Analyzing the Intensity of Complaints on Social Media

Ming Fang[1] **Shi Zong**[1] **Jing Li**[2] **Xinyu Dai**[1] **Shujian Huang**[1] **Jiajun Chen**[1]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
fangming@smail.nju.edu.cn,
{szong, daixinyu, huangsj, chenjj}@nju.edu.cn
[2]Department of Computing, The Hong Kong Polytechnic University, HKSAR, China
jing-amelia.li@polyu.edu.hk

Abstract

Complaining is a speech act that expresses a negative inconsistency between reality and human expectations. While prior studies mostly focus on identifying the existence or the type of complaints, in this work, we present the first study in computational linguistics of measuring the *intensity* of complaints from text. Analyzing complaints from such a perspective is particularly useful, as complaints of certain degrees may cause severe consequences for companies or organizations. We create the first Chinese dataset containing 3,103 posts about complaints from Weibo, a popular Chinese social media platform. These posts are then annotated with complaints intensity scores using Best-Worst Scaling (BWS) method. We show that complaints intensity can be accurately estimated by computational models with the best mean square error achieving 0.11. Furthermore, we conduct a comprehensive linguistic analysis around complaints, including the connections between complaints and sentiment, and a cross-lingual comparison for complaints expressions used by Chinese and English speakers. We finally show that our complaints intensity scores can be incorporated for better estimating the popularity of posts on social media.[1]

1 Introduction

Complaining is caused by the gap between reality and people's expectations (Olshtain and Weinbach, 1985). Brown et al. (1987) state that the purpose of complaining is not to confirm that the two parties have reached an agreement but to face-threatening acts. People use complaints to express their concerns or dissatisfaction based on the severity and urgency of situations.

Researchers from linguistics and psychology have long pointed out that people may shape their complaints to varying degrees (Olshtain and Weinbach, 1985; Jenkins and Cangemi, 1979; Trosborg, 2011). Leech (2016) classifies complaints as conflicting speech acts. Mild complaints can reach the purpose of venting emotions to promote mental health, but serious complaints can lead to hatred and even bullying behaviors (Iyiola and Ibidunni, 2013). In computational linguistics, prior studies primarily focus on building automatic classification models for identifying the existence of complaints (Preotiuc-Pietro et al., 2019). Most recently, Jin and Aletras (2021) provided a dataset annotated with different severity levels of complaints based on the theory of pragmatics, including four distinct categories "no explicit reproach", "disapproval", "accusation" and "blame".

Among these studies, we note one missing piece is to measure the *intensity* of complaints. To illustrate this point, consider two sentences from the newest annotated dataset from Jin and Aletras (2021): *can i complain to you about the coffee i just received ?* and *virgin media as usual full of lies lies lies ! ! !*. Although these two complaints may have the same type "accusation", it is clear that they are different regarding the degree of complaints. As another example, *totally not cool.* and *please reply my dm asap ! ! !* are both classified as "disapproval". However, the latter makes a stronger complaint. Analyzing different complaints levels can also be beneficial. Companies need to regularly monitor the feedback from users, as certain complaints may significantly impact the reputation of their products. Organizations or governments need to monitor people's biggest complaints to understand their urgent needs.

In this work, we analyze the intensity of complaints on social media. To the best of our knowledge, it is the first computational linguistics study that tries to automatically capture the complaints intensity from text. We present the first Chinese complaints intensity dataset, consisting of 3,103 posts

[1] Our annotated corpus is publicly available at https://github.com/nlpfang/complaint_intensity.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1742 - 1754
July 10-15, 2022 ©2022 Association for Computational Linguistics

from Weibo. We then show that the complaints intensity can be measured from text by building computational models. We further demonstrate the necessity and importance of understanding complaints intensity. This includes a detailed analysis that distinguishes the differences between our complaint intensity scores and original sentiment scores. As a pilot study for complaints in Chinese, we also perform a cross-lingual analysis to understand the differences in the complaint expressions used in Chinese and English. We have some interesting empirical findings. For example, we observe that English speakers tend to use more ironic expressions than Chinese speakers. Finally, we show how our annotated corpus can help predict the popularity of posts on social media.

2 Data

In this section, we present the first Chinese dataset that is annotated towards the intensity of the complaints reflected from text.

2.1 Data Collection

We collect data from Weibo,[2] a famous social media platform in China that is similar to Twitter. As posts about complaints only account for a minority of the total posts on Weibo, in this work we consider education domain – an area that is the primary focus for most families in China, which generally raises hot debates and complaints about current education policies. We selected a set of keywords related to complaints, including 抱怨 (*complaint*), 不公平 (*unfair*), and 举报 (*report*). We then randomly sampled 5 hashtags around these keywords and collected Weibo posts from these hashtags. We collected a total of 4,490 Weibo posts from August 2020 to May 2021.

Pre-processing. We notice that the hashtag on Weibo is usually a sentence (in "#...#" format), rather than a phrase like its Twitter counterparts. To ensure a certain amount of content generated by users, we filtered out posts with less than 10 words and more than 200 words (without hashtags). For each post, we removed the name of the author, location tags, and URLs. We also converted emoticon into text format. Finally, 3,103 Weibo posts remain for annotation. Table 1 shows the breakdown statistics in our corpus.

