International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-022-01523-9

ORIGINAL ARTICLE

Investigating human reading behavior during sentiment judgment

Xuesong Chen[1] · **Jiaxin Mao**[1] · **Yiqun Liu**[1] · **Min Zhang**[1] · **Shaoping Ma**[1]

Received: 29 June 2021 / Accepted: 17 February 2022

Abstract
Sentiment analysis is an essential task in natural language processing researches. Although existing works have gained much success with both statistical and neural-based solutions, little is known about the human decision process while performing this kind of complex cognitive task. Considering recent advances in human-inspired model design for NLP tasks, it is necessary to investigate the human reading and judging behavior in sentiment classification and adopt these findings to reconsider the sentiment analysis problem. In this paper, we carefully design a lab-based user study in which users' fine-grained reading behaviors during microblog sentiment classification are recorded with an eye-track device. Through systematic analysis of the collected data, we look into the differences between human and machine attention distributions and the differences in human attention while performing different tasks. We find that (1) sentiment judgment is more like an auxiliary task of content comprehension for humans. (2) people have different reading behavior patterns while reading microblog posts with varying labels of sentiment. Based on these findings, we build a human behavior-inspired sentiment prediction model for microblog posts. Experiment results on public-available benchmarks show that the proposed classification model outperforms existing solutions over 2.13% in terms of macro F1-score by introducing behavior features. Our findings may bring insight into the research of designing more effective and explainable sentiment analysis methods.

Keywords User behavior · Eye movement · Sentiment judgment · Machine model

1 Introduction

Sentiment analysis [12, 24, 30, 31, 41] is one of the most crucial text classification tasks and a fundamental problem in natural language processing. Plenty of proposed models optimize a function to establish the relationship between text features and label indexes. Although these models gain much success, especially with the assistance of neural models, little is known about humans' sentiment judgment process. [47] have found that eye movements are associated with emotions in video-watching settings. However, whether reading, a complicated physiological and psychological process, arouses enough emotional stimuli to affect human eye movements remains unknown. Besides, existing work shows that designing computational models inspired by human reading behavior leads to a better performance of NLP tasks [18, 22, 35, 48]. To provide insights and guidance for designing a better sentiment classification model, studying how humans accomplish such tasks and comparing the decision processes between human and machine models is necessary. Therefore, we design a user study to deal with it in this paper.

According to the reading context settings, existing human reading models can be grouped into two categories: general reading models and specific reading models under a certain context [48]. The first category includes E-Z model [33, 34], SWIFT [7] and the Bayesian reading model [4], which formalized the human reading patterns in non-contextual reading settings. The second category includes Two-Stage Examination Model [22], Reading Model in Relevance Judgment [18], and Human Behavior

✉ Yiqun Liu
yiqunliu@tsinghua.edu.cn

Xuesong Chen
chenxuesong1128@163.com

Jiaxin Mao
maojiaxin@gmail.com

Min Zhang
z-m@tsinghua.edu.cn

Shaoping Ma
msp@tsinghua.edu.cn

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Published online: 06 March 2022

Inspired Machine Reading Comprehension Model [48]. These works try to model the examination behavior in specialized task settings, e.g., on search engine result pages, during relevance judgment or comprehension tasks. All of these tasks require humans to read lengthy documents with a high cognitive load. However, sentiment analysis is a different task in which the textual content is usually not so long. Specifically, in sentiment judgment of microblogs, a post usually contains 20 to 50 words, and the reading patterns for this kind of short text remain under-investigated, which motivates our research.

In the aspect of human inspiration to models, some previous works focus on regarding eye movements as features of model input [28], regularization of machine attention layer [2] or a task of multi-task model [16, 29] to improve performance. Compared to these works, we pay more attention to analyzing human behavior during microblog sentiment judgment processes and exploring the inspiration to design classifiers. Apart from these [36], collected human-annotated words in text classification and compared them with machine attended words. After that, they revealed the similarities between attended words from both sides [5, 18]. However, their work is not based on real eye movements of users' reading behavior and doesn't fully reflect human cognitive information during judgment.

To better understand how humans read texual content and make judgment in practical sentiment judgment scenarios. We design a user study which requires assessors to read a microblog post and complete two labeling tasks: (1) Judging the blog's sentiment as *positive*, *neutral* or *negative*. (2) Judging the blog's emotion as *happiness*, *like*, *surprise*, *none*, *sadness*, *anger*, *disgust*, or *fear*. Since eye movements are tightly coupled with cognitive attention during reading in our brains [23] and may serve as a measurable indicator of the reading process, we use an eye-tracker to collect participants' eye-movements during the completion of these tasks. Based on the collected data, our study aims to answer the following research questions:

- RQ 1: How do humans make sentiment judgment while reading a microblog post?
- RQ 2: What are the differences in attention distributions between human and machine during sentiment judgment?
- RQ 3: What is the attention allocation mechanism of humans during different sentiment judgment tasks?
- RQ 4: How to improve sentiment analysis models with the findings in human judgment process?

We additionally collected users' eye movements in topic classification to better understand human reading processes when judging sentiment. The contribution of our work is three-fold:

(1) By comparing the decision processes among human making sentiment judgment, machine making sentiment judgment, and human making topic judgment, we find that both human judgment processes concentrate more on content comprehension than annotation completion. In contrast, the machine tries its best to build up the relationship between words in posts and sentiment labels.
(2) Users will dynamically adjust the attention allocation policy according to task difficulty and personal preferences during the reading process while reading blogs in different sentiments.
(3) At last, we build a sentiment predictor based on the above findings, and it achieves better performance.

The remainder of this paper is organized as follows. In Sect. 2, we review some related studies to our work. Sect. 3 describes the design of our research and the data collection procedure. Section 4 compares the differences among human sentiment judgment, topic judgment, and model decision, which addresses **RQ1** and **RQ2**. To investigate **RQ3**, we analyze reading behavior in Sect. 5. To solve **RQ4**, we build prediction models for classification, then discuss the future research directions of human-inspired models in Sect. 6. Finally, we conclude our work in Sect. 7.

2 Related work

According to our research purpose and experiment settings, we investigate three aspects of related work: Sentiment Analysis, Reading Model, and Attention-based Models.

2.1 Sentiment analysis

Sentiment Analysis (SA) is a central field of research that lies at the intersection of many fields such as text analysis, natural language processing, and biometrics. It's widely applied to social media monitoring, market research, customer service, etc. Either traditional machine learning models like Naïve Byes [26], Support Vector Machine [38], Random Forest [19], Maximum Entropy [3] and logistic regression [11] or neural models like CNNs [15], LSTMs [10, 44] and Transformers [8, 37, 39] have been proven effective in classification tasks. In recent years, more and more people like communicating, sharing, or requiring information in social media, such as Twitter, Facebook, or Sina Weibo, which attracts much sentiment analysis of microblogs to study user behavior. Compared to documents on other platforms, the blogs produced in social media own their style. The length of published blogs mostly ranges from 20 to 50 words and has a maximum limitation in some applications. The content could contain emoticons or hashtags if a poster prefers, and the

writing style is more conversational than other documents. Since emoticons and hashtags play an essential role in sentiment expressions, models taking these unique features into account could predict more accurately [20, 46].

