Predicting author age from Weibo microblog posts

Wanru Zhang, Andrew Caines, Dimitrios Alikaniotis, Paula Buttery
Department of Theoretical & Applied Linguistics
University of Cambridge
wrzhang2010@gmail.com, apc38@cam.ac.uk, da352@cam.ac.uk, pjb48@cam.ac.uk

Abstract

We report an author profiling study based on Chinese social media texts gleaned from Sina Weibo (新浪微博) in which we attempt to predict the author's age group based on various linguistic text features mainly relating to non-standard orthography: classical Chinese characters, hashtags, emoticons and kaomoji, homogeneous punctuation and Latin character sequences, and poetic format. We also tracked the use of selected popular Chinese expressions, parts-of-speech and word types. We extracted 100 posts from 100 users in each of four age groups (under-18, 19-29, 30-39, over-40 years) and by clustering users' posts fifty at a time we trained a maximum entropy classifier to predict author age group to an accuracy of 65.5%. We show which features are associated with younger and older age groups, and make our normalisation resources available to other researchers.

Keywords: Weibo, microblog linguistics, text forensics, computational sociolinguistics

1 Introduction

The area of digital text forensics continues to interest computational linguists, information engineers and web technologists alike. This interest is based on the hypothesis that demographic factors correlate with language use to some discernible extent, a hypothesis that was famously supported by William Labov's surveys in 1960s New York (Labov, 1966), and has continued to gain support from various sociolinguistic works (Trudgill, 2011; Wieling et al., 2014; Hovy, 2015).

Previous work in text forensics has for the most part focused on age and/or gender prediction via shallow text features such as word, character and part-of-speech n-grams (Mukherjee and Liu, 2010; Nguyen et al., 2011; Peersman et al., 2011). Here, we attempt to predict the age groups of Chinese microblog authors using such text features, along with features relating to non-standard use of orthography – *e.g.* emoticons, hashtags, and repeated characters.

Our data source is Sina Weibo (新浪微博), the most popular microblogging service in China with 212 million monthly active users in June 2015[1]. Sina Weibo is one of several Chinese microblogging sites, with *weibo* actually meaning 'microblog' in English; others include Tencent Weibo, Sohu Weibo, and NetEase Weibo. However, such is Sina Weibo's preeminence among its competitors that it is now commonly referred to as 'Weibo', a practice we follow henceforth.

We collected Weibo posts from users in four age groups: under-18, 19 to 29, 30 to 39, and over 40 years old. We extracted linguistic and orthographic features from these posts, trained a maximum entropy classifier, and achieved an F-measure of 65.5% by treating fifty posts from each user as a single document. The implication is that, with microblogs being characteristically so brief, there is much benefit in agglomerating many posts from a single user, for improved machine learning and classification.

[1]Source: China Internet Watch, accessed 2015-10-12. This compares to Twitter's MAU count, 316m, in the same period (source: Twitter, accessed 2015-10-12).

2 Weibo normalisation

We collected 40,000 posts from Weibo users: 100 posts from 100 users in four age groups. As has been found with data from Twitter, microblog texts are noisy (Baldwin et al., 2013; Eisenstein, 2013). Solutions to this problem fall into two types of approaches. Firstly we may adapt the NLP tools to the social media domain, an avenue that has been explored for posts from Twitter in English and Irish, for example (Foster et al., 2011; Derczynski et al., 2013; Kong et al., 2014; Plank et al., 2014; Lynn et al., 2015).

However, as Eisenstein asks, "is domain adaptation appropriate for social media?" Building on Darling and colleagues' comment "that social media is not a coherent domain at all" in the context of part-of-speech (POS) tagging, Eisenstein adds: "Twitter itself is not a unified genre, it is composed of many different styles and registers", not to mention languages (Eisenstein, 2013; Darling et al., 2012). For instance, Jørgensen and colleagues find that Twitter-adapted POS-taggers perform poorly on a non-standard variety of English – in their case, African-American Vernacular English (Jørgensen et al., 2015).

