Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230739

Ethnic minorities' mentality and homosexuality psychology in literature: A text emotion analysis with NRC lexicon

Xu Yimu

Sichuan University, No.24 South Section 1, Yihuan Road, Chengdu, Sichuan Province, 610065, P.R.China

Xuyimu1213@outlook.com

Abstract. As a significant subfield of natural language processing (NLP), text emotion analysis has been extensively researched and applied in various domains, such as media, education, and medicine. It has shown significant results in annotating blog posts that rely on an extensive corpus of short phrases. However, in interdisciplinary fields like literary pragmatics, character emotion analysis in literature becomes crucial. Despite the importance of this topic, there are fewer studies, especially for niche subjects such as ethnic minorities' mentality and homosexuality psychology. This paper examines the effectiveness of the widely used lexicon National Research Council of Canada (NRC) in detecting metaphorical words in the famous homosexual novel Maurice. To increase the accuracy of the test, we classified and cleaned the stop words using the Natural Language Toolkit (NLTK) before the analysis step. Our results indicate that the lexicon is able to demonstrate reasonable emotional changes in the story.

Keywords: Natural Language Processing, Emotion Analysis, Emotion Detection, Literature.

1. Introduction

The analysis of text emotion is an important aspect of natural language processing, with significant applications in capturing users' emotions from feedback or managing and controlling information over the internet [1-2]. Two primary methods for text emotion analysis exist classic methods that use emotional lexica or machine learning and deep learning methods [3]. Unlike sentiment analysis, text emotion analysis involves labeling emotions into more specific and fine-grained categories, such as fear, trust, surprise, and anticipation. It is, therefore, more effective in emotionally-charged data processing [4].

When applying emotional analysis to literature, the analysis can be more complex due to the volume of text and linguistic complexity, particularly for niche subjects such as homosexual novels, which often contain obscure and religious metaphors. However, understanding the main characters' emotions can provide insight into the story and the author's writing style. Unfortunately, few researchers have explored emotional analysis in traditional "Forbidden Literature," such as homosexual fiction.

This study aims to use emotional analysis methods to examine the classic homosexual novel Maurice by Edward Morgan Forster [5]. We will use the National Research Council of Canada's emotional lexicon and existing emotional analysis methods proposed in this paper, along with Python and R installers, to generate emotion-changing bar charts with plot changes.

DOI: 10.54254/2755-2721/13/20230739

This paper's organization will start with a review of related works in Sect. 2, followed by a description of related methodologies in Sect. 3. Sect. 4 will present the approaches from data cleaning to emotional analysis, and Sect. 5 will conclude and discuss the shortcomings and outlook.

2. Literature Review

It is indispensable to classify emotions for emotional analysis, the number of theories and models are proposed for emotion classification [6], and most of them are related to modern psychology [7]. French philosopher Descartes, in his book named On Emotions, argues that there are 6 primordial emotions of humans: surprise, happiness, hate, desire, joy, and sorrow, and all other emotions are branches or combinations of these six primordial emotions. Because facial micro-expressions can identify some emotions and physiological processes (e.g., heart rate, nystagmus frequency, sweating, etc.), American psychologist Ekman figures 6 basic emotions, which have a similar meaning to each word from what Descartes has mentioned [8]. They are joy, sadness, anger, fear, disgust, and surprise. In his multidimensional model of emotions, American psychologist Plutchik conceptualized eight basic bidirectional emotions represented in a wheel [9]. This includes Ekman's six basic emotions, as well as trust.

Many emotional analyses rely on emotion lexica [10-11]. Researchers have paid much attention to giving polarity to words. Some try to enlarge a significant, high-quality, word–emotion and word–polarity association lexicon [12], which is more available for sophisticated applications, for example, to build lexica based on finer-grained emotion analysis [13]. Some investigate new resources that perform best in unsupervised settings as non-domain-specific lexica [14]. Alao, static motion lexica are less useful for social media analysis because the vocabulary and context are much variable, so it is suggestible to try extracting the word-emotion lexicon from labeled corpus [15].