2.2 Reading model

Reading is a vital process to comprehend the context or make judgment by given tasks. Based on users' eye movements, the reading patterns and how language is processed can be inferred [34]. Eye movements are composed of a sequence of fixations and saccades. Eye fixations indicate periods when eyes statically land on an object, typically lasting 200 to 250 ms influenced by language, grammar, word frequency, etc. Eye saccades indicate periods when eyes are moving, typically lasting 20 to 50 ms [34]. There exist several reading models elaborating on information acquisition during the reading process. EZ Reader depicts eye-movement behavior in general reading and summarizes the four joint determinants of eye movement: word passport, visual processing, attention, and control of the oculomotor system. Based on the assumption that reading is a cognitively controlled process where the saccade to the next word is programmed, the person is cognitively processing the text available in the current fixation span. According to experiments [32, 42], users are able to identify words in the *parafoveal preview* span. In Chinese documents reading, adults' perceptual span usually covers one word on the left, and $2 \sim 3$ words on the right around the fixated word [17, 42].

There are some works modeling users' eye movement behavior into two-stage when given a specific task. Liu et al. [22] found that there usually exists a skimming step before users carefully read the search result when examining search engine result pages (SERPs), which can help estimate better relevance of search results. Li et al. [18] found that there exists a preliminary relevance judgment stage and a reading with preliminary relevance stage during the relevance judgment process. Zheng et al. [48] showed that a two-stage model also exists in Question Answer tasks. Specifically, the first stage is to search for possible answer candidates, and the second stage is to generate the final answer through a comparison and verification process. These models illustrate two-stage reading models when humans are judging high cognitive tasks. However, the cognitive process remains further investigated when the document gets shorter, or the study turns to sentiment judgment.

2.3 Attention-based models

Attention-based models have become the architectures of efficient choice for many NLP tasks, including machine translation, text classification, and question answering. Since the attention mechanism was introduced [1], the investigation of whether the attention is interpretable becomes a hot but controversial topic [5, 13, 36, 40]. Jain et al. [13] argued the explanation ability of attention architecture models. Sen et al. [36] found a significant similarity between keywords selected by the attention layer from bidirectional RNNs and human-annotated words in text classification. Bolotova et al. [5] got the same conclusion in Question Answer tasks. This paper will compare the similarity between human fixated words with eye-tracking and machine attended words with attention layer to analyze the difference and discuss model design directions.

3 Data collection

In this section, we describe the settings of our user study and the datasets we collected.

3.1 Tasks

After comparing several available Chinese Microblog[1] datasets, we finally chose the dataset from NLP&CC2013,[2] one of the most popular and challenging datasets. We sampled 1224 microblogs from it, whose sentiment includes *positive*, *neutral*, *negative*(408 blogs for each), and emotion includes *happiness*, *like*, *surprise*, *none*, *sadness*, *anger*, *disgust*, and *fear* (153 blogs for each). Next, we shuffled 1224 microblogs and divided them into six groups equally. For each group, we recruited five participants (also called users in the following paper) to judge sentiment and emotion at the same time and the order of blogs presented to each participant was randomly generated. To compare user behavior in different tasks, we randomly selected two groups of blogs and recruited five other users for each group to make topic judgment additionally. The topic includes *life*, *art*, *star*, *politic*, *science*, *sports*, *society*, and *others*.

3.2 Participants

We recruited 40 university students via online social networks and email to participant in our user study. The users include 18 males and 22 females, and their ages range from 18 to 27. All of them are undergraduate or graduate students, and their majors vary from natural science and engineering to humanities and sociology. We screened all applicants according to their visual acuity to ensure that the collected eye movements were correct. And all participants possess college-level skills in Chinese reading comprehension and skillful computer operation capability. It takes about 40 to 70

[1] https://www.weibo.com/

[2] http://tcci.ccf.org.cn/conference/2013/

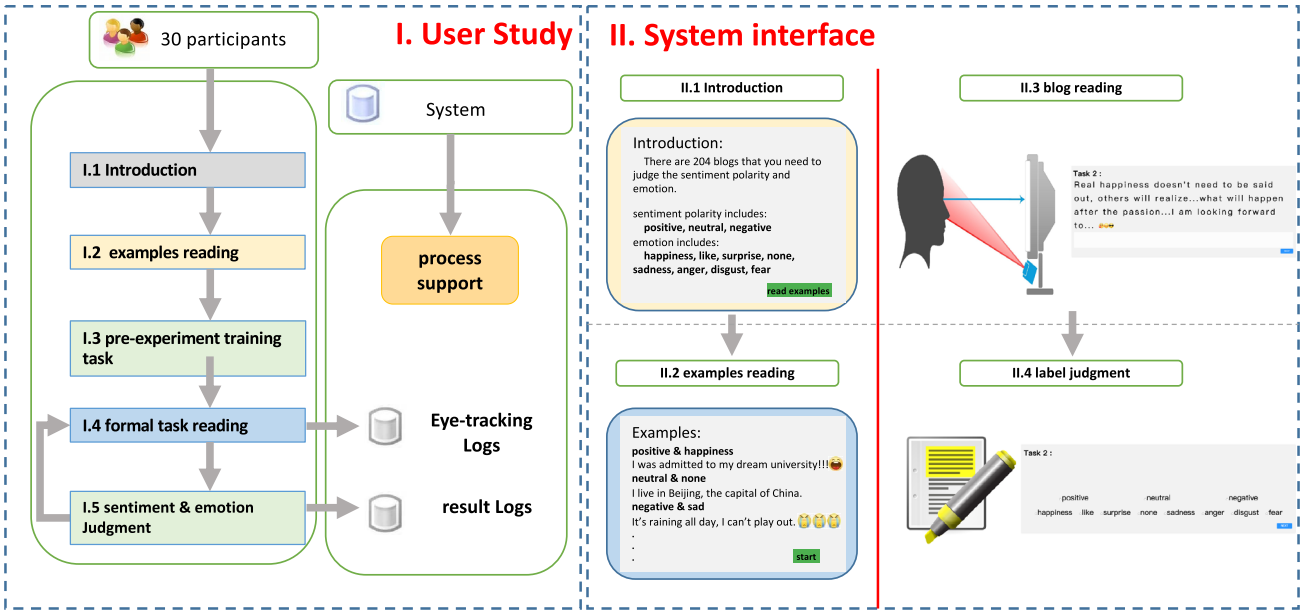

Fig. 1 User study procedure. The texts in the system interface are translated from Chinese

minutes to accomplish 204 microblogs judgments, and each participant is paid $7 or $8.6 regarding their judgment accuracy of sentiment polarity. We declared the payment policy before the experiment to encourage users to try their best.

3.3 Procedure

Our user study's procedure and system interface in sentiment judgment (topic judgment is similar) is shown in Fig. 1. Note that all the instructions, guide systems, and blogs are in Chinese. In the beginning, participants should read the introduction of our study and be told that there would be five pre-experiment training tasks and 204 formal tasks required to choose the correct sentiment polarity and emotion or topic only. After that, the system will show 24 examples to help the users be familiar with microblogs and their corresponding labels. Then, the pre-experiment tasks will help users learn the annotation process. In a single annotation task, the microblog and answer area are not presented together. At first, the system will show the microblog alone. Only after the user made a judgment in his mind, he admitted to getting the answering area by clicking a specific button, and the microblog will disappear simultaneously. Eye-tracking data is recorded during the period of the blog presented. Note that the user study's main target is collecting the natural eye movements about judgment, so we don't ask users to do any other tasks like highlighting label-related keywords, which could introduce unexpected behavior bias.

