Master Thesis

An Automatic Emotion & Purpose Classifier for Dutch Tweets Written by Members of the Dutch Parliament

Eva E. Zegelaar

an internship thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Isa Maks, Marcel Henriquez, Jacco Drabbe
2^{nd} reader: Roser Morante Vallejo

Submitted: July 16, 2020

Abstract

Twitter has a considerable body of tweets posted by politicians and research suggests that these can be influential in political affairs (Duncombe, 2019). In Red Data's *Haagsefeiten* ["Facts of The Hague"] website archive, more than 1 million tweets expressed by members of the Dutch Parliament are accessible. The goal of this thesis is to classify these tweets into meaningful emotion and purpose categories. The selection of the labels is based on an agreement study whereby trial annotation rounds took place. The emotion category resulted in a Kappa score of (0.416) and the binary proactivity category resulted in a Kappa score of (0.314). These scores are considered low, therefore a third category was introduced, polarity, resulting in a Kappa score of (0.561). The agreement study shows that there is an identifiable presence of positive and negative labels, but less so for complex emotions and proactivity labels. Two systems were implemented in this study: A baseline SVM and a state-of-the-art CNN-BiLSTM with pre-trained Dutch word embeddings. The system that performed best is the SVM with polarity labels, scoring a weighted f1-score of 0.59 and an accuracy of 0.60. The research concludes that the SVM works well with less training data and well-distributed polarity labels. To improve the performance of the SVM, feature engineering and the implementation of word embeddings are two possible solutions. The CNN-BiLSTM needs more training data. Lastly, to improve the quality of the emotion and proactivity labels, a new agreement study with possibly newly proposed labels is necessary.

Declaration of Authorship

I, Eva Evertovna Zegelaar, declare that this thesis, titled *An Automatic Emotion & Purpose Classifier for Dutch Tweets Written by Members of the Dutch Parliament* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: July 16, 2020

Signed: Eva Evertovna Zegelaar

Acknowledgments

The completion of this thesis signifies the end of my study in the Master of Linguistics, Text Mining.

Firstly, I would like to thank Isa Maks for guiding me in the entire thesis process. I would also like to thank all of the staff members of the Text Mining master for teaching me and challenging me to reach a technical level I would have never imagined to reach in the past. I would like to thank those involved in the creation of this master, as thanks to you I expanded my linguistic knowledge towards a more technical level that will help me in my future career.

I would also like to thank Marcel Henriquez and Jacco Drabbe for giving me the opportunity to do my thesis as part of an internship in Red Data. I was lucky to experience the corporate world even though the corona crisis was taking place. Red Data's positive attitude in the online meetings pushed me to work hard, efficiently and happily.

Last but not least, I would like to share my deepest gratitude to my loved ones; my friends and family for always supporting me unconditionally.

List of Figures

Contents

Chapter 1

Introduction

This research aims to find whether tweets written by members of the Dutch parliament can be classified into meaningful classes of emotions and meaningful classes of purpose categories that can be reliably annotated by humans and detectable by machines. Automatic recognition of emotions in Twitter has been an important topic in Natural Language Processing (NLP) and it continues to be a popular task in many disciplines. However, most of the research on classification systems tend to work with tweets written in English and by the general public. Therefore, working with Dutch tweets written by members of the Dutch parliament may aid new information to the current research in automatic emotion recognition.

As part of an internship with Red Data, this task is relevant for the company as they run a website called "*Haagse Feiten*"/ ["Facts of The Hague"], an online archive containing political information and news about the Dutch government. Therefore, implementing emotion and purpose detection on Dutch political tweets can be an interesting application in their website. More than 1 million of these tweets are accessible with a membership. The tweets have been a made accessible to me by saving them in a CSV file containing the content of the tweet, the author and the date posted.

Classifying emotions into discrete categories is a debated topic in psychology (Barrett et al., 2018). Rather than focusing on such debate, the thesis focuses on the emotion and purpose categories that are present in the data. The NLP task of emotion detection has shown to work well with basic sentiment; that is, positive, negative and neutral emotions. Most of the NLP research on this task has also been devoted to basic polarity. Nevertheless, expanding the polarity categories towards a more complex spectrum of emotions can yield more meaningful results and benefit disciplines such as data mining, psychology and artificial intelligence (Alswaidan and Menai, 2020). Determining political emotions in Twitter is relevant because studies have shown that tweets can both express emotions and cause emotion responses (Duncombe, 2019). In some cases, some tweets may result in outcomes that may influence political decisions (Duncombe, 2019).

Less research has been devoted to classifying purpose in political tweets; however, it is a category that is relevant in tweets expressed by politicians and therefore interesting to Red Data. For instance, it is possible that some clients such as news platforms and other important organisations are interested in such information. Tweets expressed by politicians are purposeful (Duncombe, 2019) and purpose can yield relevant information about the politicians and political parties. The thesis has a strong focus on the work by (Mohammad et al., 2015), as they successfully implemented an SVM classifier

to detect purpose and emotion categories of US electoral tweets.

1.1 Goal and Research Questions

The goal is to build a machine learning classifier that can automatically detect emotion and purpose classes in tweets written by members of the Dutch parliament. In attempt to achieve this goal, the following research questions are investigated:

Main Research Question

- *Is it possible to build an emotion and purpose classification system to detect emotion and purpose written in tweets by members of the Dutch parliament?*

Sub-question (a)

- *What annotation model is best to identify categories relevant for emotion and purpose classification of tweets?*

In order to be a able to reach the main goal and answer the main research question, sub-question (a) is unquestionably important. The creation of the emotion labels and purpose labels is a significant part of the research because identifying reliable, meaningful and representative labels requires a lot of thought, research and can be prone to subjectivity. In attempt to create the best annotation model, an agreement study has been implemented. In this agreement study, trial annotation rounds took place to calculate inter-annotator agreement for the purpose of identifying appropriate and distinguishable labels for the classifier.

Sub-question (b)

- *What machine learning methods are best to automatically classify these tweets?*

After a selection of labels have been made, implementing appropriate machine learning algorithms for this type of data is necessary. Based on recent research, two systems that have shown to have state-of-the-art results are Support Vector Machines (SVM) and Convolutional Neural Networks Bidirectional Long Short Term Memory (CNN-BiLSTM). These two systems are implemented in the thesis.

1.2 Thesis Outline

The remainder of the thesis has the following structure. Chapter 2 contains background information and a literature review of the related work about emotion detection; the classes of emotions and purpose; the related domains; and the relevant automatic classification system approaches. Chapter 3 provides a thorough agreement study of the selected labels that are based on the literature and adapted to the data. It also discusses the reasons for selecting the labels and the making of the gold data. Chapter

4 presents the training and test data and a detailed discussion of the used architectures: The SVM as a baseline and the CNN-BiLSTM as the state-of-the-art approach. Chapter 5 presents and describes the results of Precision and Recall performed by the systems. Chapter 6 describes and analyzes the errors committed by the systems; the implications of the results and possible solutions for future work. Lastly, Chapter 7 provides the central conclusions of the thesis.

Chapter 2

Related Work

Automatically recognising emotions in textual data, particularly in social media, has been a significant topic of interest in Natural Language Processing (NLP). A lot of research has been devoted to producing supervised machine learning models that can classify texts into different emotions. This chapter provides a literature review of such research and is split into two main sections. Section 2.1 is devoted to an in-depth discussion of emotion classifiers and is composed by the following subsections. Subsection 2.1.1 covers the approaches to categorising and annotating a larger spectrum of emotions. Subsection 2.1.2 discusses the common domains studied for emotion classifiers. Lastly, subsection 2.1.3 describes and compares supervised machine learning approaches for emotion classification.

2.1 Emotion Detection

Research into automatic emotion classification has a strong focus in sentiment or polarity detection, that is, classifying texts into positive, negative or neutral emotions (Nakov et al., 2016). Social media platforms such as Twitter or review sites such as IMDB, contain a lot of emotions expressed in text which attract researchers in language processing, political and social sciences (Rosenthal et al., 2017). Twitter is an emotion-rich corpora where sentiment detection is relevant and continues to be a popular task for research (Rosenthal et al., 2017). A recent example of such a system resulting in competitive results is by Rehman et al. (2019), where a state-of-the-art Hybrid Convolutional Neural Network (CNN) and Long Short-Term-Memory (LSTM) has been implemented on English IMDB and Amazon movie review data sets.

Polarity detection can yield interesting and competitive results when implemented in social media texts where a lot of negative and positive emotions are clearly classifiable. Nonetheless, there are recent developments that go beyond polarity to try and detect a more complex spectrum of emotions. Emotion detection and classification in text "refers to the task of automatically assigning an emotion to a text selected from a set of predefined emotion labels" (Alswaidan and Menai, 2020). Emotion analysis can be seen as a "natural evolution of sentiment analysis" and it is more meaningful as it captures more complex meanings in text (Seyeditabari et al., 2018). An emotion classifier which yields quality results can benefit different applications in a variety of fields such as, artificial intelligence, data mining, psychology and information filtering systems (Alswaidan and Menai, 2020). In order to produce a quality emotion classification system, several aspects which are backed up by related literature need to be

considered; these are discussed in the following subsections.

2.1.1 Classes of Emotions & Purpose-like Categories

Classes of Emotions

Classifying emotions into separate classes is an ongoing, complex debate in psychology (Barrett et al., 2018). Rather than contributing to such debate, this research focuses on studying the related work in emotion classification that is relevant in NLP tasks. More specifically, this subsection delves into examples of separating emotions into "discrete and finite sets of emotions" (Bostan and Klinger, 2018).

A phrase that contains keywords with explicit emotions is easier to classify than a longer sentence that contains implicit or context-dependent emotional expressions (Seyeditabari et al., 2018). Linguistically-speaking, the expression of an emotion is difficult to identify at the text level. This may be due to two primary reasons. Firstly, a short phrase may express more than one emotion in a way that even for humans is difficult to identify. Secondly, the metaphorical, implicit and context-dependent emotional language, makes this NLP task very challenging (Seyeditabari et al., 2018). A way of addressing this challenge, is to strongly emphasize on the emotional language use, as well as redefining annotation labels and guidelines throughout the design process of a multi-emotion classifier (Seyeditabari et al., 2018).

A lot of research on emotion classification have followed Ekman's model (Strapparava and Mihalcea, 2007; He et al., 2017; Alswaidan and Menai, 2020), in which four to six basic emotions are classified, which include *anger, disgust, fear, joy, sadness* and *surprise* (Ekman, 1992). Other research has focused on Plutchik's well-known *Wheel of Emotion* (Plutchik, 1980), which takes into consideration emotions with opposing pairs as dispayed in figure 2.1. By looking at the wheel in figure 2.1, one can identify a wide range of emotions with opposing pairs, such as *sadness-joy*. The emotions are also split by intensity: the outer layers contain more subtle emotions and it gradually increases in intensity up until the inner circle.

There is a highly relevant study where a supervised multiple emotion classifier is implemented on English tweets about the 2012 US Presidential Elections (Mohammad et al., 2015). Their study has a strong emphasis in the design of emotion categories. They implemented Plutchik's wheel and classified each tweet into one of the 8 basic emotions: *trust, fear, surprise, sadness, disgust, anger, anticipation* and *joy*. However, they believe Plutchik's wheel contains categories that are rather "coarse" (Mohammad et al., 2015) because several labels overlap. For instance, they argue that the labels of *disgust, dislike, hate, disappointment* and *indifference* belong to the *disgust* category (Mohammad et al., 2015). To make the human annotation procedure easier for the annotators, they presented a large list of 19 emotion labels [1], which were later mapped to the previously-stated basic 8 emotions (Mohammad et al., 2015).

