Emotion based text style transfer

Siwa Sardjoemissier
siwa.sardjoemissier@student.uva.nl

July 1, 2020, 43 pages

Academic supervisor: Adam Belloum, A.S.Z.Belloum@uva.nl
Daily supervisor: Michael Metternich, michael.metternich@bloomreach.com
Host organisation: Bloomreach BV, https://bloomreach.com

Universiteit van Amsterdam
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Master Software Engineering
http://www.software-engineering-amsterdam.nl

Abstract

'Sentiment' is defined as the way our emotional states, opinions, desires, feelings and concerns are reflected through our use of language. Gaining an understanding of sentiments is highly valuable - for commercial as well as academic purposes. The past few decades have seen a considerable rise in computational power and a corresponding increase in interest in programmatic sentiment analysis, specifically the detection of emotion in text.

So we can now tell, to some extent, what the emotional tone of a text is. But what if we take it a step further and attempt to change its emotional tone? Research on this topic - also called text style transfer - is scarce; most literature related to the modification of the tone of a text cover modification on a single polarity, like negative to positive. There is little research on the specific topic of emotion-based text style transfer.

In an attempt to address this gap, this research sought to answer three main questions: can we analyse texts in ways that will provide their emotional tone? Can we provide suggestions on changing the emotional tone of a text? And lastly, how can we validate the accuracy of such emotion tone changes?

Our results show that a rule-based emotion detection approach is just as accurate as existing machine-learned emotion detection approaches, with the additional benefit of providing the information needed to detect what changes need to be made to alter the emotional tone. After using this information to propose and test four types of modifications, we tested each based off two adjustment focuses. Our results showed that our approach can change the emotional tone of a text with an accuracy ranging between 39% and 72%. The code was used during this research has been made publicly available at `https://github.com/swcloud1/Thesis`.

We validated our results using a combination of grammar-checking, spellchecking and emotion detection methods. We believe this validation method to provide reliable results, but to lack in its predictive ability to accurately detect logical accuracy in sentences. We believe the introduction of human-evaluation to assess validity in conjunction with these methods to be the best approach to validate whether an input- or output sentence is valid.

Contents

Chapter 1

Introduction

Communication is a huge part of our daily lives. It is the main way we transfer information to one another. When humans communicate, they do so through language. Language is defined as "a system which consists of a set of sounds and written symbols which are used by the people of a particular country or region for talking or writing"[1]. Besides static information, we use it to express the way we feel about things. We use it to reflect our emotional state, opinions, desires and concerns. A good understanding of these sentiments is therefore valuable. For example, companies can use it to gain an understanding of the feelings of (potential) customers, politicians to gather information about the view of the public, healthcare professionals to better understand the needs of their patients and people use it to help each other.

It is therefore not surprising that throughout history people have sought to analyse the communication of others. We find evidence of this going as far back as ancient Greece; where voting was introduced as a means to gather the opinion of residents[2], as well as in the "Iliad", where Agamemnon attempted to gauge the fighting spirit of his army[3]. In more recent history, namely the 1930's, we find research on the methods used for measuring public opinion[4].

With the rising computational power of the past decades, ventures in programmatic sentiment analysis have increased. Specifically text, due to its availability as a record of human language, is often analysed. Sentiment analysis - or "opinion mining" - can therefore be defined as "a field of study that aims to extract opinions and sentiments from natural language text using computational methods"[5]

This analysis is usually done in one of two ways:

- Polarity Based - Sentiment is scored as either: Negative, Neutral or Positive. Usually scored on a range from -1 to 1.
- Beyond Polarity - Sentiment is scored using other methods, such as emotion-based [Happy, Angry, Sad, Excited, Bored, Fear] in which a score is attributed to each emotion.

Most academic research in this field has been done on polarity based sentiment analysis - so much so that some sources describe Sentiment Analysis as solely measuring positive or negative attitudes[1]. However, our communication is a lot more nuanced than solely being positive or negative. For example, one could communicate fear or anger, which would both be labelled as negative, but convey vastly different sentiment or intent. If, for example, a customer of a company conveys fear, this would require a different response from the company than if the customer were to convey anger. Hence the more recent surge in research on emotion detection, a beyond polarity approach.

1.1 Problem statement

Based on previous examples, it's safe to say there is a lot of practical value in knowing the sentiment of a text. Fortunately, due to advances in text sentiment research and the availability of computational power, we're able to tell what the sentiment of a text is. But what if, for example, we conclude that a text does not convey the sentiment we want it to convey? Do we rewrite the text-based off gut-feeling, analyse it again and hope for a more desirable outcome? Or might it be possible to alter the sentiment of a text through automatic computation as well? In the last few years, this question has spurred the research of so-called text style transfer. This is defined as "altering the stylistic elements of a text, while keeping the non-stylistic and content information intact" [6].

Most research in the field of style transfer has been done on polarity based style transfer. While valuable, it leaves us with a similar limitation as polarity based sentiment analyses, namely limited dimensions. Any style transfer can only be done along positive and negative dimensions. This greatly reduces the scope of potential application. For example; style transfer with the intent of making a text more joyful, but not more surprised, is something that could be of great value to a writer of a chatbot, but this is simply not attainable with polarity based approaches. There has been some research on beyond polarity based style transfer. However, these have been limited to single emotions and small side-experiments of larger experiments that used polarity based approaches. Most of these are discussed in chapter 6.

1.1.1 Research questions

In this research, we aim to research the viability of emotion-based text style transfer. We will explore the current landscape of emotion detection and text style transfer to see whether we can conceive a methodology for reliable emotion-based text style transfer. Based on this goal we have phrased the following research questions:

RQ1: How can we analyse text in such a way that it can provide the emotional tone of the text?

Research has been done on the detection of emotions in texts. For example, it has been shown that emotion-detection can be reliably done using a rule-based detection algorithm[7]. Other research has shown this to be possible using a machine-learning approach[8]. However, this research has been done with the intent to merely provide the emotional tone of a text, not to create a path by which each step in the detection can be traced back, providing the opportunity to later use this information to alter the emotional tone of the text. Thus, researching the requirements for a methodology that would enable alteration is a crucial step in answering our research goals.

RQ2: How can we provide suggestions on changing the current emotional tone of a text using text recommendations?

In order to alter the emotional tone of a text, we need to figure out the dimensions by which we can best approach alterations. Is this keeping the semantics intact, or achieving the desired change in emotional tone with a specific jump in intensity. We will need to use existing research on emotion detection, scarce research on emotion manipulation and other investigation in order to conceive of a compelling way to provide these suggestions.

RQ3: How can we validate the accuracy of emotional tone changes?

Once we have made an adjustment to a text, we need to know whether this change is valid. In order to do this, we need to define validity within the context of emotional tone adjustment. We will need to explore current literature alongside language processing resources to define and apply a methodology that can be used to measure validity in emotional adjustment.

1.1.2 Research method

In order to answer the research questions, we conducted various experiments; explored the current landscape of sentiment analysis and sentiment manipulation; and gathered insights and methodologies in order to conceive a model for emotion manipulation. We then tested this model using datasets that are regularly used in academic research on text sentiment and are deemed to be of high quality. We used multiple methods of validation to assess whether emotion-based sentiment manipulation is feasible with the technology we found and analyzed what hurdles are in the way of further developing this mechanism.

1.2 Contributions

Our research makes the following contributions:

1. We generate an overview of the current landscape of text sentiment manipulation, both polarity based and emotion-based.
2. We propose a model for emotion-based sentiment manipulation based on current research.

3. We analyse the feasibility of our method and assess the present difficulties and hurdles that need to be overcome in order to improve our model; or, in case our model does not meet our requirements, what is needed to create the first successful model.

1.3 Outline

In Chapter 2 we describe the background of this thesis and explain commonly used terminology. Chapter 3 describes our research methodology, design and their justification. Our results are shown in Chapter 4 and discussed in Chapter 5. Chapter 6 contains the work related to this thesis. Finally, we present our concluding remarks in Chapter 7 alongside recommendations for future research.

Chapter 2

Background

This chapter will define and provide some background information on the main concepts that are used and referred to throughout this thesis. We start with broader definitions and background information on both sentiment analysis and sentiment manipulation. We conclude with defining some of the key concepts that will return frequently. Chapter 6 covers additional work that is more directly related to our research.

2.1 Sentiment Analysis

In sentiment analysis the goal is to retrieve information on the effect of communicative output. As we can see in *Figure 2.1*, the interest in sentiment analysis has increased year by year for over a decade.

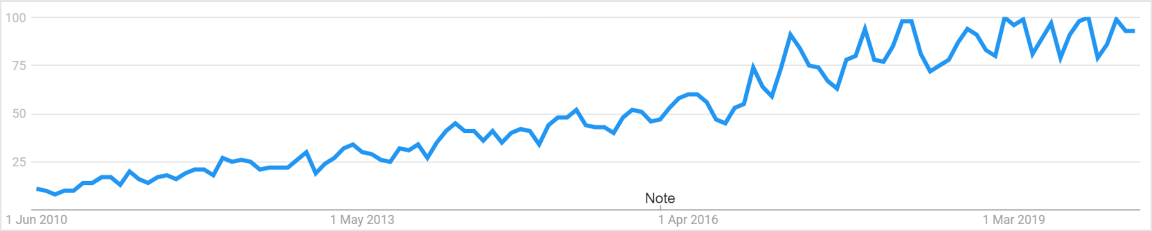


Figure 2.1: Google Trends search results for ”Sentiment Analysis” over the past 10 years.

This increase in interest coincides with the rise of social media platforms as a means of communication. These platforms provide a huge amount of accessible data on the thoughts, opinions and beliefs of both individuals and groups. Sentiment analysis of this data has a large number of applications:

- Monitoring the reputation of brands on social media[9].
- Gathering real-time sentiment information during political elections[10].
- Quantifying the response to news articles[11][12].
- Creating online marketing strategies[13].
- Creating Emotion Dictionaries[14][15][16].

There are different types of sentiment analysis. These can be divided into polarity based analysis, aspect based analysis and emotion-based analysis.

2.1.1 Polarity Based Analysis

Polarity based analysis is the most common type of sentiment analysis. In this type of analysis we quantify sentiment using a single trait, usually positive versus negative. The outcome of such an analysis is a score within a certain range, where one end represents one end of the trait and the other end represent the opposite of that trait. An example of polarity based sentiment analysis can be found in Table 2.2:

Input	*”I lost my dog, so I am sad.”*
Score	*-92*
Polarity	***Negative***

Table 2.1: Example of Polarity Based Sentiment Analysis

This method provides an easy way to label a text as being either positive or negative, but it lacks further nuance.

2.1.2 Aspect Based Analysis

If we would like to know more detailed information about our text, we can use non-polarity based approaches. For one, there is aspect-based sentiment analysis. In this type of analysis the goal is to group sentiments into different aspects. This is useful is cases like product reviews, where one might want to know the sentiments of different aspects of the products. Aspects are either predefined or automatically gathered through the analysis tool and each aspect is given a score, usually based on polarity. An example can be found in Table 2.2

Input	*”The food was great but the service was poor.”*
Found Aspects	*Food, Service*
Polarities	***Food - Positive, Service - Negative***

Table 2.2: Example of Polarity Based Sentiment Analysis

2.1.3 Emotion-Based Analysis

In emotion-based analysis the goal is to detect the intensity of the emotions. The set of emotions that are analysed differ per approach[17]. The most common are the six basic emotions defined by Paul Ekman, namely: anger, disgust, fear, happiness, sadness and surprise[18]. This analysis method often uses emotion lexicons (a list of words with the emotion they convey) or machine learning software to analyse to which extent each emotion is present in an input text. An example can be found in table 2.3:

Input	*”I lost my dog, so I am sad”*
Anger	*2.88%*
Disgust	*1.24%*
Fear	*13.97%*
Happiness	*0.55%*
Sadness	***81.28%***
Surprise	*0.70%*
Main Emotion	***Sadness***

Table 2.3: Example of Emotion-Based Sentiment Analysis

2.2 Sentiment Manipulation

Sentiment manipulation, also called text style transfer, is the act of translating a text into a different sentiment, while keeping the context intact[19]. There are different practical applications for this, such as companies that want to change a marketing text to convey a certain emotion, politicians that want to alter their speeches and developers creating chatbots to interact according to a different input.