We use a Tobii X2-30 eye tracker to record participants' eye movements, whose deviation is within the word level for the eye-tracking data. Before the experiment, there is a calibration for each participant to ensure that the eye movements' data is recorded accurately. The maximum length of the blogs is less than 150, so users can read the entire content without a scroll. We detected fixations and saccades using built-in algorithms and all default parameters from Tobii Studio. The annotation system was deployed on a 17-inch LCD monitor with a resolution of 1920×1080 pixels. In fixations heatmap (an example is shown on the top of Fig. 2), the redder the fixation point is, the longer the duration time is. At the bottom of Figure 2, every circle means a fixation, and the number on the circle represents the fixation order when the user read the current blog. With the

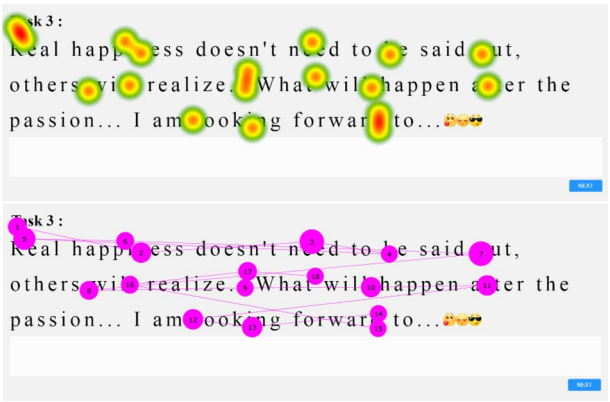

Fig. 2 Eye movements during the completion of sentiment judgment tasks. In user fixations heatmap (top), the redder the color, the longer the time. Eye movements scanpath as shown on the bottom

Table 1 The statistics of the data collected in the user study

Judgment	#Microblogs	#Groups	#Users	#Sessions
Sentiment	1224	6	30	6120
Topic	408	2	10	2040

help of scanpaths, we can observe the users' fixation transition including *Forward*, *Skip* and *Regression* [27].

3.4 Collected datasets

Through the user study, we collected two datasets. One consists of 6120 sentiment and emotion judgment sessions from 30 users, the other consists of 2040 topic judgment sessions from 10 users. The detailed statistics of the collected data as shown in Table 1.

Each blog was annotated by five users in our user study. However, the majority voting results of sentiment polarity and emotion disagreed with the given labels in NLP&CC2013 dataset. Then we reorganized a more reasonable ground truth set that considers the contents and both-side labels. Based on our ground truth, the average accuracy of 30 users is 0.796. We also measure inter-person annotation agreement by Cohen's KAPPA coefficient κ. For 3-level sentiment polarity annotation, the κ is 0.501; For 8-level emotion annotation, the κ is 0.403; For 8-level topic annotation, the κ is 0.410. All of them reach a moderate agreement level. Besides, we calculate the ratio of maximum votes greater or equal to three in emotion annotation is 0.771, while in topic annotation is 0.801. In Conclusion, the ratio and Cohen's KAPPA coefficient indicate that users are more consistent when judging topic labels, perhaps because it is easier to recognize the blogs' topic.

4 Process of sentiment judgment

This section first proposed a more reasonable method to assign human attention when reading fine-grained objects like words or characters. Then introduce three kinds of metrics to measure the similarity between different attention maps (AMs). Based on the measurement, we put forward the human reading model under this task and compare the judgment process of humans and the machine.

4.1 Human attention assignment

There are significant differences in width and morphology between Chinese and English words. The perceptual span of humans is relatively fixed, considering the width of English words is usually wider than Chinese ones generally,

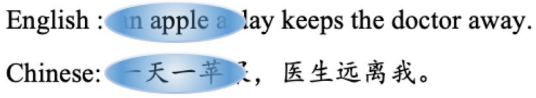

Fig. 3 A comparison between English and Chinese in morphology and width of words

which results in people perceiving more words when reading Chinese. However, existing work regarded the single word on a fixation point as the **fixated word**, which will introduce higher inconsistency with the perceived information of humans in Chinese reading because of the narrower word width. For example, when a human is fixating on the Chinese word "天" as Fig. 3, he will perceive nearby words including "一" on the left and "一", "苹果" on the right, which on the benefit of *parafoveal view* and *moving window paradigm* [42]. We named words identified in this way **adjacent words**. Compared to fixated words, adjacent words are more aligned to the information perceived in the human brain. In our work, we will compare both types of words and call them **attended words**. We suggest that using adjacent words considering the word width and perceptual span in any language may better understand human cognition in eye movement research.

4.2 Human attention

4.2.1 Human attention map (HAM)

To aggregate a generalized AM when humans are reading posts, we recruit five different people to judge the same post and record their eye movements. If a word was attended by three or more users, we will regard it as a *group-level attended word*, and all of them are made up of the human attention map (HAM) on the post.

4.2.2 Human attention agreement

We define the coefficient as below to measure human attention agreement among different users. In the definition, *#AttendedWords* is the number of group-level attended words, *#AttendedWordsUnionOfUsers* is the length of the union set of five users' attended words.

$$\rho_{agree} = \frac{\#AttendedWords}{\#AttendedWordsUnion\ Of\ Users}. \quad (1)$$

In our work, we define *task difficulty* as negatively correlated with the number of consistent annotation in a group which ranges from two to five and represents the maximum of users

Table 2 User attention agreement under different number of consistent annotation and sentiment

Objects		# consistent annotation					Sentiment			
Taxonomies		5	4	3	2	Sig.	Positive	Neutral	Negative	Sig.
# posts		466	339	355	64	\	462	352	410	\
ρ_{agree}	Fixated words	0.185	$0.210^{5\star}$	$0.215^{5\star\star\star}$	$0.230^{5\star\star}$	***	$0.187^{neg\star\star\star}$	0.209	0.216	***
	Adjacent words	0.542	$0.582^{5\star}$	$0.610^{5\star\star\star}$	$0.638^{5\star\star}$	***	$0.557^{neg\star\star\star}$	0.575	0.604	**

"*/**/***" indicates that the differences among different taxonomies in either objects are statistically significant at $p < 0.05/0.01/0.001$ level(Kruskal–Wallis (KW) H Test). "⋆/ ⋆ ⋆/ ⋆ ⋆⋆" indicates that the differences between two taxonomies in the same object are statistically significant at $p < 0.05/0.01/0.001$ level (Dunn's Post-hoc Test). "neg" is the abbreviation of "Negative" in the table

giving a consistent label. In other words, higher consistency means easier tasks. For example, if five users agreed on a post in our experiment settings, that means the sentiment is a little controversial and indicates it is an easier task to annotate. The average of ρ_{agree} under different sentiment and consistency is shown in Table 2.

Next, we use a parametric test (Pearson correlation) and a non-parametric test (Spearman correlation) to reveal the correlations between users' attention agreement and the number of consistent annotations. Regarding adjacent words as attended words, the strengths of the correlations measured by Pearson and Spearman are − 0.153 (− 0.099 for fixated words, both p-value < 0.001) and − 0.152 (− 0.108 for fixated words, both p-value < 0.001) respectively, which weakly indicates when confronting a more complicated task, the attention agreement among users tends to increase. As shown in Table 2, when users judge an easier task requiring lower cognition, especially the consistency reaches five, their attention showed a lower agreement significantly. Considering the distribution of the number of consistent annotations under different sentiments in Fig. 4, when judging a post with a sentiment polarity, either positive or negative, it is easier to reach a higher annotation agreement compared to judging neutral posts. These neutral posts may carry a little sentiment, which leads to inconsistent annotations in probability.