[1] Initial 19 emotions: *acceptance, admiration, amazement, anger, anticipation, hate, indifference, joy, like, sadness, calmness, disappointment, disgust, dislike , fear, surprise, trust , uncertainty, vigilance.*

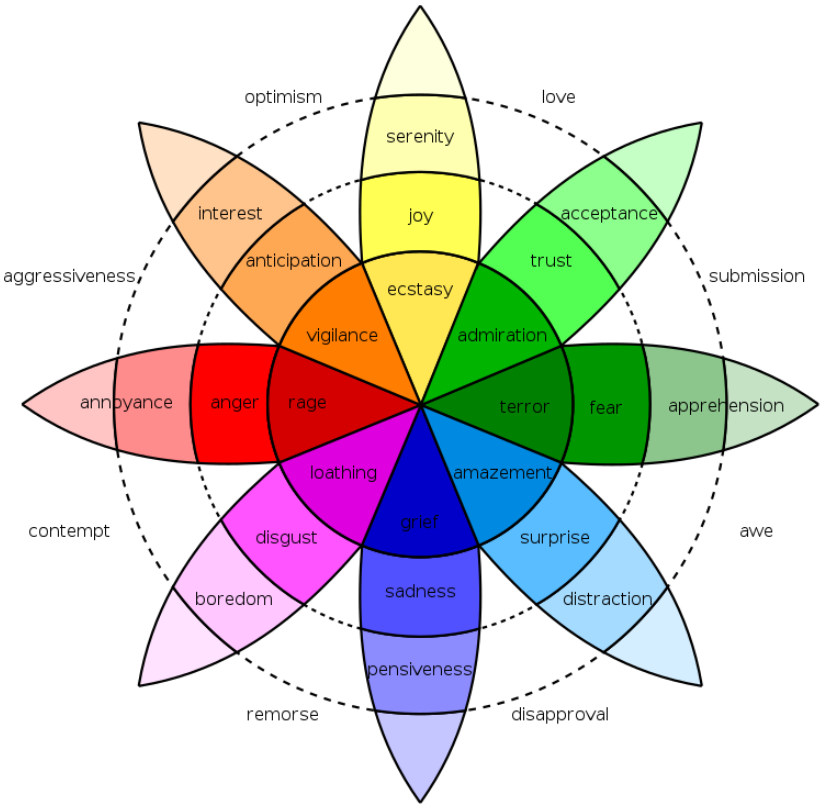

Figure 2.1: Plutchik's Wheel of Emotions (Plutchik, 1980; Donaldson, 2017)

Classes of Purpose

In the same study, they additionally implement purpose-like categories as they attempt to identify patterns and relationships between purpose and emotion in electoral tweets. Purpose can be defined as having a reason to do something or why something exists (Cambridge, 2020). Purpose is a relevant category because Twitter is a space where emotions are expressed; it is also a space where emotions can be a result of having a purpose by *provoking* emotions (Duncombe, 2019). Mohammad et al. (2015) found that although the "purpose classifier benefits from emotion features, emotion detection alone can fail to distinguish between several types of purpose" (Mohammad et al., 2015). This suggests that having an additional category of 'purpose', may provide supplementary information in the domain of political tweets that emotion alone cannot detect. For instance, the emotion of *disgust*, can be associated with different purpose categories like *criticize*, *vent* or *ridicule* (Mohammad et al., 2015). The annotators were asked to use the following 11 'purpose' labels to annotate the tweets: *to agree, to admire, to support, to point hypocrisy, to point out mistake, to disagree, to ridicule, to criticize, to vent, to provide information without emotion* or *none of the above.* (Mohammad et al., 2015). They find that their automatic classifier has an accuracy of 44.56% when tweets are classified into these 11 classes and it significantly rises to

73.91% when the 11 classes are placed into following 3 classes: *favour, oppose* or *other* (Mohammad et al., 2015). This suggests that reducing the number of classes by merging those labels that overlap is a possible method to improve system performance.

Mohammad et al. (2015) calculated inter-annotator agreement on the full sets of annotations (the average percentage of times annotators agree (IAA)) and the majority class (average probability of choosing the majority class (APMS)). For the purpose labels, IAA is 43.58 and APMS is 0.520. For the emotion labels IAA is 59.59 and APMS is 0.736 (Mohammad et al., 2015). According to Mohammad et al. (2015), this is considered to be moderate agreement. This suggests that selecting a purpose or an emotion category from a large number of labels is difficult for humans. To reduce this limitation, a minimum of three annotators had to annotate the same batches of tweets(Mohammad et al., 2015). This means that the tweet becomes part of the gold data when the same label is annotated by at least half of the annotators. If the three annotators had each annotated a tweet with a different label, the tweet was discarded from both training and validation data (Mohammad et al., 2015). This is another possible method to implement in this thesis for improving the quality of the labels.

2.1.2 Related Domains

Appropriate text sources for emotion detection tend to be blogs, news, news headlines and other forms of media (Bostan and Klinger, 2018). Twitter is a popular subject of research in text classification because it is user-friendly and "has a well-documented API"(Bostan and Klinger, 2018). Furthermore, Twitter is a platform where quick, everyday thoughts are expressed and therefore it contains a lot of "affect-related text" (Ying et al., 2019). Therefore, Twitter is a good source to study automatic emotion detection. On the other hand, it should be noted that selecting a particular domain for the system design is important because language and domain specific factors may deal better with both the implicit nature of emotions, and the context of the emotion (Seyeditabari et al., 2018). In the research by Ying et al. (2019), it was suggested that implementing domain-specific knowledge and corpora in already fine-tuned state-of-the-art systems improve multi-label emotion classification of tweets. (Ying et al., 2019).

Performance of NLP tasks does not entirely depend on choosing the appropriate machine learning algorithm, but more so on the size and the quality of the training data (Sang et al., 2017). The selection of the domain is an important part for the quality of the data. Some of the domains that have been often researched include; health, politics and the stock market (Bostan and Klinger, 2018). Twitter is a platform where a lot of politicians write tweets for different purposes such as, to express a party stance, to advertise a campaign, to get more votes, to criticize, to make a change and more (Sang et al., 2017; Duncombe, 2019; Mohammad et al., 2015).

An additional relevant paper worth mentioning is a study about the public's sentiment on the outbreak of Measles in the Netherlands from 2013 (Mollema et al., 2015). Their goal was to compare the sentiment and number of tweets expressed in Twitter in relation to the number of reported measles cases to see if there was a correlation between the public's sentiment with the disease patterns (Mollema et al., 2015). The conclusion of this paper is that monitoring Twitter can be useful for the government in deciding how to respond and inform the public about the outbreak of a disease (Mollema et al., 2015). Although this study has a different focus to emotion classification, it is a good example to show the relevance in the use of Twitter for domain-specific studies

involving the identification of emotions.

2.1.3 Automatic Classification of Emotions

In a survey paper published in 2020, several classical and state-of-the-art approaches for emotion recognition are discussed (Alswaidan and Menai, 2020). Explicit emotion recognition is more commonly researched. That is, employing emotion words that are mentioned in the text. Whereas, implicit emotion recognition remains a challenging problem that requires further study (Alswaidan and Menai, 2020). The survey demonstrates that the outperforming supervised approaches for both explicit and implicit emotion classification are: Classical learning, Deep learning approaches and Hybrid approaches (Alswaidan and Menai, 2020).

Supervised machine learning involves the implementation of a machine learning algorithm that learns patterns from training data and makes predictions on a new set of data based on what it has learnt. In other words, given a set of labels A, and a set of training examples B "which has been assigned to one of the class labels" A, the system must learn from B to predict the labels of previously unseen test data that is similar to B (Scott and Matwin, 1998).

Classical Machine Learning

In classical machine learning, the system learns from experience (Alswaidan and Menai, 2020). The most commonly applied classical approach is the linear model of Support Vector Machines (SVM) (Alswaidan and Menai, 2020; Mohammad et al., 2018). For a classification task using SVM, preprocessing of the text is required such as, the removal of stop words (Alswaidan and Menai, 2020). For an SVM to increase in accuracy, feature engineering is implemented in order to extract useful features that have the most "information gain" (Alswaidan and Menai, 2020). During training, the SVM algorithm outputs an "optimal hyperplane" and is later used to classify emotions of unseen text or test data. (Alswaidan and Menai, 2020). SVM is useful when there is not enough annotated data for training (Alswaidan and Menai, 2020). This is shown a system with a reported 89.43% accuracy using 250 annotated news headlines for training as well as implementing a dictionary of emotion words from Wordnet-Affect (Kirange and Deshmukh, 2012). In the previously discussed paper in 2.1.1 by Mohammad et al. (2015), an SVM classifier was implemented which resulted in an F score of 58.30 for emotion categories, 43.56 for an 11-class task with purpose categories and 73.91 for a 3-class task with purpose categories. The 3-class task with purpose labels scored highest.

SVM implementation for emotion classification in the SemEval-2018 Task 1 was popularly used, however, it scored a Pearson r of 52.0 for English and did not outperform the deep learning approaches (Mohammad et al., 2018). Due to the existence of deep learning algorithms with state-of-the-art results, classical machine learning such as SVM and Naive Bayes(NB) are often used as baseline methods (Wang and Manning, 2012). NB scores higher with a simple bag-of-words (BoW) approach for shorter phrases, whereas SVM with BoW scores higher for longer-length reviews (Wang and Manning, 2012). The baseline that scored highest was a combined NB and SVM approach scoring a 89.45% accuracy (Wang and Manning, 2012).

Deep Learning

Supervised Deep learning is a branch of machine learning in which the system learns from experiences by understanding the complex hierarchy of concepts and their relationships with simpler concepts (Alswaidan and Menai, 2020). Different from linear models like SVM, deep learning is inspired by the neural function of the human brain and is also known as non-linear artificial neural networks (Brownlee, 2019). "Neural Networks are powerful learning models" (Goldberg, 2016) and certain types have reported excellent, state-of-the-art results for multi-emotion classification tasks (Alswaidan and Menai, 2020).

The most commonly used model is the long short-term memory (LSTM), which is a type of recurrent neural network (RNN) (Alswaidan and Menai, 2020). When working with linguistic textual data, it is common to deal with sequences for examples, sequences of words that represent a sentence(Goldberg, 2016). RNNs enable arbitrary sentence inputs in a fixed-sized vector and pays attention to the structure and sequences of the input vector (Goldberg, 2016). However, simple RNN's suffer from the vanishing gradients problem. A vanishing gradient signifies that the changing gradient from which a neural network learns from will diminish until it remains constant and cannot learn anymore (Goldberg, 2016). The consequence of this is that it cannot learn long-range dependencies (Goldberg, 2016), meaning that the RNN cannot make a connection between previous important information with present information (Olah, 2015). In other words, it 'forgets' relevant information as a result of vanishing gradients. LSTM's solve this vanishing gradient problem because it makes use of special vectors called 'memory cells' which "preserves" (remembers) relevant "gradients across time" (Goldberg, 2016).

Two studies which make use of architectures with state-of-the-art results are by Ge et al. (2019) and Liu et al. (2020). Both studies make use of a hybrid system, which is a convolutional neural network (CNN) in conjunction with a bidirectional long short-term memory (BiLSTM) and word embeddings. Ge et al. (2019)'s experiment involved the classification of a person's utterance (in a written dialogue) into four categories, their system achieved a 0.74 F-score. Liu et al. (2020)'s use sentence level emotion classification with the same type of architecture and end up with a 0.95 F-score. What is special about these systems is that they make use of CNN's and Bi-LSTM's. CNN is used to extract local features and its benefits include: reducing the size of feature vectors, eliminates unnecessary features and inputs the features into BiLSTM model which extracts global features. To avoid overfitting, the systems make use of a dropout layer. The dropout updates the weights and eliminates some units so that it is "not dependent on the inherent characteristics of the part" (Liu et al., 2020). The BiLSTM is a regular LSTM but additionally strengthens the "bidirectional relationship between the current text frame and the next text frame" (Ge et al., 2019). In the same study by Ge et al. (2019), single RNN, single LSTM, single BiLSTM and single CNN were implemented independently of each other (as opposed to the hybrid system). The results showed that the hybrid CNN-BiLSTM outperformed the other non-hybrid architectures.

Word Embeddings

A central component of a neural-network lies in the first embedding layer, which is

composed of word embeddings. Word embeddings are low dimensional vector representations of features or words that represent semantic information by grouping words in a vector space (Goldberg, 2016). The benefit of using pre-trained word embeddings is that it provides the system with vector representations of words that may not appear in the training data (Goldberg, 2016). This allows the model to generalise better on unseen, test data (Goldberg, 2016). Implementing word embeddings in both classical and deep learning commonly result in better performances than without the use of word embeddings (Alswaidan and Menai, 2020). Common examples of word embeddings that are implemented in state-of-the-art systems are *word2vec*, *Gensim* and *Glove* (Goldberg, 2016).

In a study about comparing the use of different pre-trained word embeddings in state-of-the-art architectures has shown a excellent results for emotion detection (Polignano et al., 2019). Features directly extracted from text can be computationally costly and it requires a lot of research and linguistic work (Polignano et al., 2019). When implementing appropriate pre-trained word embeddings it can improve the informativeness of the data for the system (Polignano et al., 2019). Although most of the word-embedding implementation for emotion classification is done in English, Nieuwenhuisje (2018a) has generated Dutch word embeddings from Dutch tweets. These word representations are relevant to implement because they have been trained on a lot of political tweets, which is the selected domain for this thesis. The pre-trained word embeddings are discussed in detail in the Methodology (Chapter 4).

To conclude this chapter, a lot of research has been done on polarity detection with successful results. The expansion of polarity towards detecting a larger spectrum of emotions in a specific domain is relevant and possibly more meaningful than simple polarity. The study by (Mohammad et al., 2015) is a good example of a successful classification of two relevant classes in electoral/political tweets: Emotion and Purpose. With regards to the literature, the selection and definition of labels for this thesis is inspired by Mohammad et al. (2015)'s. This study also inspired the methods of working with the labels to improve both inter-annotator agreement and system performance. In terms of machine learning, the literature has shown that both SVM and CNN-BiLSTM seem to provide the best results for emotion detection therefore, the methodology of the thesis will make use of these two systems.

