Cultural Valence: A Comparative Analysis of Chinese and English Posts on Ukrainian War-related Topics on Reddit

Zhenni Xu (Kayn)(zx2395)
Department of Computer Science
Columbia University

Advisor: Dr. John R. Kender

May 17, 2023

Abstract

Social media, notably platforms like Reddit, has become a powerful tool for sentiment analysis, with its open environment encouraging people to freely express their emotions, especially during crisis events such as the Russian invasion of Ukraine. This study aims to apply toolkits such as VADER (Valence Aware Dictionary and sEntiment Reasoner) to examine and compare the valence and subjectivity expressed in English and Chinese Reddit posts about the invasion, filling a gap in research that currently lacks a direct sentiment analysis between these two languages and cultures about this event.

1 Introduction

1.1 The Impact of Social Media

The Internet is an expansive virtual realm where people can freely express and share their opinions, impacting various aspects of life, including marketing and communication. For instance, reviews and ratings on the web are becoming increasingly significant in shaping potential customers' perceptions when assessing products and services.[1] In this way, social media provides a favorable environment for analyzing people's sentiments due to individuals' freedom to express their emotions without inhibition. This openness and lack of self-censorship on social media make it an ideal source for sentiment analysis, allowing researchers to understand people's emotions and attitudes better.

Over the past years, social media has extensively and progressively taken on a larger role, becoming a significant alternative source of information compared to traditional media, especially in times of emergencies and disasters. Social media has gained popularity during such events and ranks as the fourth most

1. Federico Neri et al., "Sentiment analysis on social media," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (IEEE, 2012), 919–926.

favored means of accessing essential information in recent years.[2][3][4]When disasters strike, people often strongly urge to express their emotions or sentiments online since such events are often shocking and impactful, evoking feelings like fear, concern, sadness, or empathy. Social media platforms provide an accessible and immediate outlet for individuals to share their thoughts, experiences, and reactions to these disasters. Through online platforms, people can seek support, share firsthand accounts, express solidarity, or simply vent their emotions, enabling a collective expression of sentiments during challenging times.
On February 24, 2022, the Russian government issued a directive for military forces to enter Ukrainian territory, thus initiating a large-scale invasion of Ukraine. This aggressive move by Russia swiftly led to a significant escalation of the conflict, resulting in severe economic and humanitarian disasters.[5] The invasion of Ukraine by Russia has been characterized as the first "social media war."[6][7] The availability of high-quality real-time videos captured on cell phones and shared through social media networks enables people to virtually experience certain aspects of combat, regardless of their location. Additionally, social media plays a crucial role in garnering widespread support and generating sympathy from people across the globe.[8] In this way, the Ukraine war stands out as a prime subject for sentiment analysis, mainly since it is the first "social media war." With many people expressing themselves freely on various social media platforms, the conflict allows us to examine individuals' sentiments in real-time. Analyzing posts and word use online during this humanitarian disaster provides insights into people's sentiments during times of crisis and conflict.

1.2 Reddit

Reddit is a diverse and influential social media platform that fosters a multicultural meta-community, allowing individuals to express their honest thoughts and opinions on global events and news in different subreddits.[9] With its extensive user base spanning various demographics and backgrounds, Reddit is a microcosm of society, providing opportunities for people to express their honest thoughts and sentiments worldwide. In this way, Reddit is an excellent platform for analyzing sentiments due to its organized structure. The dedicated

2. Ghazaleh Beigi et al., "An overview of sentiment analysis in social media and its applications in disaster relief," in *Sentiment analysis and ontology engineering: An environment of computational intelligence* (Springer, 2016), 313–340.

3. Bruce R. Lindsay, *Social Media and Disasters: Current Uses, Future Options, and Policy Considerations*, technical report (Congressional Research Service, September 2015).

4. Alfredo Cobo, Denis Parra, and Jaime Navon, "Identifying relevant messages in a Twitter-based citizen channel for natural disaster situations," in *Proceedings of the 24th International Conference on World Wide Web Companion* (2015), 1189–1194.

5. Giuseppe Grossi and Veronika Vakulenko, "New development: Accounting for human-made disasters—comparative analysis of the support to Ukraine in times of war," *Public Money & Management* 42, no. 6 (2022): 467–471.

6. Dan Ciuriak, *The Role of Social Media in Russia's War on Ukraine*, Available at SSRN, 2022.

7. Peter Suciu, "Is Russia's Invasion Of Ukraine The First Social Media War?," March 2022, accessed May 16, 2023, https://www.forbes.com/sites/petersuciu/2022/03/01/is-russias-invasion-of-ukraine-the-first-social-media-war/?sh=c3a3fc21c5cd.

8. Suciu.

9. Carrie Moore and Lisa Chuang, "Redditors revealed: Motivational factors of the Reddit community," in *Proceedings of the 50th Hawaii International Conference on System Sciences* (2017), 2313–2322.

sections called subreddits allow people to have focused discussions and explore their interests in more detail, providing a wealth of sentiment analysis data.
Reddit has separate subreddits for English (mainly American[10]) and Chinese users, which provides access to posts from two distinct cultures. In this way, using Reddit data is advantageous for sentiment analysis as it allows us to study and analyze the sentiments expressed within different cultural contexts. Moreover, Reddit's easy-to-scrape post headlines from different subreddits, facilitated by convenient scraping tools such as the "Python Reddit API Wrapper" (PRAW), further enhance its suitability for sentiment analysis, as they offer readily available and accessible textual data that can be analyzed to understand the prevailing sentiments within specific topics or communities.[11][12]

1.3 Valence Score

Valence score, also called sentiment score or sentiment polarity, is a way to measure the emotional tone or sentiment expressed in a text. It helps determine whether the sentiment in the text is positive, negative, or neutral.
In this study, we will use a method toolkit called VADER (Valence Aware Dictionary and sEntiment Reasoner) to analyze the emotional tone of the text. The VADER collected intensity ratings from ten independent human raters for candidate lexical features, resulting in over 90,000 ratings. Ratings were obtained using Amazon Mechanical Turk (AMT), and they used a scale from -4 to +4 to rate the features' sentiment intensity. Lexical features with a non-zero mean rating and a standard deviation less than 2.5 were retained, resulting in over 7,500 lexical features with validated valence scores, indicating both the sentiment polarity and intensity on a scale from -4 to +4. For instance, "okay" had a positive valence of 0.9, "good" scored 1.9, and "great" received a 3.1, while "horrible" was -2.5, ":(" was -2.2, and both "sucks" and "sux" had a valence of -1.5.[13]
In this way, VADER is a special model developed for understanding sentiments in social media posts. It works well with short and informal texts on platforms like Reddit. By applying VADER, we can better understand the emotions and sentiments conveyed in the text data we analyze.
While many studies and reports focus on social media perspectives and public views of how English users and Chinese users react to the Russian invasion of Ukraine, such as a study highlighting the diverse opinions among the public regarding the U.S. response[14] or a study exploring the viewpoints and reactions of Chinese social media users, specifically from a feminist perspective about

10. *Alexa - Reddit Competitive Analysis, Marketing Mix and Traffic*, accessed May 16, 2023, https://web.archive.org/web/20171120163709/https://www.alexa.com/siteinfo/reddit.com#section_traffic.