Based on KW Test results in Table 2, when users read positive posts, their attention agreements are lower than reading neutral or negative posts but similar to reading easy-annotated ones. It indicates users could be quickly aware of the positive sentiment in posts and read them with lower effort. As shown in Fig. 4, there is a high percentage of consistent annotation in negative blogs, but users pay more attention to reading them like hard-annotated ones. This indicates that users have more personal preferences when reading negative posts.

4.3 Machine attention

Machine Attention Map (MAM) is exported from the words' weights of softmax attention layer in neural networks, which is Hierarchical Attention Networks [43] with BiGRU in our work. Sen et al. [36] have shown that attended words produced by the machine model with bidirectional architectures and attention mechanism are more similar to human-annotated keywords. However, these annotated results lacking eye movement data cannot reflect the human's decision-making and attention changes on the timeline. In MAM, each word's attention score is the product of the weight of the sentence it is in and the weight of the word itself. Unlike a human, a machine will allocate an attention score to every word in the post. However, we only selected out the top-n words with the highest word attention scores in MAM, where n equals the number of words in HAM.

4.4 Similarity metrics

In this section, we proposed a number of metrics to measure the similarity between different attention maps.

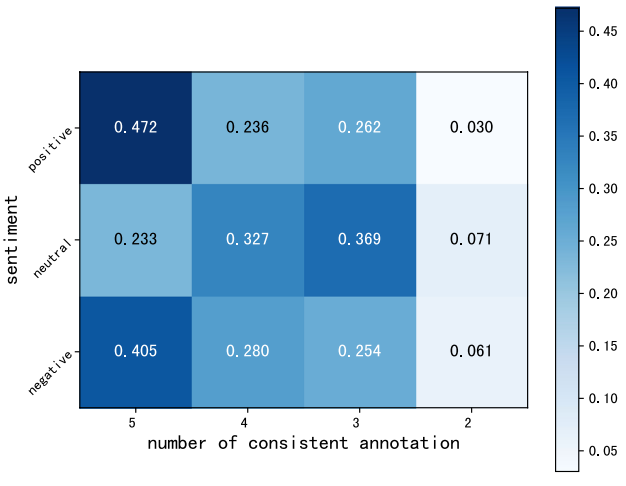

Fig. 4 The distribution of the number of consistent annotation under different sentiment

4.4.1 Recall and Precision based metrics

Inspired by definitions in Information Retrieval, we proposed two metrics to measure the utility of special category words in different judgment processes, which are also called *Recall* and *Precision*. In our work, *Recall* is the number of special category words in an AM divided by the number of total words in the AM, while *Precision* has the same numerator but divided by the number of total words in the same category existing in the post. The two metrics could be formalized in mathematical notation as follows. Moreover, the special categories we used include *Sentiment Word*, *Part-Of-Speech*, and *Word Frequency*.

$$Precision_{SpecialCategory} = \frac{\#SpecialCategoryWords_{attended}}{\#Words_{attended}}, \tag{2}$$

$$Recall_{SpecialCategory} = \frac{\#SpecialCategoryWords_{attended}}{\#SpecialCategoryWords_{post}}. \tag{3}$$

Sentiment Word (SW) Sentiment words in posts are crucial clue to help users make judgment. So we focus on the utility of sentiment words across different AMs. Firstly, we collected a large scale Chinese sentiment word dictionary.[3] Notice that when judging a post's sentiment, humans may consider positive-sentiment words in a negative post and vice versa, which means either positive or negative words are beneficial no matter the sentiment polarity of a post. So both polarity words are taken into account.

Part-Of-Speech (POS) Marimuthu et al. [25] found that lexical indicators of sentiment are commonly associated with syntactic categories such as adjective, adverb, noun, and verb. Liu et al. [20] showed that emoticons make an impressive contribution when used to infer sentiment, and they can be regarded as a special kind of POS tag.

Word Frequency (WF) Studies in psychology have shown that people read frequent words and phrases more quickly [14], thus we should consider the influence of word frequency when humans and the machine making judgments. Based on large-scale online web data [21], we divided all words into four levels according to their frequency, including high, upper-middle, lower-middle, and low. We adopted the same principles as POS to compare $Recall_{WF}$ and $Precision_{WF}$.

The *Recall* is suitable for comparing the similarity of sentiment judgment between humans and the machine because of the same length of AMs. However, the number of attended words when users making sentiment and topic judgments on the same blog is different. Specifically, the averages are 6.394 and 5.092, respectively, so it is not objective to use *Recall* to measure the behavioral similarity of HAMs. Thus we mainly adopt *Precision* to compare the AM similarity when users accomplish sentiment and topic tasks.

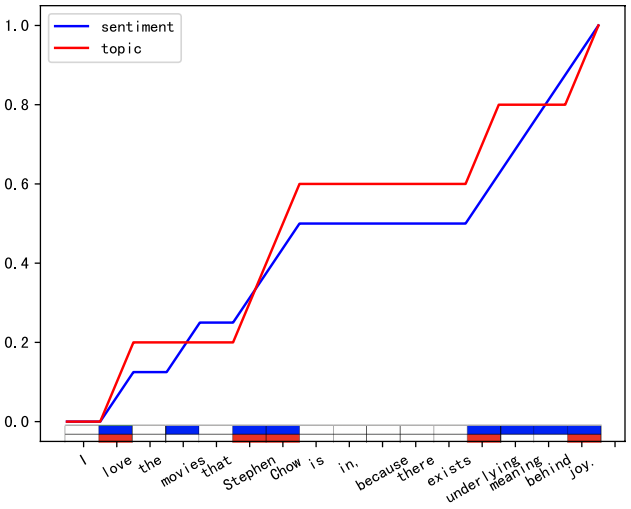

Fig. 5 An example of Attention CDF under different tasks on the same post. If the user attended a word, the block would be filled with the color corresponding to the task. The attention CDF is based on colored blocks

4.4.2 Overlap

Except for the particular category words, we use below *Overlap* considering every word to quantify agreements of two AMs, defined as the intersection size divided by the minimum size of two maps. Compared to the Jaccard similarity coefficient, ours better considers the condition that a significant size gap exists in two AMs.

$$Overlap(AM_A, AM_B) = \frac{|AM_A \cap AM_B|}{min(|AM_A|, |AM_B|)}. \tag{4}$$

4.4.3 Attention distribution

In order to weaken the variance between fixated words and human understanding, we have adopted adjacent words, as mentioned before. Additionally, we measure the similarity of attention distribution by Attention Cumulative Distribution Function (CDF) robustly. As shown in Fig. 5, Although the fixated words are clearly distinct in two tasks, both CDFs are similar, which indicates users act out similar cognitive processes when making two judgments. In our work, we use Kolmogorov–Smirnov (KS) to test whether two attention distributions are the same. The similarity of attention

[3] http://www.keenage.com/
http://nlg.csie.ntu.edu.tw/nlpresource/NTUSD-Fin/
http://nlp.csai.tsinghua.edu.cn/site2/index.php/13-sms
https://bosonnlp.com/dev/resource

Table 3 Sentiment Words comparison of attention maps

	Sentiment task	Topic task	Machine
$Precision_{SW}$	0.230	0.231	0.287***
$Recall_{SW}$	0.194	0.163*	0.228***

"*/**/***" indicates the result is significantly different at p < 0.05/0.01/0.001 level (t-test) with the result when users made sentiment judgment

Table 4 Overlap comparison of attention maps

	Machine VS human	Sentiment VS topic
*Overlap****	0.197	0.399

"*/**/***" have the same meaning as in Table 2 (t-test)

distribution weakens the impact of isolated words but emphasizes several consecutive words, which is more reasonably used to compare HAMs.