Alternatively then, we can normalise the data to better suit the existing tools and the genre of standard language they are trained upon (Han et al., 2013; Saloot et al., 2015). In the absence of domain-specific (and, ideally, dialect-versatile) NLP tools for Chinese social media we adopt a 'clean up' approach and normalise the Weibo posts to a certain extent.

To achieve this aim we designed a semi-supervised workflow controlled by an R script (R Core Team, 2015), which prompted the human operator for occasional input but otherwise ran autonomously. Supporting resources were created in the form of character

maps and word-lists. The script and supporting resources described here are made available in a GitHub repository[2].

Weibo posts are rich in the kinds of non-standard orthography that characterises microblogging in other languages. For instance, non-standard spelling, punctuation and characters, hashtags, emoticons and code-switching. We illustrate some of these features in examples (1)-(4) below.

(1) 晚上好。[握手][握手][握手]
'Good evening [shake hands] [shake hands] [shake hands]'
(2) 我的错!!!!
'My fault!!!!'
(3) 心疼 TTTTTTT
'Distressed'
(4) #2015 亚洲杯 #
'#2015AsianCup'

In (1) we see the square-bracketed format that typifies emoticons in Chinese. In (2) we find repeated punctuation, and in (3) repeated Latin characters ('T' here indicating tearfulness). A Chinese hashtag, bookended by a pair of hash characters, is demonstrated in (4). We now describe the steps taken to normalise Weibo posts.

2.1 Classical characters

Chinese may be written using classical or modern characters. We note that NLP tools may be adversely affected by the presence of classical characters. The post (5), for example, is segmented as (6)[3] and then tagged as `#VA #P #VV` (predicative verb, preposition, other verb) (7) by Stanford NLP tools (Tseng et al., 2005a; Tseng et al., 2005b)[4]. However, by replacing the classical character 无 with its modern equivalent 没 results in the correct segmentation and subsequent desired `#VE #NN #VV` (you3 as main verb[5], other noun, other verb) tagging by Stanford NLP (8).

(5) 无理由转
'No reason to turn'
(6) 无理. 由. 转
(7) 无理 #VA . 由 #P . 转 #VV
(8) 没 #VE . 理由 #NN . 转 #VV

Not all classical characters cause this sort of problem, but nevertheless we adopt a cautious approach and substitute all identified classical characters with modern versions. We selected the fifty highest ranking classical characters from Jun Da's frequency list[6] that we encountered in our dataset, and attempted to match them with modern equivalents using *The Contemporary Chinese Dictionary* (Lansheng et al., 2012).

For the majority (39) there is a context-free one-to-one mapping, and for these the subsitution was an unsupervised process; for the remaining 11 classical characters in our set which have a number of modern translations depending on context, the operator was prompted to input an appropriate modern version.

2.2 Emoticons

Whereas in other languages – English for instance – emoticons ('emotion icons') are formed by a sequence of punctuation characters through which the facial expression is represented at a 90 degree anti-clockwise rotation (*e.g.* ;-(:-/ :-p), in Chinese emoticons are alphabetic characters enclosed by square brackets and therefore do not index facial expressions but rather indicate emotions directly: *e.g.* [哈哈] 'laughter', [泪] 'tears', [偷笑] 'giggle', [爱你] 'love', [心] 'heart'.

Even though they are recognisable Chinese characters, emoticons are meta-linguistic and therefore we remove them so that they are not included in the segmentation or part-of-speech tagging process. This step involves a straightforward deletion of any pair of square brackets along with any characters they enclose.

2.3 Kaomoji

Of similar function to emoticons but of different form are so-called 'kaomoji' (from Kanji: kao 顔 ('face'), moji 文字 ('character')). These are Japanese-style emoticons designed to represent facial expressions straight on (not rotated like the emoticons of Latin alphabet languages). For example, the joyful (^ω^) in which the parentheses represent the outline of the face, the carets are the eyes, and omega represents the mouth; the surprised (⊙_⊙); or the annoyed (#><) with the hash character indicating wrinkles.