11. A. R. I. K. Kaan, "Social Media Content Review of Popular MMORPG Games: Reddit Comment Scraping and Sentiment Analysis," *Journal of Emerging Computer Technologies* 2, no. 1 (2022): 13–21.

12. Python Reddit API Wrapper Development, *PRAW: The Python Reddit API Wrapper*, accessed May 16, 2023, https://github.com/praw-dev/praw.

13. Clayton Hutto and Eric Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 1 (June 2014), 216–225.

14. Pew Research Center, *Public Expresses Mixed Views of U.S. Response to Russia's Invasion of Ukraine*, March 2022, accessed May 16, 2023, https://www.pewresearch.org/politics/2022/03/15/public-expresses-mixed-views-of-u-s-response-to-russias-invasion-of-ukraine/.

the Ukrainian war,[15] there is a lack of in-depth analysis that directly compares the sentiments between English and Chinese. Most existing research provides a general overview of public opinions without delving into the specific sentiment analysis of posts in each language. Therefore, there is a need for more research that explores the emotional responses and sentiment differences between English posts and Chinese posts about the Russian invasion of Ukraine.

2 Methodology

2.1 Data Collection

We utilize the PRAW (Python Reddit API Wrapper) API[16] to scrape data from specific subreddit channels, enabling us to gather relevant information and analyze user-generated content effectively.

For subreddit selection, we picked two English subreddits, namely "Ukraine" and "Worldnews," as they contain a significant number of posts highly relevant to the Ukrainian war, catering primarily to English-speaking (mainly American[17]) users. However, for Chinese posts, since there are limited subreddits dedicated to Chinese content, and their discussions are not solely focused on the Ukrainian war, we chose five main popular Chinese subreddit channels:
"$China\_irl$,"
"$real\_China\_irl$,"
"$LiberalGooseGroup$,"
"$KanagawaWave$,"
and "$DoubanGoosegroup$."
Among these subreddits, three are predominantly used by male users, which are China_irl," "real_China_irl," and "KanagawaWave", and two are predominantly used by female users, which are "LiberalGooseGroup" and "DoubanGoosegroup". This selection offers a broader representation of Chinese users' perspectives and discussions beyond the Ukrainian war.

For keyword selection in both English and Chinese subreddit channels, we chose "Ukrainian people," "Ukrainian war," and "Ukrainian" as the keywords. In Chinese subreddit channels, due to limited posts specifically related to the Ukrainian War, we expanded the keyword set to include "乌克兰人民" (Ukrainian people), "乌克兰战争" (Ukrainian War), "俄乌" (Russo-Ukrainian), "俄乌战争" (Russo-Ukrainian War), and "乌克兰" (Ukrainian) since these are commonly discussed topics in Chinese regarding the Ukrainian War. In total, we collected 1134 English posts and 362 Chinese posts through our scraping process from Reddit.

Initially, we scraped the headlines of the posts but noticed that some Chinese post headlines were unrelated to the Ukrainian War, while the subtexts contained relevant information. As a result, we combined the headlines and subtexts for Chinese post scraping using Chinese keywords. However, we encountered another challenge: some super long post subtexts contained garbled characters and were often forwarded from other websites' sources rather than

15. Altman Yuzhu Peng, "A Chinese feminist analysis of Chinese social media responses to the Russian invasion of Ukraine," *International Feminist Journal of Politics* 24, no. 3 (2022): 482–501.

16. Python Reddit API Wrapper Development, *PRAW: The Python Reddit API Wrapper.*

17. *Alexa - Reddit Competitive Analysis, Marketing Mix and Traffic.*

personal perspectives. To address this, we implemented a rule that for Chinese posts with exceptionally long subtexts (exceeding 1000 characters when combining headlines and subtexts), we focused solely on the headlines since they provide the main title of the post and are more reliable for analysis. For shorter Chinese posts, we combined both the headlines and subtexts together to obtain a more comprehensive understanding of the content.

```
subreddit_names = ['Ukraine', 'worldnews']
headlines_en = set()
for subreddit_name in subreddit_names:
    subreddit = reddit.subreddit(subreddit_name)

    # Define the keyword to search for: Ukraine people, Ukraine war, Ukraine
    keywords = ['Ukrainian people', 'Ukrainian war','Ukrainian']

    for keyword in keywords:
        for submission in subreddit.search(keyword, sort='new', limit=None):
            headlines_en.add(submission.title)
        print(keyword,len(headlines_en))
```

```
# Define the subreddit to search
subreddit_names = ['China_irl', 'real_China_irl','LiberalGooseGroup','KanagawaWave','DoubanGoosegroup']
headlines_ch = set()
for subreddit_name in subreddit_names:
    subreddit = reddit.subreddit(subreddit_name)

    for keyword in keywords:
        for submission in subreddit.search(keyword, sort='new', limit=None):
            headlines_ch.add(submission.title)
        print(keyword,len(headlines_ch))
```

```
# Define the subreddit to search
subreddit_names = ['China_irl', 'real_China_irl','LiberalGooseGroup','KanagawaWave','DoubanGoosegroup']
for subreddit_name in subreddit_names:
    subreddit = reddit.subreddit(subreddit_name)

    # Define the keyword to search for: Ukraine people, Ukraine war, Ukraine
    keywords_ch = ['乌克兰人民', '乌克兰战争', '俄乌', '俄乌战争', '乌克兰']

    for keyword in keywords_ch:
        for submission in subreddit.search(keyword, sort='new', limit=None):
            if isinstance(submission.selftext, str):
                sentence = submission.title+" "+submission.selftext
                if len(sentence) > 1000:
                    headlines_ch.add(submission.title)
                else:
                    headlines_ch.add(sentence)
            else:
                headlines_ch.add(submission.title)
        print(keyword,len(headlines_ch))
```