Deep learning is a prominent branch of machine learning that is often used in natural language processing, particularly for text emotion analysis based on Convolutional Neural Networks (CNNs). CNNs have been shown to effectively capture important features from microblogs and accurately classify them [16]. However, since sentences can convey a range of emotions, a multi-task CNN has been developed for text emotion analysis [17], which approach has been tested on several public text datasets and has demonstrated promising results [18].

The latest emotional analysis applications can apply to many areas. Recent advancements in the emotional analysis have made it applicable to a wide range of fields. One area where it is widely used is in short-form text analysis, with most research focusing on data from social internet platforms such as Twitter. Twitter, a popular online social networking and microblogging service, limits users' posts to 140 words, making them easily accessible as datasets. These specific datasets are then leveraged to generate personalized recommendations [19]. Some also study to identify fake news and classify different emotional patterns by proposing a Long Short Term Memory(LSTM) model that can detect unreal information [20]. However, there is still much to do for the emotion analysis method combined with the literature study. Though nowadays, researchers have paid more attention to Digital Humanities (DH), turning into a computational turn in which studies try to track plot developments and textual emotions [21]. Based on existing research, this paper aims to not only determine if existing lexica are suitable for detecting characters' emotions and how they fluctuate with changes in the storyline but also to evaluate the effectiveness of conventional emotion analysis methods in providing a proper display of the ups and downs of the plot.

3. Research Methods

Aiming to present a picture of the protagonist's changing emotions in Maurice as the plot change and validate the traditional lexicon, NRC can analyze emotion for the storyline. Figure 1 shows the framework overview.

DOI: 10.54254/2755-2721/13/20230739

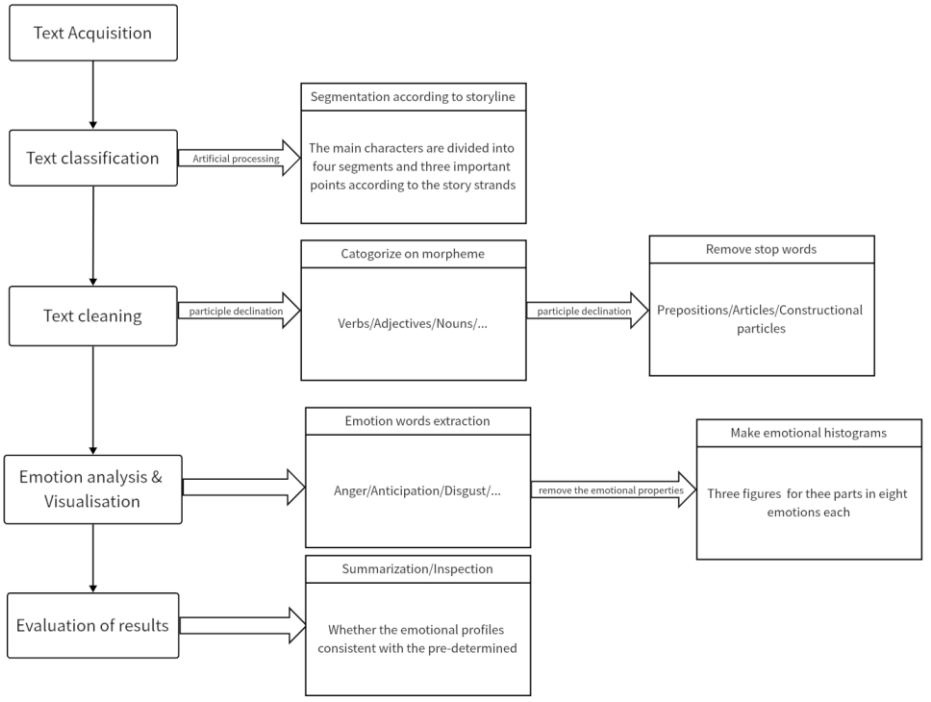


Figure 1. Framework overview.