Chapter 3

Data & Agreement Study

In this chapter the selection of data, the selection of labels and the definitions of these labels for the classification of tweets are discussed. Section 3.1 introduces the data. Section 3.2 discusses the agreement study and is composed by subsections that discuss the measurement of inter-annotator agreement (IAA), the annotation guidelines and the results of the trial rounds of tweet annotations. Lastly, section 3.3 is devoted to presenting the IAA results of the annotated training and test data; and includes subsection 3.3.1, which discusses the making of the gold labels.

3.1 Data Description and Selection

This section describes the data selection and access to the data to work with. The company Red Data runs a website called *Haagsefeiten*[1] "Facts of The Hague", where political information and political news from the Dutch parliament is presented and archived. Clients and organisations interested in political information about The Netherlands are able access this website with a membership. One of their political archives is social media, where more than 1 million tweets posted by members of the Dutch parliament is accessible. These tweets have been made accessible with the use of Twitter's API, where the tweets were extracted and saved as a CSV file. The CSV file contains the full content of the tweet, the author of the tweet, the political party of the author and the date it was posted.

The tweets that will be used are from the period of March and April 2020. The reason for this is that Red Data has only access of the *full* tweets from this period. Tweets from prior March of 2020 only contain a maximum of about 140 characters, and this limit was lifted by Twitter in around March 2020. It is necessary to work with the full tweets because the cropped tweets have important text missing for the classification. The domain of the tweets is anything that has been posted by members of the Dutch parliament during that period. I found that most tweets appear to be about politics, economics and any type of news. It must be noted that during the aforementioned period, the COVID-19 crisis was taking place, therefore, many tweets are directly or indirectly about this crisis. Furthermore, I found that most of the tweets are not as personal as tweets from the general public. For instance, most of the current tweets are written about a present, future or past situation or issue, rather than directly expressing a personal opinion about something. This results in tweets with more subtle

[1] Link to Red Data's *Haagsefeiten* website: https://haagsefeiten.nl

emotions than general tweets from the public. This is an example of a tweet written by a Dutch politician: "*Europese landen hebben financiële steun nodig, maar Hoekstra stelt de hoogste eisen van allemaal. Dit bespraken we met politiek redacteur @drskee en met Tweede Kamerlid @henknijboer.*" / [*European countries need financial support, but Hoekstra makes the highest demands of all. We discussed this with political editor @drskee and with Member of Parliament @henknijboer*". Contrastingly, the electoral tweets from the public presented in the study by Mohammad et al. (2015), look more like this: "*Mitt Romney is arrogant as hell*" (Mohammad et al., 2015). Observe that the second example is rather personal and expresses hostility, whereas the Dutch tweet uses more formal language.

Throughout the agreement study, we will continuously observe that the Dutch tweets are less personal and the defined labels will have to adhere to the present data.

3.2 Agreement Study

It is necessary to gather well-labelled training and test data for a working classifier (Sang et al., 2017). By looking at the tweets, we already know that these are generally less personal, but it does not mean that an emotion category or a purpose category is not present. The agreement study may allow us to further understand the specific emotions and the purpose of these tweets.

It should be noted that prior to providing the tweets ready for annotation, these had to be preprocessed. The most important preprocessing for the annotations was: (1) removing the author of the tweet to prevent the annotators from being influenced by their opinion of the politician or organisation posting the tweet; (2) removing duplicates of tweets; (3) and removing tweets that were not written in Dutch. The preprocessing is further discussed with more depth in the Methodology Chapter 4, because it is also important for system input.

This section is composed by the following subsections: Subsection 3.2.1 explains the way inter-annotator agreement is calculated. Subsection 3.2.2 describes the process of making appropriate annotation guidelines. Subsections 3.2.3 and 3.2.4 discuss the trial annotation rounds that have taken place to decide which labels best represent and classify the data.

3.2.1 Calculating Inter-Annotator Agreement

Producing a list of emotions and purpose labels that sufficiently represent the Dutch political tweets is as important as producing labels that result in a sufficient inter-annotator agreement Kappa score. Inter-annotator agreement (IAA) indicates that tweets can be clearly classified, which may be important for the accuracy of the system's classification. To evaluate IAA, three types of scores have been chosen: Cohen's Kappa score, percentage agreement and confusion matrices.

Cohen's Kappa is a measure of agreement between two annotators to determine which category a target (e.g. tweet) belongs to and how much agreement is not influenced by chance (Zaiontz, 2018; Landis and Koch, 1977). In other words, two annotators either agree in the category that the target is assigned to or they do not agree (Landis and Koch, 1977). Agreement percentage is another measure which is often seen in the literature on classification tasks. Unlike Cohen's Kappa, it does not take into consideration whether there is agreement by chance. Therefore, the agreement percent-

age is provided to give an additional measure to the results for a better overview. The main measure used to evaluate the labels is Cohen's Kappa. Cohen's Kappa and percentage agreement have been automatically calculated using the following online tool: *ReCal2* [2]. The Cohen's Kappa result is interpreted as follows: "values 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement" (McHugh, 2012). Furthermore, the confusion matrices provide a count of the labels where annotators agree or disagree. This gives a nice statistical summary of the data in order to identify which labels overlap or which labels are most commonly used.

3.2.2 Annotation Guidelines

A way of identifying appropriate labels and finding agreement between labels is to provide trial annotation rounds over different sets of tweets. The trial annotation rounds were also produced for sufficient training of the annotators to avoid misinterpretation of the guidelines. In the trial annotation rounds, an increase in Kappa score was expected per round until there was sufficient agreement to proceed with the final annotations for training. A series of annotation test rounds took place to test the quality of the annotation guidelines and to generate improved guidelines. A total of 5 trial annotation rounds were produced and documented, where the first 2 involved annotating different batches of around 30 tweets between at least 3 annotators. In the last 3 rounds, different batches of around 50 tweets were annotated between at least 3 annotators. In the last 2 trial rounds, agreement was sufficient to begin annotating the training and test data, and to create the gold labels to be used them for system input.

The guidelines introduced the study, provided specific instructions as to how to annotate and define the labels. The labels were defined using a combination of dictionary definitions, example scenarios and example tweets. The final annotation guidelines can be found in Appendix A. Each definition was kept simple and to the point. For each round, the guidelines were changed and improved. This was based on the results of each trial annotation round. To view an example of the way the guidelines improved, view figures 3.1 and 3.2. These figures illustrate an example of the way an emotion label is defined in the very first annotation guidelines (figure 3.1), and how it is defined in the finalized annotation guidelines (figure 3.2). The first one provides a very generalised definition. A generalised definition is prone to overlapping between labels. As opposed to the first definition, the second clarifies that the label of '*disappointment*' is a negative emotion and refers to the present or the past. It is also worth mentioning that some of the annotators had expressed that the final annotation guidelines had improved and were more easy to follow.

Disappointed: The tweet displays negativity/sadness/anger because expectations were not met; there can also be some form of surprise over someone's negative actions. E.g. *"The lockdown happened too late" or "X person should have done this instead".*

Figure 3.1: An example of the way the first annotation guidelines defined the emotion label 'disappointment'

[2]Link to access 'Recal2' tool: http://dfreelon.org/utils/recalfront/recal2/

Negative + Present/Past

disappointment: As opposed to admiration, the tweet shows negativity about a past or present situation. This label can represent both sadness and anger. For example, when expectations were not met or they are critical of someone or something currently happening. It will tend to use present or past tense in verbs. E.g. '*het was/is'*. Other examples: *"Unfortunately, the lockdown happened too late" or "Our country has been let down", "I don't like this!"*.

Figure 3.2: An example of the way the finalised (improved) annotation guidelines defined the emotion label 'disappointment'

We decided that an appropriate way to distinguish and define each emotion is by looking at two key aspects: polarity and time. Specifically, whether a tweet is positive, negative or neutral; and whether a tweet talks about the past/present or the future. Thus, the finalised labels were the following: *admiration , optimism, disappointment, scepticism , no emotion* and *other* (See figure 3.3). With regards to the second class of labels (purpose 'proactivity' labels), we decided to provide a binary categorization that would label a tweet as *proactive* or *non-proactive* (See figure 3.3). More specifically, *proactive* represents the desire or need to make a change and *non-proactive* does not represent the desire or need to make a change. The final annotation guidelines provide the tables displayed in figure 3.3. The process of deciding to select these labels will be discussed in more depth in the upcoming subsections.

	Past/Present	***Future***
Positive	admiration	optimism
Negative	disappointment	scepticism
Neutral	No emotion	

Past/present + no desire to change	***Future + wants a change***
Non-proactive	proactive

Figure 3.3: Tables included in the final guidelines to help the annotator distinguish the emotion and proactivity labels

The final annotation guidelines were provided in a written Microsoft Word document. The full annotation guidelines can be found in Appendix A. As displayed in figure 3.4 below, the tweets to label were provided in a Microsoft Excel file. The first column contains the content of the tweet. In the second column, the emotion label has to be selected from a provided list of labels. In the third column, the proactivity label is selected. In the fourth column, the annotators had the option to write comments or a second label.

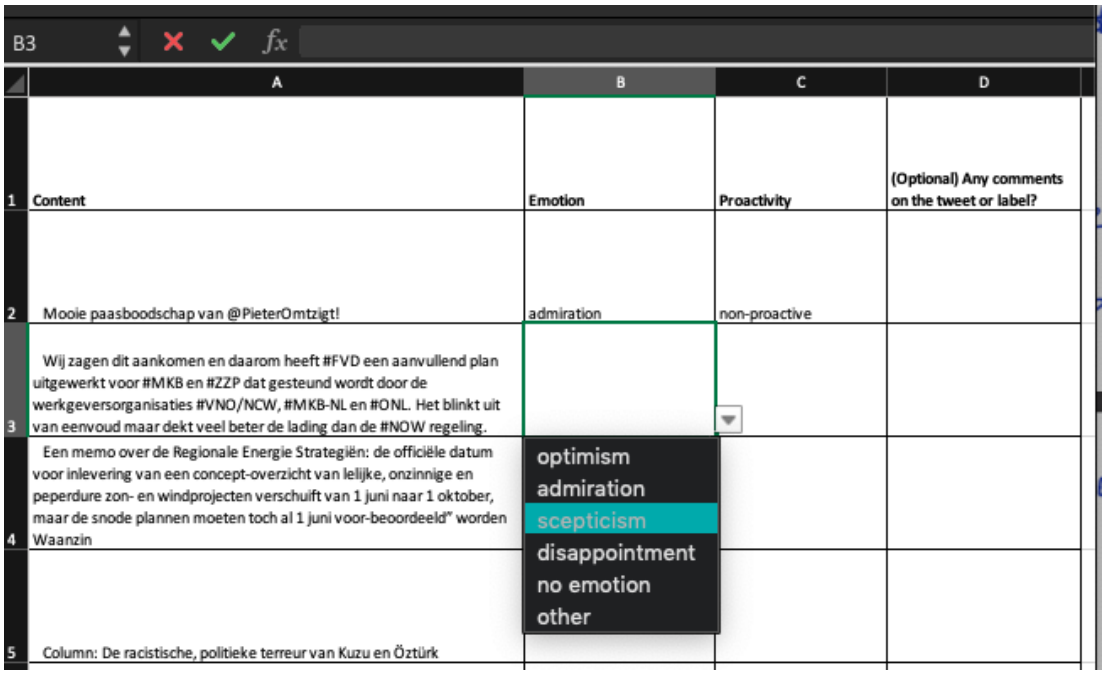

Figure 3.4: A screenshot to show a section of the layout of the tweets to be annotated in Microsoft Excel.

3.2.3 1st & 2nd Trial Annotation Rounds

The results of the 1st and 2nd trial annotation rounds are displayed in table 3.1 below. The table has been colour-coded for each round (round 1 in red and round 2 in blue). 30 tweets were annotated between 3 annotators: A, B and C. The first column of the table contains the label category and the annotator pair. The other columns contain the Kappa scores.

Mixed Labels

The first trial round consisted of annotating a mix of 8 emotions and purpose: *Optimistic, Sceptical, Surprise, Admiration, Disappointed, Proactive, No Emotion* and *Other*. This is a mix of labels because I included the label of *proactive*, which is a purpose category rather an emotion category. These labels were inspired by Mohammad et al. (2015) and by looking at the data. The labels by (Mohammad et al., 2015) were: *admiration* and *disappointed*; whereas the rest of the labels were picked by me after looking at the data. For positive emotions, Mohammad et al. (2015) had used the label *joy*, but this was not present in the data. Therefore, the label of *optimistic* and

its opposing pair *sceptical* were better choices because the tweets did seem to display these emotions. As mentioned previously, the Dutch tweets are less personal than the tweets that (Mohammad et al., 2015) worked with; therefore more subtle emotions were selected. The label of *proactive* was also included because it is a category that is interesting to Red Data and relevant in identifying tweets were politicians express the desire to make changes.