4.5 Judgment processes

We compare the judgment processes of humans and the machine in this section based on our proposed similarity metrics. Either fixated or adjacent words play the same role in the comparison, so we detail the results on fixated words as a presentative.

Firstly, we compare the utility of sentiment words in the judgment processes, as shown in Table 3. In the aspect of $Precision_{SW}$, the machine pays significantly more attention to sentiment words to judge the labels than users. Besides, if users are given a sentiment-irrelated task, like the topic judgment we used, both $Precision_{SW}$ are nearly equal to surprises us. Based on $Recall_{SW}$ considering the number of sentiment words in posts, we also find that users retrieve fewer sentiment words than the machine. This result indicates that human does not depend on the sentiment words to give the right sentiment labels. Notice that the average number of fixated words per blog in the topic task is less than sentiment, as we reported before, which results in $Recall_{SW}$ of the topic task is lower than the sentiment naturally. However, the difference at p-value < 0.05 also suggests a similar recall ratio of sentiment words somehow.

$Precision_{SW}$ and $Recall_{SW}$ are metrics to compare the task-related words utility, *Overlap* is designed to measure the general words' selection policy among AMs. We calculate the *overlap* of the machine and human judging sentiment and human judging different tasks, the results as shown in Table 4. There is a relatively lower agreement between users and the machine in general word selection policy when judging sentiment but a higher agreement when users make sentiment and topic judgment.

Based on the above observation, we can conclude that the machine and humans have different sentiment words and general words selection. When humans faced various tasks on the same blogs, they still focused on the same words.

Next, we investigate the difference in POS tags during three judgment processes. All of the POS tags are divided into three groups by us including *sentiment-related*, *difficulty-related* and *content-related*. The sentiment-related group is composed of interjection and emoticons. An interjection is used to demonstrate the emotion or feeling in the posts, like "ahh" or "eh". Emoticons are frequently used to express sentiment in an extra way when publishing posts. The words of both POS flags play a pivotal role in inferring sentiment polarity like sentiment words. As shown in Table 5, machine relies heavily on them to judge

Table 5 POS tags comparison of attention maps. $Recall_{POS}$ column presents the *Recall* of different POS tags when machine and human making sentiment judgment; $Precision_{POS}$ column presents the *precision* of different POS tags when human making judgment in sentiment and topic tasks

POS tags		$Recall_{POS}$			$Precision_{POS}$		
		Human	Machine	Difference	Sentiment	Topic	Difference
Sentiment-related	Interjection	0.073	0.380	421.60%***	0.116	0.284	144.04%
	Emoticons	0.107	0.317	195.35%***	0.373	0.303	− 18.71%
Difficulty-related	English	0.245	0.101	− 58.59%***	0.321	0.344	7.37%
	Conjunction	0.167	0.076	− 54.31%***	0.156	0.205	31.09%
	Numeral	0.155	0.075	− 51.73%***	0.188	0.255	35.61%*
	Time	0.171	0.091	− 46.78%***	0.181	0.186	2.62%
	Idiom	0.459	0.281	− 38.72%***	0.165	0.251	52.56%**
Content-related	Noun	0.254	0.179	− 29.49%***	0.291	0.336	15.30%**
	Verb	0.197	0.170	− 13.85%**	0.282	0.316	12.18%
	Prepositional	0.084	0.064	− 23.67%	0.140	0.162	15.75%
	Adverb	0.166	0.158	− 4.81%	0.180	0.226	25.06%*
	Pronoun	0.129	0.127	− 1.27%	0.195	0.229	17.43%
	Auxiliary	0.070	0.092	30.99%**	0.142	0.142	− 0.34%
	Adjective	0.181	0.261	44.17%***	0.193	0.226	17.30%

"*/**/***" have the same meaning as in Table 2 (t-test)

Table 6 Word frequency (WF) comparison of attention maps

Word frequency	$Recall_{WF}$			$Precision_{WF}$		
	Human	Machine	Difference	Sentiment	Topic	Difference
High	0.102	0.119	16.57%***	0.512	0.516	0.74%
Upper-middle	0.202	0.217	7.22%	0.223	0.267	19.53%
Lower-middle	0.214	0.209	− 2.48%	0.202	0.245	21.55%
Low	0.251	0.171	− 31.64%***	0.472	0.531	12.56%**

"*/**/***" have the same meaning as in Table 2 (t-test)

the sentiment while users not. The words in the difficulty-related group require users to pay extra attention to comprehend. Chinese speakers need to make more effort to recognize English words theoretically. Conjunctions are used to connect phrases, clauses, or sentences, which indicate the relationship between connected objects. As for numeral and time tags, they contain detailed information about the blogs. When a poster is citing idioms, he may convey the underlying or implicit meanings. Only by taking care of these words can users make clear the whole context. As shown in the $Recall_{POS}$ column in Table 5, compared to the machine, humans can realize the difficulty and attend to them during the reading process. The words in the content-related group are ubiquitous in documents. Whether humans or the machine attend to them or not depends more on their meanings. We also list $precision_{POS}$ when users judge sentiment and topic tasks. As shown in Table 5,there exist less difference in the two tasks' AMs.

Based on the observation on POS tags, machine pays more attention to sentiment-related tags during the process of sentiment judgment while human focuses more on the incomprehensible tags. Furthermore, users have no apparent POS tag preferences to judge sentiment or topic.

In the aspect of word frequency influence in sentiment judgment, we found that users pay more attention to lower frequent words than the machine, as shown in Table 6. We infer that the machine sets the target to model the relationship between words and labels, so it relies on the embeddings of higher frequent words which are well trained and carry more semantic information. In contrast, users focus on lower frequency words to comprehend the posts in both tasks.

According to the Comparison in Table 7, on half of the posts, attention distribution when users making sentiment and topic judgment does not perform significant differences. It indicates that the human attention allocation mechanism in both tasks is similar. According to the annotation agreement and reading time, topic judgment is an easier task than sentiment judgment, which indicates that reading behavior in the topic task is more like task-free reading patterns. We also observe that attention distributions in positive or easier blogs (the number of consistent annotations is four or five) are more similar to task-free (topic).

4.6 Summary

Our **RQ1** focuses on how humans make sentiment judgments during the reading process. The results show a high agreement of user behavior in sentiment and topic tasks, such as the similar recall of sentiment words or POS tags, higher overlap of attended words, and similar attention distributions. During both judgments, we think people focus on understanding the posts rather than completing the annotation task. Once he comprehends the general meaning of the posts, they will get the sentiment polarity naturally. So the eye movements are more relevant to post-reading and rarely influenced by task completion. Besides, the finding that eye movements are closely related to task difficulty also verifies the reading model we proposed makes sense in judgments.