In the following subsections, we will divide the full text into several parts based on the development timeline. We will then clean the data for each paragraph using Python, removing auxiliary nouns or verbs that are not useful for emotion analysis (e.g., "is," "have," "go," etc.). Next, we will slice the text in each section and split the words. Finally, we will use R's analysis package to analyze the histogram of the emotion changes for the three parts we divided in Sec. 4.3.

3.1. Data Set Selection and Segmentation Criteria

In order to achieve one of the main tasks of using emotional analysis to show the emotional changes of the main character in each section of the story through images, Here the research has chosen an emotional lexicon as an important foundation and basis for the analysis of emotions in this paper, that is National Research Council of Canada (NRC) [22]. It has already been embedded into "tidytext" package in R. Therefore, it would be easily deployed for analysis. With it, we can mainly detect 8 emotions throughout the full story: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, and the polarity of each word. The main reasons for choosing it are that it has already been used for several years, has also been tested on many data sets, and generalizes perfectly in other domains [23]. Besides, hopefully, it would have a good result on the test.

4. Case Study

4.1. Text Classification

To classify the novel Maurice, we have divided the story into four phases and three nodes based on three critical transitions in the main character's life. The first emotional upsurge occurs in chapter 12 when the two characters go through inner torment and eventually become lovers. Before this node, the story covers the developments of Maurice Hall's childhood and his education at Cambridge, where his emotions are relatively calm and harmonious. After this, in chapter 25, another important turning point occurs when Clive chooses to end their relationship, and Maurice's emotions plummet to the bottom. He struggles through a disillusioned and self-doubting life for years, attending Clive's wedding and seeking a doctor's help in London, among other events. In chapter 37, the story's last highlight occurs

DOI: 10.54254/2755-2721/13/20230739

when Scudder, a young servant in Clive's house, shares his body and soul with Maurice one night. The final slice of the story covers Maurice's period of tangling and quarreling with Alec, becoming lovers with each other, and confessing to Clive about his relationship with Alec. This phase is one of the easier stages for observing the emotional ups and downs of the protagonist. Figure 2 illustrates the segmentation of the story process.

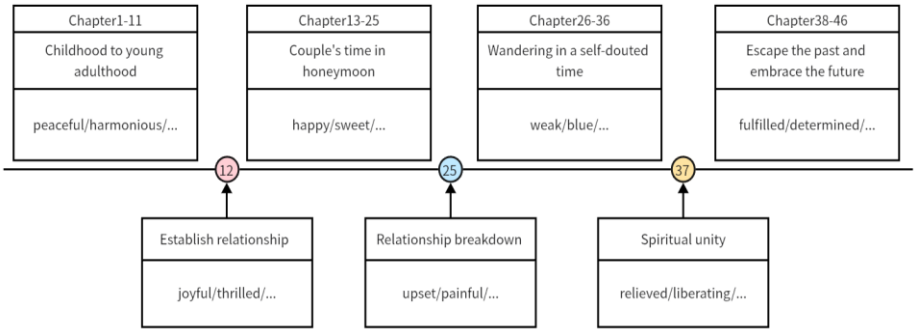


Figure 2. Sub-sections of the story.

As a result, we can create a fluctuation graph to illustrate the protagonist's emotional changes as the chapters progress. Figure 3 displays the emotional changes over the course of the novel. The horizontal axis indicates the chapters, and the vertical axis shows the two poles of emotion.