Hashtag	**Num.**
#代表建议让学生在校内完成家庭作业# (*#The representative suggested that students should complete their homework on campus#*)	762
#江苏明确教师不得用手机布置作业# (*#Jiangsu Province makes it clear that teachers are not allowed to use mobile phones to assign homework#*)	534
#院士不建议普通孩子学奥数# (*#Academician does not recommend ordinary children to learn Mathematical Olympiad#*)	627
#西安外国语大学封闭管理# (*#Close management of Xi'an International Studies University#*)	598
#人大法硕复试30余人成绩0分# (*#More than 30 people scored 0 in the postgraduate examination of law at Renmin University#*)	582
Total	3,103

Table 1: Hashtags and number of collected Weibo posts in our annotated corpus.

2.2 Data Annotation

Complaints Levels. Our goal is to measure the intensity of complaints from text. We adopt the definition from Jenkins and Cangemi (1979), which quantifies the complaints into five levels, as shown in Table 2. Higher levels indicate stronger complaints.

Level	**Description**
1	a little anxiety and disgust
2	deliberately expressing anxiety
3	actively looking for ways to solve anxiety
4	frustrated behavior
5	depression, fear, and despair

Table 2: Guideline used in annotation process for distinguishing different levels of complaints, adopted from Jenkins and Cangemi (1979).

In pilot studies, we test the feasibility of using these levels as the annotation guideline for the annotators, along with the potential mismatches between Chinese and English speakers. We observe that annotators are able to make comparison between complaints of different degrees. As discussed later, our annotations also achieve high agreement between annotators.

Best-Worst Scaling (BWS). In this work, we annotate the complaint intensity using Best-Worst Scaling, proposed by Louviere and Woodworth (1991). We choose this method as it can produce more stable and fined-grained scores than directly scoring (Kiritchenko and Mohammad, 2017). We note similar methods have been applied to various tasks, including measuring offensiveness (Hada et al., 2021) and intimacy (Pei and Jurgens, 2020) in the computational linguistic literature.

In BWS annotation, annotators are provided with

[2] www.weibo.com

4-tuples randomly generated that meet certain criteria.[3] Annotators are then asked to select the strongest complaint item and the weakest complaint item within each 4-tuple. In practice, we randomly generated $2n$ distinct 4-tuples, with n being the number of posts. This amount of tuples is considered to be sufficient for getting reliable scores from annotation (Kiritchenko and Mohammad, 2017). We assign the complaint intensity score for each post by using the percentage of strongest cases minus the weakest cases, ranging from -1 to 1.

Annotation Quality. To ensure the quality of our annotations, we manually annotated 100 posts and asked all annotators to annotate them beforehand. We removed annotators whose accuracy is less than 70% on these golden annotations. To get highly reliable results, we got each tuple annotated by 3 annotators. In total, we received more than 14,000 annotations from 15 annotators.

We follow the literature (Kiritchenko and Mohammad, 2017) and measure the quality of annotations by using score-to-half reliability (SHR). SHR score is calculated by randomly splitting all the tuples into two halves and then computing the correlation between these two groups. We repeat the above process 100 times. The average SHR score is 0.91, which indicates strong reliability.

2.3 Data Analysis

We first analyze the distribution for the annotated complaint intensity scores in our corpus. As shown in Figure 1 (Left), we observe a normal distribution for the number of posts across different complaint scores, with most of the posts having intensity of complaints within -0.2 and 0.2.

We also observe that the length of complaint posts (intensity>0) is longer than that of non-compliant posts (intensity<0) in Figure 1 (Right). By examining our data, we observe it is because stronger complaints contain more details with more aspects. For example, in bin 5 of Table A1 (in Appendix A), the target of complaint changes from 学校 (*school*) to dissatisfaction with 图书馆 (*library*) and even accuses the behavior of 门卫 (*security guard*). This is the halo effect in psychology: if something leaves a wrong impression, everything related to it becomes terrible. On the contrary, we observe that most non-complaining posts contain only plain expressions, and people will not describe too much after expressing their opinions on the matter.

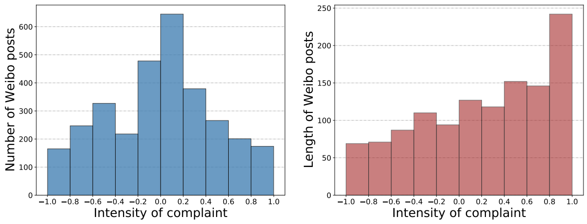

Figure 1: Distribution for the number and length of posts over complaints intensity in our corpus.

3 Predicting the Intensity of Complaints

In Section 2, we have a dataset annotated with the intensity of complaints from -1 to 1. We now build computational models for predicting the intensity of complaints of a given post.

3.1 Models

Support Vector Regression (SVR). We use support vector regression as our first baseline model. We experiment with two different input sentence representations: bag of {2,3,4}-gram features and 300-dimensional GloVe embeddings (Pennington et al., 2014). Results in Table 4 use an RBF (Radial Basis Function) kernel. We observe similar results using other kernels in practice (e.g., linear kernel).

Bidirectional LSTM. We also experiment with a bidirectional Long Short-Term-Memory (Bi LSTM) model. The LSTM and the average pooling layer concatenation are passed through a linear layer with a tanh activation, producing a score between -1 and 1. We use two sets of embedding for input layers: Glove (Pennington et al., 2014) and BERT for embedding (Li et al., 2020). The attention mechanism is also considered. Other hyperparameters for the models are a batch size of 64, a learning rate of 1e-3, 13 epochs with early stopping, and a dropout of 0.5 to avoid overfitting.