Looking back at table 3.1, the Kappa score is a combination of 'slight' (0.116) and 'fair' (0.381), which is considered very low. A plausible argument for such low agreement is that there may be too many classes of labels that overlap between annotators, see confusion matrix 3.2[3]. It seems annotators have difficulties distinguishing the labels. For example, the label of '*no emotion*' for annotator A overlaps with the label '*proactive*' for annotator B. Annotator A did not use the label '*proactive*'. This may suggest that annotator A treated '*proactive*' as '*no emotion*'. As a result, it was decided that the label of '*proactive*' should be included in a separate category.

Emotion Labels

The second trial round consisted of the same emotion labels previously selected and including *sadness*, because it seemed relevant to implement a more intense version of *disappointed*. For example, the following tweet was present: "*Mijn neefje wordt vermist sinds donderdag 9 april. Wil iedereen dit zoveel mogelijk delen!*"/[*My nephew has been missing since April 9. Everyone share this as much as possible!*]. For this tweet, the label of *sadness* is better suited than *disappointed* because a missing person is a very sad situation. Figure 3.1 shows that Kappa does not go beyond 'slight' agreement (0.35) (McHugh, 2012) and remains similar to the scores of round 1. The agreement was highest with the '*no emotion*' labels followed by the *'optimistic'* labels. Negative labels such as *anger* and *disappointment* overlapped in some cases, suggesting that it is challenging for the annotators to distinguish between different types of negative emotions in arguably more emotionally-subtle tweets such as those of this study.

Purpose Labels

The second trial round also consisted of labelling purpose categories because we noticed that this is a relevant category that is present in the tweets. The purpose labels have also been inspired by Mohammad et al. (2015)'s work on electoral tweets in combination with the insights gained from reading the raw tweets. The initial purpose labels are: *agree, commemorate, criticize, entertain, inform, proactive, question, ridicule* and *no purpose.* All of these labels are used in (Mohammad et al., 2015)'s work, except for the label of *proactive*, which is a label interesting to Red Data.

Proactive may be relevant in Dutch tweets expressed by members of the parliament and thus interesting to use. Proactivity is a type or purpose and purpose is a reason for which something is done or exists. Purpose does not need to be emotional. As previously discussed in chapter 2, tweets can represent emotions but also provoke emotions and reactions (Duncombe, 2019), because tweets are purposeful. A relevant type of purpose is whether the tweet displays a desire to make a change or not. For instance, in times of the corona crisis, members of the dutch parliament would post proactive-like

[3] B contains all of the confusion matrices for all the annotation rounds, but only the relevant ones are discussed in this chapter

tweets, for example, "*Allen samen krijgen we corona onder controle*"/["Only together will we be able to control corona"]. This tweet is proactive in the sense that only a change will be made in the current situation if people work together by following the laws of the COVID-19 lockdown. Another example is the following tweet: *"Dit moet het kabinet tijdens de lockdown doen - en dit is hoe we er weer uit komen (zo snel mogelijk)!*" / ["This is what the cabinet has to do during the lockdown - and this is how we get out again (as soon as possible)!"]. This second tweet also expresses the desire to make a change or do something.

Table 3.1 shows that the Kappa score of the purpose categories is also very low, receiving 'fair' agreement (0.31). The possible reason for this outcome is that there are too many labels for each class and the annotators are either unable to understand the definitions of the labels or there is overlap between labels. This is shown in confusion matrix table 3.3 below, as all the numbers are scattered. Interestingly, the label that was mostly used by the annotators is *inform* (8 times by annotator A), *proactive* (5 times by annotator B) and *commemorate* (5 times by annotator B).

Table 3.1: Agreement Scores for the labels of rounds 1 and 2

Label Category + Annotator Pairs	**Round 1 Cohen's Kappa**	**Round 1 Agreement**	**Round 2 Cohen's Kappa**	**Round 2 Agreement**
Mixed AB	0.116	24.1%	-	-
Mixed AC	0.206	34.5%	-	-
Mixed BC	0.381	48.3%	-	-
Emotion AB	-	-	0.102	25%
Emotion AC	-	-	0.35	54.2%
Emotion BC	-	-	0.227	37.5%
Purpose AB	-	-	0.31	37.5%
Purpose AC	-	-	0.028	37.5%
Purpose BC	-	-	0.227	37.5%

Table 3.2: Confusion Matrix of Annotators A and B in round 1

	A					
B	**admiration**	**disappointment**	**no emotion**	**optimistic**	**sceptical**	**Grand Total**
admiration	2		1	1		4
disappointment		1			2	3
no emotion		1	2	1	1	5
optimistic	2	1	1	1		5
proactive	2		3	3		8
sceptical				1	1	2
surprise		1	1			2
Grand Total	**6**	**4**	**8**	**7**	**4**	**29**

Table 3.3: Confusion Matrix of Purpose Labels for Annotators A and B in round 2

	A									
B	**agree**	**commemorate**	**criticize**	**criticize**	**entertain**	**inform**	**proactive**	**question**	**ridicule**	**Grand Total**
agree	1					1				2
commemorate		2			2		1			5
criticize			1			2				3
entertain				1	1	1		1		4
inform						2				2
proactive	1					2	2			5
question			1					1		2
ridicule									1	1
Grand Total	**2**	**2**	**2**	**1**	**3**	**8**	**3**	**2**	**1**	**24**

3.2.4 3rd, 4th & 5th Trial Annotation Rounds

In the last 3 trial annotation rounds, 50 tweets were annotated between 3 annotators. The Kappa score increased consecutively; beginning at the 3rd trial, then the 4th trial and lastly, the 5th trial. The results are shown in table 3.4, a colour-coded table representing each round in a different colour. For emotion labels, agreement increased from fair agreement up to moderate agreement. Furthermore, a binary category was introduced 'proactivity', which resulted in substantial agreement. This subsection will present the changes occurred in these last 3 trial annotation rounds which led to aforementioned results.

Purpose Labels

The purpose labels were used for the last time in round 3. Table 3.4 shows that the Kappa score is very low (0.319) . As a result, the purpose categories were removed and only binary *proactive* or *non-proactive* labels are kept.

Proactivity Labels

As previously discussed, the proactivity label is used because it is relevant for the application for Red Data. In rounds 4 and 5, the binary categories of *proactive* and *non-proactive* performed rather well, scoring up to a Kappa score of substantial agreement (0.648). This suggests that the annotators were sufficiently trained to label the training and test data for this category.

Emotion Labels

In round 3, the Kappa score was still too low, but some of the emotion categories are still relevant for application. To increase the Kappa score, the method of merging overlapping labels proposed by Mohammad et al. (2015) was used. This reduced the number of labels and were more easily distinguishable. We decided to define the labels in terms of 'polarity' and 'time'. For the positive meaning, there were two sets of labels: *admiration*, which refers to either the present or the past and *optimism*, which refers to the future. For the negative meaning, there were also two sets of labels: *disappointment* which refers to the present or the past and *scepticism* which refers to the future. Lastly, the label of *no emotion* represents neutrality. An additional label is included for those cases in which in none of the proposed labels apply, that is, *other*. The final emotion labels are: *admiration, optimism, disappointment, scepticism, no emotion* and *other*. The guidelines were more clearly defined and the annotators understood the differences

better. This was shown in the Kappa score, scoring up to moderate agreement (0.455) as presented in table 3.4.

Table 3.4: Agreement Scores for the labels of rounds 3, 4 & 5

Label Category + Annotator Pairs	**Round 3 Cohen's Kappa**	**Round 3 Agreement**	**Round 4 Cohen's Kappa**	**Round 4 Agreement**	**Round 5 Cohen's Kappa**	**Round 5 Agreement**
Emotion AB	0.3	41.7%	0.327	47.9%	0.427	55.1%
Emotion AC	0.191	31.3%	0.25	41.7%	**0.455**	**57.1%**
Emotion BC	0.26	37.5%	0.393	56.3%	0.413	55.1%
Purpose AB	0.266	43.8%	-	-	-	-
Purpose AC	0.319	47.9%	-	-	-	-
Purpose BC	0.306	47.9%	-	-	-	-
Proactivity AB	-	-	0.386	68.8%	**0.648**	**87.8%**
Proactivity AC	-	-	**0.589**	**79.2%**	0.481	79.6%
Proactivity BC	-	-	0.353	68.8%	0.169	67.3%

3.3 Inter-Annotator Agreement of the Training & Test Data

In this section, the final labels used for the training and test tweets is discussed. Firstly, the number of tweets annotated by each annotator is presented. Secondly, the categories used for system input are discussed. Lastly, the making of the gold data is explained.

A total of 1097 tweets were triple-annotated between at least 3 Dutch-speaking annotators. The tweets were annotated three times because inter-annotator agreement had to be calculated in order to evaluate the quality of the labels and create appropriate gold data. There are three classes of labels: multi-emotion labels (*admiration, optimism, scepticism, disappointment, no emotion* and *other)* ; binary labels (*proactive* or *non-proactive*) ; and polarity labels (*positive, negative* and *neutral*). The creation of the new polarity labels will be discussed in this section.

Due to time constraints, only two people were able to annotate 1096 tweets and the third annotated batch was composed by 4 annotators, with each annotating 274 different tweets. This is assumed to be okay to do because the annotators were sufficiently trained after the agreement study. In table 3.5, you can see that annotators A and B annotated the same batch X twice (1096 tweets each); annotator C did 274 tweets of batch X; annotator D did the next 274 tweets of batch X; annotator E the next; annotator F the next. The result is the same 1096 tweets annotated three times: the first time by one person, the second time by another person and the third time by 4 people.

Table 3.5: Number of tweets annotated by each annotator

Annotator	Batch ID	Number of annotated tweets by single annotator
A	X	1096
B	X	1096
C	X	274
D	X	274
E	X	274
F	X	274

Emotion Labels

The Kappa score of the *emotion* reached moderate agreement (0.415) and the score was relatively balanced between each annotator (0.405, 0.416 and 0.415) and 54.1% percentage agreement (see table 3.6). Annotators seemed to agree more often with the meaning of polarity and less so with the meaning of time. This is illustrated in the bold numbers in the confusion matrix 3.7 below. Between annotators A and C, it is often the case that labels where annotators do not agree, the labels referring to the past or present overlap with the labels referring to the future. In other words, only the polarity meaning of the tweet is correctly captured. For instance, *disappointment* for annotator C tends to overlap for annotator A 49 times with the label *scepticism*. The label of *admiration* for annotator C overlaps 47 times with the label *optimism* for annotator A. A possible argument is that time may be difficult to identify in the tweets. For example, in some of the longer tweets, both past and future were used. In the example tweet below, the first two sentences refers to the future, as it uses future verbal cues such as "*gaat nog zeker*" ["will certainly"], thus this part would be labelled as *optimism*. However, the last two sentences describe the present using present verbal cues such as "*mooi om te zien*" ["it is nice to see"]. The last two sentences would be labelled as *admiration*. This illustrates that it is difficult to classify a longer tweet with one label in terms of time. In this case, a decision between *optimism* and *admiration* needs to be made.

- *"Nederland gaat nog zeker tot 28 april door met verstandig afstand tot elkaar houden. En de rechtspraak sluit daarbij aan. Naast urgente zaken, worden steeds meer andere zaken via de digitale weg behandeld. Mooi om te zien dat ook in crisistijd de rechtspraak blijft functioneren." /* ["*The Netherlands will certainly continue to keep a distance from each other until 28 April. And the case law is consistent with this. In addition to urgent matters, more and more other matters are handled digitally. It is nice to see that the judiciary continues to function even in times of crisis.*"]

Polarity Labels

Since the meaning of polarity seemed to be easier to detect for the annotators, more than the meaning of time in the emotion category, we decided to implement an additional category of basic polarity (*positive, negative* and *neutral*). This means all the *admiration* and *optimism* labels, were replaced with the label *positive*; the labels that were *disappointment* and *scepticism* were replaced with the label *negative* and all the labels that were *no emotion* and *other* were replaced with the label *neutral*. Although the label *other* may have contained some emotion in the tweet, the annotators very rarely used that label therefore, it was decided to merge it with the *neutral* label. Table 3.6 illustrates that the Kappa score for polarity is higher, reaching almost substantial agreement with a kappa score of 0.561 and a percentage agreement of 70.6%. Although this is not perfect agreement, the system trained on the data with polarity labels outperformed the system containing the multi-emotion labels (this is discussed further in chapter 6).

Proactivity Labels

Lastly, the binary labels of *proactive* and *non-proactive* drastically decreased in with a Kappa score of '0.314', yet a maximum percentage agreement score of 73% (see table 3.6). Although the agreement score is rather high, the Kappa score depicts the amount of agreement that has occurred not due to chance/luck. This result is rather paradoxical because in the trial annotation round number 5, the kappa score is substantial (0.648). It is difficult to depict the reason for this, but this factor will be considered when discussing the results in chapters 5 and 6.

