.

ACKNOWLEDGMENTS

The authors thank Cathy Larson for the proofreading and editorial suggestions.

REFERENCES

- [1] Sartaj Ahmad and Rishabh Varma. 2018. Information extraction from text messages using data mining techniques. *Malay. J. Mat. S.* 1 (2018), 26–29. DOI : <https://doi.org/10.26637/MJM0S01/05>
- [2] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2012. Mining social emotions from affective text. *IEEE Trans. Knowl. Data Eng.* 24, 9 (2012), 1658–1670. DOI : <https://doi.org/10.1109/TKDE.2011.188>
- [3] Steven Bedrick, Russell Beckley, Brian Roark, and Richard Sproat. 2012. Robust kaomoji detection in twitter. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM'12)*. Association for Computational Linguistics, 56–64.
- [4] Ray L. Birdwhistell. 1952. *Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture*. Department of State, Foreign Service Institute.
- [5] Ray L. Birdwhistell. 2010. *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press.
- [6] Brita. 2011. Sept. 19, 1982: Can't You Take a Joke? :-). Retrieved from <https://www.wired.com/2011/09/0919fahlman-proposes-emoticons/>.
- [7] Zengjie Cao and Jing Ye. 2009. Attention savings and emoticons usage in BBS. In *Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology (ICCIT'09)*. IEEE, 416–419. DOI : <https://doi.org/10.1109/ICCIT.2009.112>
- [8] Jianmei Chen. 2009. *The Construction and Application of Chinese Emotion Word Ontology*. Master's thesis. Dalian University of Technology, China.
- [9] Kuo-chan Chiu. 2007. *Explorations in the Effect of Emoticon on Negotiation Process from the Aspect of Communication*. Ph.D. dissertation. National Sun Yat-sen University.
- [10] Geeta G. Dayalani and B. K. Patil. 2014. Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets. *Int. J. Eng. Res. Gen. Sci.* 2, 6 (2014).
- [11] Daantje Derks, A. E. R. Bos, and Jasper Von Grumbkow. 2007. Emoticons and social interaction on the internet: The importance of social context. *Comput. Hum. Behav.* 23, 1 (2007), 842–849. DOI : <https://doi.org/10.1016/j.chb.2004.11.013>
- [12] Paul Ekman. 1999. Basic emotions. In *Handbook of Cognition and Emotion*, Tim Dalgleish and Mick Power (Eds.). Wiley Online Library, Chapter 3, 45–60.
- [13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newslett.* 11, 1 (2009), 10–18.
- [14] Alexander Hogenboom, Daniella Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. *J. Web Eng.* 14, 1–2 (2015), 22–40.
- [15] Faliang Huang, Shichao Zhang, Jilian Zhang, and Ge Yu. 2017. Multimodal learning for topic sentiment analysis in microblogging. *Neurocomputing* 253 (2017), 144–153. DOI : <https://doi.org/10.1016/j.neucom.2016.10.086>