Sentiment manipulation is generally more difficult than sentiment analysis, because of the added dimension of having to keep the context of the sentence intact. Sentiment manipulation can be done through different mechanisms.

2.2.1 Polarity Based Manipulation

Polarity based manipulation, just like polarity based analysis, relies on variations of a single trait. This trait is often positive vs. negative. An example of polarity based sentiment manipulation can be found in Table 2.4.

Input	*”I will never go to this restaurant again.”*
Goal	*More Positive*
Output	***”I will definitely go this restaurant again.”***

Table 2.4: Example of Polarity Based Sentiment Manipulation

Research has known this type of manipulation to be possible. Santon and Padhi showed it was possible to transform offensive language into non-offensive language[20]. There are additional use cases, like detecting when an e-mail composer is using a lot of negativity and then suggesting alterations to change the tone of a text. However, these applications have similar limitations as polarity based sentiment analysis: a lack of nuance. A writer might want to make their text more sad, but not less angry. This approach would be too limited to allow for this level of control.

2.2.2 Aspect Based Manipulation

In aspect based sentiment manipulation the goal is to take a certain trait or group of traits in an input text and to modify them. This type of manipulation has been done before. In 2019 Yun et al used aspect based sentiment manipulation to alter the level of formality in sentences. An example of which can be found in Table 2.5.

Input	*”That is if you truly adore them.”*
Goal	*Formal → Informal*
Output	***”That is if u truly luv them”***

Table 2.5: Example of Aspect Based Sentiment Manipulation

2.2.3 Emotion-Based Manipulation

Emotion-based sentiment manipulation attempts to change the emotional tone of a text, while keeping the context intact. An example of this can be found in Table 2.6.

Input	*”The experience in this hotel was awful”*
Goal	*More Surprised*
Output	***”The experience in this hotel was remarkable”***

Table 2.6: Example of Emotion-Based Sentiment Manipulation

This type of manipulation has a host of potential applications. Despite these use cases, the topic of emotion-based sentiment analysis has not been researched much as of yet. There have been some small findings, but, according to our literature research, there has been no research on the feasibility of emotion-based sentiment manipulation. Which is why this research was conducted.

2.3 Approaches

There are different ways to approach both the act of analysis and manipulation, each with their own benefits and drawbacks. In this section we discuss two approaches: Rule-based emotion detection and Machine-learned emotion detection.

2.3.1 Rule-Based Approach

In a rule-based design, both the analysis and manipulation are done according to human-made rules. The rule-based approaches have the benefit of providing information on the reason a particular score is given. An example of a rule-based approach can be found in a research paper by Asghar and Khan[7], in which they propose a model that is able to detect basic emotions with a high degree of accuracy in input text using linguistic tools like tokenization and emotion lexicons. Additional research using rule-based emotion detection can be found in chapter 6.

With a rule-based design, it would be possible to use an approach that uses predefined rules, emotion lexicons and other lexical resources to detect emotions in an input text. Research on rule-based emotion detection has shown that, at least with basic emotions, it is possible to detect emotions with an accuracy of 74% within one rule-based emotion detection approach [7] and 67% in another[21].

To alter the emotional tone, one would be able to trace back the rules and detect which elements contribute to the emotions. These could then be altered using the same rules and lexical resources in order to alter the emotional tone of the text.

A downside to this approach is the fact that, while accuracy in rule-based detection can be as high as 74%[7], it is usually still less accurate than detection that relies on machine-learning models where accuracy as high as 88% is found in simple sentiment analysis[22]. This might be due to the fact that a rule-based approach only allows for detection using predefined variables. It will not be able to discover new contributing elements or discount the contribution of others.

A second downside comes from the fact that rule-based emotion detection approaches rely on lexical resources like emotion lexicons, and emotion word-nets. If there are errors in these resources these will be hard to detect as there is no training happening using real-life data to create verified correlations.

A third downside to this approach is that a rule-based system can only account for the variables that it is given by its designers. It is unable to find potential undiscovered correlations, which may lie outside of the variables provided by the designers. This increases the odds that a rule-based system produces less reliable results than a machine-learned model.

2.3.2 Machine-Learned Approach

In a machine-learned approach, a model is trained using correctly identified data. As it is fed more data, it looks for features that attribute to the labels of the input data. A model can be trained until a high degree of accuracy is achieved. This model can find related factors that no one else has thought of. This has the potential of achieving very high degrees of accuracy. Additional research using machine-learned emotion detection can be found in chapter 6.

Most approaches to emotion detection rely on the use of machine-learned models. As described above, the approaches tend to have higher accuracy than rule-based approaches[22]. They are also able to find additional elements, like additional emotions[23]. While this approach is good for the detection of emotions, it is difficult to use when the goal is to alter the emotional tone of a text. This is due to the so-called black-box phenomenon[24]. It is extremely difficult to extrapolate a fundamental understanding of the underlying mechanism. This makes it difficult when the goal is to trace back the steps to discover what elements contributed to the emotional tone and to what extent. Because this information is not available, we cannot use it to alter the text.

Chapter 3

Research Setup

3.1 Mixed-Model Approach

As can be seen in Figure 2.3.1, in order to alter the emotional tone of a text, we first need to know the current emotional tone. The second step is to make the desired emotional tone adjustments, and the final step is to validate the output. We propose a model based on the strengths and weaknesses of both the rule-based approach and machine-learned design described in sections 2.3.1 and 2.3.2. We call this model the mixed-model approach. The code that accompanies this approach has been made publicly available at `https://github.com/swcloud1/Thesis`.

There are two different kinds of approaches to detect the emotions in a text, each with their own benefits and drawbacks pertaining to their subsequent ability to adjust the emotional tone.

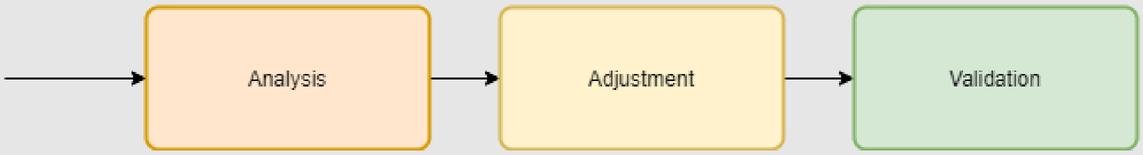


Figure 3.1: The three steps in the Mixed-Model Approach

We propose a combined approach in which the base emotional tone is calculated using both a rule-based approach and a machine-learned software approach. We will then use the rules to attempt to make a change in the emotional tone of the text. Upon completion, we will validate the change. Validation consists of checking whether the sentence is still correct and whether the desired emotional change has occurred using the rule-based analysis method. The model design of the Mixed-Model Approach is explored in-depth in section 3.2.

3.2 Mixed-Model Design

Our final design consist of three main steps: *Analysis*, *Adjustment* and *Validation*[3.1]. Each of these steps is explained in the following sections.

3.2.1 Analysis

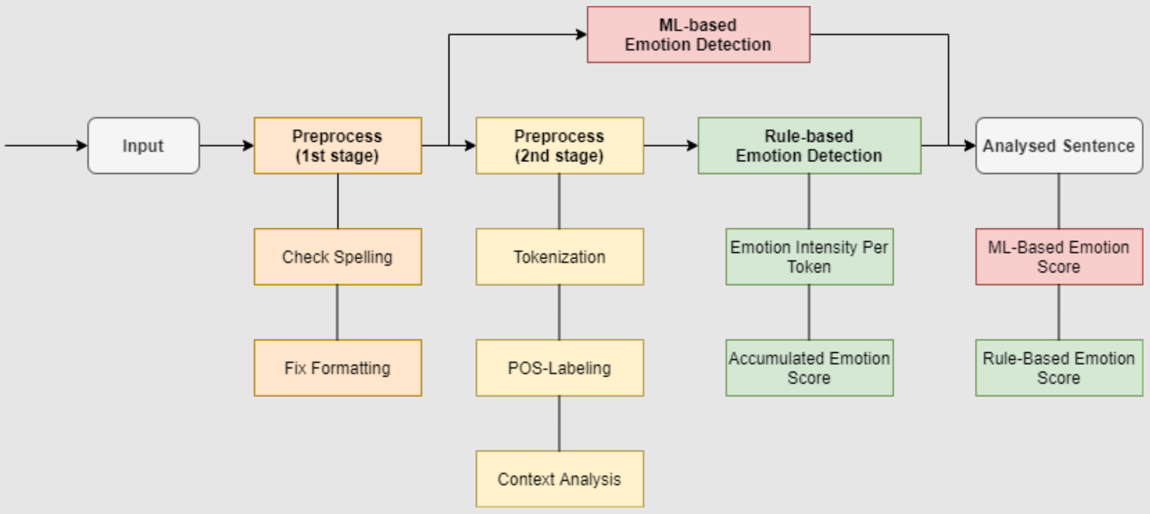


Figure 3.2: Diagram Explaining The Emotion Analysis Flow

During the analysis step the emotional tone of an input sentence is analysed using both rule-based and machine-learned based analysis. In order to do either of these analysis, we first need to prepare the text to be analysed - so-called ”preprocessing” of the text. This preprocessing is done in two stages. In the first stage (see figure 3.2 - 1st stage), we check the spelling of the text and make sure the text is formatted in such a way that both analysis methods can read the text. This is done using automated spell-checking and by removing elements such as emoticons, unknown words and punctuation. An example of this can be found in table 3.1:

Input	*“smh! I don’t have anyything to do, I feel lost ☹, I’m not happy”*
Correct Spelling	*“I don’t have anything to do, I feel lost ☹, I’m not happy”*
Formatting	*“I don’t have anything to do, I feel lost, I’m not happy”*

Table 3.1: The 1st stage of preprocessing where spelling and formatting are checked

Once the first stage of preprocessing has finished, we use our machine-learned approach to detect the emotions. Details of this approach are explained in section 3.4.1. We then move on to the rule-based emotion detection. In order to do this we need a second stage of preprocessing (see figure 3.2 - 2nd stage). This stage consists of natural language processing techniques that attempt to understand the meaning of each word and label it for further analysis. This second stage of preprocessing is mainly done using the Python library Natural Language Toolkit (NLTK), which will be expanded upon in section 3.4.3. The second stage of preprocessing consists of three steps: *Tokenization*, *Part-Of-Speech labeling* and *Context Analysis*. An example of each step can be found in table 3.2:

Input	*“I don’t have anything to do, I feel lost, I’m not happy”*
Tokenization	*[“I”,“don’t”,“have”,”anything”,”to”,”do”,”I”,”feel”,”lost”,”I’m”,”not”,”happy”]*
POS-Tagging	*[(“I”, pronoun), (“don’t”, verb) ....]*
Context Analysis	*[(“Not Happy”, negation)]*

Table 3.2: The 2nd stage of preprocessing where language processing is used to prepare for rule-based analysis

Once we’re finished with the second stage of preprocessing we can start executing the steps for the rule-based emotion detection (see figure 3.2 - Rule-based Emotion Detection). We do this by first analysing the emotion intensity per token. We do this using the NRC Emotion Intensity Lexicon[25], which is publicly available emotion lexicon resource which will be described more in section 3.3.4. This gives us an emotion alongside an intensity for that emotion. An example of which can be found in table 3.3:

Input	*“I am scared”*
Emotions	*[(“I”, none), (“am”, none), (”scared”, fear:0.734)]*

Table 3.3: An example of the rule-based emotion detection output

We now know the emotion for each token in the preprocessed input. We sum the emotional score for each token in order to get an accumulated score for each emotion (see table 3.4).

Input	*“I am scared”*
Accumulated Emotion Score	
’anger’	0.0
’disgust’	0.0
’fear’	**0.734**
’joy’	0.0
’sadness’	0.0
’surprise’	0.0

Table 3.4: The total score per emotion for an input sentence after rule-based analysis

We conclude our analysis step with an analyzed sentence that contains a score using our rule-based methodology and our machine-based methodology (see table 3.5)..