Variation within and around this basic structure has given rise to many thousands of kaomoji. We obtained a list of kaomoji from an online resource[7] and selected a subset of highly frequent list of 748 which we would identify and remove from Weibo texts.

2.4 Hashtags

In social media of the 'western world' – Facebook, Instagram, Twitter, *etc* – hashtags begin with the hash character (#) and proceed in a rightwards fashion until whitespace, line-end or a non-alphanumeric character are encountered. In Chinese, where whitespace is less frequently used to segment words, the social media convention is to bind the hashtag in a pair of hash characters (*e.g.* #2015 亚洲杯 # '#2015AsianCup').

[2] `https://github.com/cainesap/sino-nlp`

[3] In this and following segmented examples we insert a full stop (period) and parenthetical whitespace characters for clarity.

[4] Part-of-speech tags are of the format, hash character plus part-of-speech, according to the conventions and tagset from the Penn Chinese Treebank (Xia, 2000).

[5] The tag `#VE` is reserved for 没 'not have' along with 有 'have' as possessive or existential you3, which analysis is somewhat debated, and therefore these tokens may easily be extracted from a corpus (Xia, 2000).

[6] Source: `http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=CL`, accessed 2016-02-15.

[7] Source: `http://kaomoji.ru/en`; accessed 2016-02-14.

To normalise the texts we removed any such hash character parentheses including the tag they enclose.

2.5 Latin characters

Weibo users make use of characters from the Latin alphabet to supplement Chinese characters. We remove all Latin characters and in the process note any occurrence of homogenous consonant clusters. These are sequences of a repeated consonant not mixed with any other Latin character, such as 'hhh' and 'ttt', which represent laughter ('h' coming from 'haha'), and tears respectively (for the visual similarity of the letter 'T' to a falling tear; both uppercase and lowercase 't' are used).

2.6 Punctuation

Finally, we remove all punctuation clusters from the Weibo posts, and reduce them to single sentence boundary markers where applicable, while again tracking their occurrence as a feature extraction process. For example, this means that we reduce punctuation sequences such as !!!!!, ????, and 。。。。。 to !, ?, and 。 .

2.7 Normalisation: overview

By processing the Weibo posts with the steps described above, a post such as (9), with a hashtag, emoticon and punctuation sequence (boxed), may be transformed into a normalised version (10).

(9) # 春晚'# . [爱你] . 春晚. 很. 好看 ~~
'#SpringFestivalGala [love] The Spring Festival Gala is very good ~~'
(10) 春晚. 很. 好看

Or from (11), with its kaomoji and punctuation, we reach (12):

(11) 哇. 哦. ⊙ω⊙ . 帅呆了 ~
'Oh wow :-) awesome ~'
(12) 哇. 哦. 帅呆了

Finally (13) contains the classical character 莫 'do not', and a sequence of Latin 'h' characters (for laughter); the processed version is shown in (14):

(13) 我. 莫 . 名. 觉得. 好心. 酸. 啊 hhhhhhhhhh
'I feel so inexplicably sad [laughter]'
(14) 我. 不. 要. 名. 觉得. 好心. 酸. 啊

2.8 Evaluation

What difference do the above normalisation measures make to the processing of Weibo texts? In this section we consider this question in two steps, firstly examining non-standard orthography across our Weibo user age groups (section 2.8.1.), before an analysis of segmentation and POS-tagging errors before and after normalisation (section 2.8.2.).

2.8.1. Age group differences

As one might expect given that in social media we are working with a genre more closely associated with younger age groups[8], and that the non-standard devices we are dealing with here are associated with social media, usage of the six orthographic features introduced in section 2 is higher among younger age groups. Figure 1 shows how many of the six features are used at least once in any given post (with no post using all six features at once). It is apparent that the two younger age groups (under-18 and 19-29 years) compose a higher proportion of Weibo posts with one or more of our chosen features, whereas the age groups over 30 have higher proportions of posts containing none of the selected non-standard features.