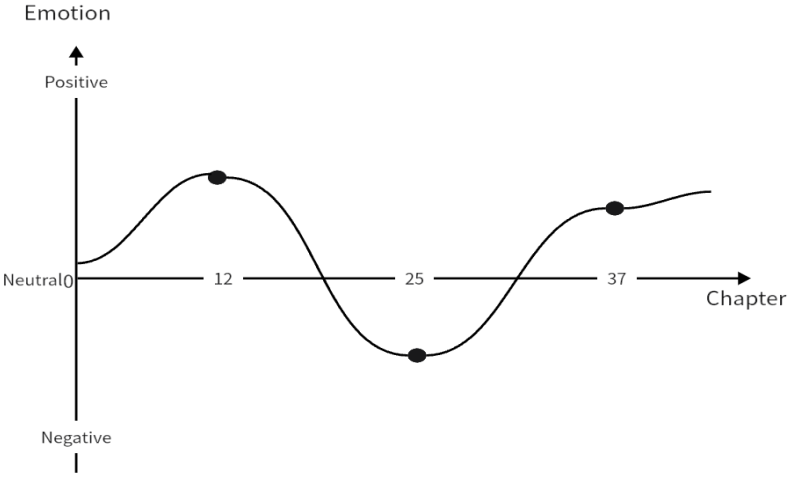


Figure 3. Graph of the protagonist's emotional changes.

4.2. Text Cleaning

Cleaning up high-frequency short function words or phrases is an essential step in data preprocessing for emotional analysis. It reduces the amount of data for subsequent text processing and minimizes interference with the accuracy of the emotion analysis. In our research, we removed stop words from the Natural Language Toolkit (NLTK), such as quantitative words (e.g., "one," "many," "lot"), text linkers or prepositions (e.g., "and," "but"), alternatives to uncertainty (e.g., "someone," "who," "every"), frequency words (e.g., "always," "usually," "rarely"), and articles (e.g., "the," "an," "a"). We also removed descriptive phrases, such as character backgrounds or short conversations (e.g., "How old are you?").

The whole novel contains approximately 60,000 words. After the word separation and de-stopping process, the final materials for analysis were reduced to approximately 5,000 words. To vividly represent the frequency of words, we tested the effect of the de-stopping process by creating a word

DOI: 10.54254/2755-2721/13/20230739

cloud of the preprocessed text in Figure 4. The word cloud displays the high-frequency words in the text, which include names of key individuals (e.g., Maurice, Clive, Durham), nouns with special meaning (e.g., man, boy, love), and verbs or conjunctions that play a vital role and cannot be removed (e.g., look, call, without).

Figure 4. word cloud diagram of the story.

After removing useless words, the next step is to convert the data into a structured format that R can use for emotion analysis. First, we use Python's Panda package to segment the text, which involves using regular expressions to remove blank lines and organize the text into data boxes. Then, we split the complete text into lines using line breaks and added line numbers to each line. Finally, we convert the data to CSV format so that R can read and process it. Table 1 displays the first three rows of the array conversion from Chapter 12 into a data frame using the summary function.

Table 1. the frame of the first three rows of chapter 12.

	LINE	TEXT
0	1	Clive had suffered little from bewilderment as…
1	2	At first, he thought God must be trying him, an…
2	3	His sixteenth year was ceaseless torture. He…
3	4	These terrors had visited Maurice, but dimly…
4	5	The boy had always been a scholar, awake to…

4.3. Emotion Analysis and Visualization

RStudio provides an interactive environment that allows us to execute R commands and receive immediate feedback on the results. This study used a lexicon-based approach mentioned in Sec. 3 to analyze emotion. The following steps were taken to implement the method:

1. After importing the data in CSV format, we installed the "dplyr," "tidytext," "tidyr," and "ggplot2" packages and loaded them.
2. We separated the sentences into words using the "tidytext" package, and the "unnest_token" statement was executed. We retained the original line numbers to see which line each word comes from, which facilitated our analysis of lines and even paragraph units below. Table 2 provides an example of splitting words in the first line of Chapter 12.

DOI: 10.54254/2755-2721/13/20230739

Table 2. example of splitting words in the first line of Chapter 12.