We investigate the differences between humans and the machine making sentiment judgment to answer **RQ2** and find that the judgment process of neural networks is saliently different from humans. We consider the machine models with attention mechanisms to mainly optimize the matching function between input words and labels. They pay more attention to task-related words roughly, while humans rely more on content meaning but isolated words. We suggest a better classification model should focus on document understanding rather than on a rough matching approach. Besides, the high attention similarity of humans in sentiment and topic tasks inspires us to design

Table 7 Comparison of attention distribution similarity when users make sentiment and topic judgments

		Non-significant	Significant
Polarity	Positive	56.49%	43.51%
	Neutral	46.67%	53.33%
	Negative	45.52%	54.48%
Consistency	4.5	51.94%	48.06%
	2.3	46.67%	53.33%

The similarity is measured by the KS test, if p-value < 0.05, indicates a significant difference between attention distributions. Otherwise, it indicates a non-significant difference. This table shows that when users judge sentiment and topic, the proportion of posts where users' attention distributions have a non-significant or significant difference in different sentiment polarity or number of consistent annotations

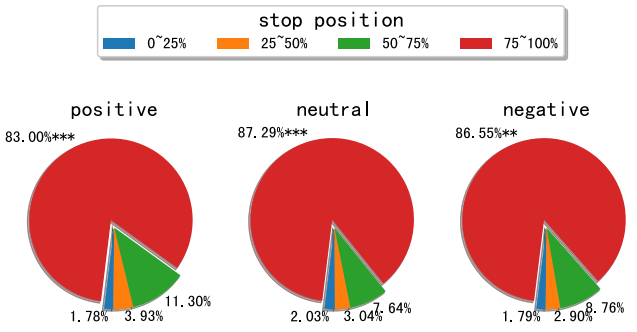

Fig. 6 The distribution of early stopping position in different sentiment blogs. "*/**/***" have the same meaning as in Table 2 (t-test)

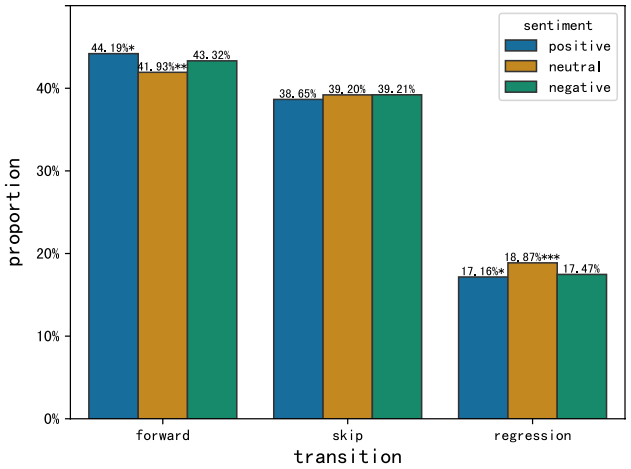

Fig. 7 Fixations transitions on the posts. "*/**/***" have the same meaning as in Table 2 (t-test)

Table 8 Reading behaviors in different posts position in three sentiment blogs

Metrics	Position	Positive	Neutral	Negative	Sig.
Fixation rate	0~25%	0.220	0.217	0.237	
	25 ~ 50%	0.214	0.212	0.206	
	50 ~ 75%	0.198	0.221	$0.224^{pos\star\star}$	**
	75 ~ 100%	0.189	$0.222^{pos\star\star\star}$	$0.219^{pos\star\star\star}$	***
Time per word	0 ~ 25%	42.91	41.45	46.78	
	25 ~ 50%	40.61	40.23	39.29	
	50 ~ 75%	37.69	41.58	$42.51^{pos\star\star}$	**
	75 ~ 100%	37.05	$42.74^{pos\star\star\star}$	$42.80^{pos\star\star\star}$	***

Higher average fixation rate and longer average reading time per word indicate that the users put more attention in this part of the posts. "pos" is the abbreviation of "Positive" in the table. "*/**/***" and "$\star$/$\star\star$/$\star\star\star$" have the same meaning as in Table 2 and Significance Tests are also as same as in Table 2

human-like models considering the transfer capability that may perform well. We will discuss this in more detail later.

5 Attention allocation mechanism

To address **RQ3**, we analyze the attention allocation mechanisms in users' reading process during sentiment judgment. In the first part of the section, we study attention decay and early stopping in the process. Then, we investigate attention allocation in three kinds of polarity posts by point-wise, line-wise, and phrase-wise eye movement features.

5.1 Attention decay and early stopping

Attention decay and early stopping are common phenomena in human reading behavior when given tasks, which is ubiquitous in the examination of the search results on SERPs [6, 9], and long document annotation tasks [18, 48]. As shown in Table 8 and Fig. 6, when users read posts and make sentiment judgments, there is no obvious attention decay, and users tend to stop reading at the end position of the majority of blogs except for positive posts. As we discussed in Sect. 4, these positive posts are relatively easier for users to understand. Thus, users will have higher confidence in judging them and stop reading earlier. Based on Table 8, users like putting more attention to the end position of neutral and negative posts than positive ones. We infer that when users reach the end of these two sentiment blogs, they would not be as confident as positive ones, then read slowly to make the right decisions.

5.2 Eye movement features

Eye movements are a series of gaze actions composed of fixation and saccade. We divide all metrics about eye movement features into three types: point-wise, line-wise, and phrase-wise. Point-wise features like pupil size and the duration or number of fixations or saccades generated from independent actions. Line-wise features like the distance or speed of the saccade and the number of regressions generated from two adjacent actions. Phrase-wise features like the start or end phrase of eye movements in a blog are generated from several continuous actions.

Firstly, we analyze the fixation transitions on the posts and split them into three categories based on [27] including *Forward*, *Regression* and *Skip*. Percentages of the three transitions are shown in Fig. 7, which are similar in all sentiment polarities. Most transitions are forwards, followed by skips, and then regressions. There are fewer forwards and more regressions transitions in the neutral posts, indicating users may have less

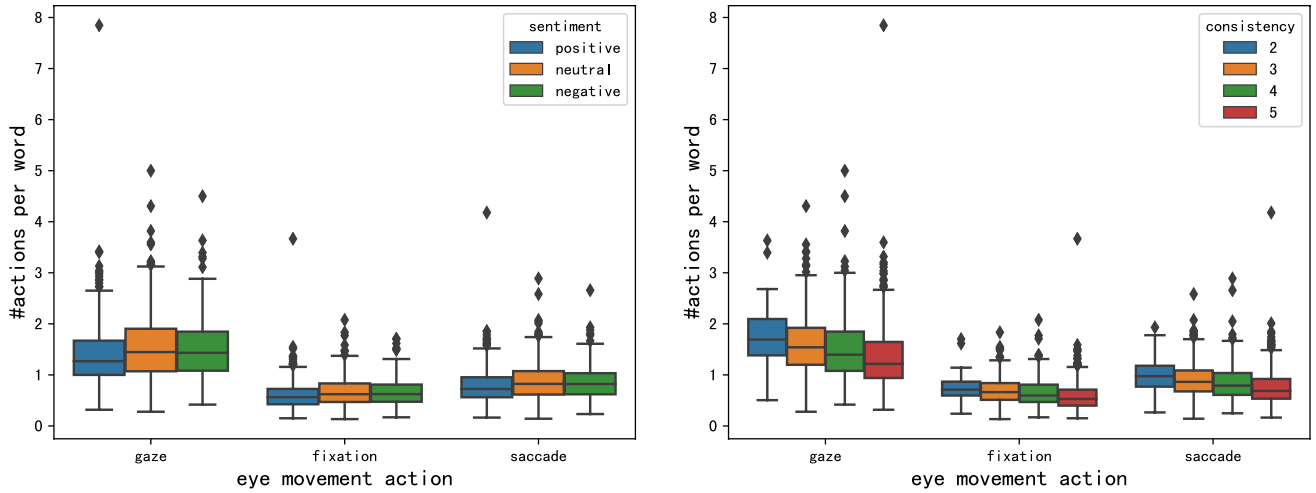

Fig. 8 The number of gaze/fixation/saccade actions per word in different sentiment and consistency blogs

interest in reading the details. There exist more forward and fewer regressions transitions in positive, which verifies the conclusion we proposed that positive posts are relatively easier for users to understand.