2.8.2. Error analysis

To assess the effect of our normalisation measures we randomly sampled one hundred Weibo posts from each of four age groups: under 18 years, 19 to 29, 30 to 39, and over 40. With these 400 posts, we considered whether there were any segmentation or POS-tagging errors if Stanford NLP tools were applied to the original texts 'as is' (Tseng et al., 2005a; Tseng et al., 2005b). We then checked whether the text post-normalisation still contained any segmentation or POS-tagging errors.

Table 1 shows the results of this sampling process as percentages. It transpires that the normalisation process does not improve the segmentation error rate: indeed the error rate slightly increases. POS-tagging on the other hand is more noticeably improved by our normalisation steps, with the overall error rate decreasing from 58% on original texts to 53% on cleaned texts.

The other notable pattern to emerge from evaluation of NLP output errors is that error rates are highest for the under-18 age group: this outcome suggests that the youngest Weibo users write in a fashion the furthest removed from the data on which NLP tools have been trained. This manifests itself in a greater use of non-standard orthographic devices (Figure 1) and results in a higher error rate in NLP (Table 1), a pattern consistent with previous findings that the older readership and authorship of the data typically used to train NLP "puts older language users at an advantage" (Hovy and Søgaard, 2015).

3 Segmentation and tagging

All Weibo texts were passed to Stanford NLP tools for segmentation and part-of-speech tagging. Chinese text is not normally segmented into words by whitespace – as it is in, for instance, Latin writing systems — and thus in order to identify parts-of-speech, we need to segment the unbroken Weibo posts according to a tool that has been reported to achieve an F-measure of 0.828 (Tseng et al., 2005a). The segmented posts were subsequently passed to Stanford's tagger, which is reported to achieve 84.8% accuracy on unknown words (Tseng et al., 2005b).

[8]Source: Pew Research Center, accessed 2016-02-16.

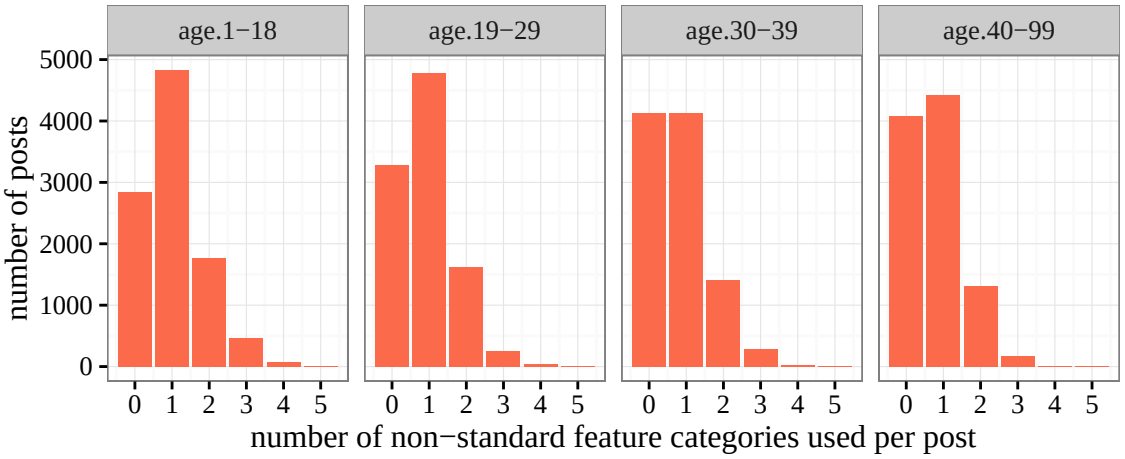

Figure 1: Usage of six non-standard orthographic features (classical characters, emoticons, kaomoji, hashtags, Latin characters, punctuation clusters) in Weibo posts by age: count of how many features occurred in each post (no post contained all six features).