- [16] Shanshan Jia, Shuling Di, and Tongrang Fan. 2013. Text emotion analysis model based on emoticons and emotional words. *J. Hebei Acad. Sci.* 30, 2 (2013), 11–15.
- [17] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2016. Catching synchronized behaviors in large networks: A graph mining approach. *ACM Trans. Knowl. Discov. Data* 10, 4, Article 35 (Jun. 2016), 27 pages. DOI: <https://doi.org/10.1145/2746403>
- [18] Meng Jiang, Peng Cui, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2014. Scalable recommendation with social contextual information. *IEEE Trans. Knowl. Data Eng.* 26, 11 (2014), 2789–2802. DOI: <https://doi.org/10.1109/TKDE.2014.2300487>
- [19] Aditya Joshi, A. R. Balamurali, Pushpak Bhattacharyya, and Rajat Mohanty. 2011. C-feel-it: A sentiment analyzer for micro-blogs. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations (HLT'11)*. Association for Computational Linguistics, 127–132.
- [20] Kun-lin Kl Liu, Wu-jun Wj Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the Annual AAAI Conference on Artificial Intelligence (AAAI'12)*. 1678–1684.
- [21] Xiangxi Liu. 2015. *The Linguistic Analysis of Chinese Emoticon*. Ph.D. dissertation. University of Massachusetts, Amherst.
- [22] Jack M. Maness. 2008. A linguistic analysis of chat reference conversations with 18-24 year-old college students. *J. Acad. Libr.* 34, 1 (2008), 31–38. DOI: <https://doi.org/10.1016/j.acalib.2007.11.008>
- [23] Junpei Nakamura, Takeshi Ikeda, Nobuo Inui, and Yoshiyuki Kotani. 2003. Learning face marks for natural language dialogue systems. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 180–185. DOI: <https://doi.org/10.1109/NLPKE.2003.1275891>
- [24] Liqiang Nie, Yi-Liang Zhao, Xiangyu Wang, Jialie Shen, and Tat-Seng Chua. 2014. Learning to recommend descriptive tags for questions in social forums. *ACM Trans. Inf. Syst.* 32, 1 (Jan. 2014), 1–23. DOI: <https://doi.org/10.1145/2559157>
- [25] Mingdong Ou, Peng Cui, Fei Wang, Jun Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Comparing apples to oranges: A scalable solution with heterogeneous hashing. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. ACM, 230–238. DOI: <https://doi.org/10.1145/2487575.2487668>
- [26] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*. 466–475.
- [27] S. Poongodi and N. Radha. 2013. Classification of user opinions from tweets using machine learning techniques. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 3, 9 (2013), 1061–1065.
- [28] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka, and Kenji Araki. 2010. CAO: A fully automatic emoticon analysis system based on theory of kinesics. *IEEE Trans. Affect. Comput.* 1, 1 (2010), 46–59. DOI: <https://doi.org/10.1109/T-AFFC.2010.3>
- [29] Xiaojun Quan, Qifan Wang, Ying Zhang, Luo Si, and Liu Wenyin. 2015. Latent discriminative models for social emotion detection with emotional dependency. *ACM Trans. Inf. Syst.* 34, 1 (Jul. 2015), 1–19. DOI: <https://doi.org/10.1145/2749459>
- [30] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*. Association for Computational Linguistics, 43–48. DOI: <https://doi.org/10.3115/1628960.1628969>
- [31] Landra L. Rezabek and John J. Cochenour. 1998. Visual cues in computer-mediated communication: Supplementing text with emoticons. *J. Vis. Lit.* 18, 2 (1998), 201–215.
- [32] Yuita Arum Sari, Evy Kamilah Ratnasari, Siti Mutrofin, and Agus Zainal Arifin. 2014. User emotion identification in twitter using specific features: Hashtag, emoji, emoticon, and adjective term. *J. Ilmu Komput. Inf.* 7, 1 (2014), 18–24.
- [33] Qi Su, Xinying Xu, H. Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. 2008. Hidden sentiment association in chinese web opinion mining. In *Proceeding of the 17th International Conference on the World Wide Web*. ACM, 959–968. DOI: <https://doi.org/10.1145/1367497.1367627>
- [34] Nobuo Suzuki and Kazuhiko Tsuda. 2006. Express emoticons choice method for smooth communication of e-business. In *Proceedings of International Conference on Knowledge-Based and Intelligent Information and Engineering Systems 2006, Part II*, LNAI 4252. Springer, Berlin, 296–302. DOI: https://doi.org/10.1007/11893004_38
- [35] Yuki Tanaka, Hioya Takamura, and Manabu Okumura. 2005. Extraction and classification of facemarks with kernel methods. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*. 28–34. DOI: <https://doi.org/10.1145/1040830.1040847>
- [36] Jie Tang, Sen Wu, and Jimeng Sun. 2013. Confluence: Conformity influence in large social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. ACM, 347–355. DOI: <https://doi.org/10.1145/2487575.2487691>
- [37] Mike Thelwall. 2017. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*. Springer, Cham, 119–134. DOI: https://doi.org/10.1007/978-3-319-43639-5_7