Pre-trained Models. We finally experiment with pre-trained models, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ERNIE (Zhang et al., 2019). For all pretrained models, we add a linear layer as a regression layer to the model. We then fine-tune these models using a mean square error loss objective. We set the batch size to be 16 and learning rate to be 2e-5. The model is trained for 3 epochs. All hyperparameters are selected using a held-out dev set.

[3] Requirements are: (1) no two 4-tuples are the same; (2) no two posts within a 4-tuple are identical; (3) each post appears approximately in the same number of 4-tuples; (4) each pair of posts appears approximately in the same number of 4-tuples.

Bin	Weibo posts	Scores
1	以前在没有手机的年代，孩子们都是自己记作业。我觉得有助于形成自我管理能力 (*In the era when there were no mobile phones, children kept their homework by themselves. I think it helps to form self-management ability.*)	-1
2	最好的解决办法就是没有家庭作业，对老师对家长都好 (*The best solution is to have no homework, which is good for teachers and parents.*	-0.56
3	现在的学生作业为啥都得用手机做? (*Why do students have to use mobile phones to do their homework now?*)	+0.12
4	我是真的觉得用手机交作业很烦 (*I find it really annoying to hand in homework with my mobile phone.*)	+0.4
5	气死了！食堂涨价，超市关门，就没人管理嘛? (*Mad! The price of the canteen increases, and the supermarket closes, no one manages it?*)	+1

Table 3: Sample posts with complaints intensity scores. We divide our scoring scale (from -1 to 1) into 5 bins of size 0.4 (i.e., bin 1 refers to scores ranging from -1.0 to -0.6, bin 2 from -0.6 to -0.2, etc.). More examples are provided in Table A1 in Appendix A.

3.2 Experiments

We evaluate the model performances for our complaint intensity prediction task in two settings: (1) mix hashtag, where we combine Weibo posts from different hashtags together, and (2) cross hashtag, where the posts for train, dev and test sets are separately from different hashtags. We use Pearson's correlation and MSE (Mean Square Error) as metrics for all our experiments.

Mix Hashtag. We combine Weibo posts from different hashtags together and then split them into 80% for training, 10% for validation, and 10% for test. Results are shown in Table 4. We observe that RoBERTa outperforms all other models and reaches Pearson up to 0.79, followed by the LSTM model. The SVR model has the worst performance.

Cross Hashtag. We choose four of total five collected hashtags as the train and development set, and hold out the rest one for test. We report the average value after five experimental runs in Table 4. We observe under cross hashtag setting, models get comparable performances to mix hashtag setting. It indicates models seem to learn common linguistic cues between different hashtags.

Models	Mix Hashtag		Cross Hashtag	
	r	MSE	r	MSE
SVR ({2,3,4}-gram)	0.36	0.46	0.35	0.46
SVR (GloVe)	0.49	0.36	0.47	0.38
LSTM (GloVe)	0.69	0.24	0.65	0.27
LSTM Attn (Glove)	0.72	0.22	0.70	0.25
LSTM (BERTembed)	0.76	0.15	0.75	0.16
ERNIE	0.76	0.14	0.76	0.14
BERT	0.77	0.20	0.75	0.23
RoBERTa	0.79	0.11	0.78	0.11

Table 4: Pearson's r and Mean Square Error (MSE) on two datasets for predicting the intensity of complaints.

3.3 Error Analysis

We perform an error analysis to shed light on the limitations of our best-performing model. A prediction is defined to be wrong when the difference between ground truth and the predicted score is greater than 0.5. We randomly sample 100 errors and manually inspect them. All errors are divided into three categories: 43% of errors are because of irony expression in complaints, 29% are due to implicit expressions, and 28% are due to the insufficient and vague expressions.

Ironic Expressions. We observe that most errors happen when the posts contain ironic expressions. Users use positive words such as "perfect" or "great" to express dissatisfaction, which are misleading models to ignore the implicit complaints. A typical example is as follows: 食堂的涨价消息比封校政策来的快, 这学校真好 (*The news of price increase in the cafeteria is coming faster than the school closure policy. This school's management is really good*).[4]

Implicit Expressions. The model struggles with complaints expressed in more subtle ways. These complaints do not contain any prominent negative words, but through other means like strike a chord or entrust. In the following example, the user hopes that managers can personally experience the status quo to understand the user's dissatisfaction. Therefore, predicting them correctly requires more contextual understanding: 真的非常极其的希望校领导也能来感受一下封校的生活 (*I really hope that school leaders can also come and experience the life of the closed school*).

[4] During COVID-19 pandemic, Chinese universities restrict students from going outside and limit their activities within campus. It thus causes students' complaints.

Vague Expressions. We observe that the model is likely to be confused by vague or incomplete expressions in the posts. Consider the following example: 赶上这破事，如果教育真的公平不如直接取消复试吧 (*Encountered this shit. If the education is really fair, it is better to cancel the exam*). The hypothetical relationship using 如果 (*if*) is an expression of uncertainty (Wei et al., 2018) and there are no prominent emotional words in the text. Thus, our model fails to understand the speaker's intention well.

4 Complaints as an Emotion

From Table 3, we notice stronger complaints seem to be associated with negative emotion words. Prior studies also point out that complaints can be treated as an influential emotional dimension (Iyiola and Ibidunni, 2013). Then a natural question to ask is whether existing sentiment models have already been able to predict complaints intensity scores and our annotation efforts are actually not needed?