Analyzed Sentence	*“I am scared”*
Rule-Based Emotion Score	
’anger’	0.0
’disgust’	0.0
’fear’	0.734
’joy’	0.0
’sadness’	0.0
’surprise’	0.0
Machine-Learning-Based Emotion Score	
’anger’	0.0365
’disgust’	0.0
’fear’	0.450
’joy’	0.0
’sadness’	0.464
’surprise’	0.0

Table 3.5: The final output of the analysis step containing rule-based emotion analysis and ML-based emotion analysis

3.2.2 Adjustment

In the adjustment step, we will attempt to alter the emotional tone of the sentence we have analyzed in the analysis step described in section 3.2.1. There are many possible desired directions to alter the tone of a text; one might want to make a text that is mainly *angry* mainly *surprised*; or want to remove all *happiness* from a text while keeping all other emotions intact; one might want to make a *fearful* text even more *fearful*; or a host of different adjustments or a combination of different adjustments.

While it is possible to account for as many different adjustments as possible, it is worthwhile to define the underlying logic of these adjustments in order to define a smaller set of adjustment types that are the most common types of adjustment that take place in these examples. This will allow us to focus on a relatively small set of adjustments that can be used, possibly in combination, to achieve many different types of desired adjustments.

Based on this conviction we have generated 4 methods of adjustments:

Adjust Main Emotion Intensity

Intensity adjustment can be defined as the intensification or lessening of the most prominent emotion in a text. This type of adjustment involves identifying the most prominent emotion in a text and then lessening or intensifying that emotion. Examples of which are found in table 3.6:

Adjust Main Emotion Intensity → Intensify	
Input	*"I am scared"*
Main Emotion	Fear
Output	***"I am horrified"***

Adjust Main Emotion Intensity → Lessen	
Input	*"I am scared"*
Main Emotion	Fear
Output	***"I am concerned"***

Table 3.6: An example of main emotion intensity adjustment

The chart below[3.3] shows the flow of the adjustment model.

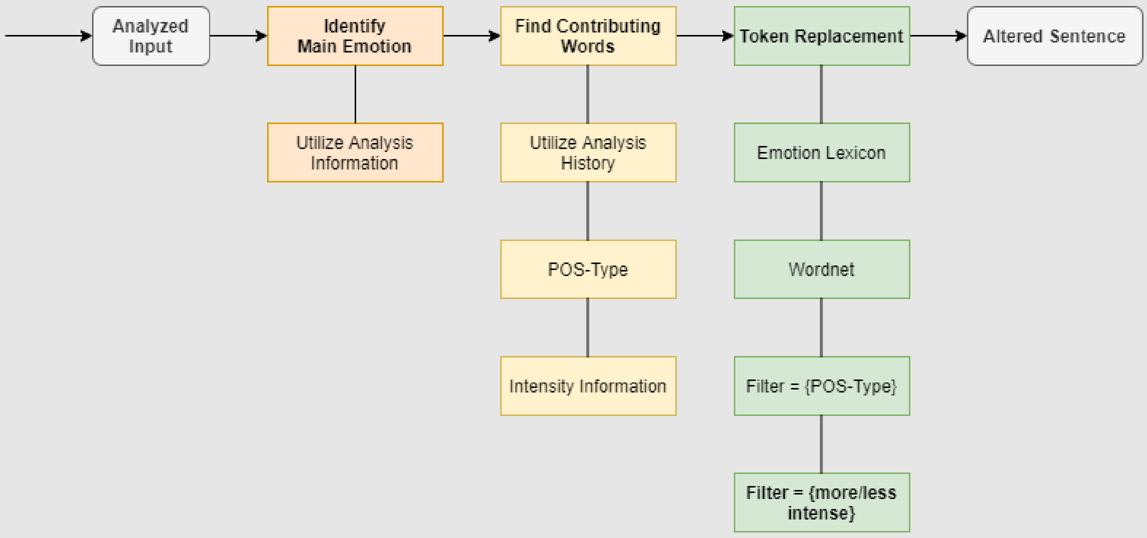


Figure 3.3: Data Flow: Intensity Adjustment Main Emotion

In this chart we start with an analyzed sentence, which means we have meta-data available from the analysis step. We use the output of the analysis to find the most prominent emotion, after which we cycle through the tokens in order to find out to what extend each word contributes to this main emotion. Because we're using an analyzed sentence, we know the part-of-speech(POS)-type and the main emotion intensity of each token. We use this information to replace these tokens by using emotion lexical resources as described in section 3.3.4 to find out what words convey that same emotion. We then analyse these words using natural language processing as described in section 3.4.3 in order to filter them by their POS-type. From these results we then filter by our desired adjustment in intensity, be it less or more intense. We conclude this step with an altered sentence, which we will validate in the final step.

Adjust Specific Emotion Intensity

The process of adjusting the intensity of a specific emotion is similar to that of the main emotion. The key differences are found in the fact that an additional parameter is provided, namely the desired emotion to be adjusted. This also means that the flow changes slightly, as detecting the main emotion in an analysed sentence is not a part, but instead looking for the emotion that is provided as additional parameter. The chart below[3.4] shows the flow of the adjustment model.

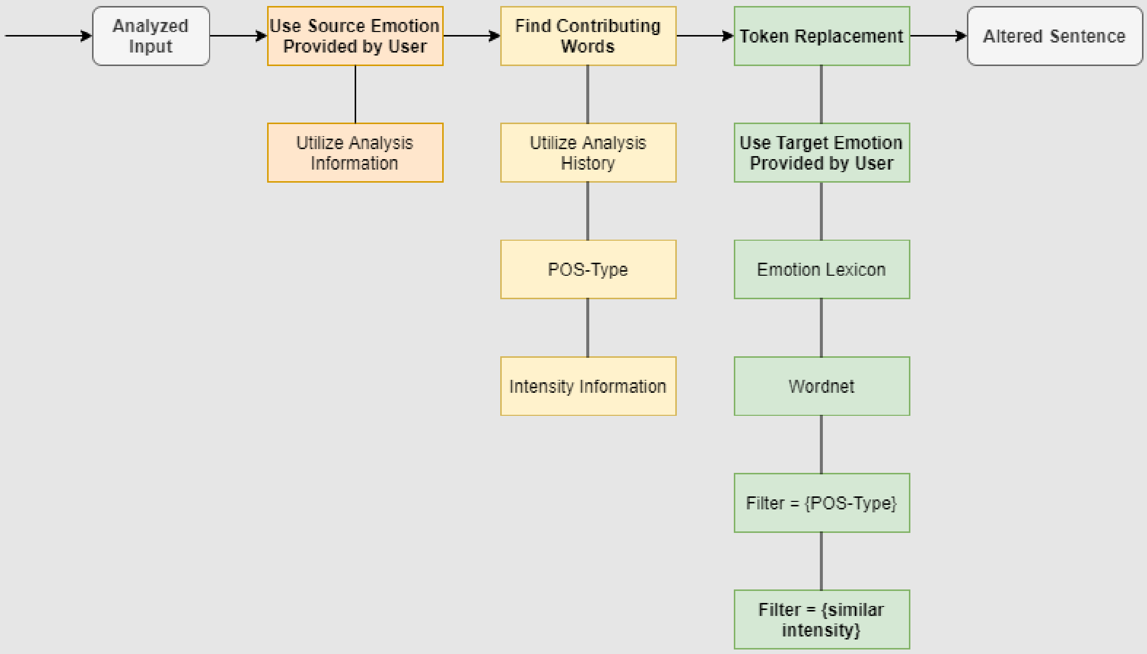


Figure 3.4: Data Flow: Intensity Adjustment Specific Emotion

Emotion Replacement

In the emotion implementation adjustment we provide two additional parameters; the emotion we would like to change and the emotion we would like to change it to. We first identify all the tokens that have the emotion to be replaced. We then need to identify their POS-type and their intensity. For this we use the information provided from the analysis step. Once this is done we move to token replacement, where we use our lexical resources to find words that have the same POS-type and intensity as the source emotion tokens, but are scored as having the target emotion instead of the emotion to be replaced. The flow of this process can be found in the following chart.

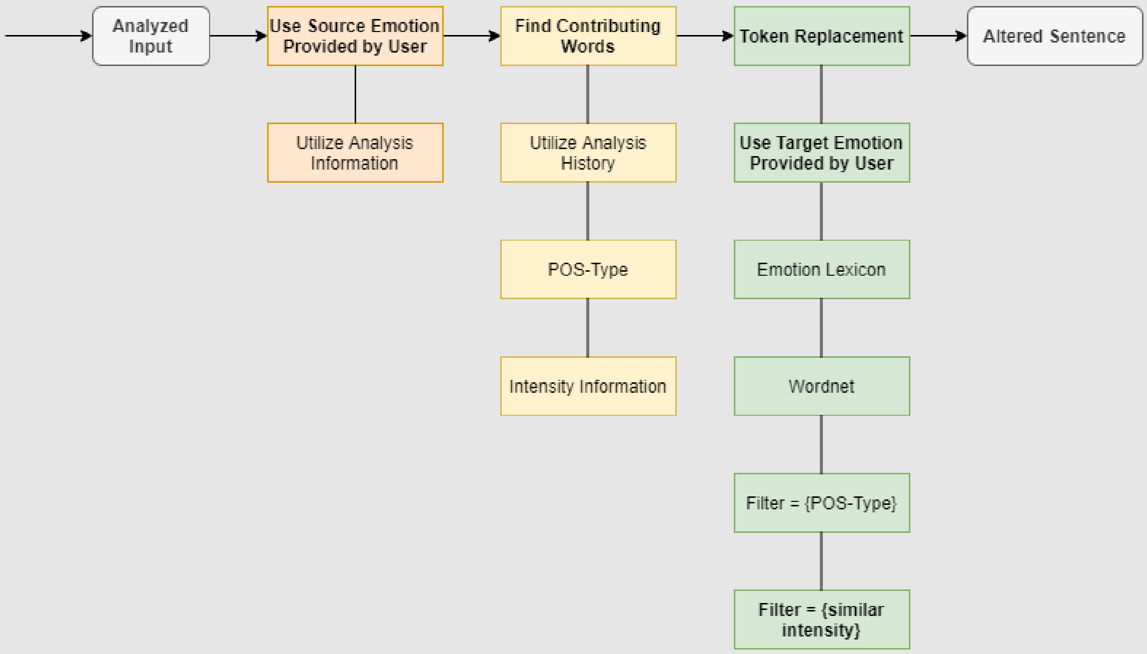


Figure 3.5: Data Flow: Emotion Replacement

Emotion Removal

Our goal during the emotion removal step is to find an emotion provided by a user in an analysed sentence and to neutralize that emotion. The process is similar to that of the previous types of alteration, but differs on the base of aiming to find all contributing tokens to a given emotion and to replace these tokens with a token of the same POS-type, but with a relatively neutral score across all emotions. The NRC Emotion Intensity Lexicon[3.3.4] provides an emotion to each word by default, so we define a neutral word as one that has a very low correlation regardless of the emotion it is attributed to. The flow of emotion removal is described in the chart below.

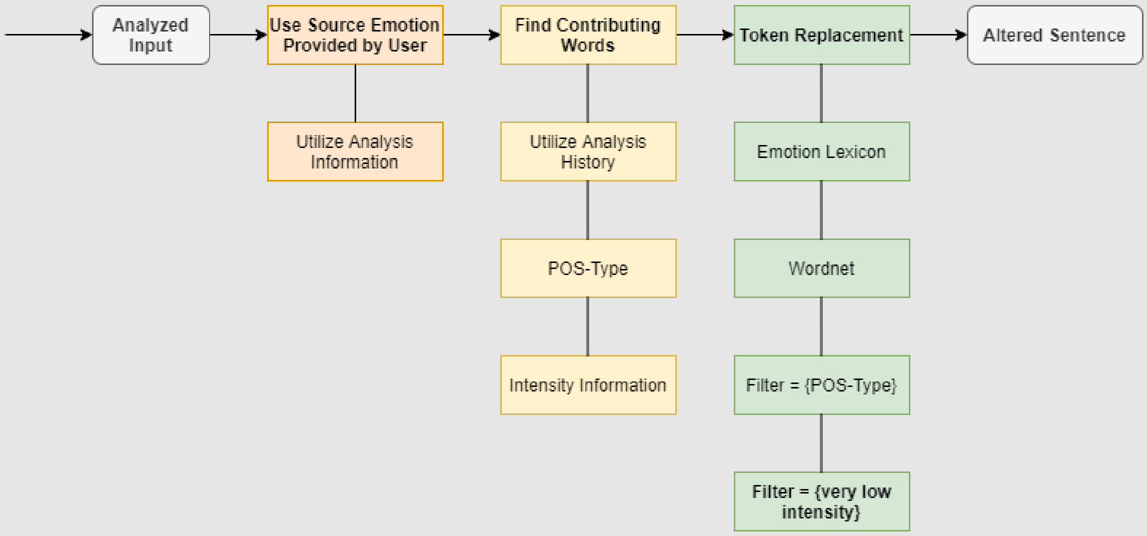


Figure 3.6: Data Flow: Emotion Removal

3.2.3 Validation

After we have analysed and adjusted a sentence we move on to validation. In this step, we first need to make sure our sentence is still a logical one. To do this we make use of automated proofreading, grammar control and additional lexical validation as described in section 3.4.4. Once we have done this and the output is that the sentence is valid we move on the next step in which we perform the entire initial analysis as described in section 3.2.1 in order to get a Rule-Based emotion score. We compare these against the adjustment goal from the adjustment step as described in section 3.2.2 in order to see whether we have achieved our goals. If so, we mark this as successful emotion manipulation.