	LINE	WORD
1	1	clive
2	1	had
3	1	suffered
4	1	little
5	1	from
6	1	bewilderment

3. We invoked the NRC lexicon embedded in the "tidytext" package. Throughout the process, the NRC lexicon was used to give proper sentiments for each word. For example, in Table 3, we can see different sentiments of emotional words extracted from the first ten lines of Chapter 12. Interestingly, the same word "god" appears in three separate lines: 8, 9, and 10, and is analyzed for three different emotions: anticipation, fear, and joy. This meticulous analysis of the same word in different contexts reflects Clive's obscure emotions about religious belief and constraint. Therefore, NRC's grasp of the emotional expression of the same word in different contexts is quite adequate.
4. We used the "ggplot" package to visualize each part of emotions. Since analyzing emotional changes at a 1-line unit granularity is too fine, and the entire story contains several thousand lines of text, we used 10 or 15 lines as the base unit for analysis. We used the index to take the original line numbers and process them into paragraphs using the "%/%" symbol for integer division. Then, we used the "geom_col" command to draw the bar charts. Different moods were indicated in different colors.

Table 3. sentiments of words in first ten lines of chapter 12.

	LINE	WORD	SENTIMENT
1	1	bewilderment	fear
2	1	bewilderment	surprise
3	1	sincere	positive
4	1	sincere	trust
5	1	sense	positive
6	1	wrong	negative
7	1	damned	negative
8	1	god	anticipation
9	1	god	fear
10	1	god	joy

Typically, positive and negative sentiments correlate negatively in the same process. Therefore, if there is a positive correlation, we need to examine the emotional words that are positively and negatively correlated and add them to the stop words list. Even though deactivation is handled in Sec 4.2, and the "tidytext" package provides a list of stop words, there is still a possibility that some stop words are undefined and may disturb the analysis results. For instance, we noticed that "Mother" is classified as both a positive and negative word. Therefore, revising the stop words list is necessary. Using the "bind_rows" command in R, we can add stop words to the base list of pre-populated stop words.

DOI: 10.54254/2755-2721/13/20230739

Table 4. Examples of miss-classified 'Mother'.

JOINING WITH BY = JOIN_BY(WORD)		
	word	n
1	mother	55
2	words	23
3	wrong	15

(a) Example of classifying 'Mother' into the negative word

JOINING WITH BY = JOIN_BY(WORD)		
	word	n
1	mother	55
2	love	33
3	friend	30

(b) Example of classifying 'Mother' into the positive word

5.We added the stop words list and removed the emotional properties using a filter. We analyzed emotions, not sentiments. As in Sec 4.1, the story was separated into four periods with three timing nodes. We generally combined period 1 and node 1 into one graph, period 2 with node 2 into the second one, and the last part became the third graph. Those three parts correspond to the figures below. We then used the "facet_wrap" function to split up the different emotions and draw them separately for those three parts, which was convenient for identifying emotions.

4.4. Evaluation

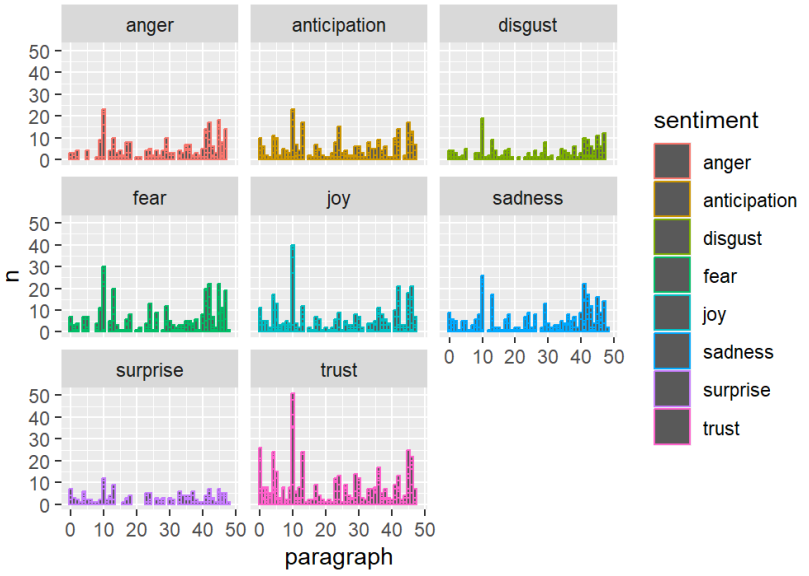
The above results show distinct changes in the characters' emotions throughout the story, and they appear to have different emotional tendencies. Each graph shows eight basic emotions included in NRC, and their columns appear in different colors. The horizontal axis represents the number of paragraphs, while the vertical axis shows the number of emotional words.