In this section, we demonstrate the necessity of building corpus annotated with complaints intensity, by showing the model trained on standard sentiment datasets fails to do well in our complaints intensity prediction task. We also show that analyzing complaints can be a useful complement for sentiment analysis.

4.1 Differences between Complaints and Sentiment

In sentiment analysis, models normally output a score between 0 and 1, indicating how likely a post is to express negative emotion. Here we make the assumption that the most negative emotion may lead to the strongest complaint. We first examine if these probability scores from sentiment models can be used as intensity scores for measuring complaints intensity.

Setup. To ensure a fair comparison, we select a newly developed dataset on COVID-19, collected also from Weibo using hashtags related to COVID-19 by Lyu et al. (2020). The dataset contains 21,174 posts with fine-grained emotion annotations.[5]

In our experiments, we follow the same steps in Section 2.1 to pre-process this dataset. As our pre-processing steps result in a category imbalance issue, we merge categories with similar emotions. Specifically, we merge labels “fear”, “anger”, “disgust”, and “sadness” into “negative” category and merge labels “gratitude”, “surprise”, and “optimism” into the “positive” category. Finally, we have 8,783 posts in the negative category and 8,336 posts in the positive category.

We use BERT to train a sentiment model using the above COVID-19 data. We also sample 80% of our corpus for developing our BERT-based complaints model. The performances of both models are compared on the left 20% annotated posts. During evaluation, the sentiment scale from 0 to 1 is linearly mapped to our complaints intensity interval from -1 to 1.

Results. Results are shown in Table 5. We observe that using the probability scores from sentiment models shows decent performance on our complaints intensity prediction task. It indicates a clear connection between complaints and emotions. We also observe that models trained on our annotated corpus outperform sentiment model, demonstrating the necessity of building such corpus for complaints intensity estimation.

Model	Pearson's r	MSE
Complaint	0.76	0.20
Sentiment	0.71	0.24

Table 5: Performances of sentiment model and complaint model for complaints intensity prediction task.

Valence and Arousal. We also quantitatively studied the correlation between complaints and sentiment through Valence-Arousal. Valence can be positive or negative and corresponds to the standard dimension of sentiment analysis; Arousal, which can be low or high and express the degree (Vorakitphan et al., 2020). We use the VA score annotated by Xu et al. (2021), which contains 11,310 simplified Chinese words. The valence and arousal ratings include scores -3 to +3 for valence rating and scores 0 to 4 for arousal rating.

We identify sets of words in the Valence-Arousal lexicon that have high valence scores (>2), low valence scores (<-2), high arousal scores (>3), and low arousal scores (<2). A similar approach is used in Hada et al. (2021). We average the scores of tokens from the above four dimensions in each post and calculate the correlation with our complaints intensity. Results in Table 6 show low valence and high arousal are more correlated with complaints intensity compared to the other two dimensions.

[5] To the best of our knowledge, there does not exist a Chinese Weibo dataset annotated with continuous sentiment intensity. We note some datasets with discrete sentiment levels, like Douban movie short comments dataset (Ma et al., 2011).

Dimension	Pearson's r
High valence	0.02
Low valence	0.31
High arousal	0.18
Low arousal	0.05

Table 6: Pearson's correlation between the complaints intensity scores and emotion dimensions.

4.2 Complaints Help Sentiment Analysis

We now show that analyzing complaints could be helpful for the binary sentiment analysis task.

Models. We still use the COVID-19 dataset discussed in Section 4.1 for the binary sentiment classification task. We experiment with the SVM and BiLSTM-Attention models. The complaints score is added as an additional feature input to the model.

Results. Table 7 shows the results of the models on the sentiment classification task. Overall, we observe that the models with the complaint feature perform better than the original model. It demonstrates that a simple add-on can boost the prediction accuracy of sentiment classification for non-neural and traditional neural models. We also provide the performance of BERT for reference in Table 7.

Models	P	R	F1
SVM	0.51	0.49	0.50
+ complaint	0.53	0.50	0.51
BiLSTM-Att	0.72	0.70	0.71
+ complaint	0.74	0.71	0.72
BERT	0.79	0.76	0.77

Table 7: Results for binary sentiment prediction. F1 score of models with complaint feature is significantly better than the original model (p-value < 0.01, t-test).

4.3 Case Study

We are interested in what types of tokens sentiment model and complaint model try to capture. We thus take the BiLSTM-Attention model trained for sentiment classification task in Section 4.2 and our complaints model for comparison. We visualize the attention weights extracted from the above two models for the following example: 准备这么久的考试推迟真是绝了呵呵 (*After preparing for so long, the exam is now postponed. It's absolutely speechless. Hmm. How interesting.*). We observe that sentiment model assigns high attention weights for tokens 绝了 (*speechless*) and 呵呵 (*Hmm. How interesting.*), both expressing emotions. However, our complaint model puts high weights on tokens 推迟 (*postponed*) and 考试 (*exam*). These are tokens that reflect the reasons for complaining. These differences again demonstrate the need for building a specific dataset for complaints intensity.