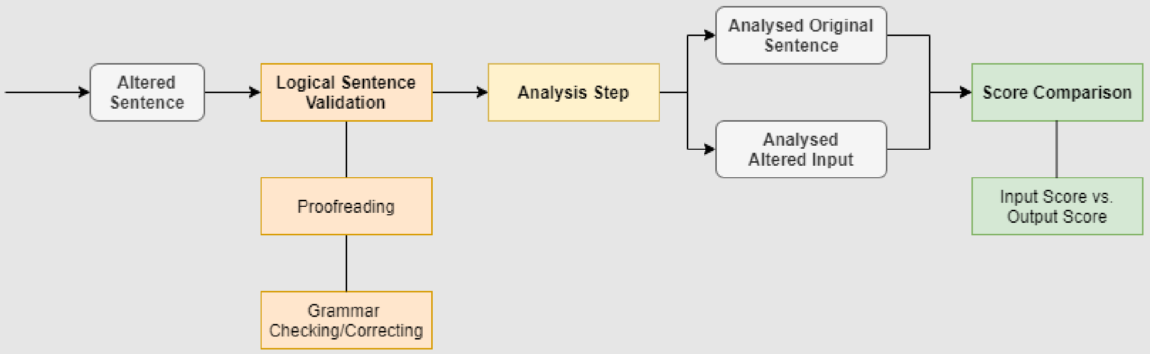

Altered Sentence
Logical Sentence Validation
Proofreading
Grammar Checking/Correcting
Analysis Step
Analysed Original Sentence
Analysed Altered Input
Score Comparison
Input Score vs. Output Score

Figure 3.7: Validation Flow

3.2.4 Adjustment Focuses

All adjustment methodologies require the modification, alteration or removal of existing tokens in the sentence. Each of these methodologies differ to some extent in the requirements they put on these adjusted tokens. One might need to contain a certain emotion while another methodology requests the removal of an emotion. Despite these differences, there is another dimension to these replacements; the extent to which we require the meaning of the word to stay similar to the original word versus the extent to which we want to achieve our goals with regard to achieving a different emotional tone. We believe this choice to be a subjective one to be decided during the practical application of the methodologies as they will be discussed during this research. Because of this, we have defined two types of adjustment focus which will both be used during the research experiments:

Semantic Retention Focus

With this focus, the goal is to achieve the desired adjustment whilst keeping the replacement words as similar in meaning to the original words. This approach makes use of synonym sets and word associations to create a list of acceptable replacement words that are syntactically related to the word that is replaced and has a known emotional association. It also takes into account the grammar, spelling, part-of-speech, requested intensity and requested emotion. This list is created using a combination of NLTK tools: Part-of-speech tagging and WordNet[26] synonym sets and the NRC Emotion Intensity Lexicon[27]. After the list is generated the word with the highest similarity is returned.

The exact methodology of generating these sets relies heavily on nested association. The Python implementation found in 3.8 shows the step by step logic.

```python
def getMostFittingReplacementWord(self, source_word, matching_words):
    matching_similar_words = {}
    source_word_synsets = self.get_wordsets(source_word.word, source_word.pos)
    for source_word_synset in source_word_synsets:
        source_word_synset_vars = self.getLemmas(source_word_synset)
        for source_word_synset_var in source_word_synset_vars:
            source_word_synset_var_synsets = self.get_wordsets(source_word_synset_var, source_word.pos)
            for source_word_synset_var_synset in source_word_synset_var_synsets:
                source_word_synset_var_synset_vars = self.getLemmas(source_word_synset_var_synset)
                for source_word_synset_var_synset_var in source_word_synset_var_synset_vars:
                    for matching_word in matching_words:
                        if source_word_synset_var_synset_var in self.synonyms(matching_word, source_word.pos):
                            similarity = source_word_synset.wup_similarity(source_word_synset_var_synset)
                            if similarity and source_word_synset_var_synset_var != source_word.word:
                                matching_similar_words[matching_word] = similarity
```

Figure 3.8: Python implementation for generating replacement candidates with the semantic retention focus

Emotion Modification Focus

With this focus the goal is to achieve the desired adjustment with as large a change in emotional tone as possible while retaining correct grammar. This focus is less computationally intensive as no similarity sets need to be generated. The list generated by the emotion modification focus does take into account the grammar, spelling, part-of-speech, requested intensity and requested emotion. It then returns the word that produces the desired change with the greatest change in score.

Because of this approach, all words that are produced by the semantic retention focus are contained in the list generated by the emotion modification focus. This will likely result in higher correctness for the emotion modification focus during the experiments.

3.3 Lexical Datasets

In order to adjust the emotional tone of a text, we need to understand to what extent the elements that make up the text contribute to each emotion. One method to do this is to use datasets like an emotion lexicon alongside a sentiment wordnet, like was done in recent research where the goal was to create a rule-based emotion detection system[7]. In order to find the right dataset to use, we will first look at the most commonly used datasets and academic research. We will then discuss their benefits and drawbacks as pertaining to our research goals. We will conclude with our selected dataset alongside our justification.

3.3.1 SentiWordNet

SentiWordNet is described as a "publicly available lexical resource for opinion mining"[28]. It scores a set of most commonly used words using what the authors describe as "PN-polarity", which is the extent to which each word is correlated with a positive opinion and a negative opinion. The scores are gathered using semi-supervised synset classification. The results are produced by a set of eight ternary classifiers. The dataset is available for research goals.

SentiWordnet is one of the most used lexical resources for opinion mining. On Google Scholar alone, SentiWordNet is referenced in over 10,000 results. It has been used as a key resource in many well-cited papers. For example, it was used to analyse sentiment in micro-blogging[29], used to predict the sentiment of movies based on reviews[30] and used to improve word-of-mouth sentiment classification[31].

SentiWordNet, while proven reliable as a lexical resource, only provides polarity based information. If the goal is to detect emotion, we need a reliable way to link the words and their polarization to a set of emotions.

3.3.2 SenticNet

SenticNet is described as a ”semantic resource for sentiment analysis based on conceptual primitives”[32]. Most lexical datasets rely solely on single words as polarized items, affect words and their frequencies. However, sentiment can vary with combinations of words that are commonly grouped together. For example, the word *’accidental’* can often have a strong negative correlation and so does the word *damage* as we can see in the data below, taken from the SenticNet resource, whereas the words when used together as *’accidental damage’* have a less strong impact (see table 3.7):

SenticNet Scores		
accidental	negative	-0.81
damage	negative	-0.97
accidental damage	negative	-0.66

Table 3.7: An example of SenticNet emotion scores

Having a broader labelled lexicon increases the ability to accurately assess the sentiment scores of a text. This approach however, just like the SentiWordNet, only provides us with polarization scores of words, not emotion. In order to do this, we need to be able to link the words to an emotion as well.

3.3.3 NRC Emotion Lexicon (EmoLex)

The National Research Council Emotion Lexicon (called the EmoLex) is a large list of over 24,000 variations of English words, labelled with which of eight emotions and/or two polarisations they convey. The emotions are anger, fear, anticipation, trust, surprise, sadness, joy, and disgust and the polarisation is negative and positive[33]. It was created using Amazon Mechanical Turk Service[34] in 2010 and has been used for several research purposes, for example to identify emotions in social media[35], or as part of a study with the goal of identifying personality[36] and to track emotions in novels and fairy tales[37].

While it is a valuable resource, it can be considered too blunt of an instrument with regard to our research goals as it only labels whether or not a word is associated with a (set of) emotions, not the extent to which it is. An example is the word *’scare’*, which intuitively is mostly linked to fear. According to EmoLex, this word is associated with: fear, anger, anticipation, surprise and negativity without a means to check for variation in the height of the correlations.

What is needed then is a data-set that combined the benefits of both SenticNet/SentiWordnet and EmoLex, one that includes the polarity of each word and information on the emotion that each word is correlated with.

3.3.4 NRC Emotion Intensity Lexicon (NRC-EIL)

The NRC-EIL is the first affect intensity lexicon that not only categorises words with an emotion, but also provides a quantitative score of association[27]. Like EmoLex, it was designed by the same researchers at the National Research Council Canada. It is a very recent lexicon, only released in 2017, but has been used in much academic research since. In 2017, it was used to gather information on the emotion intensities in tweets[38]. In 2019 researchers used it to predict the degree of suicide risk in Reddit posts[39]. In that same year, it was used to check whether emotion cognizance improved fake news detection[40] and to leverage emotional signals to detect credibility[41].

The most recent version, released March 1st 2020, covers nearly 10,000 unique English words and scores them based on eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Each entry consists of a word, emotion and correlation. An example can be seen in table 3.8.

Word	Emotion	Intensity
disliked	anger	0.359
inviting	anticipation	0.578
unhealthy	disgust	0.367

Table 3.8: An example of NRC Emotion Intensity Lexicon scores

The scores were calculated by having 50 to 75 native English speakers rate which of 4 randomized emotions was the most likely match. The words and emotions were presented in different orders and across all English speaking participants. Inter-annotator agreement was used to determine the reliability of the results.

Based on this information this is the best lexical resource to use in this research. It will be used to analyse input text and as a dictionary to get new words from.

3.3.5 Emotions dataset for NLP (emoNLP)

In order to test the accuracy of the rule-based emotion detection and the machine-learned emotion detection systems, a set of input data is required. A requirement for this input data is that we have labelled emotions for each sentence, so we have a label to test our accuracy against. We choose the Emotions dataset for NLP[42]. It contains 2000 sentences that were labelled using CARER: a semi-supervised, graph-based algorithm that is evaluated using various emotion recognition tasks.

3.4 Tools

3.4.1 ParallelDots

ParallelDots is a licensed software product that allows for the detection of emotions using a machine-learned model[43]. They manually labelled a set of words and used Convoluted Neural Networks to implement emotion classification. They provide an API that can be used with a licence. A free licence provides the user with a 1000 API requests per day, which equates to emotional analysis on a 1000 sentences. There is a maximum of 50 requests per minute. Higher tier licences start at $69 per month with 6000 API requests and go up to $399 per month for 2 million API requests. ParallelDots analyses sentences for the following emotions: anger, fear, joy, disgust, sadness and boredom.

This API is used to validate the rule-based emotion detection accuracy. It is also used to check whether a desired adjustment of emotional tone is achieved after transformation.

3.4.2 Rule-Based Emotion Detection

Our rule-based emotion detection is inspired by research done using SenticNet and EmoLex[7]. However, as described in the section 3.3, we will use the NRC Emotion Intensity Lexicon (NRC-EIL). At the time of the rule-based emotion detection paper, the NRC-EIL was not available. It is based off EmoLex, but differs in the added benefit of providing a quantifiable association for the emotion it is most correlated with. This provides more precise information when we try to manipulate the emotional tone.

3.4.3 Natural Language Toolkit

In order to execute language processing techniques like tokenization and POS-tagging, we will use the Natural Language Toolkit (NLTK)[44]; a free-to-use Python package that allows for a more convenient method of language processing. NLTK is also a heavily optimized library, reducing the computational load caused by a custom implementation of the same language processing techniques. It also allows for more time to perform additional experiments, because less time is taken up by the development of these techniques. Initial experiments showed it to be able to accurately turn a sentence into different tokens. It can also identify the parts-of-speech for the basic sentences that are used during the experiments. Furthermore, it incorporates WordNet libraries, which allow us to analyse the similarity of words and get related words to an input word.