- [38] Yuki Urabe, Rzepka Rafal, and Kenji Araki. 2013. Emoticon recommendation for japanese computer-mediated communication. In *Proceedings of the 2013 IEEE 7th International Conference on Semantic Computing (ICSC'13)*. IEEE, Los Alamitos, CA, 25–31. DOI: <https://doi.org/10.1109/ICSC.2013.15>
- [39] Joseph B. Walther and Kyle P. D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Soc. Sci. Comput. Rev.* 19, 3 (2001), 324–347. DOI: <https://doi.org/10.1177/089443930101900307>
- [40] Senzhang Wang, Xia Hu, Philip S. Yu, and Zhoujun Li. 2014. MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, 1246–1255. DOI: <https://doi.org/10.1145/2623330.2623728>
- [41] Yaqiong Wang, Hongfu Liu, Hao Lin, Junjie Wu, Zhiang Wu, and Jie Cao. 2013. SEA: A system for event analysis on chinese tweets. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. ACM, 1498–1501. DOI: <https://doi.org/10.1145/2487575.2487719>
- [42] Alecia Wolf. 2000. Emotional expression online: Gender differences in emoticon use. In *CyberPsychology & Behavior*. Vol. 3. Mary Ann Liebert, Inc., 827–833. DOI: <https://doi.org/10.1089/10949310050191809>
- [43] Taichi Yamada, Seiji Tsuchiya, Shiongo Kuroiwa, and Fuji Ren. 2007. Classification of facemarks using N-gram. In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'07)*. IEEE, 322–327. DOI: <https://doi.org/10.1109/NLPKE.2007.4368050>
- [44] Yuki Yamamoto, Tadahiko Kumamoto, and Akiyo Nadamoto. 2015. Multidimensional sentiment calculation method for Twitter based on emoticons. *Int. J. Perv. Comput. Commun.* 11, 2 (2015), 212–232. DOI: <https://doi.org/10.1108/IJPC-03-2015-0019>
- [45] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 275–278. DOI: <https://doi.org/10.1109/WI.2007.51>
- [46] Ken Yasumoto-Nicolson. 2007. :-). Turns 25, But How Old Are Japanese Emoticons (?_?). Retrieved from <https://whatjapanthinks.com/2007/09/19/turns-25-but-how-old-are-japanese-emoticons/>.
- [47] Takeru Yokoi, Mizuki Kobayashi, and Roliana Ibrahim. 2015. Emoticon extraction method based on eye characters and symmetric string. In *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2979–2984. DOI: <https://doi.org/10.1109/SMC.2015.518>
- [48] Shuo Yu, Hongyi Zhu, Shan Jiang, and Hsinchun Chen. 2014. Emoticon analysis for chinese health and fitness topics. In *Smart Health*. Springer International Publishing, 1–12. DOI: https://doi.org/10.1007/978-3-319-08416-9_1
- [49] Zhiwen Yu, Zhitaowang, Liming Chen, Bin Guo, and Wenjie Li. 2016. Featuring, detecting, and visualizing human sentiment in chinese micro-blog. *ACM Trans. Knowl. Discov. Data* 10, 4, Article 48 (May 2016), 23 pages. DOI: <https://doi.org/10.1145/2821513>
- [50] Zheng Yuan and Matthew Purver. 2012. Predicting emotion labels for chinese microblog texts. In *Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD'12)*. 40–47.
- [51] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 2012. MoodLens: An emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. ACM, 1528–1531. DOI: <https://doi.org/10.1145/2339530.2339772>

Received September 2017; revised March 2018; accepted December 2018