Figure 1: Code for Scraping Data from English and Chinese Subreddits

After examining the sentiment analysis algorithms of vaderSentiment.py for analyzing the sentiment of texts in other languages in the VADER model, we discovered that it relies on a My Memory Translation Service API with certain usage limitations and less robust performance than the latest OpenAI model. In this way, to translate Chinese posts into English for sentiment analysis, we decided to utilize the OpenAI API with the model "gpt-3.5-turbo." As a result, we chose to use the OpenAI API directly for translating Chinese posts into English. This approach enables us to utilize the translated posts for sentiment analysis effectively.

```
openai.api_key = os.getenv("OPENAI_API_KEY")
def translate(sentence):
    # Set up the translation prompt
    prompt = "Translate '"+sentence+"' to English. Please only return the translated English."

    # Generate the translation using ChatGPT
    completion = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {
                "role": "user",
                "content": prompt,
            },
        ]
    )
    # Extract the translation from the response
    for choice in completion.choices:
        return choice.message.content
```

```
translated_headlines = list()
for i in range(len(headlines)):
    res = translate(headlines[i])
    print(i,res)
    translated_headlines.append(res)
print(translated_headlines)
```

Figure 2: Code for Translating Data from Chinese and English

2.2 Sentiment Analysis

To analyze the sentiment of each post, we utilize VADER in Python. By applying VADER, we obtain sentiment scores categorized into four parts: compound score, "pos" score, "neu" score, and "neg" score.

The compound score is a single measure of sentiment for a sentence, computed by summing the valence scores of each word in the lexicon and normalizing it between -1 (most negative) and +1 (most positive). According to developers, we should classify sentences as positive, neutral, or negative based on standardized thresholds (compound score >= 0.05 for positive sentiment, between -0.05 and 0.05 for neutral sentiment, and compound score <= -0.05 for negative sentiment). The "pos", "neu", and "neg" scores represent the proportions of text falling into each sentiment category, with their sum adding up to 1.[18]

```
h_en = pd.read_csv('headlines_en.csv')
headlines_en = h_en['headlines'].tolist()
```

```
h_neg_en = list()
h_neu_en = list()
h_pos_en = list()
h_com_en = list()
for i, row in h_en.iterrows():
    h_sentence = row['headlines']
    sentiment_dict_h = obj.polarity_scores(h_sentence)
    h_neg_en.append(sentiment_dict_h['neg'])
    h_neu_en.append(sentiment_dict_h['neu'])
    h_pos_en.append(sentiment_dict_h['pos'])
    h_com_en.append(sentiment_dict_h['compound'])
```

```
df_sen_en = pd.DataFrame(list(zip(headlines_en, h_neg_en, h_neu_en, h_pos_en, h_com_en)))
df_sen_en
```

Figure 3: Code for Obtaining Valence Scores of English Posts

18. Hutto and Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text."

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

```
trans_ch = pd.read_csv('headlines_translation.csv')
headlines_ch = trans_ch['headlines'].tolist()
trans = trans_ch['translation'].tolist()
```

```
obj = SentimentIntensityAnalyzer()
```

```
h_neg_ch = list()
h_neu_ch = list()
h_pos_ch = list()
h_com_ch = list()
for i, row in trans_ch.iterrows():
    h_sentence = row['translation']
    sentiment_dict_h = obj.polarity_scores(h_sentence)
    h_neg_ch.append(sentiment_dict_h['neg'])
    h_neu_ch.append(sentiment_dict_h['neu'])
    h_pos_ch.append(sentiment_dict_h['pos'])
    h_com_ch.append(sentiment_dict_h['compound'])
```

```
df_sen_ch = pd.DataFrame(list(zip(headlines_ch, trans, h_neg_ch, h_neu_ch, h_pos_ch, h_com_ch)))
df_sen_ch
```

Figure 4: Code for Obtaining Valence Scores of Chinese Posts

headlines	neg	neu	pos	com
Canada announces additional sup…	0.287	0.515	0.199	-0.296
Ukrainian woman said her Russian…	0.361	0.639	0.0	-0.7845
Ukraine Revolt: sticky post	0.0	1.0	0.0	0.0
A compilation of the destruction of t…	0.163	0.837	0.0	-0.5719
6:13 EET; The Sun is Rising Over …	0.082	0.825	0.093	0.1007
Russians use smartphone app that…	0.0	1.0	0.0	0.0
Ukraine war: Joe Biden to meet Uk…	0.309	0.691	0.0	-0.8555
During a Propaganda football matc…	0.05	0.882	0.068	0.1779
"You actually can feel this battlefiel…	0.385	0.615	0.0	-0.7845
1.5 million Ukrainian children at ris…	0.404	0.596	0.0	-0.6249
Need help with finding a Ukrainian …	0.0	0.69	0.31	0.4019
Russia Is Concentrating Its Forces …	0.191	0.736	0.074	-0.5499
Russia drops bomb on Ukrainian s…	0.262	0.738	0.0	-0.4939
Why Ukraine does not want to surr…	0.133	0.867	0.0	-0.0572

Figure 5: Examples of English Posts and Corresponding Valence Scores in Dataframe

headlines	translation	neg	neu	pos	com
乌克兰战争：招募罪犯参战的俄罗斯雇佣军"瓦格…	"Who is the leader of the Russian mercenary group 'Wagner' t…	0.309	0.691	0.0	-0.8225
普京在人民日报上称习最亲密朋友 习在俄罗斯报…	"Putin calls Xi his closest friend in the People's Daily; Xi calls …	0.0	0.776	0.224	0.5994
Day 234 俄乌战争总结 - 冬季来临之前，赫尔松城…	Day 234 Summary of the Russo-Ukrainian War - Before the a…	0.14	0.86	0.0	-0.5994
弗拉基米尔·普京曾在2014年发出警告，声称他可…	Vladimir Putin once issued a warning in 2014, claiming that h…	0.094	0.906	0.0	-0.34
【网络民议】"unacceptable"不是"可以接受"的意…	[Online opinion] Does "unacceptable" not mean "not acceptab…	0.28	0.56	0.16	-0.3156
民主党这是没活了?开始称呼基辅为纳粹，美国的…	"Is the Democratic Party dead? They are now calling Kiev Na…	0.245	0.657	0.099	-0.6124
为什么俄罗斯一直没能融入自由主义国际秩序?	"Why has Russia been unable to integrate into the liberal inte…	0.0	1.0	0.0	0.0
IMF总裁：中国疫情将进一步打击经济	"IMF President: China's epidemic will further hit the economy."	0.0	1.0	0.0	0.0
25岁乌克兰士兵舍身炸掉了一座桥 (乌克兰"董存…	"A 25-year-old Ukrainian soldier sacrificed himself to blow up …	0.0	1.0	0.0	0.0
Day 159 俄乌战争总结 - 乌军否认了对克里米亚半…	"Day 159 Summary of the Russo-Ukrainian War - Ukrainian …	0.299	0.701	0.0	-0.8689
俄篓好歹也有1亿多人，怎么感觉这次军队动员能…	"At least Russia has over 100 million people, why does it see…	0.146	0.77	0.084	-0.4313
大乌克兰宣言	The Great Ukraine Declaration	0.0	0.423	0.577	0.6249
乌克兰战争：乌克兰女性在战争中是如何生存的?	"What is it like for Ukrainian women to survive in the war in U…	0.212	0.652	0.136	-0.34
乌克兰大反攻：俄军败退会让普京付出代价吗	"Ukraine's major counterattack: Will the Russian military retre…	0.099	0.775	0.127	0.1027