3.4.4 Logical Sentence Validation

After a change has been made to a text, we need to validate whether the sentence is still syntactically and grammatically intact. In order to do this, we will use automated proofreading and grammar checking. For this we will use a combination of the grammar functions the NLTK mentioned above as well as LanguageTool, an open-source technology that is free to use for academic research and that allows for contextual spell checking, advanced style checking and intelligent grammar checking[45] that powers widespread software like OpenOffice. Initial experiments showed LanguageTool to correctly identify incorrect use of ”im” instead of ”I am”, providing a message of missing capitalization and incorrect contraction.

Chapter 4

Results

In this chapter, we present the results of our experiments. The chapter is divided into sections reflecting the research questions as stated in the introduction. We start by explaining the steps we followed to validate our rule-based emotion detection system, we will then discuss the results we found using each of the four emotion adjustment types. We conclude by describing our validation findings.

4.1 Rule-Based Emotion Detection Validation

4.1.1 Validation

In order to validate our rule-based emotion detection system, we use a labeled dataset called "Emotions Dataset for NLP" as described in section 3.3.5. By using our rule-based detection approach to check whether we detect the same main emotion as the emotion that is labeled to the sentence, we can derive a percentage of correct emotion detections as a percentage of the total amount of sentences analysed. Finally we compare this score to the percentage of correct emotion detection using an established machine-learned approach. In order to make sure the scoring has been done fairly, only the sentences that are labeled with emotions that are used in the Emotion Dataset for NLP, the NRC-EIL (which the emotion lexicon the rule-based system uses) and the ParallelDots API (machine-learned model) are used. These are the emotions: *anger*, *fear* and *sadness*. The results are show in the table 4.1.

	Total Sentences	**Rule-Based**	**ML-Based**
Anger	275	119 (**43%**)	86 (**31%**)
Fear	212	128 (**60%**)	71 (**33%**)
Joy	704	503 (**71%**)	352 (**50%**)
Sad	550	267 (**49%**)	369 (**67%**)
Overall Correctness	1741	914 (**53%**)	878 (**50%**)

Table 4.1: Results of rule-based and ML-based emotion detection accuracy experiment.

As we can see from the results the correctness of both the custom made rule-based approach and the machine-learned approach are quite similar, with the rule-based approach accurately detecting the emotion in a sentence on average in **53%** of the trials and the machine-learned approach doing so in an average of **50%** of the trials. Using this result, we move on to our next results with the assumption that the rule-based approach is an equally valid method to detect emotion in a sentence as a machine-learned approach, whilst providing us with the possibility to trace back the elements that lead to the analysis, which we do not have with a black-box that is created by the machine-learning approach. Based on the assumption of equal accuracy and greater possibility to analyze, we show the outcomes of the experiments on emotion adjustments in the following sections using the rule-based emotion detection approach.

4.1.2 Generated Dataset

Based on the validation as described, a dataset is formed that will allow us to more accurately measure the accuracy of the following mechanisms. The dataset is formed by taking the sentences that are accurately predicted by both the rule-based approach and the machine-learned approach.

ValidationSet = {sentence — sentence from emoNLP, CorrectRB(sentence), CorrectML(sentence)}

This final validation set contains **526 sentences** that contain the four emotions: anger, fear, sadness and joy. Subsequent results were obtained using this dataset. While keeping sentences that can correctly be identified by both analysis approaches ensures that deviations in output emotional score are not due to deviation in the input emotional scores, it may add a concern for validity. It might be the case that certain properties of language are not accounted for by one of the approaches. This may lead to a validation set that is not representative of the entire language domain. This concern is explored further in Section 5.2.3.

While having a dataset that is as close to the common language domain as possible is always desired, the adjustment types rely primarily on associated emotions and the extent to which they are associated. A more representative dataset may influence detection accuracy, but would not change the mechanisms of adjustment. Based on this conclusion we continue with the assumption that the current dataset meets the requirements for a valid method to test the adjustment types while acknowledging the added value of checking the dataset for representation of the language domain.

4.2 Adjust Main Emotion Intensity

Tables 4.3 and 4.4 show the results for the adjustment of the main emotions. A successful adjustment consists of the following steps:

- Correct identification of the most prominent emotion in an input sentence.
- Adjusting the sentence in such a way that the output sentence scores higher on the emotion that was identified as the most prominent emotion in the input sentence.
- Doing the adjustment while creating no more grammatical errors and spelling errors.

An example result of a successful adjustment, in which the main emotion is more intensive, is shown in table 4.2.

Input Sentence	”im feeling quite sad and sorry for myself but ill snap out of it soon”
Input Emotion Scores	{’anger’: 0.48, ’fear’: 0.949, ’joy’: 0.359, ’sadness’: 1.61}
Most Intense Emotion	’sadness’
Input Grammatical Errors	[’lowercase start’,’incorrect I’m’]
Adjustment Type	**Main emotion more intense**
Output Sentence	”im feeling quite traumatic and sorry for myself but ill snap out of it soon”
Output Emotion Scores	{’anger’: 0.882, ’fear’: 1.793, ’joy’: 0.359, ’sadness’: 1.625}
Output Grammatical Errors	[’lowercase start’,’incorrect I’m’]
Result	More intense main emotion, no change in grammar

Table 4.2: Example of Main Emotion Intensification

4.2.1 Semantic Retention Focus

Table 4.3 contains the results for lessening and increasing the intensity of the main emotion in a text with the focus on achieving the desired goals while keeping the text as semantically close to the original text as possible.

With regards to intensifying the main emotion of a sentence, Table 4.3 shows that **170 (32%)** of the **526** input sentences had a higher score for their main emotion after the adjustment. When making the main emotion less intense, this goal was achieved for **212 (40%)** of the 526 input sentences. The sentences that were identified as **angry (33% and 50%)** and **sad (36% and 50%)** had the highest number of correct adjustments.

	Total Sentences	**More Intense**	**Less Intense**
Anger	48	16 **(33%)**	24 **(50%)**
Fear	30	4 **(20%)**	5 **(17%)**
Joy	283	90 **(32%)**	101 **(36%)**
Sad	165	60 **(36%)**	82 **(50%)**
Overall Correctness	526	170 **(32%)**	212 **(40%)**

Table 4.3: Results of main emotion intensity adjustment experiment (SRF)

4.2.2 Emotion Modification Focus

Table 4.4 contains the results for lessening and increasing the intensity of the main emotion in a text with the focus on achieving the desired emotional adjustment goal with as large a delta as possible.

In both the adjustment focuses to make the main emotion more intense and the adjustment to make the main emotion less intense there was a greater number of correct adjustments than in the Semantic Retention Focus. In the case of making the main emotion more intense, **447 (85%)** of the **526 sentences** are more intense and in the case of making the main emotion less intense this number is **455 (85%)** of the 526 sentences. Just like the Semantic Retention Focus, the sentences that were identified as **angry (88% and 98%)** and **sad (95% and 97%)** had the highest number of correct adjustments.

	Total Sentences	**More Intense**	**Less Intense**
Anger	48	42 **(88%)**	47 **(98%)**
Fear	30	20 **(67%)**	19 **(63%)**
Joy	283	229 **(81%)**	229 **(81%)**
Sad	165	156 **(95%)**	160 **(97%)**
Overall Correctness	526	447 **(85%)**	455 **(85%)**

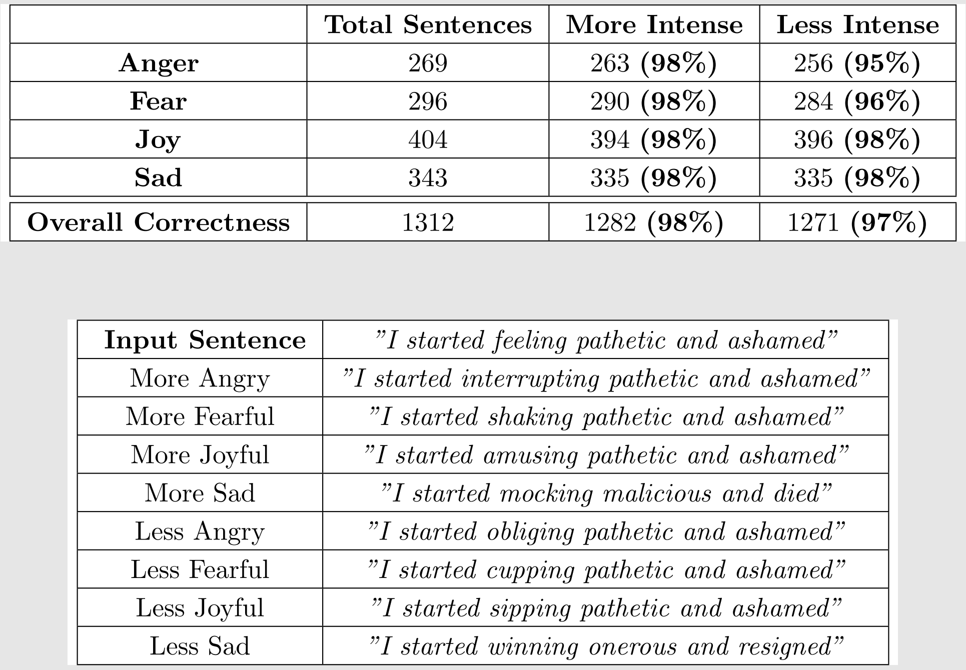
Table 4.4: Results of main emotion intensity adjustment experiment (EMF)

4.3 Adjust Specific Emotion Intensity

In the adjustment of the intensity of a specific emotion we used the same dataset consisting of 526 sentences spanning the four emotion anger, sadness, joy and fear. For each sentence we attempted to increase and decrease the intensity of each of the four emotions. The experiments were done twice using both the Semantic Retention Focus and the Emotion Modification Focus.

4.3.1 Emotion Modification Focus

As we can see in Table 4.5 a more intense version of the sentence was created for all four categories in **98%** of the cases. Where the goal is to lessen the intensity of a specific emotion, this number drops to **95% for anger** and to **96% for fear**. Results for both **joy** and **sadness** had an accuracy of **98%**, similar to the accuracy of the intensification experiment.



	Total Sentences	**More Intense**	**Less Intense**
Anger	269	263 **(98%)**	256 **(95%)**
Fear	296	290 **(98%)**	284 **(96%)**
Joy	404	394 **(98%)**	396 **(98%)**
Sad	343	335 **(98%)**	335 **(98%)**
Overall Correctness	1312	1282 **(98%)**	1271 **(97%)**

Input Sentence	*”I started feeling pathetic and ashamed”*
More Angry	*”I started interrupting pathetic and ashamed”*
More Fearful	*”I started shaking pathetic and ashamed”*
More Joyful	*”I started amusing pathetic and ashamed”*
More Sad	*”I started mocking malicious and died”*
Less Angry	*”I started obliging pathetic and ashamed”*
Less Fearful	*”I started cupping pathetic and ashamed”*
Less Joyful	*”I started sipping pathetic and ashamed”*
Less Sad	*”I started winning onerous and resigned”*

Table 4.5: Results of specific emotion intensity adjustment experiment (EMF)

4.3.2 Semantic Retention Focus

Table 4.6 shows the results for specific emotion intensity adjustment, while keeping the sentence semantically as similar to its original as possible. Similarly to the results of adjusting the main emotion intensity, the scores are lower with the Semantic Retention Focus on both the intensification **(73%)** and lessening **(58%)** than the emotion detection focus. This is expected, as the Semantic Retention Focus does not return a replacement if it does not find a relation to the word it is replacing.