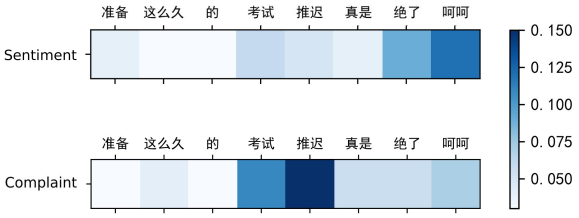

Figure 2: Attention weights for the sample sentences in sentiment model and our complaint model.

5 Cross-lingual Analysis

Our newly collected complaints intensity dataset is written in Chinese, while current existing datasets (Preotiuc-Pietro et al., 2019; Jin and Aletras, 2021) contain English tweets. It provides us an opportunity to understand the linguistic differences for complaints made by Chinese and English speakers on social media.

Our cross-lingual analysis is performed in the following way. First, we evenly sample 200 complaint tweets released in Jin and Aletras (2021) from their defined four categories. Similarly, we divide our annotated data with intensity greater than 0 (as complaint posts) into 4 bins for sampling. We ask 5 in-house annotators to mark and then use majority voting to decide if a post makes a direct or indirect complaint, along with the strategy used to make complaints.

Direct and Indirect Complaints. According to Boxer (2010), the speech of complaint can be divided into direct and indirect complaints. In social media, direct complaints are addressed to a complainee who is held responsible for the complaint action, and the addressee is fully or partially responsible for this behavior. Here is an example for a direct complaint: 学校的管理怎么这么不人性化? (*Why is the management of the school so inhumane?*) While indirect complaints refer that the recipient is not primarily responsible for perceived complaints and is more about the evaluation of the target or event, such as: 这样做让我们的压力很大,真无语 (*This puts us under a lot of pressure, so speechless*).

We compare the percentages of direct and indirect complaints for two languages based on our annotations. Results show that indirect expressions are more likely to be used in Chinese posts (80%

are indirect complaints). On the contrary, 91% of English tweets make complaints in a direct way. This finding seems to be consistent with the study of Deng et al. (2019), which demonstrates that Chinese people tend to use indirect expressions.

Strategy. In pragmatics, strategy is an appropriate countermeasure adopted to achieve the purpose of language communication. House and Kasper (1981) conducted a comparative study of English and German complaints from the direct degree and emotional markers. Anna (2008) discussed the choice of Chinese and English complaining strategies and proposed an explicit-implicit strategy. Implicit complaints are very subtle and even use metaphors to express complaints about the target, which requires more semantic information to capture. In contrast, explicit complaints can be further divided into types of with-redress and without-redress. With-redress is the strategy of request for repair; without-redress usually contains complaint targets or objects, which can be easily identified by recognizing an obvious complaint word or phrase.

Table 8 shows that strategy varies across languages. We find that the Chinese are more inclined to without-redress strategy, while the most frequent strategy used by Americans is with-redress strategies. It provides some empirical support for findings in Anna (2008).

Strategy	Chinese	English
Implicit	65%	12%
With-redress	13%	78%
Without-redress	22%	10%

Table 8: Percentages of strategies across languages.

Irony. Irony implies the opposite of the literal meaning. Dealing with non-literal means is a challenging task. On Twitter, Reyes et al. (2012) used specific hashtags as gold labels to detect irony in a supervised learning setting, such as *#irony* and *#sarcasm*. Attardo (2013) observed that native speakers are usually able to process the meaning of sarcasm automatically, but the ability of second language learners to infer meaning from context varies greatly. In Section 3.3, we observe irony counts for the majority of errors made by our model.

We analyze the number of complaints using irony. To detect ironic expressions, we separately use the Chinese irony dataset of Tang and Chen (2014) and the English dataset of Van Hee et al. (2018) to train the Bi-LSTM model. Results showed that 10% of Chinese data contained irony, and 26% of English data contained irony. It shows that English speakers use ironic expressions more often compared to Chinese speakers. Further, we conduct part-of-speech analysis on these ironic expressions. Table 9 shows that Chinese irony has the highest proportion of nouns, followed by verbs; while in English irony, verbs are the most, followed by nouns. In addition, there are more adjectives and adverbs in English than in Chinese.

Part of Speech	Chinese	English
Nouns	31.2%	27.9%
Verbs	21.8%	35.2%
Adjectives	3.1%	10.7%
Adverbs	9.9%	11.9%

Table 9: Percentages for POS tags across languages.

Limitations. We note a few limitations for our cross-lingual analysis. One limitation is domain mismatch. Our Chinese posts are collected from education domain, while English posts are collected from domains including food or online service. People may exhibit different behaviors when making complaints. We also note that the sample size for making comparisons is rather small, due to budget issues for experts annotations. In future work, we will perform a large-scale comparison by using the data collected from the same domains and utilizing crowdsourcing annotations.

6 Predicting Post Popularity

Finally, we demonstrate that complaint intensity scores from our computational models can help estimate the post popularity on social media. We envision incorporating these scores into existing social media monitoring systems to improve their prediction accuracy.

Task. Predicting the popularity of content on social media has been extensively studied in literature (Szabo and Huberman, 2010; Hong et al., 2011; Bao et al., 2013; Carta et al., 2020). Our task is to predict the popularity of a Weibo post. Specifically, given the popularity prediction $p(t_{i-1})$ at time t_{i-1}, we wish to predict the popularity $p(t_i)$ at next time step t_i. The popularity $p(t)$ is measured by the number of blog posts under the topic at time t.