Figure 6: Examples of Chinese Posts and Corresponding Valence Scores in Dataframe

We then conduct a comparative analysis of the compound scores and positive versus negative sentiments, and we perform statistical analysis to explore the cultural differences in valence between Chinese and English posts.

Furthermore, we analyzed the TextBlob subjectivity scores[19] in conjunction with the previously computed valence scores. This exploration aimed to understand the connection between subjective content and perceived valence in Chinese and English Reddit posts. The subjectivity score, which ranges from 0.0 (highly objective) to 1.0 (highly subjective),[20] provided valuable insights

19. Steven Loria, *TextBlob*, accessed May 16, 2023, https://github.com/sloria/TextBlob.
20. Steven Loria, "TextBlob Documentation," *Release 0.15* 2, no. 8 (2018).

into the degree of personal bias or opinion. This analysis would enable us to better understand how subjectivity may be related to the valence within these language communities on Reddit.
By examining these aspects, we aim to understand better how sentiments are expressed and perceived within these two language communities, shedding light on potential variations in emotional responses and cultural nuances.

2.3 Word Segmentation

In addition to the valence analysis of the entire posts, we treat the scraped data in English and Chinese as separate sets. We employ word segmentation techniques to identify the top five most frequently occurring words in each language related explicitly to discussions about the Ukrainian War.
For English posts, we utilize the word segmentation method from the WordCloud library.[21] This allows us to break down the text into individual tokens. Then, we employ the Counter from python, which enables us to count the occurrences of each tokenized word and find the top occurring words and their corresponding frequencies.

```
# define some sample text
# create the wordcloud object with default settings
wc = WordCloud()

text = ' '.join(full_text_en)

# process the text and get the word frequencies
word_freqs = wc.process_text(text)

# get the top words and their corresponding frequencies
top_words = Counter(word_freqs).most_common(300)

# create a dictionary of the top words and their frequencies
freq_dict = {}
for word, freq in top_words:
    freq_dict[word] = freq

print(freq_dict)
```

Figure 7: Coding Example: Word Segmentation for English Posts

For Chinese posts, we use PKUSEG-Python for word segmentation. PKUSEG-Python is a toolkit developed for multi-domain Chinese word segmentation (CWS). It addresses the challenge of word segmentation in different domains by providing separate models for specific domains, such as the web, medicine, and tourism.[22]
In our selection process, we specifically choose the web domain from the PKUSEG-Python toolkit, which is optimized for handling web-based text segmentation tasks for our Chinese Reddit data. This domain adopts a pre-training technique[23] to effectively process data originating from the Weibo dataset pro-

21. Andreas Mueller, *WordCloud: A Python package to generate word clouds*, accessed May 16, 2023, https://github.com/amueller/word_cloud.
22. Ruixuan Luo et al., "PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation.," *CoRR* abs/1906.11455 (2019), https://arxiv.org/abs/1906.11455.
23. Jingjing Xu and Xu Sun, "Transfer learning for low-resource Chinese word segmentation with a novel neural network," *CoRR* abs/1702.04488 (2017).

vided by the NLPCC-ICCPOL 2016 Shared Task.[24]
We selected PKUSEG as the preferred toolkit for tokenization among other Chinese word segmentation toolkits because it consistently outperforms them in cross-domain testing across different datasets,[25] making it a reliable and effective choice for word segmentation.

```
# define some sample Chinese text
text = ' '.join(full_text_ch)

# create a PKUSeg object for tokenizing Chinese text
seg = pkuseg.pkuseg(model_name='web')

# tokenize the text using PKUSeg
words = seg.cut(text)

# count the frequency of each word using the Counter class
word_counts = Counter(words)

# get the top most frequent words
top_words = word_counts.most_common(300)

# print the top most frequent words and their counts
for word, count in top_words:
    print(word, count)
```

Figure 8: Coding Example: Word Segmentation for Chinese Posts

After performing word segmentation, the next step involves applying certain techniques to refine the results. In this step, we use stopwords to eliminate frequently used words and filter out proper nouns, such as "Ukraine," "Russian," and "Zelenskyy." Subsequently, the selection process involves manually choosing the top 5 words with a non-neutral compound score, indicating their sentiment-related significance.

3 Results

3.1 Sentiment Analysis

3.1.1 Overall Visualization

In terms of the compound score, the analysis reveals that English posts have a mean score of -0.194 with a standard deviation of 0.569. On the other hand, Chinese posts exhibit a mean score of -0.223 with a slightly lower standard deviation of 0.557. These findings mean that the average valence in both English and Chinese posts is generally more on the negative side.

After briefly analyzing the overall compound scores for both English and Chinese posts, we take a step further to analyze the data deeply. We perform a t-test, a statistical method used to compare the averages. Specifically, we compare the positivity, neutrality, and negativity scores between English and Chinese posts. By calculating the p-value, we are able to measure the strength of our results. The p-value helps us understand whether the differences we see in scores between

24. Xipeng Qiu, Peng Qian, and Zhan Shi, "Overview of the NLPCC-ICCPOL 2016 Shared Task: Chinese word segmentation for micro-blog texts," in *NLPCC/ICCPOL*, vol. 10102, Lecture Notes in Computer Science (Springer, 2016), 901–906.

25. Luo et al., "PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation."

English and Chinese posts are statistically significant.

	t-test	p-value
Positivity:English vs. Chinese	-0.285	0.776
Neutrality:English vs. Chinese	2.145	0.032
Negativity:English vs. Chinese	-2.075	0.038

Table 1: Statistical Comparison of Valence Scores between English posts and Chinese Posts: t-test and p-value Results.