Table 3.6: Agreement results in 1096 annotated tweets for training & test data

Label category	**Annotator Pair**	**Cohen's Kappa**	**Percent Agreement**
Emotion	A & B	0.405	53.2%
Emotion	A & C	**0.416**	**54.1%**
Emotion	B & C	0.415	54.1%
Proactivity	A & B	0.26	67%
Proactivity	A & C	0.253	73%
Proactivity	B & C	**0.314**	**67.9%**
Polarity	A & B	0.513	68.1%
Polarity	A & C	0.467	63.5%
Polarity	B & C	**0.561**	**70.6%**

Table 3.7: Confusion Matrix of the annotated training & test data for emotion labels between annotators A and C

	C				
A	admiration	disappointment	no emotion	optimism	scepticism
admiration	155	14	24	**26**	4
disappointment	15	135	24	7	**19**
no emotion	34		176	21	16
optimism	**47**	3	16	39	1
other	16	9	28	8	6
scepticism	13	**49**	13	5	33

3.3.1 Making The Gold Data

The Kappa score did not reach substantial agreement for emotion (0.416) labels and polarity (0.561) labels. To make the data more "correct" and transform it into gold data, we decided that the tweets where at least 2 out of 3 annotators used the same label is used for training and test data. If 3 out of 3 annotators labelled a tweet with 3 different labels, then the tweet is discarded from the training data. This method of transforming the annotations to gold data might result in labels that are more accurate and it is also a method used in other research (Mohammad et al., 2015). Some of the tweets where there was no agreement between annotators was used in the test data, in order to have a balance between 'easier' and 'difficult' tweets for the classifiers. Although there was a loss in some of the training data, it can be argued that losing

some training data but having well-labelled data is better than having more data with incorrect labels.

To conclude this chapter, several annotation rounds took place in order to achieve the highest possible inter-annotator agreement and train the annotators to acquire well-labelled training data. Based on the Kappa scores, it was decided that the most appropriate labels to implement for this study were three different classes: (1) the multiple emotion labels that differentiate between polarity and time (*admiration*, *optimism, disappointment, scepticism, no emotion, other*); (2) the polarity labels by merging and replacing the previously mentioned labels into *positive, negative, neutral*; and lastly, (3) the binary purpose labels of *proactive, non-proactive*. Lastly, tweets where at least 2 out of 3 annotators used the same label is used as the gold data, ready for training and testing. The next chapter discusses the methodology.

Chapter 4

Methodology

In this chapter, the methods used to create the classification systems are discussed. Section 4.1 discusses the gold labels generated after the agreement study, which are splitted into training and test data. Section 4.2 explains the preprocessing of the tweets ready for both the annotators and system input. Section 4.3 delves into the system architectures: The Baseline SVM and CNN-Bi-LSTM. This section also discusses the experimental process. That is, three experiments per system architecture due to having three different category labels. Section 4.4 discusses the way in which the systems will be evaluated. Lastly, section 4.5 illustrates the proposed system pipeline for the company Red Data.

4.1 Training and Test Data

The data for system input will be split into training and validation data. Training data is used to let the system learn from the data. Based on what it has learnt, the unseen validation or test data is implemented into the machine learning algorithm to evaluate how well it performed after training (Gonfalonieri, 2019). There is a total of 787 annotated training tweets for the emotion category; 897 annotated training tweets for the purpose category; and 899 annotated training tweets for the polarity category. Around 20% of the rest of the annotated tweets per label category is used as unseen test data. Tables 4.1, 4.2, 4.3 below display a count of the number of labels and the grand total number of labelled tweets in the training data. As illustrated, there is an imbalance in the number of labels in the emotion category (table 4.1) and in the purpose category 4.2. Due to having less than 1000 annotated tweets suitable for training, it was not possible to balance the number of labels by removing the tweets where a label was used often. Otherwise, the amount of data would be insufficient for the systems to learn. Fortunately, the polarity category has a more balanced proportion of labels (see table 4.4). The proportion of labels used in the training is something that is taken into consideration during the experiments and in the evaluation of systems in later chapters.

Due to time constraints, it was not possible to annotate data for development data. Development data is used to fine-tune the machine learning system to prevent over-fitting, under-fitting and improve its performance (Gonfalonieri, 2019). Therefore, the experimentation phase does not have an emphasis on fine-tuning, but rather on the way it performs on two separate systems using different categories of labels.

Table 4.1: Number of Tweets and Labels in the Training Data for the Emotion Classifier

Label name	Training Labels	Test Labels
no emotion	307	64
admiration	195	39
disappointment	171	48
optimism	53	21
scepticism	35	18
other	26	3
Total	787	193

Table 4.2: Number of Tweets and Labels in the Training Data for the Proactivity Classifier

Label name	Training Labels	Test Labels
non-proactive	658	146
proactive	239	54
Total	897	200

Table 4.3: Number of Tweets and Labels in the Training Data for the Polarity Classifier

Label name	Training Label	Test Label
neutral	362	69
positive	269	67
negative	268	62
Total	899	198

4.2 Data Preprocessing

Preprocessing was necessary for both the annotators and system input. For the annotators, filtering of the tweets that were not written in Dutch was done using the langdetect[1] package. A lot of the tweets were written in English and in other languages. Since the system only works with Dutch data, tweets written in other languages had to be removed. Secondly, the author's username of the tweet was removed using regular expressions [2]. This was done because the presence of the author's username may influence the way in which the annotators label the tweet. To make the labeling process as objective as possible, the annotator has to label the content of the tweet without being influenced by the identity of the author and put their world knowledge behind. Thirdly, tweets that contained less than 25 characters were also filtered. Based on discussions that took place with the annotators during the agreement study (Chapter 3), it appeared to be often difficult for the annotators to identify the specific emotion in shorter tweets. Furthermore, cleaning up Unicode characters, emojis and URL's was also done as these get in the way of the written tweet and are unnecessary information for system input.

For system input, additional preprocessing was necessary. First of all, all of the stopwords had to be removed. Stopwords are the most common words used in language. These are words such as "the", "is", "in", etc (Singha, 2020). For text classification,

[1]Link to langdetect site: https://pypi.org/project/langdetect/
[2]Link to regular expressions site: https://docs.python.org/3/library/re.html

particularly emotion and purpose classification, these words do not seem to add any valuable information to the system's algorithm. The large amount of stopwords in the data can become noisy and can negatively interfere with the system's training. The removal of stopwords may improve system performance by focusing on important content words that can help identify the emotion or purpose of a tweet. Moreover, stopword removal can decrease the time it takes for a system to learn during training (Singha, 2020). Stopword filtering was implemented using the NLTK stopword identifier for Dutch[3]. Just like stopwords, punctuation can also be noisy in the data. Punctuation was filtered by creating a list of all possible punctuation and eliminating these using regular expressions. Additionally, all of the words were transformed to lower case. This is because the system might treat two words that have the same meaning differently (e.g. Cat, cat). Lastly, a lot of tweets are repeated due to the presence of "retweets". The repetition of tweets was removed by creating a python set of all the tweets.

After all of the unwanted items in the data were filtered, the data was ready to be transformed into a list of nested lists, where each nested list represents a sequence of word tokens. A sequence of word tokens represents a tweet. This is done for two reasons. Firstly, it is important to create a format that represents a tweet. In this case, the system can recognize a sequence of words of an ordered list as being a single tweet and can thus assign a label to it. Secondly, this sequence format is needed because these are transformed into vector numerical representations so that a tweet is readable for the system. These vector representations are the input of the SVM baseline and the CNN-BiLSTM. The vector inputs are slightly different in both systems and this is discussed further in section 4.3. The example below demonstrates the way a tweet originally looks before preprocessing (the original tweet) and after preprocessing (a sequence of word tokens):

1. **Tweet before preprocessing:** "@ZihniOzdil : Column: De racistische, politieke terreur van Kuzu en Öztürk"

2. **Tweet after preprocessing:** ['column', 'racistische', 'politieke', 'terreur', 'kuzu', 'öztürk']

4.3 System Architecture

In this section, the machine learning architectures are discussed. Firstly, 4.3.1 discusses the baseline system, which is a classical machine learning Support Vector Machines (SVM) classifier. Secondly, 4.3.2 is about the hybrid Deep Learning system, a Convolutional Neural Network Bidirectional Long-Short-Term-Memory classifier (CNN-BiLSTM). Thirdly, section 4.3.3 explains the 6 experiments that are performed, that is 3 experiments for the SVM classifier and 3 experiments for the CNN-BiLSTM classifier.

4.3.1 Baseline SVM

Given that Neural Networks are currently state-of-the-art, particularly, CNN-BiLSTM for text classification problems such as emotion detection, (Liu et al., 2020; Ge et al., 2019; Alswaidan and Menai, 2020), it was decided to implement a baseline system would be used for system comparison of the system discussed in 4.3.2. A baseline is used as a "point of comparison for the more advanced methods that are evaluated

[3]NLTK documentation: https://www.nltk.org

later" (Brownlee, 2020). Therefore, the implementation of a baseline is relevant and important.

Recent studies where state-of-the-art neural network systems are implemented, use baseline SVM's for system comparison (Alswaidan and Menai, 2020). The motivation for implementing an SVM as baseline is that it is known to have provided state-of-the-art results in past years and it is simple to use through *Sklearn* [4]. As discussed in Chapter 2, SVM works well when there is relatively few annotated data and that is the case for this study (Alswaidan and Menai, 2020; Kirange and Deshmukh, 2012). As a result, it was decided that a basic SVM is an appropriate baseline method to implement.

An SVM is a discriminative classification model that uses a hyperplane that separates datapoints with their corresponding label. In other words, given the labelled training data, the SVM algorithm "outputs an optimal hyperplane which categorizes" new, unseen data (Patel, 2017). For the multiple emotion labels and polarity labels of the training data, the SVM will classify the tweets using multiple hyperplanes because we are dealing with more than 2-dimensions (more than 2 labels). Whereas for the binary purpose labels of 'proactive'/'non-proactive', a single hyperplane will classify this 2-dimensional dataset.

The input of the system is the preprocessed tweets (discussed in section 4.2) which are transformed into machine-readable vector representations. Representing words into numbers is a key aspect in Text Mining and Sklearn provides code to transform a textual dataset into a vector representation for system input. In this case, the input of the SVM will be the Tf-idf Vectorizer, which stands for Term Frequency-Inverse Document Frequency. The Tf-idf vectorization not only places the words into a matrix, but it calculates how relevant/often a word appears in a document, text or sentence (Stecanella, 2019). For example, tweets with similar words will have similar vector representations and the SVM algorithm can predict patterns with these similarities (Stecanella, 2019).

The SVM is being trained whilst fitting x (the vectorized tweets) and y (the labels), and once it has gone through all the training dataset, the training is finalized. To evaluate the system performance, the trained SVM predicts the labels of unseen validation data which will be later compared with the gold annotations to calculate Precision, Recall and F1-score; and an error analysis will be provided. The evaluation of the systems is discussed in detail in section 4.4.

4.3.2 CNN-BiLSTM Architecture

The state-of-the-art system architecture used in this thesis is a hybrid neural network: a CNN-BiLSTM. It is known as a 'hybrid' because it contains two different types of Neural Netoworks: A Convolutional Neural Network (CNN) and a Bidirectional Long-Short-Term-Memory Neural Network (BiLSTM) which is a type of Recurrent Neural Network (RNN). The architecture is selected because as discussed in the literature review in Chapter 2, studies have shown that it has state-of-the-art results and outperforms single CNNs and single LSTMs (Ge et al., 2019; Liu et al., 2020; Alswaidan and Menai, 2020).

Although this is discussed in greater depth insection 2.1.3 of Chapter 2, here I provide a reminder of the main advantages of implementing this hybrid neural network.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.htmlsklearn.svm.LinearSVC

Firstly, using a BiLSTM instead of a traditional RNN or LSTM, solves the problem of exploding gradients (Liu et al., 2020). Secondly, BiLSTMs are bidirectional and take into consideration the relationships and semantics before and after each word (Liu et al., 2020). Thirdly, by combining a CNN with a BiLSTM, the classifier benefits from extracting local features by the CNN and global features by the BiLSTM (Liu et al., 2020; Ge et al., 2019). This solves the issue of the CNN not considering the semantics of words because it does not have a memory cell like the BiLSTM (Liu et al., 2020). By merging both local and global features, it can result in good accuracy when classifying text (Liu et al., 2020).

An important aspect of the CNN-BiLSTM is the use of pre-trained word embeddings. Word embeddings are vector representation of words that represent semantic information by putting together words that are similar into a vector space; or by separating words that are less similar from each other in that same vector space (Goldberg, 2016). The use of word embeddings will give the system more features or words that may not possibly appear in the training data, thus allowing the system to perform better on unseen data and generalise well (Goldberg, 2016).

The word embeddings used for this system are open-source pre-trained Dutch word embeddings trained using a Word2Vec model on a large Dutch corpus in social media [5] (Nieuwenhuisje, 2018b). More specifically, the model has been trained using around 600 million individual Dutch social media messages and 36 million messages of Dutch news, blogs and fora posts (Nieuwenhuisje, 2018a). The corpus dates from 01/01/2017 up to 31/12/2017 (Nieuwenhuisje, 2018b). Moreover, the embeddings are trained on a lot of political tweets and conversations, which is highly relevant for the political tweets to work with. In figure 4.1, a query is shown where a cluster of the terms that are most similar to the political party 'vvd' is provided. The term 'pvda' is most similar to the term 'vvd' with a distance of almost 0.9 and 'vvder' is a little less similar with a distance of 0.75 Additionally, the word embeddings contain Dutch slang, jargon and hashtags. Social media such as Twitter uses informal language and most of the hashtags can be treated as significant content words.