In Figure 5(a), all emotions are undulating, and the emotional word count hovers between 20 and 30. Trust and joy are the most prominent emotions, which climb up to 40 and 50 words in one phrase. This suggests Maurice's teenage years are generally peaceful, tranquil, and light-hearted but ignorant.

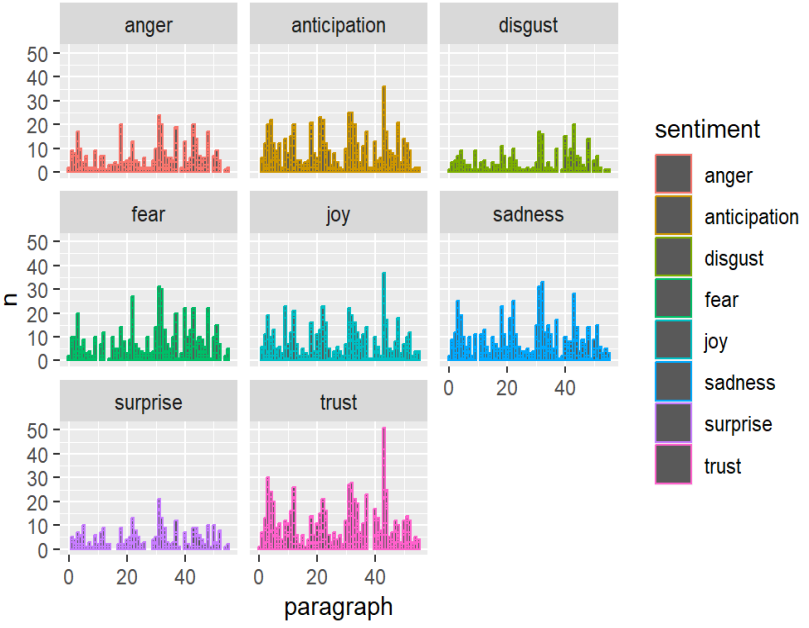
Regarding Figure 5(b), emotional ups and downs become apparent. From chapter 13 to chapter 36, Maurice experiences a rollercoaster ride of emotional change, from sweet couple time to the end of their intimate relationship and the terrible blow of Clive's marriage. Distinctly, all emotional histograms appear to have a U-shape tendency in this part. Most of the emotions can still be captured correctly, and negative emotions like sadness surge to about 35 words in about phrase 35, while the emotion of joy shows a downward trend.

In Figure 5(c), the magnitude of emotional change is somewhere between Figure 5 (a) and Figure 5 (b), smoother than Part 2 but more undulating than Part 1. Notably, all the emotions peak to a high level, even up to 50 emotional words in ten lines. This is because the story ends with an emotional climax for both main characters, Maurice and Clive. Maurice gets the happiness he has desired for years, a lifetime of companionship with a loved one, while Clive suddenly realizes that he has lost Maurice forever. Therefore, sadness and fear come as unnoticed.