	Total Sentences	**More Intense**	**Less Intense**
Anger	269	210 (**78%**)	91 (**34%**)
Fear	296	229 (**77%**)	226 (**76%**)
Joy	404	277 (**69%**)	295 (**73%**)
Sad	1312	242 (**71%**)	150 (**44%**)
Overall Correctness	1312	958 (**73%**)	762 (**58%**)

Input Sentence	*”I started feeling pathetic and ashamed”*
More Angry	*”I started uprising pathetic and ashamed”*
More Fearful	*”I started suffering pathetic and ashamed”*
More Joyful	*”I started giving pathetic and ashamed”*
More Sad	*”I started suffering miserable and ashamed”*
Less Angry	*”I started feeling pathetic and ashamed **(no change)**”*
Less Fearful	*”I started swelling pathetic and ashamed”*
Less Joyful	*”I started receiving pathetic and ashamed”*
Less Sad	*”I started feeling nonsensical and ashamed”*

Table 4.6: Results of specific emotion intensity adjustment experiment (SRF)

4.4 Emotion Replacement

Tables 4.7 and 4.8 show the results for emotion replacement. In this experiment for each of the **526 sentences** each emotion, if present, was attempted to be changed to each of the other emotions. For both adjustment focuses a total of **3936 adjustments** were attempted. The tables show the total number of trials for each input emotion alongside the number of correct adjustment for each target emotion. For both focuses, we display an example result. These results were hand-picked based on their ability to accurately display the attempted changes. The assessment was done using a combination of experiment results and subjective judgement.

4.4.1 Emotion Modification Focus

With the Emotion Modification Focus the focus is on achieving the desired emotional adjustment goal with a large as delta as possible. The results in Table 4.7 show that in **3523 (90%)** of the **3936 attempted changes**, the source emotion was replaced with the target emotion. Sentences containing the emotion **joy, with an accuracy of 98%**, were able to replace joy with a different emotion most often. Sentences containing **anger, with an accuracy of 81%**, were able to replace anger with a different emotion least often.

		Target Emotion				
Source Emotion	**Total**	**Anger**	**Fear**	**Joy**	**Sadness**	**Correct**
Anger	807		211	230	212	653 **(81%)**
Fear	888	273		259	236	768 **(86%)**
Joy	1212	403	390		399	1192 **(98%)**
Sad	1029	312	284	314		910 **(88%)**
Total Trials	3936					3523 **(90%)**

Input Sentence	*”I feel inadequate because it prompts comparison”*
Replace sadness with anger	*”I feel veryangry because it prompts comparison”*
Replace sadness with fear	*”I feel stranger because it prompts comparison”*
Replace sadness with joy	*”I feel celebrations because it prompts comparison”*

Table 4.7: Results of emotion replacement experiment (EMF)

4.4.2 Semantic Retention Focus

With the Semantic Retention Focus, the goal is to keep the meaning of the sentence as close to the origin as possible while simultaneously achieving the desired adjustment goal. The results in Table 4.8 show that in **1379 (35%)** of the **3936 attempted changes**, the source emotion was replaced with the target emotion. Sentences containing the emotion **anger, with an accuracy of 48%**, were able to replace anger with a different emotion most often. Sentences containing **joy, with an accuracy of 30%**, were able to replace joy with a different emotion least often.

		Target Emotion				
Source Emotion	**Total**	**Anger**	**Fear**	**Joy**	**Sadness**	**Correct**
Anger	807	0	79	154	157	390 **(48%)**
Fear	888	61	0	170	50	281 **(32%)**
Joy	1212	141	76	0	141	358 **(30%)**
Sad	1029	147	72	131	0	350 **(34%)**
Total Trials	3936					1379 **(35%)**

Input Sentence	*”I feel a lil bit gloomy”*
Replace sadness with anger	*”I feel a lil bit forbidding”*
Replace sadness with fear	*”I feel a lil bit ghastly”*
Replace sadness with joy	*”I i feel a lil bit gentle”*

Table 4.8: Results of emotion replacement experiment (SRF)

4.5 Emotion Removal

Tables 4.9 and 4.10 show the outcomes of the removal of emotions from sentences. Below each table is a sample result as a visual reference to the output of each adjustment. Because the lexical database is set up in a way that every word is associated with an emotion to some extent, we consider an emotion to be removed when its correlation to an emotion is below a threshold value. In order to determine the threshold value, we looked at the correlation which has around **90%** of all possible words above it. This way we capture the **10%** of the most neutral words in the lexicon. In this case, this required a threshold association value of **0.3**, with correlations ranging between **0 and 1.0**.

4.5.1 Emotion Modification Focus

With the Emotion Modification Focus the focus is on achieving the desired emotional adjustment goal with as large a delta as is possible. The results in Table 4.9 show that in **706 (73%)** of the **970** cases the selected emotion was successfully removed from a sentence. Sadness, with an accuracy of **135 (64%)** of **211 sentences**, was the most difficult emotion to remove. Anger and fear, both with an accuracy of **79%**, had the highest accuracy of removal.

Source Emotion	**Total**	**Correct**
Anger	114	90 **(79%)**
Fear	256	203 **(79%)**
Joy	389	278 **(71%)**
Sad	211	135 **(64%)**
Total Trials	970	706 **(73%)**

Input Sentence	*”I feel incredibly lucky just to be able to talk to her”*
Removed Joy	*”I feel incredibly unhappy just to be able to talk to her”*

Table 4.9: Results of emotion removal experiment (EMF)

4.5.2 Semantic Retention Focus

With the Semantic Retention Focus, the goal is to keep the meaning of the sentence as close to the original as possible while simultaneously achieving the desired adjustment goal. The results in Table 4.10 show that in **159 (16%)** of the **970 cases** the selected emotion was successfully removed from a sentence. Similarly to the Emotion Modification Focus, **sadness**, with an accuracy of **26 (12%)** of **211 sentences**, was the most difficult emotion to remove. **Anger**, with an accuracy of **28 (25%)** out of a **114 sentences**, had the highest accuracy of removal.

Source Emotion	**Total**	**Correct**
Anger	114	28 **(25%)**
Fear	256	56 **(22%)**
Joy	389	49 **(13%)**
Sad	211	26 **(12%)**
Total Trials	970	159 **(16%)**

Input Sentence	*”I feel pathetic for lying if i say no”*
Removed Anger	*”I feel pathetic for dwelling if i say no”*
Removed Sadness	*”I feel ridiculous for lying if i say no”*

Table 4.10: Results of emotion removal experiment (SRF)

4.6 Comparing Adjustment Types

We took the overall correctness result for each experiment in order to compare the correctness of each adjustment type with the other adjustment types. We grouped together the results in tables based on the adjustment focus. As hypothesized in section 3.2.4 the correctness for the Emotion Modification Focus **(88%)** is higher than the correctness for semantic modification focus **(42%)** for all adjustment types.

The average was calculated by adding the accuracy of each experiment and dividing this by the total number of adjustments. The number of trials were not used to weigh the scores, because they would disproportionately skew the results. For example, the emotion replacement adjustment attempts to replace each found emotion in a sentence with all the other emotions in the set of included emotions. This increases the number of trials by an order of magnitude, while only analysing a single sentence.

4.6.1 Emotion Modification Focus

With the Emotion Modification Focus the focus is on achieving the desired emotional adjustment goal with as large a delta as is possible. On average, the Emotion Modification Focus has an accuracy of **(88%)**. The highest correctness **(98%)** is found in the adjustment in which the main emotion is intensified. The lowest accuracy is found in the removal of emotions **(73%)**. Therefore, the overall correctness of emotion adjust with Emotion Modification Focus lies between the range of **(73%)** and **(98%)** with an average of **(88%)**.

Adjustment	**Trials**	**Correct**	**Accuracy**
More Intense Main	526	447	**(85%)**
Less Intense Main	526	455	**(85%)**
More Intense Specific	**1312**	**1282**	**(98%)**
Less Intense Specific	1312	1271	**(97%)**
Emotion Replacement	3936	3523	**(90%)**
Emotion Removal	**970**	**706**	**(73%)**
Average			(88%)

Table 4.11: Compared results of all emotion adjustment types (EMF)

4.6.2 Semantic Retention Focus

With the Semantic Retention Focus, the goal is to keep the meaning of the sentence as close to the origin as possible while simultaneously achieving the desired adjustment goal. On average, the Semantic Retention Focus has an accuracy of **(42%)**. As with the Emotion Modification Focus, the highest correctness **(73%)** is found in the adjustment in which the main emotion is intensified. Also similarly to the Emotion Modification Focus, the lowest accuracy is found in the removal of emotions **(16%)**. Therefore the overall correctness of emotion adjust with Semantic Retention Focus lies between the range of **(16%)** and **(73%)** with an average of **(42%)**. The interval with which the results lie in is more than twice as large in the Semantic Retention Focused results **(54%)** than in the Emotion Modification Focused results **(24%)**.

Adjustment	**Trials**	**Correct**	**Accuracy**
More Intense Main	526	170	**(32%)**
Less Intense Main	526	212	**(40%)**
More Intense Specific	**1312**	**958**	**(73%)**
Less Intense Specific	1312	762	**(58%)**
Emotion Replacement	3936	1379	**(35%)**
Emotion Removal	**970**	**159**	**(16%)**
Average			**(42%)**

Table 4.12: Compared results of all emotion adjustment types (SRF)

Chapter 5

Discussion

In this chapter, we discuss the results of the experiments on the emotion detection system, emotion adjustment systems and validation systems. We explore these results within the context of the research question they aim to answer. We conclude by discussing some threats to the validity of the research.

5.1 Answers to Research Questions

5.1.1 RQ1: How can we analyse text in such a way that it can provide the emotional tone of the text?

In this study, we proposed the rule-based approach described in section 3.2.1 to analyse the emotional tone of texts. In order to do this, we relied on emotion intensity lexicons, part-of-speech tagging, grammar- and spellchecking, tokenization and similarity analysis. With the emotion detection mechanism, we were able to analyse the extent to which four emotions (anger, fear, joy and sadness) were present in an input text. In our experiment, we used a publicly available dataset to test the accuracy of our detection mechanism and compared this with a leading commercially available machine-learned emotion detection software system.

We hypothesised that the machine-learned approach would outperform our rule-based emotion detection approach by a small margin. Our results show a 6% overall improvement in accuracy over the machine-learned control. While we did not expect the rule-based approach to score higher, we do believe this relatively small difference still lies within a margin of error and that a different dataset might yield results in the opposite direction. We have also found the overall accuracy of both the rule-based and machine-learned approach to be relatively low (53% and 50% respectively). As stated in section 2.3.1, accuracy for competing rule-based emotion detection mechanisms can be found as high as 74%. Which may be proof that this analysis methodology is the best fit for emotion detection.

However, in order to further investigate the cause of this relatively low accuracy, there are two possible variables to be considered; the accuracy of the dataset and the accuracy of the machine-learned control, both of which are explored more in-depth in section 5.2. By varying both of these factors, we should be able to detect whether the cause of the lower score is the rule-based approach, the machine-learned approach, the dataset or a combination of these.

5.1.2 RQ2: How can we provide suggestions on changing the current emotional tone using text recommendations?

In order to answer this question we first had to define change within the context of emotional tone adjustment. We looked at multiple examples of altered sentences and compared them with the intent of finding out which alterations formed the building blocks that, when used (in combination), lead to most of the changes in the altered sentences. Based on our findings, we have defined four types of adjustments: adjusting the main emotion intensity, adjusting the intensity of specific emotions, replacing one emotion with another and removing an emotion entirely.

In our experiments, we used a set of labelled sentences that were known to be correctly identified by both the rule-based and machine-learned approach. We then made the adjustments and analyzed the sentences again to see whether we'd achieved our goal. We perform each adjustment according to two adjustment focuses, the rationale of which is explained more in depth in section 5.1.3. They consist of a semantic retention focus in which we aim to keep the output similar in meaning to the input; and an emotion modification focus in which we aim to achieve the desired change in emotion with as great a delta as possible.

Results of the experiments show that for all adjustment types, valid adjustments can be made. We find that for all adjustment types, the emotion modification focus greatly outperforms the semantic retention focus. On average, it produces a correct output 2-4 times as often. A large difference between the two approaches is expected however, as the semantic retention focus is more conservative by design. This means it will only make an adjustment if the words that are being adjusted are sufficiently similar to their replacements.

There are limitations; for one, the system as-is does not implement affective intensity modifiers. This means the text *"very sad"* has a similar emotional tone as *"sad"*. Additionally, the current system does not account for negation, which means *"not sad"* and *"sad"* are detected as having a similar emotional tone. However, we believe this does not impact the reliability of the adjustment, but rather impacts the validity of the emotional analysis because the adjustments do not rely on the emotion-type, but rather on changing the output score relative to the input score.