All of the other eye movement features demonstrate a similar annotation behavior as follows. When users judge a post, eye movements are more influenced by task difficulty rather than sentiment. Besides, the judgment processes of positive blogs are much easier, which reflects in higher reading speed, fewer gaze actions per word, etc. Moreover, the features are more similar in neutral and negative posts reading. We will not show details on these features and give a pair of examples shown in Fig. 8.

5.3 Summary

This section investigates the attention allocation mechanism in sentiment judgment to answer our research question **RQ3**. We found that eye movements are more related to task difficulty than sentiment. Among blogs of three sentiment polarities, users have higher confidence to judge a positive blog and lower interest in neutral blogs. Besides, users prefer to read the whole text in the sentiment judgment processes of post-type documents.

6 Sentiment prediction

In this section, we attempt to make use of the observations to design models. First, we use pure eye movements to predict the sentiment and consistency, then combine the users' attention with textual features to predict sentiment in different ways to investigate **RQ4**.

6.1 Behavior features

We organize the eye movement features as model input and perform fivefold cross-validation to evaluate classical models' prediction performance of sentiment and consistency. On the one hand, Table 9 shows that eye movement features are not efficient enough to train models which could achieve high accuracy in 3-class classification. The result supports our proposed reading model that sentiment judgment is an auxiliary task to content comprehension. On the other hand, we find that models using these features predict better in 4-class consistency tasks, which is associated with that eye movements are more related to task difficulty than sentiment.

6.2 Behavior and text features

As shown in Table 2, attention agreement in word-level is lower especially in easier tasks, and a large percentage of sentiment annotations are easy tasks according to Fig. 4. Humans have better word association and prediction skills in reading [33], so the attended words, even adjacent words we proposed, may still not align with users' perceived context perfectly enough. Then we introduce the *sentence-level*

Table 9 Prediction accuracy of sentiment and consistency by classic models using only eye movement features

	LR	SVM	KNN	RF	GBDT
Sentiment (3-class)	0.393	0.402	0.402	0.417	0.423
Consistency (4-class)	0.410	0.426	0.434	0.430	0.435

Table 10 A simple validation of model design based on our observations in long and short sentences settings

	Regularization		Log loss	Accuracy	Macro-f1
	Word	Sentence			
Long	\	\	0.839	0.627	0.609
	Fixated	\	0.891	0.564	0.518
	Adjacent	\	0.845	0.616	0.593
	\	Fixated	0.858	0.605	0.585
	\	Adjacent	0.861	0.598	0.576
short	\	\	0.835	0.623	0.604
	Fixated	\	0.84	0.623	0.602
	Adjacent	\	**0.829**	0.609	0.592
	\	Fixated	0.839	0.617	0.599
	\	Adjacent	0.831	**0.636**	**0.622**

Bold values indicate the best performance than other models, and underlined values indicate the second-best performance

attention to capture human attention distribution on posts better. The definition is the number of attended words divided by the total words in the sentence. There are two kinds of sentence-level attention corresponding to fixated words and adjacent words. The general sentence segment method is based on the full stop punctuation in the paragraph. However, considering the task difficulty, users could make sentiment judgments without reading the entire sentence. Then, we cut a long sentence into several short sentences by commas to better catching users' attention.

Hierarchical Attention Networks [43] are used in our experiment to validate the utilities of our findings. As shown in formula 5, the loss of the model consists of two parts, i.e., the cross-entropy loss of the ground truth labels to the predicted labels and the l2 loss to regularize the attention distribution of the model to the human in the word-level and sentence-level. In the loss function, α, β, and γ are used for trading off different parts of the loss, and their sum is constrained to one in our settings. The fivefold cross-validation results are shown in Table 10. We have to admit that 1224 samples in our study are a relatively small dataset. The model incorporating the short-sentence level attention calculated from adjacent words achieves a better performance than others. To a certain extent, the results show that a better sentiment analysis model should reasonably use user behavior.

$$\begin{cases} \mathcal{L} = \alpha\mathcal{L}_{label} + \beta\mathcal{L}_{sentence} + \gamma\mathcal{L}_{word} \\ \mathcal{L}_{label} = CrossEntropy(\hat{y}, y) \\ \mathcal{L}_{word|sentence} = \|HAM - MAM\|_2 \end{cases} \tag{5}$$

6.3 Discussion

Inspired by the observations in our study, we believe that there are three directions of model design for sentiment classification or even text classification that are worth investigating. We will describe three directions as follows, including *Continuous attention*, *Auxiliary tasks*, and *Dynamic policy*.

6.3.1 Continuous attention

Single-layer attention in classifiers tries to find the keywords related to labels, and multi-head attention in transformers takes the relationship between the current word and others into account. Both popular methods underestimate the particularity of near words. However, several words are processed together by humans during the reading process, which helps humans better perceive the documents' meanings. The current attention mechanism pays attention to the discrete words, which is more like solving the classification problem by a matching method rather than an understanding way. If a model adopted continuous attention mechanisms, it might be equipped with better association and understanding. Fixing a fixed attention window size or dynamically resizing it by sentence structure is also under-investigated when adopting continuous attention into models.

6.3.2 Auxiliary tasks

In our user study settings, participants are asked to annotate the sentiment polarity and emotion together, which means that users behave similarly in these dependent tasks. According to the questionnaires from users, they told us that they could judge sentiment and emotion simultaneously, which supports our assumption. As for independent tasks, users also act out high agreement behaviors across sentiment and topic annotation tasks. Existing work has shown that training sentiment classification models with auxiliary tasks, like POS tag or subjectivity extraction, will achieve better results. [29]. However, the principle of selecting valuable auxiliary tasks for the main task is unknown. There are several dependent tasks like emotion classification, sentiment word recognition, and independent tasks like POS recognition, named entity recognition, topic judgment for the sentiment classification task. Nevertheless, the characteristics of efficient auxiliary tasks remain to be studied.

6.3.3 Dynamic policy

During the reading processes, users could perceive the task difficulty and dynamically adjust the reading policies, like

forward, skip, regression, or early stopping. These policies depend on the distance between users' comprehension and the target of their tasks. However, most models ignore the distance and input a complete document, then make judgments as we know. Incorporating these policies into decision processes may help models better understand text meanings. As far as we know, [45] shows that training the reinforcement learning model for text classification through some of these user actions could achieve better performance. However, the action difference between their model and humans' decision remains unknown. It is still a long way to reconsider the human decision processes to build effective classifiers.