DOI: 10.54254/2755-2721/13/20230739

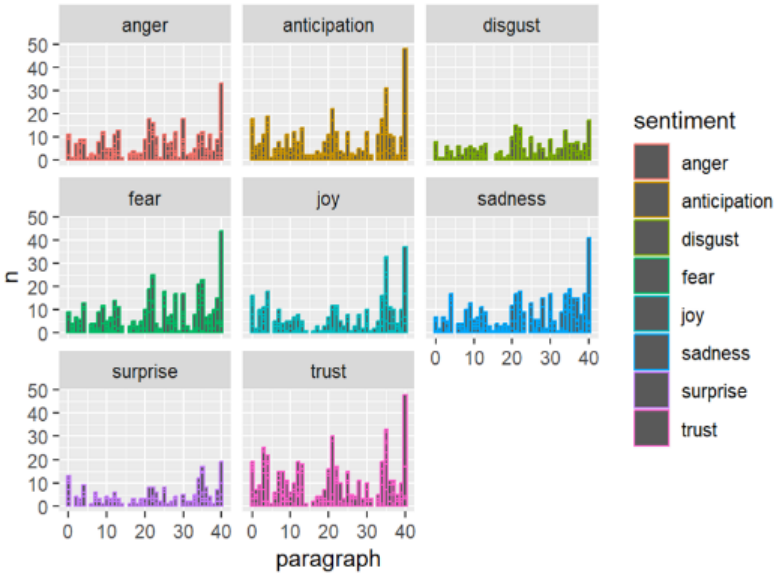


(a) Emotional histogram of part1



(b) Emotional histogram of part2

DOI: 10.54254/2755-2721/13/20230739



(c) Emotional histogram of part3

Figure 5. Emotional histograms of three parts.

5. Conclusion

In our study, we employed the classic emotion analysis method based on the NRC lexicon to analyze the entire content of Forster's novel, Maurice. We followed a series of steps, including text classification, cleaning, and emotional visualization using the "tidytext" approach. By adding a list of exclusive stop words for literature in combination with the list already in the NRC, we could approximately demonstrate the variations in the characters' emotions throughout the storyline using multidimensional tangents plotted with "ggplot". However, we also discovered that the lexicon might lead to inaccurate emotion capture when the parts to be processed become longer or when the emotions of the same word change in different contexts, as exemplified by the word "mother" in Sec. 4.3.

It is important to note that this work is limited to only one book, so the study sample size is relatively narrow. Furthermore, the character's emotions in the text are sketched in Forster's style, and our results may not necessarily indicate that all niche novels can have the same effects. In addition, we did not compare the efficiency of commonly used lexica, which is a potential avenue for future research. It is worth mentioning that handling the stop words in Sect. 4.1 and emotional sub-sections of the content are primarily based on our subjective judgment.

As we look towards future work, we propose developing a lexicon trained with English-language-based literature data sets. This would involve paying more attention to the logic and model algorithm building for psychological depiction in literary texts while incorporating religious metaphors, author's literary genre, or other deep semantics into the conditions to be considered. Additionally, there is still much progress to be made when using deep learning methods to interpret emotions, which could be beneficial in literature, pedagogy, and psychology.

References

[1] Q. Dong and R. Fang, 'A Deep Learning-Based Text Emotional Analysis Framework for Yellow River Basin Tourism Culture', *Mobile Information Systems*, vol. 2022, pp. 1–9, Sep. 2022, doi: 10.1155/2022/6836223.

[2] F. Li, H. Tang, Y. Zou, Y. Huang, Y. Feng, and L. Peng, 'Research on information security in text emotional steganography based on machine learning', *Enterprise Information Systems*, vol. 15, no. 7, pp. 984–1001, Aug. 2021, doi: 10.1080/17517575.2020.1720827.

[3] R. Feldman, 'Techniques and applications for sentiment analysis', *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013, doi: 10.1145/2436256.2436274.

[4] W. Medhat, A. Hassan, and H. Korashy, 'Sentiment analysis algorithms and applications: A survey', *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[5] E. M. Forster, *Maurice*, Repr. London: Penguin, 1993.