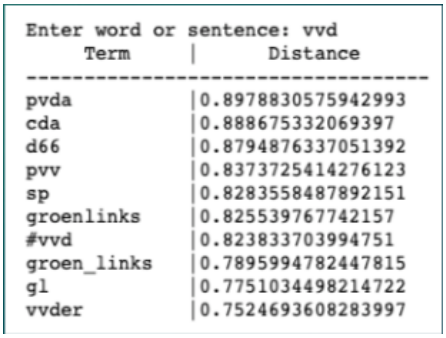

Figure 4.1: Query of the word 'vvd' (a political party in The Netherlands)

Now that the CNN-BiLSTM has been introduced and the word embeddings to use have been discussed, the system architecture will be discussed. To build the code of this system, Keras with Tensorflow backend is used. The code is therefore inspired by an example code provided by Keras (Keras, 2020) [6] as well as the CNN-BiLSTM created

[5] Github page access to word embeddings: https://github.com/coosto/dutch-word-embeddings
[6] Code example CNN-BiLSTM: https://keras.io/examples/nlp/bidirectional$_lstm_imdb$/

by (Liu et al., 2020). Using Keras, the 'sequential' module is used, which allows you to easily add layers to create the neural network. Below, a step-by-step guide is provided which can be followed by looking at figure 4.2 to see how the system works.

1. The first input of the system is a sequence of word tokens x. Figure 4.2 shows the following sequence as input: ['samen', 'krijgen', 'we', 'corona', 'onder', 'controle']. This sequence of tokens represents one tweet. Before the system can "read" this tweet, two steps need to be done. Firstly, the tweets need to be placed into a vector space and padded. Padding is necessary because not every sentence has the same number of word tokens but every tweet needs to have the same dimension. Therefore, those tweets that are shorter will contain zeros to maintain the same vector size across every tweet. The code for this is inspired by (Dandge, 2019). Secondly, the padded and vectorized tweets need to be linked with the word embeddings. The pre-trained Dutch word embeddings to be used are 300 dimensions. The code for loading and linking the word embeddings is inspired by (Dandge, 2019; Kaggle, 2017). The linking is done by indexing every word token of the training data with the pre-trained word embeddings. As for the labels y, these had to be encoded into numerical values using *Sklearn* built-in function *Label Encoder*[7].

2. Once the tweets are vectorized and linked to the word embeddings, the first layer of the architecture is added, that is the embedding layer.

3. The next layer is the CNN layer which is a 1D convolution. In this layer, the maxpooling operation takes place. Here, the local features are extracted and the maxpooling takes the most important features and eliminates the redundant features (Liu et al., 2020). This is followed by a dropout layer which drops some of the learnt features. A dropout layer is implemented as a method to prevent overfitting. The dropout layer is assumed to be necessary here because there is few training data, which is an aspect that may lead to overfitting.

4. The next layer is the BilSTM. The input is the output of the CNN layer. The LSTM contains two hidden layers. The input of these layers is the bidirectional sequence of the tweet, which at this point is represented as the important features learnt by the CNN. The memory cell of the LSTM 'remembers' the features learnt of both directions. Finally, the output of the hidden layers are joined together to get the output of the BiLSTM.

5. Lastly, the classification takes place with the use of the Softmax activation function, which is the probability of classifying x into class y given the features learnt (Liu et al., 2020). According to the Tensorflow documentation, the Softmax activation function is appropriate for categorical probabilities (Katariya, 2020).

6. As an additional step after training, we want to predict the labels for unseen tweets of the test data and use these for the evaluation. To do that, it was necessary to implement the Keras function *model.predict()* and convert the encoded labels back to their original label using the *Sklearn* built-in function *inverse.transform()*. The predicted labels are concatenated with their corresponding tweet and ready for evaluation.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

Parameters: As mentioned previously, there is no development data in this study. The study focuses on creating appropriate labels rather than on fine-tuning. As a result, the parameter values used are the default values from the Keras documentation (Keras, 2020).

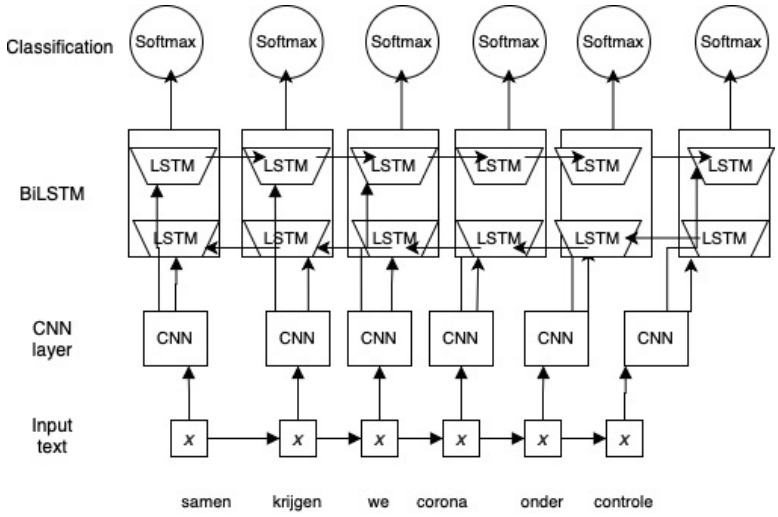

Figure 4.2: CNN-BiLSTM Architecture

4.3.3 Experiments Performed

Given that this thesis has a strong focus on investigating the labels that work best for classifying tweets as well as which model works best with this data, the following experiments are appropriate. 6 experiments were performed: 3 experiments for the SVM baseline and 3 experiments for the CNN-BiLSTM. In other words, the same two models were trained using 3 different types of labels as shown in the list below:

1. **Experiment 1:** Baseline SVM with training data annotated with 6 emotion labels: *optimism, admiration , scepticism, disappointment, no emotion* or *other.*

2. **Experiment 2:** Baseline SVM with training data annotated with 2 purpose labels: *proactive* or *non-proactive.*

3. **Experiment 3:** Baseline SVM with training data annotated with 3 polarity labels: *positive, negative* or *neutral.*

4. **Experiment 4:** CNN-BiLSTM with training data annotated with 6 emotion labels: *optimism, admiration* , *scepticism*, *disappointment*, *no emotion* or *other.*

5. **Experiment 5:** CNN-BiLSTM with training data annotated with 2 purpose labels: *proactive* or *non-proactive.*

6. **Experiment 6:** CNN-BiLSTM with training data annotated with 3 polarity labels: *positive, negative* or *neutral.*

4.4 Evaluation of Systems

Two main methods are used to evaluate the performance of the two systems: Calculating Precision, Recall and f1-score, and an error analysis.

Precision, recall and f1-score are conventional methods used in automatic text classification, to compare the gold data with the system output. This method has been used in most of the literature discussed in Chapter 2. Some studies make use of the percentage accuracy measure as a form of evaluation however, this measure does not take into consideration false classification(Shung, 2020). Precision and Recall identify the true positives, true negatives, false positives and false negatives, thus giving us a more reliable measure of system performance (Shung, 2020). The f1-score is the weighted average between precision and recall (Shung, 2020). Precision, Recall and f1-score are the primary measures used for evaluation however, percentage accuracy will also be displayed to provide an additional viewpoint. To aid to the precision, recall and f1-measures confusion matrices will be presented to compare the individual labels between system ouput and gold. Furthermore, an error analysis will be carried out by looking at some of the labelled tweets by the system to check for errors. Lastly, the discussion of results will take into consideration the agreement study result of 3, the results in chapter 5, the confusion matrices and error analyses in chapter 6.

4.5 Final Pipeline for Red Data

For the purpose of producing a product for Red Data, the pipeline illustrated in figure 4.3 shown below has been created. First, the CSV of the tweets are preprocessed. The preprocessed data enters two classifiers: the multi-label classifier and the binary classifier. Lastly, the polarity label and the proactivity labels are concatenated with the original tweet, and these are placed into a JSON dictionary file. The python script containing the system is made readily available to Red Data. The classifier used in this pipeline is the SVM classifier for polarity labels as it performed the best. However, it still needs a lot of improvement. Moreover, the SVM classifier for proactivity labels is also implemented because it is a label that is interesting to the company and is relevant with the domain of the tweets. However, the performance of this system is poor but there are possibilities for improvement. The performance of the systems is discussed further in chapters 5 and 6. Although the systems need improvement, the proposed pipeline is an appropriate product for Red Data in future work.

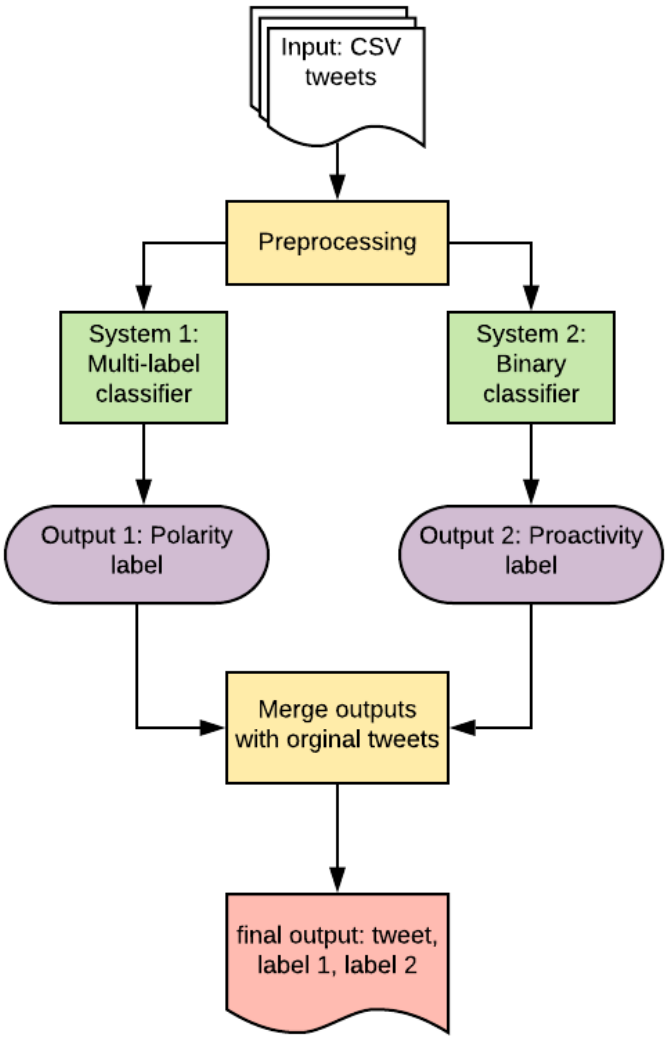

Figure 4.3: Final Pipeline product delivered for Red Data containing two classification systems

To conclude this chapter, two classifiers with three different classes of labels will be implemented: A baseline SVM with the emotion, proactivity and polarity labels; and the CNN-BiLSTM, with those same 3 categories. The systems will be evaluated using Precision, Recall and F1-score with additional confusion matrices and error analyses. Lastly, although the classifiers need improvement, a pipeline has already been constructed so that Red Data can make use of two classifiers. The pipeline intends to automatically preprocess, classify tweets and output tweets with two sets of labels.

Chapter 5

Results

In this chapter, the results of the baseline SVM classifier and the CNN-BiLSTM are presented. Firstly, the overall performance of both systems is introduced. Then, the chapter is divided into the following sections: Section 5.1 displays the results of the SVM classifier, which is composed by subsections 5.1.1, 5.1.2 and 5.1.3, where the precision, recall and f1-scores of each label is presented. These metrics tell us how good or how bad the method performed on different classes. Section 5.2 displays the results of the CNN-BiLSTM classifier and is structured in the same way as the aforementioned section 5.1.

Table 5.1 below illustrates an overview of the accuracy and weighted precision, recall, f1-scores of the system outputs of 6 experiments. The baseline SVM system outperformed the CNN-BiLSTM in all label categories, scoring an f1-score of 0.40 for emotion labels, 0.68 for proactivity labels and 0.59 for polarity labels. The CNN-BiLSTM scored an f1-score of 0.29 for emotion labels, 0.64 for proactivity labels and 0.32 for polarity labels.

The label category that performed best in the SVM system is the polarity category with a slightly higher precision than recall (P=0.63, R=0.60). The binary proactivity labels seem to have a high precision (P=0.70, R=0.74), but an error analysis is necessary to judge appropriately its performance. In comparison to the SVM, the CNN-BiLSTM scored in all three categories low precision with higher recall and both systems with the emotion labels performed poorly. Generally, precision and recall performance is higher and more balanced in the SVM baseline.