Group	Original		Normalised	
	segmentation error	POS-tagging error	segmentation error	POS-tagging error
under-18	30	58	32	54
19-29	27	55	29	46
30-39	35	65	37	62
over-40	40	52	40	51
Total	*33*	*58*	*35*	*53*

Table 1: Error rates (%) in Weibo post segmentation and POS-tagger outputs from the original and normalised texts.

For example, post (a) was segmented as in (b), after which the text can be augmented with parts-of-speech as in (c). In the output, part-of-speech is indicated after hash characters with a tagset from the Linguistic Data Consortium Chinese Treebank (Xia, 2000). In the case of (c), the tags are NN 'other noun', AD 'adverb', and VA 'predicative adjective'.

(a) 春晚很好看
'The Spring Festival Gala is very good'

(b) 春晚. 很. 好看

(c) 春晚 #NN 很 #AD 好看 #VA

4 Text features

In the process of cleaning up and processing the texts we extracted a number of features about each Weibo post, listed below and exemplified in Table 6:

i. Length of post in characters;

ii. Presentation in poetic format: *i.e.* if punctuated, split the post and check whether resulting segments are of equal length – 5, 7 and 8 character line lengths being typical in this format;

iii. Use of fourteen popular expressions gathered from Weibo annual reports and Wikipedia[9] – see Table 2;

iv. Use of classical characters from a pre-defined list of fifty frequent characters;

v. Use of emoticons as signified by square brackets parentheses, *e.g.* [哈哈] 'laughter', [泪] 'tears', [偷笑] 'giggle', [爱你] 'love', [心] 'heart';

vi. Use of 'kaomoji' from a pre-defined list of 748 of these Japanese-style emoticons;

vii. Occurrence of hashtags, which in Chinese are enclosed by a pair of hash characters, *e.g.* #2015 亚洲杯 # '#2015AsianCup';

viii. Occurrence of homogenous Latin character clusters, such as 'hhh', 'ttt', and so on (which represent laughter – the 'h' coming from 'haha' – and for the visual similarity of the letter 'T' to a falling tear);

ix. Use of homogeneous punctuation sequences such as !!!!! and ????

[9]Source: Weibo report 2013, Weibo report 2014, `http://zh.wikipedia.org/zh/`中国网络流行语列表; accessed 2016-03-07.

x. Counts of nouns, pronouns, adverbs, verbs, subordinating conjunctions;

xi. Lexical types.

Chinese	English
醉了	whatever (lit. 'drunk')
心塞	sad
拼	work hard
泥垢	shut up (lit. 'dirt')
任性	capricious
萌	cute (lit. 'sprout')
哒	adjective forming part.
duang	collision (onomatopoeic)
小伙伴	little partner
卧槽	WTF
吐槽	complain
傻逼	idiot
牛逼	awesome
赞	like (in a social media sense)

Table 2: List of popular expressions sought out in Weibo posts

5 Age classification

With our target age labels associated with each post (under-18, 19 to 29, 30 to 39, over-40), we trained a number of classifiers on the Weibo data, varying the number of posts considered at once, from 1 to 100: from a single Weibo post as an instance, to each Weibo user's collected 100 posts as a combined instance with all features summed. We refer to this as the 'clustered posts' count, with each cluster being a 'document'. In post clusters, the feature types listed in Table 6 were respected: Boolean features remained Boolean, scalar features were summed.

A maximum entropy classifier (MaxEnt) outperformed naive Bayes and support vector machine classifiers in preliminary tests (Figure 2), and so we present MaxEnt accuracies averaged over ten-fold cross-validation in Table 3.

It is apparent in Table 3 that mean document length goes from 38 characters considering one post at a time, to 3817 when considering 100 posts at a time. Peak performance, however, comes at the 50-post mark, when the documents are just under 2000-characters long, on average. There is clearly some benefit to considering many posts as a single document, but perhaps a 'sweet-spot' in terms of learning from these agglomerated data which then falls away at 100 posts (*cf.* Figure 2: the sweet-spot for naive Bayes is found at 4, and for the support vector machine at 10).