5.1.3 RQ3: How do we validate accuracy of emotional tone changes?

In order to validate the accuracy of emotional tone changes, we assessed the requirements for sentence validity. Programmatically we found that in order to check whether a sentence is valid we would need to make sure the grammar is correct and that there are no errors in spelling. However, we found that a sentence can be grammatically correct, contain no spelling errors, but still be considered illogical by a human reader.

Most of these illogical results are due to the fact that words do not seem to fit the context of the sentence. A possible solution to this is to implement semantic analysis on each word to make sure the replacement word is similar to the original word. However, if the goal is to change the emotions in a sentence, this limits the extent to which we can deviate. This can lead to possible fitting words being omitted from the results.

The question then becomes: to what extent are we willing to lose similarity in order to achieve our adjustment goals? We believe this answer to be a subjective one linked to the intended practical use, instead of a decision to be made by us. Based on this we decided to perform each experiment with two focuses that represent both extremes on this matter: the focus on retaining as much of the semantics as possible (called the semantic retention focus) and the focus on achieving the adjustment goals with as large a delta as possible (called the emotion modification focus). By performing each experiment with both focuses, we feel we can accurately represent the upper- and lower-bounds of the correctness of our adjustment methodologies.

5.2 Threats to validity

5.2.1 Emotion detection validation dataset

In order to test the accuracy of our rule-based emotion detection system, we collected a dataset repository from `kaggle.com`. According to its publishers, this dataset was created and verified using multiple machine-trained models (some of which were trained using human-evaluation). However, the eventual dataset was not verified by human-readers, which may have affected the accuracy of the results. This makes it harder to determine the absolute accuracy of the rule-based approach. By finding or creating a dataset of sentences that are each labelled for their main emotion by human readers, we should be able to account for this threat to validity.

5.2.2 Machine-learned emotion detection validity

During the validation of the rule-based emotion detection (as described in section 4.1), we only compared the sentences that were labelled with emotions that are found in both the rule-based approach and machine-learned approach. However, the machine-learned approach checks for two additional emotions (boredom and excitement). Whenever results that primarily contain these emotions were found, we took the highest value of the set of emotions that are shared across all analysis methodologies and datasets (anger, joy, fear and sadness). This may have negatively impacted the reliability of the results; if there had been no check for the two additional emotions, the distribution amongst the shared four would have been different, thus providing different results.

A way to account for this risk would be to use approaches that check for the exact same set of emotions, but during this research, no such publicly available resource was found. It would be possible to build a custom one, using different datasets as training data for a machine-learned model. A machine-learned model that uses the same emotions as the rule-based model might also be available in the future.

5.2.3 Representation of language domain by validation dataset

The dataset that was used as the input for the experiments was generated by gathering sentences from the emoNLP dataset (described in: 3.3.5) that were correctly identified by both the rule-based and machine-learned approach. While this assured us of a reliable baseline for validation, it did raise a concern: It could be possible they were correctly identified by both approaches due to the fact that these sentences lacked linguistic properties that are present in common language but are not detected by either approach. If this is the case, then the dataset might not be an accurate representation of the language domain.

5.2.4 Emotion intensity modifiers and negation handling

As discussed in section 5.1.2, we did not implement detection for emotion intensity modifiers and negation. This means that a sentence containing *"not really happy"* and *"happy"* are analysed as being equally joyful. This affects the validity, because a sentence can have a deviating distribution across emotions if it was to be accounted for. However, we do believe that the mechanisms for adjustment and the framework for analysis and validation are not affected by the validity of emotion detection output; these mechanisms solely rely on altering a tone relative to the tone of an input. Because of this, the findings are not less reliable but do contain aspects that should be investigated further by future iterations to improve validity.

5.2.5 Lexical Resource Accuracy

In order to provide recommendations to the adjustments of emotions, we required lexical resources that provide us with words and the extent to which they relate to a given emotion. The resource we ended up using is the NRC Emotion Intensity Lexicon[27], which has been built on top of existing lexicons that have been used in various academic papers (as described in section 3.3.4). We believe in the validity of the proposed emotions and the levels of their association. However, the data was gathered using social media posts and was not tested for existence in the English dictionary. This means that phrases like *"veryangry"* and *"sohappy"* made it into the lexicon. While most human readers might understand these phrases, some of the text analysis processes like part-of-speech tagging, grammar-checking and spellchecking have difficulty in processing these technically incorrect words. A lexicon that has been filtered using dictionary words and their grammatical conjugations might lead to more valid results.

Lastly, the lexicon listed an emotion and their corresponding level of association for each word. This meant that we would always - to some extent - introduce new emotions when removing emotions from a sentence through replacement. We worked through this limitation by setting a threshold score. If a word scored under this threshold for each emotion, the word was considered neutral. However, the lexicon was created with the intent to detect emotions, so it might have excluded words that were considered completely neutral. Therefore there may have been more neutral words available, which would have increased the accuracy of the emotion removal mechanism.

5.2.6 Programmatic validation vs. human validation

In order to validate our results, we used a combination of similar software solutions and programmatic validation tools. We believe these decisions to have provided reliable results. However, because of the complex nature of language and emotions and the limitations of current software solutions, we believe that human-validation would increase the validity of the results. It may be the case that some sentences would have been labelled as mainly conveying a certain emotion, while human evaluation might label another. Because text and emotion are mainly created by humans for humans, we feel there is more merit to human evaluation than programmatic evaluation.

Chapter 6

Related work

In this chapter, we present studies related to our research. We have divided the related work into two categories: sentiment analysis and sentiment manipulation. Almost all related work on sentiment manipulation contains sections on sentiment analysis, so in order to differentiate we divide by the main focus of each work. A table containing an overview of the works that will be discussed can be found below. It shows whether the paper-covered sentiment analysis(SA), sentiment manipulation(SM), polarity(P) and/or emotions(E).

	SA	**SM**	**P**	**E**	**Title**
[7]	x			x	Sentence-level emotion detection framework using rule-based classification
[46]		x	x		A syntax-aware approach for unsupervised text style transfer
[47]	x	x	x	x	Delete, retrieve, generate: A simple approach to sentiment and style transfer
[48]	x	x	x		Fighting offensive language on social media with unsupervised text style transfer
[49]	x			x	Learning to identify emotions in text
[50]	x	x	x		UPAR7: A knowledge-based system for headline sentiment tagging

Table 6.1: Overview of related works with regards to sentiments analysis(SA) and sentiment manipulation(SM). Sorted by polarity-based(P) and emotion-based(E)

6.1 Sentiment Analysis

6.1.1 Rule-based Emotion Detection Framework

A study by Asghar et al.[7] proposes a solution for the limited coverage of emotions by current supervised and unsupervised methods. They describe a rule-based framework to detect emotion-based sentiment. This framework detects negation, intensity shifters, emoticons and emotions using various lexical resources like SentiWordNet[28], SenticNet[32] and NRC Emotion Association Lexicon[33]. Their framework was tested against competing emotion detection methodologies and showed greater accuracy in correctly detecting the emotional tone of the text.

The emotion detection system as used in our system is inspired by their framework. However, there are two major differences. Firstly, in order to make any adjustments to the emotional tone of the text we need to be able to remember and trace back what elements in an input sentence contributed to the overall score and to what extent. In the framework proposed by Asghar et al., the score is calculated by performing additions, leaving no trace for further analysis. Secondly, The framework proposed in the paper relies on the use of the NRC emotion Association Lexicon combined with a polarized wordnet like SenticNet and SentiWordNet. The combination of these resources connects one or multiple emotions to a polarity score, but does not provide information on the extent to which they do so. This makes it hard to determine whether a word is more strongly related to one emotion than another if it correlated to multiple emotions. The creator of the NRC Emotion Association Lexicon created a new lexicon based on this finding, named the NRC Emotion Intensity Lexicon[27]. This lexicon powers the emotion detection framework used in our research.

6.1.2 Learning to Identify Emotions in Text

In this paper Strapparava et al.[49] performs experiments regarding the automatic detection of emotions in news headlines. They looked at six basic emotions: anger, disgust, fear, joy, sadness and surprise. They used different algorithms of varying complexity on a custom-made dataset in order to see what the effect on the accuracy of emotion detection was. Their results suggest it is possible to accurately detect emotion in text using both supervised and unsupervised methods.

Our research also uses different methodologies in order to evaluate the accuracy of emotion detection. Similarly to the paper by Strapparava et al., a dataset with labelled emotions is used to assess the accuracy of the custom emotion detection approach which the rest of our research questions are tackled with. The concept of varying complexity algorithms is also used in our research, but is applied to our emotion adjustment, instead of the detection. We do believe that there is value in some of the complex emotion detection approaches proposed by Strapparava et al., and want to see if this value can be translated to emotion manipulation. Lastly, four of the six emotions that form the basis of this paper (anger, joy, fear, sadness) return in our research as they are the most common emotions used across various detection methods, lexical resources and validation datasets.

6.1.3 University Paris 7 (UPAR7)

Research from the University of Paris by François-Régis Chaumartin[50] on the detection of emotions in news headlines showed that by using a combination of Part-of-Speech tagging and both polarity and affect lexicons they are able to accurately predict the emotional tone of a short sentence. Their main objective was to show that it is possible to accurately detect emotions in a short text, as long as the appropriate combination of resources is leveraged.

We used some of the same techniques in our detection. We relied heavily on part-of-speech tagging, language processing and lexical resources in order to detect the emotional tone. However, we also apply these mechanisms to our adjustment system. Thus, besides attempting the confirm the value of these processing mechanisms with the intent to analyse, we also attempt to prove this in the context of modification. We also expand this set of mechanisms with spell-checking, grammar-analysis and similarity detection.

6.2 Sentiment Manipulation

6.2.1 Syntax-Aware Unsupervised Text Style Transfer

In a 2019 paper by Ma et al. [46] the researchers attempt to change the style of a text, while preserving the contextual information. They aim to do this not just by using word affect, but also syntactic knowledge about the words. Their model is referred to as the Syntax-Aware Style Transfer (SAST) model. Their research shows that their model is effective in changing the emotional tone. Some example results can be found in Table 6.2:

Change	**negative → positive**
Input	*”i just received a delivery order from them and essentially wasted my money.”*
Output	***i just received a delivery order from them and essentially love this place!***

Change	**formal → informal**
Input	*”that is if you truly adore them.”*
Output	***that is if u truly luv them***

Table 6.2: Example result of SAST model. Transferring the style from negative to positive and from formal to informal

While the SAST model shows promising results, it is only capable of polarity based style transfer. They've included variations in positivity and formality. Our research aims to take this style of adjustment to a set of emotions and see whether it can effectively identify and modify these emotions in an input text.

6.2.2 A Simple Approach to Sentiment and Style Transfer

Li et al[47] proposed an approach to sentiment style transfer by gathering attributes from sentences in an unsupervised way. According to their results, which are based on human evaluation, their model is 22% more accurate in transferring style than the last best model. They used three different datasets: reviews from Yelp, reviews from Amazon and a set of image captions. For both the Yelp and Amazon datasets, they altered the sentiment between negative and positive, a polarity based approach. With the image caption dataset, they altered the captions to be more humorous or romantic. They demonstrated that in principle, style transfer is possible using emotional tones. Our research builds on this finding, by incorporating mechanisms and tools with the goals of researching whether this finding can be applied to sets of emotions with style transfers between each.

6.2.3 Fighting Offensive Language with Unsupervised Text Style Transfer

In a study by Santos et al.[48], a new method for text style transfer is proposed. This is done by presenting a new method of training encoder-decoders and by focusing on preserving the content while transferring the style. They test their new method on datasets from Twitter and Reddit using three metrics: classification accuracy, content preservation an perplexity. The results show that their system outperforms competing solutions on accuracy and content preservation. An example results can be seen in Table 6.3:

Input	*”for f**k sake , first world problems are the worst”*
Output	***”for hell sake , first world problems are the worst”***

Table 6.3: Example result of offensive to non-offensive text style transfer

Like research by Ma et al.[46] and Li et al[47], Santos et al. provide evidence for accurate text style transfer. In our research, we want to find out whether these findings hold for emotion adjustment with the intent of expanding our current knowledge on the possibilities with text style transfer.