Methods. We follow Szabo and Huberman (2010) and consider the following baseline that only uses early prevalence for prediction:

$$\ln \widehat{p}(t_i) = \alpha_1 \ln p(t_{i-1}) + \alpha_2,$$

where α_1 and α_2 are learnable coefficients. It is justified as a strong baseline in Bao et al. (2013) that a strong correlation exists between logarithmically transformed popularity and early popularity.[6]

The popularity of posts on social media can be measured by multiple dimensions (Bao et al., 2013). To show the effectiveness of our complaint scores, we add in the complaint intensity as a new term to estimate the final logarithmic popularity:

$$\ln \widehat{p}(t_i) = \beta_1 \ln p(t_{i-1}) + \underbrace{\beta_2 d_c(t_{i-1})}_{\text{complaints density}} + \beta_3,$$

where $d_c(t_{i-1})$ is the complaints density at time t_{i-1}, calculated by the ratio between the sum of the complaint intensity of blog posts per unit time to the number of all blog posts. β_1, β_2 and β_3 are learnable new coefficients from data.

Setup. We collected a new set of 4,973 posts under 8 hashtags on Weibo from March 2021 to November 2021, which are shown in Table A2 (in Appendix B). We pre-process these posts using the same steps in Section 2.1. Within each hashtag, 80% of the posts are used for training, and the left 20% are used for testing. We set the time step to be two hours. RMSE (root mean square error) and MAE (mean absolute error) are used to evaluate predicted results.

Results. We first examine the relationship between complaints density and post popularity as a sanity check. Results show a strong positive correlation with an upper cluster slope of 0.95.

We report our post popularity prediction results in Table 10. We observe our method that combines complaint density outperforms the baseline method. In Figure 3 we also show the comparison between our predictions and real values for a specific hashtag: #巨人教育宣布倒闭# (*#JuRen Education Group announces its bankruptcy#*). We observe that adding complaints scores help better estimate the post popularity, especially in the early stages. It is probably because complaints are likely to draw users' attention to engage in discussions and hence boost the popularity of events.

Method	RMSE	MAE
Baseline	0.41 ± 0.01	0.35 ± 0.01
+ complaints density	0.35 ± 0.02	0.33 ± 0.01

Table 10: RMSE and MAE for popularity prediction.

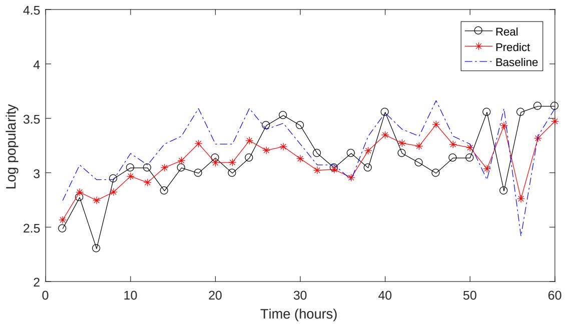

Figure 3: Comparison between actual post popularity and our predictions for hashtag *#JuRen Education Group announces its bankruptcy#*. RMSE = 0.32, MAE = 0.30.

7 Related Work

There have been various studies for complaints in linguistics, economics, and public opinion research. In Olshtain and Weinbach (1985), complaints are defined as what happened does not meet people's expectations, making people dissatisfied and blaming others. Kolodinsky (1995) explained the characteristics of consumers' complaining behavior from an economic point of view. Liu and Yen (2016) analyzed complaints about public transportation. Complaints on social media have also drawn great attention in recent years. Andreassen and Streukens (2013) focused on the study of the difference between social media complaints and traditional complaints, and argue that social media complaints are a unique way for consumers to express dissatisfaction. Balaji et al. (2015) studied the causes of complaints and found most of the complaints occur after a double deviation caused by dissatisfaction with the last solution. Motivated by these, in this paper, we collect complaints from Weibo, a widely used social media application.

In the area of linguistic studies on computational sociology, Meinl and Ebba (2010) studied the complaints act sequence in eBay reviews through 200 annotated English and German reviews. Ganesan and Zhou (2016) collected 2,500 reviews from Yelp and Walmart about commodity, then manually categorized them into 5 categories: negative only, complaint, positive only, raise, and irrelevant. Preotiuc-Pietro et al. (2019) focused on binary classification between complaints and non-complaints in various domains, such as food, car, online service, e-commerce. Jin and Aletras (2021) categorized complaints into 4 categories: no explicit reproach, disapproval, accusation, and blame. In this work,

[6] Bao et al. (2013) further proposed to use link density and diffusion depth for popularity prediction. Above methods were tested on WISE 2012 challenge dataset. We tried our best but are not able to have access to this dataset. In our own collected dataset, we do not have link or diffusion information.

we present the first study of estimating the intensity of complaints from text.

Our work is also related to prior work on emotion detection and sentiment intensity estimation (Mohammad and Bravo-Marquez, 2017; Cortis et al., 2017). Kiritchenko and Mohammad (2017) created a Twitter dataset annotated with sentiment intensity. We have discussed the connections between complaints and sentiment in detail in Section 4.

8 Conclusion

In this paper, we present the first study of measuring the intensity of complaints from text. We build a corpus of 3,103 Chinese Weibo posts about complaints, annotated with complaints intensity scores using Best-Worst Scaling method. We then demonstrate that our corpus supports the development of automatic computational models for accurate complaints intensity predictions. Furthermore, we study the connections between complaints and sentiment, and perform a cross-lingual comparison for complaints expressions between Chinese and English. We finally show that our complaints intensity scores help better estimate the posts popularity on social media.