In Table 4 we present precision, recall and F-measure (harmonic mean of precision and recall) for the optimal configuration from Table 3: a MaxEnt classifier over 50 posts per instance. It is apparent that the extremes of the 4-class age labels are more distinguishable from the middle classes: under-18 with an F-measure of 0.69 and over-40 with an F-measure of 0.74.

We hypothesised that use of non-standard orthographic devices, such as emoticons, kaomoji and repeated characters, would correlate with decreasing age, whereas the use of classical characters and poetic format would correlate with increasing age. To test these hypotheses we performed linear regression using R and the `lme4` package (R Core Team, 2015; Bates et al., 2015), with age in years as the dependent variable, and our various Weibo features as the independent variables. Results are shown in Table 5, reporting R-squared for the model overall (R^2), along with coefficients (B), standard error (SEB), *beta*- and P-values for each variable.

As can be seen in Table 5, the relationships are negative and significant (*i.e.* correlates with decreasing age) for the use of popular expressions, emoticons, kaomoji, hashtags and repeated characters, whereas the relation is significantly positive (*i.e.* correlates with increasing age) for the use of poetic format and classical characters. In terms of parts-of-speech, nouns and verbs are associated with increasing age, whereas pronouns and adverbs are associated with decreasing age, indicating that older Weibo users employ more standard lexico-syntactic structures in their posts.

6 Conclusion

We have shown that by simultaneously 'cleaning up' and extracting features from non-standard orthography in texts from Sina Weibo, then passing those cleaned up texts to NLP tools to obtain further features, we are able to classify texts by author age to a reasonable degree of accuracy: 65.5% F-measure at most. This level of accuracy is comparable to the naive Bayes and support vector machine classifiers reported in Li et al. (2013), and the accuracy of their most successful classifier, a decision tree, on their smallest training set of twenty thousand users (which our own dataset does not come close to in size)[10]. The outer classes in our age labels – under-18 and over-40 – are more accurately classified than the middle classes (19-29 and 30-39 years). The features most strongly associated with the younger age groups are the use of homogeneous Latin character sequences, hashtags and popular expressions; the use of poetic format and classical characters are most strongly associated with the older age groups.

7 Acknowledgements

We thank the three anonymous reviewers for their helpful comments. The second and last authors are supported by Cambridge English Language Assessment. The third author is supported by the Onassis Foundation.

[10]Precision of 60% for NB, 65% for SVM, 64% for decision tree with a 20,000 user training set; recall not reported (Li et al., 2013).