Chapter 7

Conclusion

In this thesis, we explored the topic of emotion-based text style transfer. We wanted to know how we could analyse text in such a way that we can assess its emotional tone. We also wanted to how we can provide suggestions on changing the emotional tone of a text. Lastly, we wanted to know how we could validate the accuracy of emotional tone changes.

We proposed a rule-based emotion detection system that would allow us to trace back the analyses and provide us with insights into the adjustments that could be made in order to alter the emotional tone. These changes could then be validated using a combination of grammar-checking, spellchecking and emotion analyses.

Our rule-based emotion detection mechanism was built using a publicly available emotion intensity lexicon. When compared with ParallelDots, a commercially available machine-learned emotion detection model, on a dataset of sentences with labelled emotions, it achieved an accuracy of 53%, while the machine-learned emotion detection model achieved an accuracy of 50%. We believe based of these results that the rule-based emotion detection as described in the paper is at least equally as accurate as the competing machine-learned model at detecting emotions in text, whilst providing us with the possibility to trace back the elements that lead to the analysis so that we can use this information to adjust the emotional tone. This possibility is not present in a machine-learned model due to the fact that with the machine-learned model a black-box is created. However, We did find that with an absolute accuracy of 53%, there is progress to be made in the rule-based emotion detection approach. This might be achieved by implementing additional detections for language markers like: negation, intensifiers and context.

We defined four types of emotional tone adjustments: adjusting the main emotion intensity, adjusting a specific emotion intensity, replacing an emotion and removing an emotion. We ran experiments on all four or these, splitting each experiment by focus on keeping the meaning of a sentence intact versus achieving the change in emotional tone with as large a delta as possible. In all of our experiments focusing on keeping the meaning of a sentence intact resulted in lower accuracy. On average, with the semantic retention focus, we were able to change the emotional tone of a sentence 39% of the time. With the emotional tone adjustment focus, this accuracy was on average 72%. Based on this result, we believe that using the proposed adjustment methodologies can work when the goal is to adjust the emotional tone of a sentence. However, there are different factors to take into consideration when adjusting the emotional tone. For one, as the focus shifts more towards changing the emotional tone than to keeping the meaning similar, the proposed approach becomes a better, more reliable, solution. Also, the type of adjustment plays a role. Adjusting the extent to which an emotion is present more often leads to correct adjustments than replacing one emotion with another.

Finally, in order to validate sentences we used a combination of automated grammar checking and spellchecking methods prior to each adjustment and after each adjustment. While this method was reliable for detecting whether the numbers of grammatical errors changed due to the modification, we found that a lot of the output sentences that might be deemed illogical by human readers would pass the validation tests. Therefore we believe this validation method is not a sufficient way to validate adjustments made to the emotional tone of text.

7.1 Future work

With this research, we feel we have taken some key steps in exploring and proving the viability of emotion manipulation. However, there is more work that needs to be done. In this section, we cover various directions future research can take to further explore this topic and to make sure findings are validated to a greater extent.

7.1.1 (Partial) Sentence Replacement

The main mechanism by which analysis, adjustments and validation took place was focused on the scope of single words. However, parts of the emotional tone of a text might be attributed to a combination of words or even the context of entire sentences. Future research can explore the effect of taking these various scopes into account during analysis, adjustment and validation to research to what extent they contribute to emotional tone. One possible resource that could be used is the SenticNet Semantic Resource [32]. It incorporates groups of associated words and assesses their polarity both separately and in context to each other. While this resource is solely polarity based, it might be used in combination with an emotional lexicon to attribute emotions to these word groups.

7.1.2 Incorporating Human Control

Only some of the mechanisms that created the datasets and the machine-learned mechanism were created using human validation. By incorporating human control after validation for the rule-based detection and each of the adjustment styles we can more accurately measure what the emotional effect of an adjusted sentence is on a person.

7.1.3 Context Handling

Text sentiment analysis is complicated. There are many different factors that influence the style of a text. For example, a sentence can be interpreted in a different way depending on the sentence that preceded it or the one that follows it. Other factors that can influence the perceived emotional tone of a sentence are emotion intensifier like ”very”, ”a little” and ”a bit”. Other factors are punctuation, negation and capitalization. Future research can incorporate more of these factors in other to reduce as many threats to validity as possible.

7.1.4 Machine Learning

In the current research, a rule-based mechanism is proposed to alter the emotional tone of text. However, there can be value in researching a machine-learning-based method. It could use human-validated input sentence to train a model and then use either supervised or unsupervised learning to adjust and validate adjustments. Adjustment types that are similar to the ones could be used. Results might vary, leading to a possibility of higher accuracy, specifically with a focus on semantic retention.

Acknowledgements

I'd like to start off by thanking both my supervisors: Michael Metternich from the host company Bloomreach and Adam Belloum for the University of Amsterdam. Michael provided me with the opportunity to discuss my progress daily and to really discuss the contents of my thesis each week. He also provided key insights when faced with tough decisions throughout the entirety of the research process. Adam was an incredibly valuable resource in providing information and advice on the technical aspects of the research and in providing information on the process and details regarding the execution of the research. Both supervisors were always willing and able to provide feedback and information. From the University of Amsterdam, I'd also like to thank Ana Oprescu, who provided a lot of guidance during the research preparation phase and set up weekly digital meetings in order to gauge my progress and make sure I was doing well. I'd also like to thank my friend Julia Schaap for helping me with spell checking, her help is greatly appreciated.

Most of all I'd like to thank my mother, Maltie Sardjoemissier. It is because of her countless sacrifices and emotional support that her son was even able to perform this research.

Bibliography

[1] M. ONeill, *Collins English dictionary.* Collins, 2019.

[2] J. Thorley, *Athenian democracy.* Psychology Press, 2004.

[3] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers", *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[4] D. D. Droba, "Methods used for measuring public opinion", *American Journal of Sociology*, vol. 37, no. 3, pp. 410–423, 1931.

[5] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press, 2015.

[6] H. Gong, S. Bhat, L. Wu, J. Xiong, and W.-m. Hwu, "Reinforcement learning based text style transfer without parallel training corpus", *arXiv preprint arXiv:1903.10671*, 2019.

[7] M. Z. Asghar, A. Khan, A. Bibi, F. M. Kundi, and H. Ahmad, "Sentence-level emotion detection framework using rule-based classification", *Cognitive Computation*, vol. 9, no. 6, pp. 868–894, 2017.

[8] M. Z. Asghar, F. Subhan, M. Imran, F. M. Kundi, S. Shamshirband, A. Mosavi, P. Csiba, and A. R. Varkonyi-Koczy, "Performance evaluation of supervised machine learning techniques for efficient detection of emotions from online content", *arXiv preprint arXiv:1908.01587*, 2019.

[9] D. Arora, K. F. Li, and S. W. Neville, "Consumers' sentiment analysis of popular phone brands and operating system preference using twitter data: A feasibility study", in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, IEEE, 2015, pp. 680–686.

[10] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle", in *Proceedings of the ACL 2012 system demonstrations*, Association for Computational Linguistics, 2012, pp. 115–120.

[11] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news", *arXiv preprint arXiv:1309.6202*, 2013.

[12] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news", *World Wide Web*, vol. 17, no. 4, pp. 723–742, 2014.

[13] A. Micu, A. E. Micu, M. Geru, and R. C. Lixandroiu, "Analyzing user sentiment in social media: Implications for online marketing strategy", *Psychology & Marketing*, vol. 34, no. 12, pp. 1094–1100, 2017.

[14] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis", in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, Association for Computational Linguistics, 2011, pp. 142–150.

[15] Y. Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining", *Information Sciences*, vol. 266, pp. 90–100, 2014.

[16] A. Bandhakavi, N. Wiratunga, S. Massie, and D. Padmanabhan, "Lexicon generation for emotion detection from text", *IEEE intelligent systems*, vol. 32, no. 1, pp. 102–108, 2017.

[17] 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, 2011.

[18] P. Ekman, "Basic emotions", *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[19] Y. Cheng, Z. Gan, Y. Zhang, O. Elachqar, D. Li, and J. Liu, *Contextual text style transfer*, 2020. [Online]. Available: `https://openreview.net/forum?id=HkeJzANFwS`.

[20] C. N. D. Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer", *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018. DOI: `10.18653/v1/p18-2031`.

[21] E. Tromp and M. Pechenizkiy, "Rule-based emotion detection on social media: Putting tweets on plutchik's wheel", *arXiv preprint arXiv:1412.4682*, 2014.

[22] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents", *Expert Systems with applications*, vol. 34, no. 4, pp. 2622–2629, 2008.

[23] A. Agrawal and A. An, "Unsupervised emotion detection from text using semantic and syntactic relations", in *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, IEEE, vol. 1, 2012, pp. 346–353.

[24] C. Nugent and P. Cunningham, "A case-based explanation system for black-box systems", *Artificial Intelligence Review*, vol. 24, no. 2, pp. 163–178, 2005.

[25] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon", *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[26] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran, "Indexing with wordnet synsets can improve text retrieval", *arXiv preprint cmp-lg/9808002*, 1998.

[27] S. M. Mohammad, "Word affect intensities", in *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018.

[28] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining.", in *LREC*, Citeseer, vol. 6, 2006, pp. 417–422.

[29] H. Hamdan, F. Béchet, and P. Bellot, "Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging", in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 455–459.

[30] A. Firmanto, R. Sarno, *et al.*, "Prediction of movie sentiment based on reviews and score on rotten tomatoes using sentiwordnet", in *2018 International Seminar on Application for Technology of Information and Communication*, IEEE, 2018, pp. 202–206.

[31] C. Hung and H.-K. Lin, "Using objective words in sentiwordnet to improve word-of-mouth sentiment classification", *IEEE Intelligent Systems*, no. 2, pp. 47–54, 2013.

[32] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining", in *2010 AAAI Fall Symposium Series*, 2010.

[33] S. M. Mohammad and P. D. Turney, "Nrc emotion lexicon", *National Research Council, Canada*, 2013.

[34] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon", in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA: Association for Computational Linguistics, Jun. 2010, pp. 26–34. [Online]. Available: `https://www.aclweb.org/anthology/W10-0204`.

[35] E. Kušen, G. Cascavilla, K. Figl, M. Conti, and M. Strembeck, "Identifying emotions in social media: Comparison of word-emotion lexicons", in *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, IEEE, 2017, pp. 132–137.

[36] S. Mohammad and S. Kiritchenko, "Using nuances of emotion to identify personality", in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[37] S. Mohammad, "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales", *arXiv preprint arXiv:1309.5909*, 2013.

[38] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets", *arXiv preprint arXiv:1708.03696*, 2017.

[39] A. Zirikly, P. Resnik, O. Uzuner, and K. Hollingshead, "Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts", in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 24–33.

[40] K. Anoop, P. Deepak, and V. Lajish, "Emotion cognizance improves fake news identification", *arXiv preprint arXiv:1906.10365*, 2019.

[41] A. Giachanou, P. Rosso, and F. Crestani, “Leveraging emotional signals for credibility detection”, in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 877–880.

[42] Praveen, *Emotions dataset for nlp*, Apr. 2020. [Online]. Available: `https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp`.

[43] ParallelDots, *Emotion detection using machine learning*, May 2017. [Online]. Available: `https://medium.com/@ParallelDots_67026/emotion-detection-using-machine-learning-706ddceaa1`.

[44] E. Loper and S. Bird, “Nltk: The natural language toolkit”, *arXiv preprint cs/0205028*, 2002.

[45] *Spell and grammar checker*. [Online]. Available: `https://languagetool.org/`.

[46] M. Yun, C. Yangbin, M. Xudong, and L. Qing, “A syntax-aware approach for unsupervised text style transfer”, *ICLR 2020*, 2019.

[47] J. Li, R. Jia, H. He, and P. Liang, “Delete, retrieve, generate: A simple approach to sentiment and style transfer”, *arXiv preprint arXiv:1804.06437*, 2018.

[48] C. N. d. Santos, I. Melnyk, and I. Padhi, “Fighting offensive language on social media with unsupervised text style transfer”, *arXiv preprint arXiv:1805.07685*, 2018.

[49] C. Strapparava and R. Mihalcea, “Learning to identify emotions in text”, in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.

[50] F.-R. Chaumartin, “UPAR7: A knowledge-based system for headline sentiment tagging”, in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 422–425. [Online]. Available: `https://www.aclweb.org/anthology/S07-1094`.