

Emotional Language usage across Genres

Mattias Kraak, S5407427

<https://github.com/mattias767/IRM-final-project>

Abstract

In a book, the use of words that contain emotion is often connected to the genre of the book. For example, horror books are connected to fear and suspense, while romantic books are more connected to joy and affection. This report investigates the difference in emotional language usage across two very different genres. We will investigate this difference using the books *Frankenstein* (Shelley, 1818) and *Pride and Prejudice* (Austen, 1813). This project focusses on the frequency and distribution of the emotion words in the books. The words are classified based on the lexicon created by Mohammad and Turney, 2013. The report shows how the emotion words can be quantitatively examined and how the genre of a book can be determined by this.

1 Introduction

Emotion plays an incredible important role in literature. It defines how the story is told, how the characters are and how the reader engages with the story. From fear and anger that is present in a horror novel to the joy and love in a romantic novel. Authors use emotional language to convey certain feelings and meanings to the readers.

This report investigates how emotional language usage differs between two genres of novels; horror and romantic fiction. We compare the books *Frankenstein* by Shelley and *Pride and Prejudice* by Austen. These two novels are published around the same time, but belong to the two mentioned genres. By focussing on the distribution of the emotion words across predefined emotional categories, this report tries to explore

the influence of these words on the genre of the book. By understanding this influence and to what extent each word in a certain category influences the genre, we can later use this to automatically classify books on genre.

The research question for this report is:

"How does the usage of emotional language differ between a horror and a romantic novel, and to what extent do certain emotion word categories occur in each genre?"

To answer this question the, we use a crowd-sourced lexicon to categorize the emotion words. It is expected that the horror genre contains significantly more negative-emotion words, especially words that imply anger and fear. Whereas the romantic genre will have significantly more positive-emotion words, such as words that imply joy and happiness.

2 Related Work

2.1 Word-Emotion Association Lexicon

For this report the way the emotion words are categorized is very important. Mohammad and Turney (2013) introduced the NRC Emotion Lexicon, which associates words with eight basic emotions; joy, sadness, anger, fear, trust, disgust, anticipation and surprise. Mohammad and Turney crowd-sourced this lexicon with more than 2,000 annotators. They provided a lexicon that is well suited to quantify emotional language in a more objective way than our own subjective interpretation of the words. The lexicon will be used to identify and count all the emotion words in both books, but it can be used for any text.

2.2 Relation between Genre and Emotional Plot Development

Kim, Padó, and Klinger (2017) investigate the relationship between literary genres and emotional plot development. Their study focuses not only on the presence of emotion words in texts, but also on the emotional arcs. These arcs represent how emotional intensity rises and falls throughout the text. They found that genres often show characteristic emotional arcs: for example, horror tends to show a sustained negative emotional pattern, while romance shows a more positive development in the story.

This research supports the idea the genre influences the emotional expression and therefore the emotion words in a text. This could mean that the amount and distribution of emotion words in a text are indicative of the genre.

3 Data

3.1 Data selection

The data that will be used in this research is first of all the two books: *Frankenstein* (Shelley, 1818) and *Pride and Prejudice* (Austen, 1813), obtained from <https://gutenberg.org/>. Then for the analysis of the texts, the NRC Emotion Lexicon will be used to categorize the words.

The variables used in this research can be found in table 1 The independent variables are the genres of the book. The dependent variables will be negative-emotion words and positive-emotion words as annotated in the Lexicon.

	Horror	Romantic
Negative-emotion words		
Positive-emotion words		

Table 1: Dependent and independent variables

3.2 Preprocessing

Before processing the texts they should first be cleaned up. The titles, index, headers, etc. have to be removed. Then the texts will be tokenized, converted to lowercase, punctuation removed and the words will be lemmatized to their base form. This way the emotion word counts are not influenced by differences in spelling or formatting.

3.3 Method

First each token from the text will be matched against the lexicon, if a match is found the word will increase the counter for either negative or positive emotion words. The amount of emotion words will be normalized, meaning it will be divided by the total amount of words, so you get the amount of negative or positive emotion words per 1000 words. At last the amount of emotion words will be compared between the genres.

4 Predicted Results

Table 2 shows the results from the research carried out. It shows the amount of positive and negative emotion words per 1,000 words in each of the novels.

	Horror	Romantic
Negative	87	34
Positive	23	103

Table 2: Amount of emotion words per genre (per 1000 words)

Discussion These results show that this novel of the horror genre has significantly more negative emotion words than the romantic novel. It also shows that the romantic novel has significantly more positive emotion words. An interesting note is that the romantic novel has relatively more positive words than the horror novel has negative words. The romantic novel still has some negative words, however the horror novel had relatively less positive words.

5 Conclusion

This study aimed to show the influence of emotion words on the genre of a book. We did this by getting two book of very different genres from the same time period and counted their emotion words using a crowdsourced lexicon. It showed that the horror indeed has more negative emotion words and the romantic novel had more positive emotion words. This could mean that the amount of emotion words can be a good indication for what the genre of a certain book is.

In this research the words from the text are matched to the words in the lexicon and we only look whether the word is more positive or more negative. This approach works as long as there

are no more than two genres and you want to know which one is more positive or negative compared to the other one. It does not predict the genre of the book by analyzing the text. In future research the words could be matched to one of the 8 emotions also in the NRC Emotion Lexicon, this can give a better distinction between the texts since it can have more words of one or more emotions than others.

References

- Austen, J. (1813). *Pride and Prejudice*. T. Egerton.
- Kim, E., S. Padó, and R. Klinger (2017). Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, pp. 17–26. Association for Computational Linguistics.
- Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29(3), 436–465. 436.
- Shelley, M. (1818). *Frankenstein; Or, The Modern Prometheus*. Lackington, Hughes, Harding, Mavor Jones.