7 Conclusion

To compare the differences between machine and human making decisions in sentiment annotation tasks and further shed light on the model design, we investigate the humans' reading behavior during making sentiment judgment in this paper. By conducting a carefully designed eye-tracking experiment, we observe that users pay more attention to content comprehension than task completion, while attention models are trying to build the latent relationship between the sentiment clue in content and the labels. Besides, we found that task difficulty and users' preferences could significantly influence the attention allocation policy in reading. Users are able to understand the text in greater depth than current models, and we discuss three directions: Continuous Attention, Auxiliary Tasks, and Dynamic Policy for sentiment classification design to improve the comprehension ability of the machine.

Nevertheless, our work has some limitations, which remain to be considered in future work. (1) We only focus on the reading behavior on microblogs during sentiment judgment. While humans are reading other documents with different task difficulties, lengths, or text styles, the behavior may change. (2) The reading patterns are observed on Chinese microblogs tasks, so the generalization ability in other languages is still under-investigated. (3) The effectiveness of the proposed directions in model design is waiting for further verification and investigation.

References

1. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473
2. Barrett M, Bingel J, Hollenstein N, Rei M, Søgaard A (2018) Sequence classification with human attention. In: Proceedings of the 22nd conference on computational natural language learning, pp 302–312
3. Berger A, Della Pietra SA, Della Pietra VJ (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–71
4. Bicknell K, Levy R (2010) A rational model of eye movement control in reading. In: Proceedings of the 48th annual meeting of the association for computational linguistics, pp 1168–1178
5. Bolotova V, Blinov V, Zheng Y, Croft WB, Scholer F, Sanderson M (2020) Do people and neural nets pay attention to the same words: studying eye-tracking data for non-factoid qa evaluation. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 85–94
6. Craswell N, Zoeter O, Taylor M, Ramsey B (2008) An experimental comparison of click position-bias models. In: Proceedings of the 2008 international conference on web search and data mining, pp 87–94
7. Engbert R, Nuthmann A, Richter EM, Kliegl R (2005) Swift: a dynamical model of saccade generation during reading. Psychol Rev 112(4):777
8. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, Wu XC, Durbin EB, Doherty J, Stroup A et al (2021) Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 25:3596–3607
9. Granka LA, Joachims T, Gay G (2004) Eye-tracking analysis of user behavior in www search. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp 478–479
10. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
11. Hosmer DW Jr, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, vol 398. Wiley, Hoboken
12. Jain PK, Quamer W, Pamula R, Saravanan V (2021) SpSAN: Sparse self-attentive network-based aspect-aware model for sentiment analysis. J Ambient Intell Humaniz Comput 1–18
13. Jain S, Wallace BC (2019) Attention is not explanation. arXiv preprint arXiv:190210186
14. Just MA, Carpenter PA (1987) The psychology of reading and language comprehension. Allyn & Bacon, Boston
15. Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:14085882
16. Klerke S, Goldberg Y, Søgaard A (2016) Improving sentence compression by learning to predict gaze. arXiv preprint arXiv:16040 3357
17. Li X, Pollatsek A (2020) An integrated model of word processing and eye-movement control during Chinese reading. Psychol Rev 127(6):1139
18. Li X, Liu Y, Mao J, He Z, Zhang M, Ma S (2018) Understanding reading attention distribution during relevance judgement. In: Proceedings of the 27th ACM international conference on information and knowledge management, pp 733–742
19. Liaw A, Wiener M et al (2002) Classification and regression by randomforest. R News 2(3):18–22
20. Liu KL, Li WJ, Guo M (2012) Emoticon smoothed language models for twitter sentiment analysis. Aaai Citeseer 12:22–26
21. Liu Y, Chen F, Kong W, Yu H, Zhang M, Ma S, Ru L (2012) Identifying web spam with the wisdom of the crowds. ACM Trans Web TWEB 6(1):1–30
22. Liu Y, Wang C, Zhou K, Nie J, Zhang M, Ma S (2014) From skimming to reading: a two-stage examination model for web search. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pp 849–858
23. Liversedge SP, Findlay JM (2000) Saccadic eye movements and cognition. Trends Cognit Sci 4(1):6–14
24. Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of

the 49th annual meeting of the association for computational linguistics: human language technologies, pp 142–150
25. Marimuthu K, Devi SL (2012) How human analyse lexical indicators of sentiments-a cognitive analysis using reaction-time. In: Proceedings of the 2nd workshop on sentiment analysis where AI meets psychology, pp 81–90
26. McCallum A, Nigam K, et al. (1998) A comparison of event models for Naive Bayes text classification. In: AAAI-98 workshop on learning for text categorization, Citeseer, vol 752, pp 41–48
27. McDonald SA, Shillcock RC (2003) Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. Vis Res 43(16):1735–1751
28. Mishra A, Dey K, Bhattacharyya P (2017) Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 377–387
29. Mishra A, Tamilselvam S, Dasgupta R, Nagar S, Dey K (2018) Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior. In: Thirty-second AAAI conference on artificial intelligence
30. Osman NA, Mohd Noah SA, Darwich M, Mohd M (2021) Integrating contextual sentiment analysis in collaborative recommender systems. PLoS One 16(3):e0248695
31. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070
32. Rayner K (2009) The 35th sir Frederick Bartlett lecture: eye movements and attention in reading, scene perception, and visual search. Q J Exp Psychol 62(8):1457–1506
33. Reichle ED, Pollatsek A, Fisher DL, Rayner K (1998) Toward a model of eye movement control in reading. Psychol Rev 105(1):125
34. Reichle ED, Rayner K, Pollatsek A (2003) The ez reader model of eye-movement control in reading: comparisons to other models. Behav Brain Sci 26(4):445
35. Salmerón L, Delgado P, Mason L (2020) Using eye-movement modelling examples to improve critical reading of multiple webpages on a conflicting topic. J Comput Assisted Learn 36(6):1038–1051
36. Sen C, Hartvigsen T, Yin B, Kong X, Rundensteiner E (2020) Human attention maps for text classification: Do humans and neural networks focus on the same words? In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 4596–4608
37. Sergio GC, Lee M (2021) Stacked debert: all attention in incomplete data for text classification. Neural Netw 136:87–96
38. Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
39. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
40. Wiegreffe S, Pinter Y (2019) Attention is not not explanation. arXiv preprint arXiv:190804626
41. Yadav RK, Jiao L, Granmo OC, Goodwin M (2021) Human-level interpretable learning for aspect-based sentiment analysis. In: The thirty-fifth AAAI conference on artificial intelligence (AAAI-21). AAAI
42. Yan G, Wang L, Wu J, Bai X (2011) A study on eye movements of different grade students' reading perception span and parafoveal preview (in Chinese). Acta Psychol Sin 03:249–263
43. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489
44. Yayla M, Demirkol MD, Alqaraleh S (2021) Cnn vs. lstm for turkish text classification. In: 2021 international conference on innovations in intelligent systems and applications (INISTA), IEEE, pp 1–6
45. Yu K, Liu Y, Schwing AG, Peng J (2018) Fast and accurate text classification: skimming, rereading and early stopping. In: International conference on learning representations (ICLR)
46. Zhao J, Dong L, Wu J, Xu K (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, pp 1528–1531
47. Zheng WL, Liu W, Lu Y, Lu BL, Cichocki A (2018) Emotionmeter: a multimodal framework for recognizing human emotions. IEEE Trans Cybern 49(3):1110–1122
48. Zheng Y, Mao J, Liu Y, Ye Z, Zhang M, Ma S (2019) Human behavior inspired machine reading comprehension. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 425–434