Our p-value results indicate that the positivity scores between English and Chinese posts aren't significantly different. This means that posts in both languages tend to exhibit similar levels of positivity.
However, when it comes to neutrality and negativity scores, it is a different story. The p-value results for these scores are less than 0.05, which suggests a significant difference between English and Chinese posts. For these two aspects, the English and Chinese posts don't seem to mirror each other as closely as they do in terms of positivity.
The t-test results reveal some interesting differences between English and Chinese posts statistically. It appears that English posts tend to be significantly more neutral than Chinese posts, as indicated by the positive t-test result when comparing English and Chinese scores. On the other hand, Chinese posts are significantly more negative than English posts, as demonstrated by the negative t-test result in the same comparison.
Our boxplots further illustrate these findings. In the boxplot for positivity scores, we observe no significant differences between English and Chinese posts, indicating similar levels of positivity in both languages. However, the boxplot for neutrality scores reveals that the lower bound of the first quartile for English posts is higher than that for Chinese posts, which suggests that English posts tend to express neutrality more often. Conversely, the boxplot for negativity scores shows that the quartile box for Chinese posts is higher than for English posts, indicating that Chinese posts tend to express negativity more frequently."

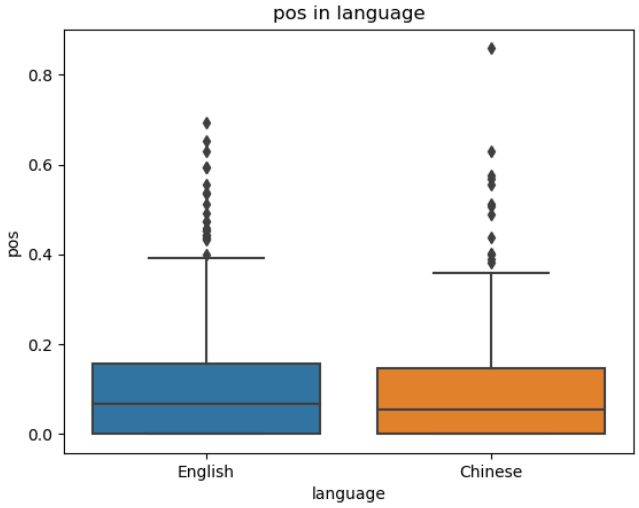

Figure 9: Boxplot of Positivity Scores: A Comparison between English and Chinese Posts

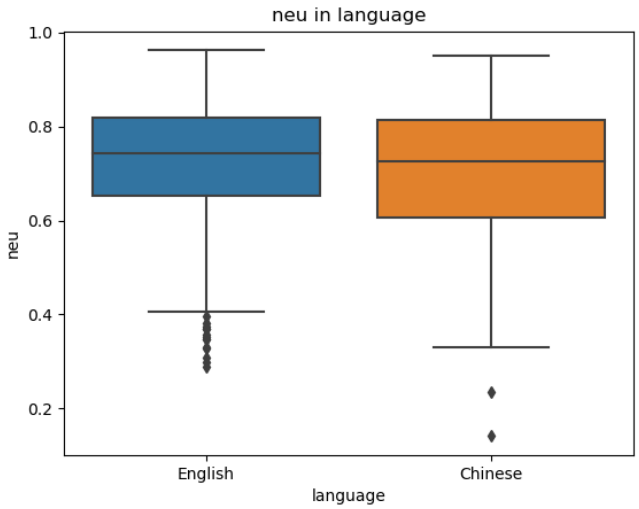

Figure 10: Boxplot of Neutrality Scores: A Comparison between English and Chinese Posts

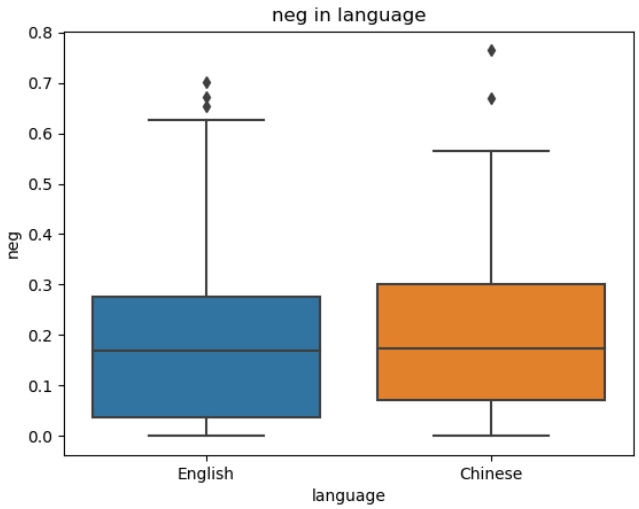

Figure 11: Boxplot of Negativity Scores: A Comparison between English and Chinese Posts

Overall, these interesting differences in positivity, neutrality, and negativity scores between English and Chinese posts are parts that we cannot determine solely by looking at the compound scores or the boxplot for compound scores.

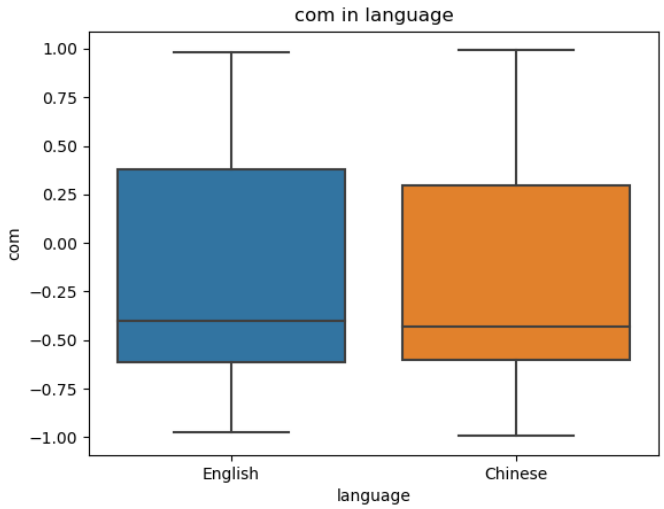

Figure 12: Boxplot of Overall Compound Valence Scores: A Comparison between English and Chinese Posts

3.1.2 Positivity vs. Negativity

To better evaluate the positivity versus negativity score, we constructed a histogram. However, some posts in both the English and Chinese categories were very neutral, resulting in a significant disruption at the bottom left corner of the graph. To mitigate this issue, we removed the neutral posts, specifically those with a compound score ranging from -0.05 to 0.05. After this adjustment, we were left with 893 non-neutral English posts and 292 non-neutral Chinese posts for further analysis.