[6] G. K. Verma and U. S. Tiwary, 'Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals', *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014, doi: 10.1016/j.neuroimage.2013.11.007.

[7] J. Hofmann, E. Troiano, K. Sassenberg, and R. Klinger, 'Appraisal Theories for Emotion Classification in Text'. arXiv, Nov. 03, 2020. Accessed: Feb. 19, 2023. [Online]. Available: http://arxiv.org/abs/2003.14155

[8] J. L. Tracy and D. Randles, 'Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt', *Emotion Review*, vol. 3, no. 4, pp. 397–405, Oct. 2011, doi: 10.1177/1754073911410747.

[9] E. Tromp and M. Pechenizkiy, 'Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel'. arXiv, Dec. 15, 2014. Accessed: Feb. 18, 2023. [Online]. Available: http://arxiv.org/abs/1412.4682

[10] J. Staiano and M. Guerini, 'DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News'. arXiv, May 07, 2014. Accessed: Feb. 20, 2023. [Online]. Available: http://arxiv.org/abs/1405.1605

[11] L. De Bruyne, P. Atanasova, and I. Augenstein, 'Joint emotion label space modeling for affect lexica', *Computer Speech & Language*, vol. 71, p. 101257, Jan. 2022, doi: 10.1016/j.csl.2021.101257.

[12] S. M. Mohammad and P. D. Turney, 'CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON', *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, Aug. 2013, doi: 10.1111/j.1467-8640.2012.00460.x.

[13] H. Li and F. Ren, 'The study on text emotional orientation based on a three-dimensional emotion space model', in *2009 International Conference on Natural Language Processing and Knowledge Engineering*, Dalian, China, Sep. 2009, pp. 1–6. doi: 10.1109/NLPKE.2009.5313815.

[14] O. Araque, L. Gatti, J. Staiano, and M. Guerini, 'DepecheMood++: A Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques', *IEEE Trans. Affective Comput.*, vol. 13, no. 1, pp. 496–507, Jan. 2022, doi: 10.1109/TAFFC.2019.2934444.

[15] A. Bandhakavi, N. Wiratunga, D. P, and S. Massie, 'Generating a Word-Emotion Lexicon from #Emotional Tweets', in *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, Dublin, Ireland, 2014, pp. 12–21. doi: 10.3115/v1/S14-1002.

[16] D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W. Shi, 'Deep learning based emotion analysis of microblog texts', *Information Fusion*, vol. 64, pp. 1–11, Dec. 2020, doi: 10.1016/j.inffus.2020.06.002.

[17] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang, 'Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network', in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Jul. 2018, pp. 4595–4601. doi: 10.24963/ijcai.2018/639.

[18] S. Wang, M. Huang, and Z. Deng, 'Densely Connected CNN with Multi-scale Feature Attention for Text Classification', in *Proceedings of the Twenty-Seventh International Joint*

Conference on Artificial Intelligence, Stockholm, Sweden, Jul. 2018, pp. 4468–4474. doi: 10.24963/ijcai.2018/621.

[19] S. M. Mohammad and S. Kiritchenko, 'Using Hashtags to Capture Fine Emotion Categories from Tweets: USING HASHTAGS TO CAPTURE FINE EMOTION CATEGORIES', *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, May 2015, doi: 10.1111/coin.12024.

[20] B. Ghanem, P. Rosso, and F. Rangel, 'An Emotional Analysis of False Information in Social Media and News Articles', *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–18, May 2020, doi: 10.1145/3381750.

[21] E. Kim and R. Klinger, 'A Survey on Sentiment and Emotion Analysis for Computational Literary Studies'. Jul. 11, 2022. doi: 10.17175/2019_008.

[22] S. M. Mohammad and P. D. Turney, 'Nrc emotion lexicon', 2013.

[23] X. Zhu, S. Kiritchenko, and S. Mohammad, 'NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets', in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 443–447. doi: 10.3115/v1/S14-2077.