Ethical Concerns

In this paper, we collect a complaint dataset from Weibo. The tools we use to collect posts comply with Weibo's terms of service. We will follow Weibo's policy for content redistribution to release our annotated corpus. Specifically, we will not release any user information or demographic data, including the authors' names, ages, and origins.

We recruited part-time research assistants for our annotation task. Annotators were warned that the complaint posts might contain offensive or upsetting content. Annotators were shown only anonymized posts and agreed not to make attempts to de-anonymize them. We did not collect any personal data from the annotators before, after, or during the annotation task. Moreover, we pay them 15.7 USD/hour and at most 14 hours per week.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback. This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/25200821) and CCF-Baidu Open Fund (No. 2021PP15002000).

References

Tor W Andreassen and Sandra Streukens. 2013. Online complaining: understanding the adoption process and the role of individual and situational characteristics. *Managing Service Quality: An International Journal*.

Anna. 2008. *Comparative Analysis of Complaint Strategies in Direct Complaint of Chinese and American College Students*. Ph.D. thesis, Dalian University of Technology.

Salvatore Attardo. 2013. Intentionality and irony. *Irony and Humor: From pragmatics to discourse*, pages 39–58.

MS Balaji, Subhash Jha, and Marla B Royne. 2015. Customer e-complaining behaviours using social media. *The Service Industries Journal*, 35(11-12):633–654.

Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. 2013. Popularity prediction in microblogging network: a case study on sina weibo. In *Proceedings of the 22nd international conference on world wide web*, pages 177–178.

Diana Boxer. 2010. How to gripe and establish rapport. *Speech act performance: Theoretical, empirical and methodological issues*, pages 163–178.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Salvatore Carta, Alessandro Sebastian Podda, Diego Reforgiato Recupero, Roberto Saia, and Giovanni Usai. 2020. Popularity prediction of instagram posts. *Information*, 11(9):453.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Xinmei Deng, Sieun An, and Chen Cheng. 2019. Cultural differences in the implicit and explicit attitudes toward emotion regulation. *Personality and Individual Differences*, 149:220–222.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kavita Ganesan and Guangyu Zhou. 2016. Linguistic understanding of complaints and praises in user reviews. In *Proceedings of NAACL-HLT*, pages 109–114.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for english reddit comments. *arXiv preprint arXiv:2106.05664*.

Liangjie Hong, Ovidiu Dan, and Brian D Davison. 2011. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58.

J. House and G. Kasper. 1981. *Politeness Markers in English and German*. Rasmus Rask Studies in Pragmatic Linguistics, Volume 2, Conversational Routine.

OO Iyiola and OS Ibidunni. 2013. The relationship between complaints, emotion, anger, and subsequent behavior of customers. *IOSR Journal of Humanities and Social Sciences*, 17(6):34–41.

William M Jenkins and Joseph P Cangemi. 1979. Levels of intensity of dissatisfaction: A model. *Education*, 99(4).

Mali Jin and Nikolaos Aletras. 2021. Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2264–2274, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Jane Kolodinsky. 1995. Usefulness of economics in explaining consumer complaints. *Journal of Consumer Affairs*, 29(1):29–54.

Geoffrey Leech. 2016. *Principles of pragmatics*. Routledge.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Weng-Kun Liu and Chia-Chun Yen. 2016. Optimizing bus passenger complaint service through big data analysis: Systematized analysis for improved public sector management. *Sustainability*, 8(12):1319.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Xiaoting Lyu, Zhe Chen, Di Wu, and Wei Wang. 2020. Sentiment analysis on chinese weibo regarding covid-19. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 710–721. Springer.

Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 287–296.

Meinl and M. Ebba. 2010. Electronic complaints: An empirical study on british english and german complaints on ebay.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

E Olshtain and L Weinbach. 1985. Complaints: A study of speech act behavior among native and non-native speakers of hebrew. the prag-matic perspective.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Daniel Preotiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. *arXiv preprint arXiv:1906.03890*.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88.

Yi-jie Tang and Hsin-Hsi Chen. 2014. Chinese irony corpus construction and ironic structure analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1269–1278.

Anna Trosborg. 2011. *Interlanguage pragmatics: Requests, complaints, and apologies*, volume 7. Walter de Gruyter.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Vorakit Vorakitphan, Marco Guerini, Elena Cabrio, and Serena Villata. 2020. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *The 28th International Conference on Computational Linguistics (COLING'2020)*.

Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 2018. An empirical study on uncertainty identification in social media context. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 79–88. World Scientific.

Xu Xu, Jiayin Li, and Huilin Chen. 2021. Valence and arousal ratings for 11,310 simplified chinese words. *Behavior Research Methods*, pages 1–16.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A More Sample Posts

We provide more sample posts in Table A1 from our dataset grouped according to the 5 bins defined in Table 3.