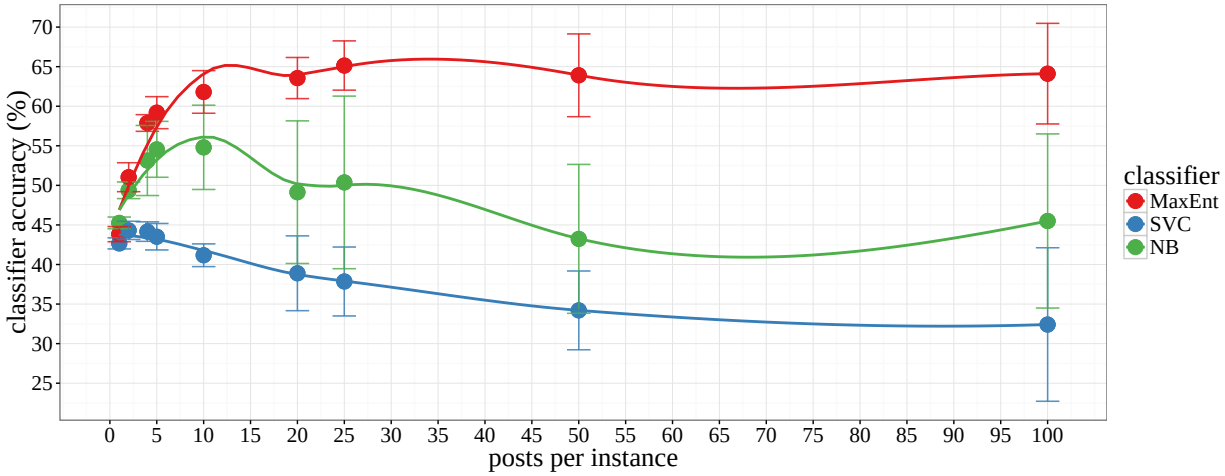

Figure 2: Classifying author age (4-class) in Weibo posts, varying cluster size ('posts per instance'), using maximum entropy, support vector machine and naive Bayes classifiers (MaxEnt, SVC, NB), and showing mean accuracy over 10-fold cross-validation (points), 1 standard deviation (error bars), and local polynomial regression lines (loess)

Clustered posts	**Number of documents**	**Mean document length (chrs)**	***F*-measure (%)**
1	40,000	38.2	43.8
2	20,000	76.4	51.0
5	8000	190.9	59.2
10	4000	381.7	61.8
50	800	1908.7	65.5
100	400	3817.5	64.1

Table 3: Maximum entropy classification accuracy of author age group, varying cluster size, using ten-fold cross-validation.

Age	**Precision**	**Recall**	***F*-measure**
under-18	0.636	0.764	0.690
19-29	0.528	0.498	0.507
30-39	0.566	0.488	0.529
over-40	0.739	0.760	0.740

Table 4: Classifying author age (4-class) in Weibo posts, 50 posts per instance, maximum entropy.

	R^2	B	SEB	*beta*	P
Model	0.14				$< .001$
Constant		29.5	0.09		$< .001$
post length		-0.01	0.001	-0.03	.225
poetic format		5.25	0.5	0.05	$< .001$
popular expr.		-3.04	0.24	0.1	$< .001$
classical chrs		1.25	0.07	-0.14	$< .001$
emoticons		-1.05	0.05	-0.05	$< .001$
kaomoji		-2.1	0.22	-0.03	$< .001$
hashtags		-3.14	0.22	-0.07	$< .001$
rep letters		-10.8	1.01	-0.05	$< .001$
rep punct		-0.5	0.04	-0.07	$< .001$
nouns		0.7	0.03	0.3	$< .001$
pronouns		-0.53	0.04	-0.08	$< .001$
adverbs		-0.36	0.04	-0.1	$< .001$
verbs		0.16	0.03	0.1	$< .001$
subord conj		0.45	0.25	0.01	.070

Table 5: Regressions of Weibo features as predictors of user age.

8 Bibliographical References

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., , and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*.

Bates, D., Maechler, M., Bolker, B., and Walker, S., (2015). *lme4: linear mixed-effects models using Eigen and S4*. R package version 1.1-9.

Darling, W. M., Paul, M. J., and Song, F. (2012). Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic Bayesian HMM. In *Proceedings of EACL Workshop on Semantic Analysis in Social Media*.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of Recent Advances in Natural Language Processing*.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT*.

Foster, J., Çetinoğlu, Ö., Wagner, J., Roux, J. L., Nivre, J., Hogan, D., and van Genabith, J. (2011). From news to comment: resources and benchmarks for parsing the language of Web 2.0. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*.

Han, B., Cook, P., and Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):5.

Hovy, D. and Søgaard, A. (2015). Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jørgensen, A. K., Hovy, D., and Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Labov, W. (1966). *The social stratification of English in New York City*. Center for Applied Linguistics, Washington D.C.

Jiang Lansheng, et al., editors. (2012). *The Contemporary Chinese Dictionary*. The Commercial Press, Beijing, 6th edition.

Li, Y., Liu, T., Liu, H., He, J., and Du, X. (2013). Web Information Systems Engineering –- WISE 2013. In Xuemin Lin, et al., editors, *Responses to language endangerment. In honor of Mickey Noonan*. Springer-Verlag, Berlin.

Lynn, T., Scannell, K., and Maguire, E. (2015). Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*.