The histogram results reveal that Chinese posts tend to express more distributions in negative sentiments and the same level of positivity compared to English posts. It is noteworthy that the bottom left block of the histogram doesn't represent "neutral" sentiments but rather "light positive" or "light negative" ones, given that we've already removed the neutral posts from the analysis. Our findings align with our previous statistical analysis, reinforcing the patterns we had already identified.

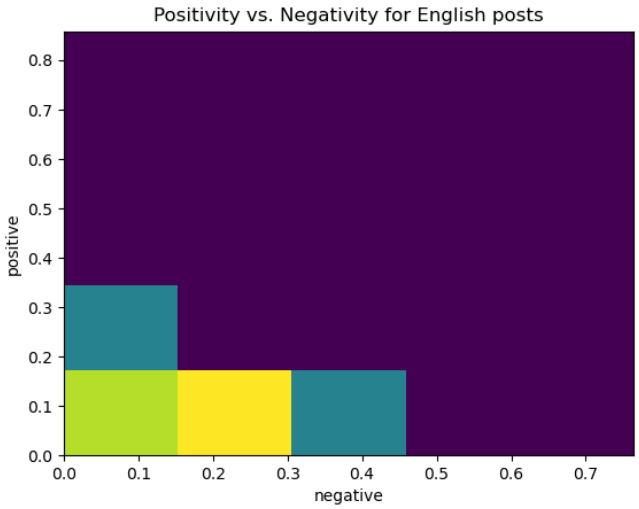

Figure 13: Positivity vs. Nagativity Plot for English Posts

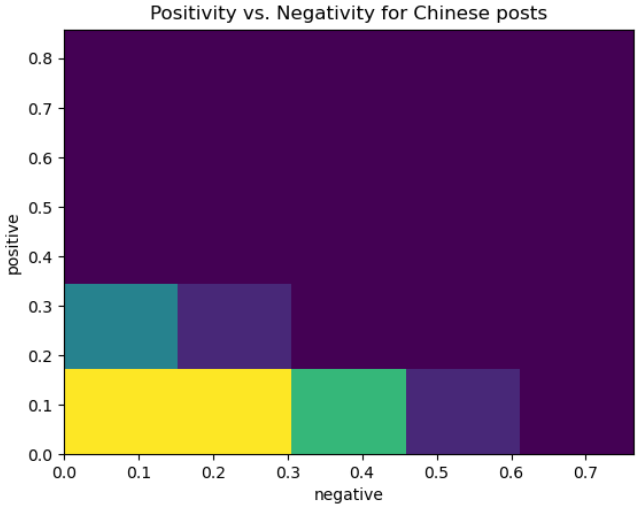

Figure 14: Positivity vs. Nagativity Plot for English Posts

We uncover some fascinating details when we adjust the number of bins in the 2D histogram to 10. Although the primary positivity and negativity ranges are similar for English and Chinese posts(with Chinese posts displaying more negative sentiments), the range for Chinese posts is broader. The Chinese histogram shows a lightly populated block reaching up to 0.5 in complete positivity and another up to 0.6 in complete negativity. Additionally, several lightly populated blocks in the Chinese histogram indicate more extreme positive and negative sentiments. These extreme sentiment expressions, as represented by these less populated blocks, are not observed in the English histogram, suggesting that there are unique distributions of extreme sentiments in Chinese posts not mirrored in English."

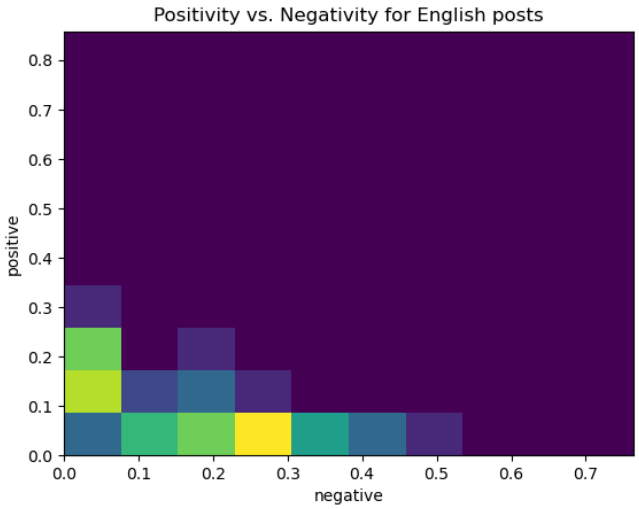

Figure 15: Positivity vs. Nagativity Plot for English Posts with bin=10

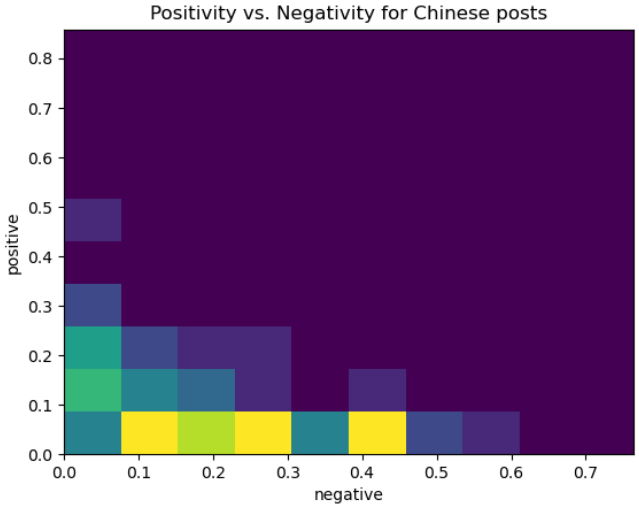

Figure 16: Positivity vs. Nagativity Plot for Chinese Posts with bin=10

3.1.3 Valence vs. Subjectivity

We're curious about how valence and subjectivity relate to each other in posts from these two languages, so we've also created a hist2d histogram to explore this connection visually. In exploring the relationship between valence and subjectivity scores, we analyzed the distribution and density of histogram blocks for both Chinese and English posts. Our findings indicate that the Chinese posts graph has notably more blocks on the subjective side of the histogram than American posts. This finding suggests that, compared to English posts, many Chinese posts tend to express more subjective feelings or opinions.
Additionally, we observe a particular pattern in the Chinese posts: they are

often more subjective when expressing slightly positive or extremely negative sentiments. A similar pattern also appears in English posts, though to a lesser extent. This finding implies that both English and Chinese language users tend to express personal feelings or opinions more intensely when their posts are lightly positive or highly negative, with this pattern being more noticeable among Chinese posts.
In summary, the majority of the posts in both languages tend to be objective. However, Chinese posts display a wider distribution on the subjective side, indicating a greater use of personal feelings or opinions in their expression compared to English posts.