Table 5.1: Table displaying weighted averages in precision, recall, f1-score and accuracy of all system experiments

System	Label category	Precision	Recall	f1-score	Accuracy
SVM	**1. emotion**	0.49	0.47	0.40	0.47
	2. proactivity	0.70	0.74	0.68	0.74
	3. polarity	**0.63**	**0.60**	**0.59**	**0.60**
CNN-BiLSTM	**4. emotion**	0.27	0.35	0.29	0.35
	5. proactivity	0.63	0.71	0.64	0.71
	6. polarity	0.32	0.34	0.32	0.34

5.1 Results of SVM Baseline

5.1.1 Experiment 1: SVM + Emotion Labels

Experiment 1 consisted of predicting the multiple emotion labels using the SVM classifier (*admiration, disappointment, no emotion, optimism, other* or *scepticism*). A total of 787 annotated tweets were used for training and 193 tweets were used for testing. In this experiment, the overall weighted f1-score is 0.51 and a system accuracy of 0.48. Looking at table 5.2 below, the precision, recall and f1-scores for each separate label is provided.

Labels with a high precision and low recall are *disappointment* (P = 0.55, R = 0.41) and *admiration* (P = 0.53, R=0.31). High precision and low recall means that although returning few results, most of the predicted values are correct. This is the same for the label of *optimism* (P=1, R=0.05) however, it was only able to find one tweet as being correctly 'optimistic' and there is an imbalance between precision and recall.

Table 5.2: Results of baseline SVM with multi-emotion labels

	precision	**recall**	**f1-score**	**support**
admiration	0.53	0.41	0.46	39
disappointment	0.55	0.38	0.44	48
no emotion	0.43	0.86	0.57	64
optimism	1.00	0.05	0.09	21
other	0.000	0.000	0.000	3
scepticism	0.000	0.000	0.000	18
accuracy			0.47	193
macro avg	0.42	0.28	0.26	193
weighted avg	0.49	0.47	0.40	193

5.1.2 Experiment 2: SVM + Proactivity Labels

Experiment 2 consisted of predicting the binary proactivity labels using the SVM classifier (*proactive* or $non - proactive$). A total of 897 annotated tweets were used for training and 200 tweets were used for testing. In this experiment, the overall weighted f1-score is 0.68 and a system accuracy of 0.74. Table 5.3 shows that *Non-proactive* labels has a very high scores in precision and recall (P=0.76, R=0.95) whereas the *proactive* labels have low precision and recall (P=0.56, R=0.17).

Table 5.3: Results of baseline SVM with binary 'proactivity' labels

	precision	**recall**	**f1-score**	**support**
non-proactive	0.76	0.95	0.84	146
proactive	0.56	0.17	0.26	54
accuracy			0.74	200
macro avg	0.66	0.56	0.55	200
weighted avg	0.70	0.74	0.68	200

5.1.3 Experiment 3: SVM + Polarity Labels

Experiment 3 consisted of predicting the polarity labels using the SVM classifier (*positive*, *negative* or *neutral*). A total of 899 annotated tweets were used for training and 198 tweets were used for testing. In this experiment, the overall weighted f1-score is 0.59 and a system accuracy of 0.60. The two labels that performed best in terms of high precision and low recall are *positive* (P=0.76, R=0.50) and *negative* (P=0.61, R=0.49). Conversely, the neutral label had a low precision and high recall (P=0.52, R=0.78). This suggests that the system found it more difficult to label correctly *no emotion* as opposed to the other two categories. In terms of precision and recall, this experiment displays the best results of all.

Table 5.4: Results of baseline SVM with polarity labels

	precision	**recall**	**f1-score**	**support**
negative	0.61	0.49	0.55	67
neutral	0.52	0.78	0.63	69
positive	0.76	0.50	0.60	62
accuracy			0.60	198
macro avg	0.63	0.59	0.59	198
weighted avg	0.63	0.60	0.59	198

5.2 Results of CNN-BiLSTM

5.2.1 Experiment 4: CNN-BiLSTM + Emotion Labels

Experiment 4 consisted of predicting the emotion labels using the CNN-BiLSTM classifier. The same training and test data was used as for the SVM classifier. In this experiment, the overall weighted f1-score is 0.29 and a system accuracy of 0.35. These results show that this was the lowest-performing system in comparison to the other 5 experiments. The two labels with a high precision and low recall are *disappointment* (P=0.33, R=27) and *admiration* (P=0.27, R=0.23). *No emotion* had a low precision and high recall (P=0.40, R=0.72). Lastly, the last three labels that scored 0 are *optimism*, *other* and *scepticism*.

Table 5.5: Results of CNN-BiLSTM with multi-emotion labels

	precision	**recall**	**f1-score**	**support**
admiration	0.27	0.23	0.25	39
disappointment	0.33	0.27	0.30	48
no emotion	0.40	0.72	0.51	64
optimism	0.000	0.000	0.000	21
other	0.000	0.000	0.000	3
scepticism	0.000	0.000	0.000	18
accuracy			0.35	193
macro avg	0.17	0.20	0.18	193
weighted avg	0.27	0.35	0.29	193

5.2.2 Experiment 5: CNN-BiLSTM + Proactivity Labels

Experiment 5 consisted of predicting the proactivity labels using the CNN-BiLSTM classifier. The same training and test data was used as for the SVM classifier. In this experiment, the overall weighted f1-score is 0.64 and a system accuracy of 0.71. The *proactive* label scored a higher precision than recall but is still considered low (P=0.36, R=0.09). The *non-proactive* label scored a singificantly higher precision score but had an even higher recall (P=0.74, R=0.94).

Table 5.6: Results of CNN-BiLSTM with binary 'proactivity' labels

	precision	**recall**	**f1-score**	**support**
non-proactive	0.74	0.94	0.83	146
proactive	0.36	0.09	0.15	54
accuracy			0.71	200
macro avg	0.55	0.52	0.49	200
weighted avg	0.63	0.71	0.64	200

5.2.3 Experiment 6: CNN-BiLSTM + Polarity Labels

The final experiment number 6, consisted of predicting the polarity labels using the CNN-BiLSTM classifier. The same training and test data was used as for the SVM classifier. In this experiment, the overall weighted f1-score is 0.32 and a system accuracy of 0.34. This is considered to be very low . The labels with a higher precision and lower recall are negative (P=0.30, R=0.25) and positive (P=0.28, R=0.16). Conversely, the neutral label had a low precision and high recall (P=0.38, R=0.58).

Table 5.7: Results of CNN-BiLSTM with polarity labels

	precision	**recall**	**f1-score**	**support**
negative	0.30	0.25	0.28	67
neutral	0.38	0.58	0.46	69
positive	0.28	0.16	0.20	62
accuracy			0.34	198
macro avg	0.32	0.33	0.31	198
weighted avg	0.32	0.34	0.32	198

To conclude this chapter, the SVM baseline performed better than the CNN-BiLSTM in all three classes of labels. The SVM showed reasonable results in all three polarity labels. The emotion labels performed poorly, especially the labels of *scepticism*, *optimism* and *other.* For the binary proactivity classifier, the system generally performed poorly as the *proactive* label had a poor result but the *non-proactive* label performed very well. Finally, the CNN-BiLSTM performed poorly in all categories.

Chapter 6

Discussion

This chapter analyses the results and addresses the implications of these results. In section 6.1, the confusion matrices of the 6 experiments are illustrated and described. In section 6.2, an error analysis on the SVM with polarity labels is performed. Lastly, in section 6.3, a final discussion is presented where ideas for future work are also considered.

6.1 Confusion Matrices of System & Gold Labels

In this section, the confusion tables from all 6 experiments that compare system labels against the gold labels are presented and described. These tables display patterns which can indicate where the errors originate from.

6.1.1 Confusion Matrix 1: SVM + Emotion Labels

The confusion table in figure 6.1 shows us that *no emotion* (P=0.43, R=0.86) is the category that was used the most by the SVM classifier, but in most cases it was not used correctly. For example, the classifier labelled tweets with the *admiration* label tweets as *no emotion* 19 times and it labelled tweets with *disappointment* label as *no emotion* 29 times. This indicates that the SVM failed to distinguish *no emotion*, from both the labels *admiration* and *disappointment*. This indicates that there is some bias to the *no emotion* label. Furthermore, the SVM also confused the label of *scepticism* with the label of *disappointment* 7 times and *scepticism* was not used by the classifier (P=0, R=0). The matrix also indicates that there is less confusion between *admiration* and *disappointment*.

Overall, the results show that the SVM performed poorly with the emotion labels.

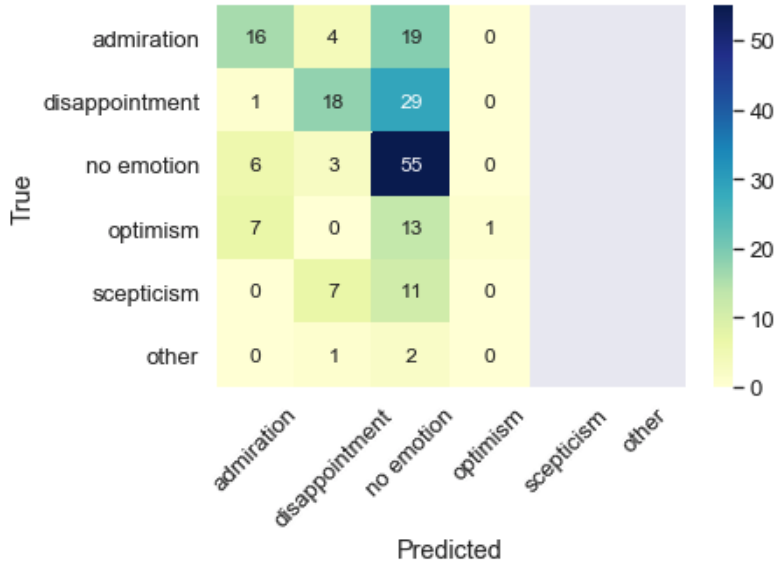

Figure 6.1: Confusion Matrix Comparing SVM Predicted Emotion Labels (x axis) with the True/Gold Labels (y axis)

6.1.2 Confusion Matrix 2: SVM + Proactivity Labels

The confusion table in figure 6.2 shows that there is a bias for the *non-proactive* label (P=0.76, 0.96). The SVM system made use of this label on almost every tweet of the test data and it had 45 errors, as those tweets should have been labelled as *proactive*. As a result, the *proactive* label (P=0.56, R=0.17) was almost never used and therefore we cannot conclude on the quality of this label.

The clear bias towards the *non-proactive* label shows that the SVM performed poorly with this binary category.

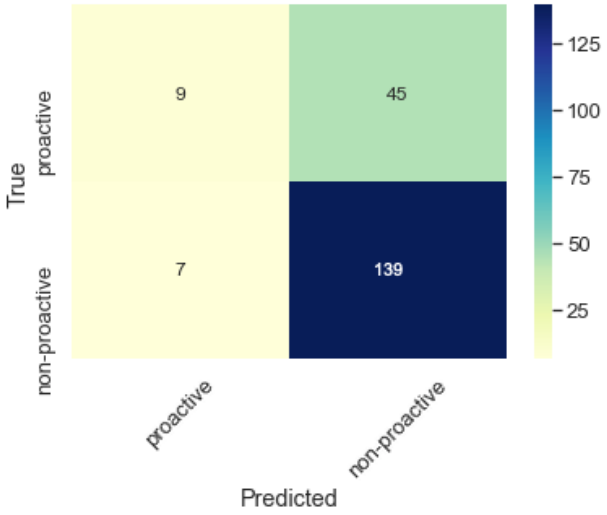

Figure 6.2: Confusion Matrix Comparing SVM Predicted Proactivity Labels (x axis) with the True/Gold Labels (y axis)

6.1.3 Confusion Matrix 3: SVM + Polarity Labels

The confusion table in figure 6.3 shows as a first observation that most of the labels labels have been used correctly and with similar frequencies, specifically the *positive* (P=0.76, R=0.50) and *negative* (P=0.61, R=0.49) labels. The *neutral* label (P=0.52, R=0.78) has been used the most and has the most errors in comparison to the other two. This indicates that there is bias towards the neutral label. There seems to be more confusion between the *negative* and the *neutral* labels , and *positive* and *neutral* labels . There is much less confusion between the *negative* and *positive* labels.

The SVM with the polarity labels performed best out of all 6 experiments.