Mukherjee, A. and Liu, B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nguyen, D., Smith, N. A., and Rosé, C. P. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Peersman, C., Daelemans, W., and Vaerenbergh, L. V. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC)*.

Plank, B., Hovy, D., McDonald, R., and Søgaard, A. (2014). Adapting taggers to Twitter with not-so-distant supervision. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*.

R Core Team, (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Saloot, M. A., Idris, N., Shuib, L., Raj, R. G., and Aw, A. (2015). Toward tweets normalization using maximum entropy. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*.

Trudgill, P. (2011). *Sociolinguistic Typology: social determinants of linguistic complexity*. Oxford University Press, Oxford.

Tseng, H., Chang, P., Galen, A., Jurafsky, D., and Manning, C. (2005a). A conditional random field word segmenter. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*.

Tseng, H., Jurafsky, D., and Manning, C. (2005b). Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 32–39.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*, 93:669–692.

Xia, F., (2000). *The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0)*. University of Pennsylvania IRCS Technical Report 00-07.

Table 6: List of features extracted from Weibo posts.

Feature	Type	Description	Example
post length	scalar	Number of characters in post, including punctuation (maximum 140)	祝贺阿森纳。希望枪迷有好的心情过春节。‘Congratulations to Arsenal. Gunner fans have a good mood for the New Year.’ (count = 19)
poetic format	Boolean	If post is delimited by punctuation or whitespace, are the segments of fixed length?	没办法真睡不着。没有比赛的日子, 我夜里都不醒的。‘Have no way, really cannot sleep. No game day, I do not wake up at night.’
popular expressions	scalar	Count number of popular expressions found in post from pre-defined list	黑顺毛哈哈‘Black smooth hair haha’ (count = 1)
classical characters	scalar	Count number of classical characters found in post from pre-defined list	也用以自勉‘Also use it to encourage myself’ (count = 1)
emoticons	scalar	Count number of square-bracket-enclosed emoticons in post	晚上好。[握手][握手][握手] ‘Good evening [shake hands] [shake hands] [shake hands]’ (count = 3)
kaomoji	scalar	Count number of kaomoji found in post from pre-defined list	一起加油 (^ω^) ‘Work hard together’ (count = 1)
hashtags	scalar	Count number of hashtags in post, as indicated by hash-character pairs	勇敢而坚强，仁慈而善良！# 蓝天别走 # ‘Brave and strong, benevolent and kind! #blueSkyDoNotGo’ (count = 1)
repeated Latin characters	scalar	Count of homogeneous character clusters in post	心疼 TTTTTTT ‘Distressed’ (count = 1)
repeated punctuation	scalar	Count number of homogeneous punctuation clusters in post	滚蛋辣!!!! 我的错!!!! Piss off!!!! My fault!!!!’ (count = 1)
nouns	scalar	Count number of nouns in tagged post	的 #DEG 也 #AD 保佑 #VV 案件 #NN 能够 #VV 早日 #AD 破案 #VV 乘客们 #NN 都 #AD 能够 #VV 回家 #VV (count = 2)
pronouns	scalar	Count number of pronouns in tagged post	幸福 #VA 的 #DEG 事情 #NN 你们 #PN 都 #AD 有 #VE 吗 #SP (count = 1)
adverbs	scalar	Count number of adverbs in tagged post	为啥 #AD 这 #PN 姿势 #NN (count = 1)
verbs	scalar	Count number of verbs in tagged post	不 #AD 该 #VV 抱怨 #VV 都 #AD 是 #VC 自己 #PN 找 #VV 的 #DEC 这 #PN 叫 #VV 活该 #VV (count = 6)
subordinating conjunctions	scalar	Count number of subordinating conjunctions in tagged post	如果 #CS 久久 #AD 惦记 #VV 的 #DEC 拥有 #VV 了 #AS 会 #VV 怎么样 #VA 呢 #SP (count = 1)
lexical types	Boolean	‘True’ value for each token in segmented post	没, 我, 喜欢, 的, 明星, 不, 看