Bin	Weibo posts	Scores
1	突然感觉农业大学做的还不错 (*Suddenly I feel that the Agricultural University is doing pretty well.*)	-1.00
1	校内完成作业挺好的，回家可以做自己喜欢的事 (*It’s good to finish homework in school, you can do what you like when you go home.*)	-0.80
2	为了疫情防控而封闭管理其实是没有问题的。但是既然实行这项制度，就要真正把好关，校外人员不能随意进出学校 (*In fact, there is no problem with closed management for the prevention and control of the epidemic. However, since this system is implemented, it is necessary to truly ensure that people outside the school cannot enter and leave the school at will.*)	-0.50
2	让学生在校内写完作业这不怎么可能实现吧? 对家长来说确实挺好的，因为很多题都不会，也没办法辅导。 (*Isn’t it possible for students to finish their homework in school? It’s really good for parents, because many problems parents don’t know how to solve them, and they can’t help students.*)	-0.33
3	从来没有人去教家长要如何做一个合格的学生家长 (*No one has ever taught parents how to be a qualified student parent.)*	-0.15
3	这不太好吧，所谓家庭作业不是在家完成么，在校内完成的不是校内作业或者课堂作业吗，真是这样那干脆不要布置家庭作业罢了 (*This is not so good. Isn’t the so-called homework done at home? Isn’t it done in school or classwork? If it’s true, then just don’t assign homework.*)	+0.17
4	老师也是人他们虽然是服务行业但也需要自己的生活吧小学作业也没有那么多根本不用老师熬夜去批那初中毕业学年和高中呢作业量多难度大一个老师交两三个班一百多号人学生写完作业都要十点多了难道还让老师在学校通宵批完没有效率出现错误又说是老师不负责? (*Teachers are also humans. Although they are in the service industry, they also need their own lives. There are not so many primary school homework. There is no need for teachers to stay up late to correct them. The middle and high school homework is a lot and difficult. One teacher teaches two or three classes. For many students, it’s more than ten o’clock when the students finish their homework. Could it be that the teacher is allowed to finish the correction at school overnight? This is not efficient. If there is a mistake, the teacher will be accused of being irresponsible.*)	+0.40
4	为什么都是0分啊? 是作弊被抓到了吗? 还是怎么样? 还是根本就没来考试啊? 浪费这机会，有那机会给我多好，我想上还上不了呢 (*Why are their scores all 0 points? Was it caught for cheating? Or what? Or didn’t you come to the exam at all? The two of them wasted this opportunity, how good it is for me to have that opportunity, I want to go to school but I don’t have the opportunity.*)	+0.57
5	学校偏僻，所以西安这些学校的职工都是从隔壁村子随便找的? 隔壁西电，门卫满嘴官话实际怠惰工作，保洁在图书馆大声唠嗑，合着招职工没有限制应聘即上岗?西外职工都敢拖行女生了，原来职工素质低不是我校特例啊西安高校，你有事吗? (*Due to **the school**’s remoteness, the employees of these schools in Xi’an are all looking for them from the neighboring village? At the Xidian University next door, **the guards** are lazy, and **the cleaners** babble loudly in the library. Is it possible to recruit staff to take up jobs without restrictions? The school’s **security guards** are very rude to girls. It turns out that the low quality of staff is not a special case of our school. The problems in Xi’an colleges and universities are severe.*)	+0.92
5	现在的学校，真恶心，老师们拿的有工资啊，双休寒暑假，还美其名曰：家校共育这本来没错，但，能不能不要让家长充当老师的角色? ？？？作业，回家写可以，家长还要拍照、打卡、发视频交作业，还要批改作业，这都是老师的工作好吗? ？这部分工作家长做了，老师在干啥? 这部分工作的工资，学校给家长发了吗? ？没有啊！！那就请老师们，完成你们份内的工作！！别说什么辛苦之类的，拿着那份工资与待遇，就要干好那份工作在其位不谋其职！！！！ (*The current school is really disgusting. The teachers are paid, and they have two winter vacations and summer vacations. They also have a good name: family-school co-education is not wrong. But can we not let parents act as teachers? ? ? ? Homework can be written at home. Parents have to take photos, check in, send videos to hand in homework, and also correct homework. This is the teacher’s job. If the parents do this part of the work, then what does the teacher do? Has the school sent the salary for this part of the work to the parents? ? No! ! Then please teachers, finish your job! ! Not to mention that the work is very hard. With the salary and benefits, you must do the job well. Don’t be irresponsible in this position! ! ! !*)	+1.00

Table A1: More sample posts for each of the 5 bins. Words in bold are some points of concern.

B Data Used for Post Popularity Prediction

We collected blog posts under 8 topics from Weibo to verify the relationship between complaint density and popularity. Table A2 shows the hashtag contents, along with the number and time of collection.

Hashtag	Number	Start From
#巨人教育宣布倒闭# (*#JuRen Education announces bankruptcy#*)	1,663	2021/8/31
#官方回应开学典礼学生晕倒无人扶# (*#The official responded to the situation where the student fainted and no one helped at the opening ceremony#*)	331	2021/9/2
#芝加哥大学24岁中国留学生被枪杀# (*#A 24-year-old Chinese student at the University of Chicago was shot dead#*)	179	2021/11/11
#教育部将抑郁症筛查纳入学生体检# (*#Ministry of Education incorporates depression screening into student physical examination#*)	1,023	2021/10/31
#江苏一建停考# (*#Jiangsu Province Level One Architect Examination Suspended#*)	286	2021/8/24
#计算机二级证书有必要吗# (*#Is it necessary to take the second-level computer certificate#*)	242	2021/11/9
#西安外国语大学封闭管理# (*#Closed management of Xi'an International Studies University#*)	287	2020/9/19
#大连理工大学支教# (*#Supporting Teaching at Dalian University of Technology#*)	962	2021/3/8

Table A2: The hashtag and its number used in the application.