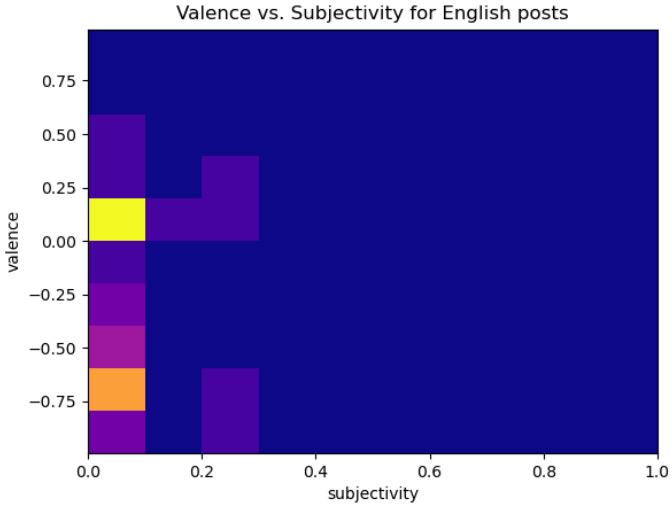

Figure 17: Valence vs. Subjectivity Plot for English Posts with bin=10

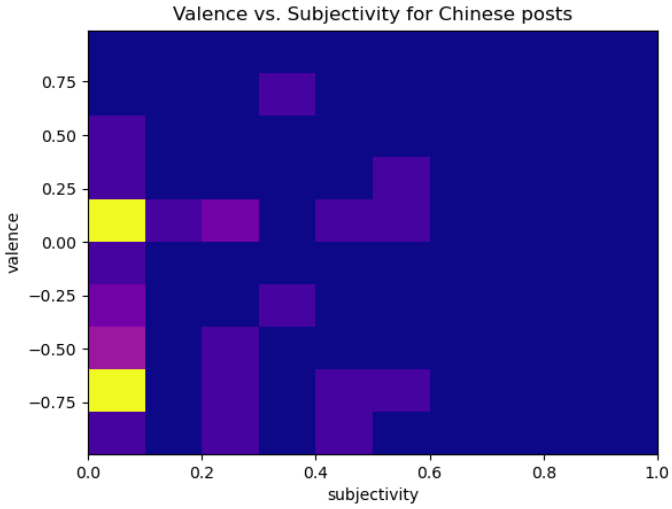

Figure 18: Valence vs. Subjectivity Plot for Chinese Posts with bin=10

3.2 Top Occurring Non-neutral Words Analysis

In our analysis, we found the top five most frequently used non-neutral words in both English and Chinese posts. For English posts, these words are "Discussion Charities," "help," "Fight," "support," and "killed." On the other hand, the Chinese posts frequently feature the words "支持" (support), "加入" (join), "反攻" (counter attack), "打" (fight), and "冲突" (conflict).

3.2.1 Word Cloud

We created two word-clouds, each representing the top 5 most frequent words in English and Chinese posts. We added an interesting twist to these word clouds by color-coding the words based on their valence scores. Here is how it works: the words with a more positive valence score appear in shades of red, while those with a more negative valence score are represented in shades of blue. This setting allows an immediate visual understanding of the word's frequency and its associated sentiment.

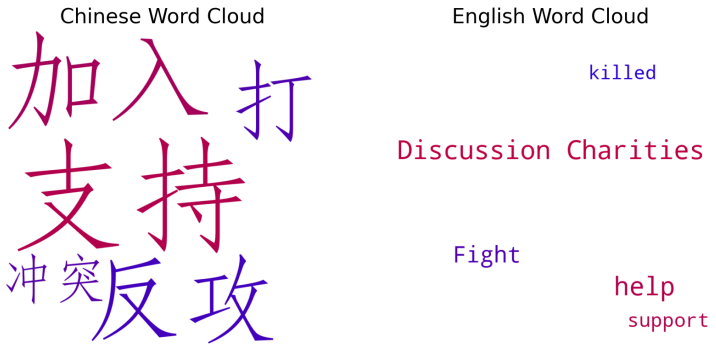

Figure 19: Comparative Word Clouds: Top 5 Most Frequent Non-neutral Words in English and Chinese Posts

3.2.2 Scatter Plot

We created a scatter plot to gain more insights into the valence scores of the top words and their frequency of occurrence. This visualization helps us understand the relationship between a word's sentiment and how often it appears. Given that we collected fewer Chinese posts than English posts, it is expected that the frequency of words from Chinese posts is generally lower.

From the scatter plot, we can see that for positive words, the three positive English words tend to have slightly higher valence scores than the two positive Chinese words. This finding suggests a trend of stronger positive sentiments in English posts among 5 top words.

On the negative side, there is an interesting observation. Even though one English word has the lowest valence score, it has the fewest occurrences among the top five English words. There are three negative words in Chinese compared to only two in English, which indicates a tendency for Chinese posts to express more negative sentiments.

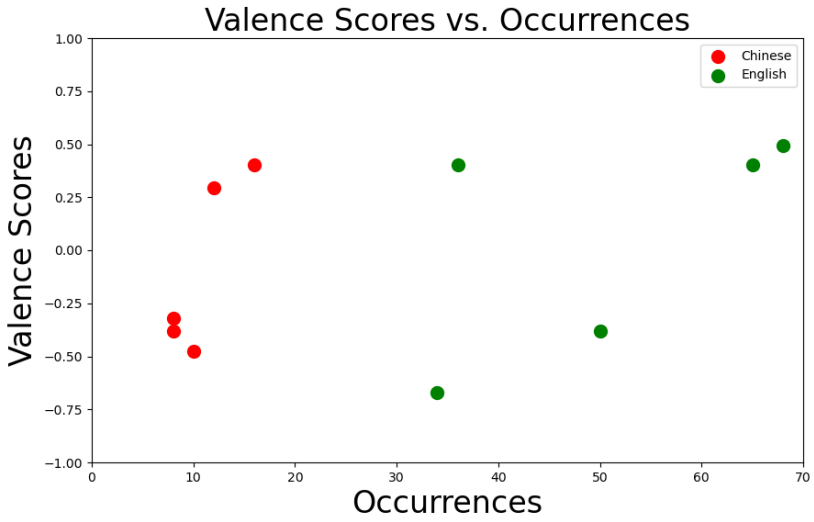

Figure 20: Occurrences vs. Valence Score for Top 5 Most Frequent Non-neutral Words in English and Chinese Posts

4 Conclusion

In conclusion, our research shows that people who write in English and Chinese about the war in Ukraine have different sentiments. Generally, English posts are more neutral in valence, while Chinese posts lean towards more negative valence. Chinese posts, in particular, were found to be more subjective than English posts.

Despite the cultural distance from the center of the conflict, the emotional landscape expressed in Chinese posts on Reddit is diverse and complex. Chinese posts display a mixture of sentiments, with negative emotions such as aggression and frustration language usage being more prevalent. On the other hand, English posts present a more neutral sentiment and tend to be slightly more positive than Chinese posts. This difference could be attributed to how English-language discussions on platforms such as Reddit often approach news events. Instead of adopting an aggressive stance, these discussions are typically characterized by a peaceful and measured exploration of the issues.

Moreover, a noticeable trend in English posts is the expression of support for Ukrainian citizens. Among the challenging circumstances of war, many users writing in English demonstrate empathy and solidarity. They often highlight humanitarian aspects, calling for aid and international intervention to relieve the suffering of those affected by the war. In contrast, Chinese posts are primarily debated around political aspects of the conflict. A substantial number of Chinese Reddit users strongly criticize the Russian leadership, attributing the human disaster in Ukraine to their actions.

Overall, the divergence in responses between English and Chinese posts underscores the influence of cultural and linguistic contexts on the expression of sentiment for global events like the war in Ukraine.

5 Future Work

Indeed, the sentiment analysis conducted in this study must be understood within a broader context. Notably, social media platforms often harbor unique political standpoints and user demographics, which can significantly influence the sentiment and content of the posts. For instance, most English and Chinese users data we scrapped from Reddit express support for Ukraine in the conflict. This observation could be different on other platforms, and a comparative analysis between different social media sites could be a fruitful direction for future research.

From a linguistic perspective, it's also crucial to acknowledge that the sentiment expressed in posts doesn't always mirror the user's actual emotions. For instance, a user could be discussing an entirely unrelated negative event in their daily life[26] that occurred before they engaged in a conversation about the Ukrainian war, or they might typically use language that carries negative sentiment even when their emotional state is neutral or positive. This behavior is especially noticeable among some Chinese users, who frequently use words with negative connotations, such as "conflict," "fight," and "attack," without necessarily having negative emotions in that post.

Future research could further explore these intriguing aspects of sentiment analysis. It could dive deeper into the intersection of cultural, platform-specific, and individual linguistic factors that shape the sentiment expressed online, providing a better understanding of how people worldwide react to international conflicts such as the Ukrainian war.

References

Alexa - Reddit Competitive Analysis, Marketing Mix and Traffic. Accessed May 16, 2023. https://web.archive.org/web/20171120163709/https://www.alexa.com/siteinfo/reddit.com#section_traffic.

Beigi, Ghazaleh, Xia Hu, Ross Maciejewski, and Huan Liu. "An overview of sentiment analysis in social media and its applications in disaster relief." In *Sentiment analysis and ontology engineering: An environment of computational intelligence*, 313–340. Springer, 2016.

Ciuriak, Dan. *The Role of Social Media in Russia's War on Ukraine.* Available at SSRN, 2022.

Cobo, Alfredo, Denis Parra, and Jaime Navon. "Identifying relevant messages in a Twitter-based citizen channel for natural disaster situations." In *Proceedings of the 24th International Conference on World Wide Web Companion*, 1189–1194. 2015.

Grossi, Giuseppe, and Veronika Vakulenko. "New development: Accounting for human-made disasters—comparative analysis of the support to Ukraine in times of war." *Public Money & Management* 42, no. 6 (2022): 467–471.

26. Maite Taboada, "Sentiment analysis: An overview from linguistics," *Annual Review of Linguistics* 2 (2016): 325–347.

Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Proceedings of the international AAAI conference on web and social media*, 8:216–225. 1. June 2014.

Kaan, A. R. I. K. "Social Media Content Review of Popular MMORPG Games: Reddit Comment Scraping and Sentiment Analysis." *Journal of Emerging Computer Technologies* 2, no. 1 (2022): 13–21.

Lindsay, Bruce R. *Social Media and Disasters: Current Uses, Future Options, and Policy Considerations*. Technical report. Congressional Research Service, September 2015.

Loria, Steven. *TextBlob*. Accessed May 16, 2023. https://github.com/sloria/TextBlob.

———. "TextBlob Documentation." *Release 0.15* 2, no. 8 (2018).

Luo, Ruixuan, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. "PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation." *CoRR* abs/1906.11455 (2019). https://arxiv.org/abs/1906.11455.

Moore, Carrie, and Lisa Chuang. "Redditors revealed: Motivational factors of the Reddit community." In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2313–2322. 2017.

Mueller, Andreas. *WordCloud: A Python package to generate word clouds*. Accessed May 16, 2023. https://github.com/amueller/word_cloud.

Neri, Federico, Carlo Aliprandi, Federico Capeci, and Montserrat Cuadros. "Sentiment analysis on social media." In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 919–926. IEEE, 2012.

Peng, Altman Yuzhu. "A Chinese feminist analysis of Chinese social media responses to the Russian invasion of Ukraine." *International Feminist Journal of Politics* 24, no. 3 (2022): 482–501.

Pew Research Center. *Public Expresses Mixed Views of U.S. Response to Russia's Invasion of Ukraine*, March 2022. Accessed May 16, 2023. https://www.pewresearch.org/politics/2022/03/15/public-expresses-mixed-views-of-u-s-response-to-russias-invasion-of-ukraine/.

Python Reddit API Wrapper Development. *PRAW: The Python Reddit API Wrapper*. Accessed May 16, 2023. https://github.com/praw-dev/praw.

Qiu, Xipeng, Peng Qian, and Zhan Shi. "Overview of the NLPCC-ICCPOL 2016 Shared Task: Chinese word segmentation for micro-blog texts." In *NLPCC/ICCPOL*, 10102:901–906. Lecture Notes in Computer Science. Springer, 2016.

Suciu, Peter. "Is Russia's Invasion Of Ukraine The First Social Media War?," March 2022. Accessed May 16, 2023. https://www.forbes.com/sites/petersuciu/2022/03/01/is-russias-invasion-of-ukraine-the-first-social-media-war/?sh=c3a3fc21c5cd.

Taboada, Maite. "Sentiment analysis: An overview from linguistics." *Annual Review of Linguistics* 2 (2016): 325–347.

Xu, Jingjing, and Xu Sun. "Transfer learning for low-resource Chinese word segmentation with a novel neural network." *CoRR* abs/1702.04488 (2017).