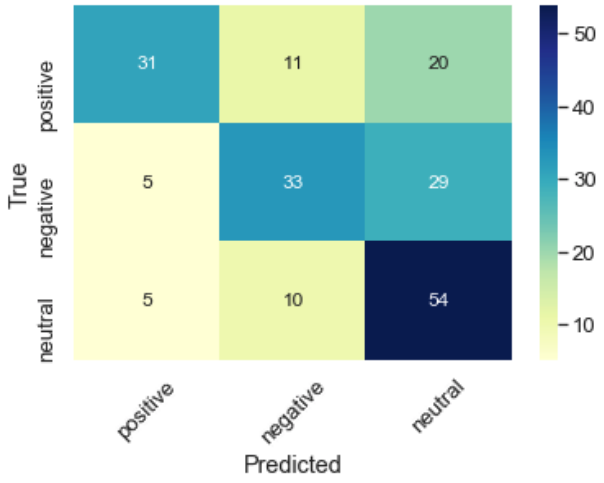

Figure 6.3: Confusion Matrix Comparing SVM Predicted Polarity Labels (x axis) with the True/Gold Labels (y axis)

6.1.4 Confusion Matrix 4: CNN-BiLSTM + Emotion Labels

The confusion table in figure 6.4 shows that there is confusion between most of the emotion labels and the *no emotion* label, specifically between the following pairs: *admiration* and *no emotion*; *disappointment* and *no emotion*; *optimism* and *no emotion*; and *scepticism* and *no emotion*. This strongly indicates a bias with the *no emotion* category. The *optimism* label is almost never present in the system output. The *scepticism* and *other* labels are not present in the system output. The *admiration*, *disappointment* and *no emotion* labels scored low precision and recall. The *optimism*, *scepticism* and *other* labels scored 0 for precision and recall.

The results demonstrate that the CNN-BiLSTM with the emotion labels performed very poorly.

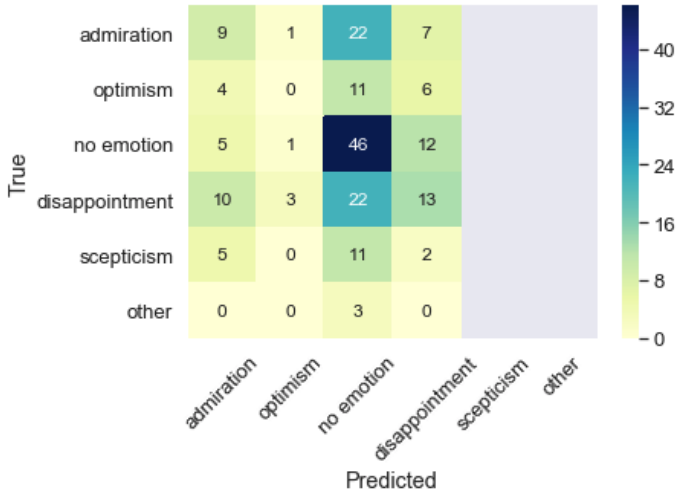

Figure 6.4: Confusion Matrix Comparing CNN-BiLSTM Predicted Emotion Labels (x axis) with the True/Gold Labels (y axis)

6.1.5 Confusion Matrix 5: CNN-BiLSTM + Proactivity Labels

The confusion table in figure 6.5 also shows a bias towards the *non-proactive* label (P=0.74, R=0.94) because that category was used by the system multiple times and has a 47 errors. The *proactive* label (P=0.36,R=0.09) is also present but it was classified correctly 7 times and incorrectly 18 times. Hence, there is clearly confusion between both labels.

The CNN-BiLSTM with proactivity labels performed poorly.

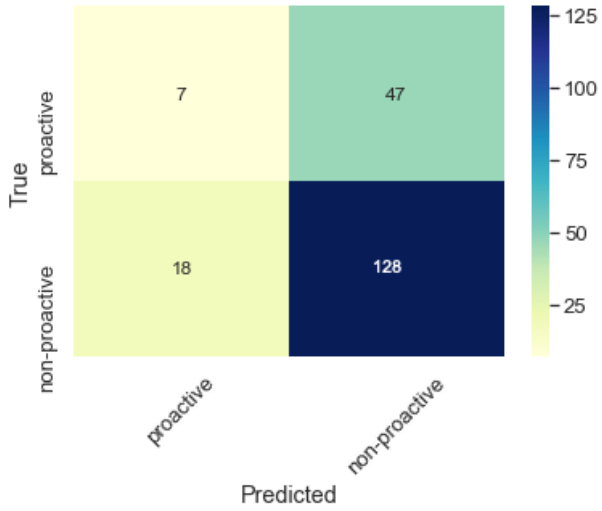

Figure 6.5: Confusion Matrix Comparing CNN-BiLSTM Predicted Proactivity Labels (x axis) with the True/Gold Labels (y axis)

6.1.6 Confusion Matrix 6: CNN-BiLSTM + Polarity Labels

Lastly, the confusion table in figure 6.6 shows a lot of confusion between neutral and negative; followed by neutral and positive. There is also confusion between the negative and positive emotions. The system does seem to pick up correctly some of the emotions from time to time, but overall, there is too much mismatching happening between all classes. The system also had low precision and recall.

As a result, the CNN-BiLSTM with polarity labels performed poorly.

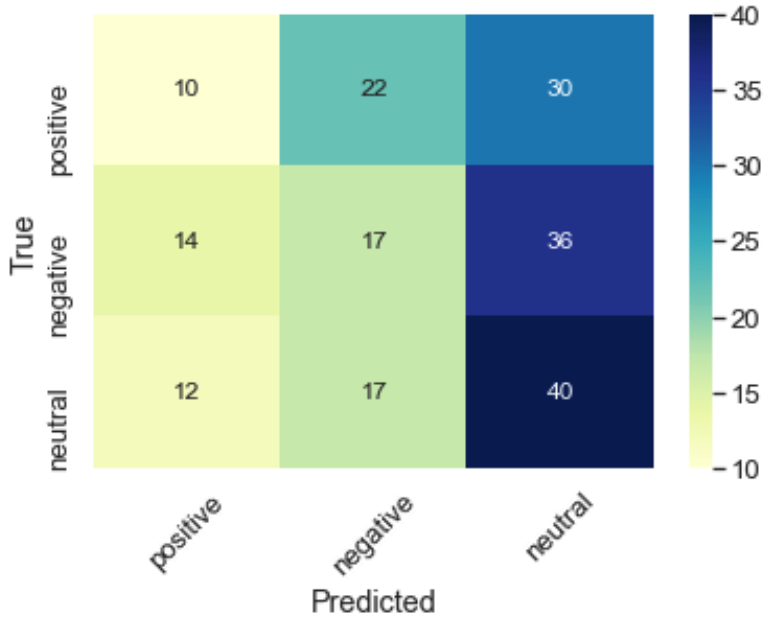

Figure 6.6: Confusion Matrix Comparing CNN-BiLSTM Predicted Polarity Labels (x axis) with the True/Gold Labels (y axis)

To conclude section 6.1, the confusion matrices reinforce that the SVM baseline with the polarity labels performed reasonably out of all experiments. The CNN-BiLSTM always performed inadequately as there is a lot of confusion between all of the used labels from the system output and the gold labels. Furthermore, although the SVM baseline is more successful in distinguishing between *positive* and *negative* tweets, the matrix in figure 6.3 shows that there is confusion between the *neutral* and *negative* labels and the *neutral* and *positive* labels.

6.2 Error Analysis

In this section an error analysis is performed in which some of the misclassified tweets will be analysed. The error analysis will focus on the errors performed by the best-performing system in this experiment: The SVM on polarity labels. The analysis will not consider the experiments with the emotion labels because there is too much confusion between most of the labels. It will also not consider the binary proactivity classifier because there is too much bias towards the non-proactive label. Providing an error analysis on systems that perform poorly will not contribute to identifying specific errors that are comparable to the correct labels. However, we see much less

confusion between *positive* and *negative* labels, indicating that the SVM can identify basic sentiment. An error analysis in this case may provide information for finding solutions to improve the system's accuracy.

As a reminder of the SVM's specific results, the polarity labels had the best precision and recall over *positive* labels (P=0.76, R=0.50) and *negative* labels (P=0.61, R=0.49). The confusion matrix previously described in figure 6.3, shows the least confusion between *positive* and *negative* labels and more confusion with the *neutral* labels.

For example, in several tweets where a positive or negative emotion is evident, the system labelled it as *neutral* e.g. "*Mooie paasboodschap van @PieterOmtzigt*" / ["*Beautiful easter message @PieterOmtzigt*"]. This tweet is explicitly positive, but it seems the SVM was unable to capture the positive cue of "Mooie". In a similar example, the following negative tweet "*Column: De racistiche, politieke terreur van Kuzu en Ozt urk" / ["Column:The racist, political terror of Kuzu and Ozturk*"] was mislabelled as *neutral*. It seems the system did not identify clear negative cues such as "racist, political terror". A possible solution for this is implementing word embeddings trained on social media or a Dutch sentiment lexicon, so that the SVM can capture semantic information that is relevant for emotion detection in these tweets. Another possibility is to implement feature engineering, by feeding the system with linguistic information that best represent the emotions.

In some cases, the system also had some confusion with labelling positive tweets when these should have been negative: "*Die rook is vast heel schoon, zo uit zo'n duurzame subsidiefabriek*"/["*That smoke is probably very clean, from such a sustainable subsidy factory*"]. Although this tweet seems to depict a positive emotion, the gold label for this tweet is *negative*. It seems the system correctly identified some positive cues such as "heel schoon" or "duurzame", indicating that the SVM is sometimes successful in identifying explicitly positive tweets but less so when the emotion is implicit as this tweet may be sarcastic. For instance, the tweet may be actually referring to a polluting factory but sarcasm is very difficult to identify. Another possible reason why annotators labelled *negative* is that the they failed to put their world knowledge and opinions behind for this particular case. Although time costly, a possible solution to this is to ask annotators to give a short reason why certain tweets are *positive* or *negative*. Another possible solution is to provide features that represent sarcasm. In this case, a possible feature for this tweet is to classify it in terms of topic. The topic is about an issue with a factory therefore, tweets associated with this topic may be most of the time negative. This may be a way of letting the SVM deal with correlating features of emotion and sarcasm. Word embeddings trained on social media from the same scope as the tweets may also provide semantic information when a topic is associated with negative or positive meaning.

6.3 Discussion

There are several important implications and challenges shown in both the results and error analysis that need to be addressed in order to answer the research question introduced Chapter 1, that is: **Is it possible to build an emotion and purpose classification system to detect the emotion and purpose written in tweets by members of the Dutch parliament?**. In section 6.3, the annotation model's challenges and possible solutions is discussed. In section 6.4, the automatic classification challenges are also discussed.

6.3.1 The Annotation Model

The main research question is composed by the following sub-question (a): **What annotation model is best to identify categories relevant for emotion and purpose classification of tweets?** In the inter-annotator agreement study in Chapter 3, there were two main goals: (1) To select emotions and purpose labels that are interesting to Red Data and representative of the tweets; (2) To train the annotators in order to achieve a sufficient inter-annotator agreement. To achieve these goals, a set of 5 trial annotation rounds were conducted to test the quality of the labels and the annotation guidelines. The labels were initially inspired by another study by Mohammad et al. (2015), where the selection of emotions and purpose labels seemed appropriate for political and electoral tweets.

The results in the first trial annotation rounds showed that the annotators had difficulties distinguishing between so many labels, especially between those labels that had similar meanings. In the last trial annotation rounds, agreement sufficiently reached between moderate and substantial agreement when the labels reduced in number and were more-clearly distinguishable. That is, differentiating emotions in terms of polarity and time; and proposing a binary classification for the purpose label which differed in whether a tweet is proactive or non-proactive. Therefore, the best annotation model to identify these categories is one that contains clearly defined guidelines, clearly distinguishable labels and the possibility to train the annotators over a period of time so that they can learn the meaning of the labels as good as possible.

The main challenge of this annotation model is that agreement results are not always guaranteed to increase after a set training time. Classifying emotions appeared to be difficult for humans and this is shown in the Kappa scores and confusion matrices in chapter3 . The final agreement results of the annotated training data decreased for the emotion categories down to fair-moderate (about 0.4 Kappa) and down to slight-fair (about 0.2 Kappa) for the binary categories. In attempt to reduce this limitation, two things were done. Firstly, 3 sets of annotations were generated for adjudication purposes. Secondly, an additional category was introduced: polarity. In other words, regardless of time, the proposed six emotions labels were merged into positive, negative or neutral. In doing so, an additional limitation was reduced: the fair distribution of annotated labels. This was a challenge in the purpose and emotion categories, where some labels occurred much more frequently than other labels. This means that the polarity labels were well balanced in the training and test data.

In answering sub-question (a), for this particular study, the best annotation model involved the following aspects: (1) combining annotator training with the inclusion of well-written guidelines and distinguishable labels; (2) selecting tweets where there was higher inter-annotator agreement; and (3), creating an additional category of polarity which reduced the number of labels.

On the other hand, the initial goal for this study was to automatically identify more complex emotions rather than simple polarity. The goal of identifying proactivity was also a challenge. This is because there is too much confusion between the emotion labels and proactivity labels and their distribution is very unbalanced. The poor distribution is what may have resulted in the bias of labels like *non-proactive* or *no emotion* in the system's output. However, labels which appeared less also suggest that some of the emotions are simply not prominent or relevant in the data. A first possible solution is to implement a new agreement study for a longer period of time, with other emotion labels and more detailed guidelines over a larger crowd. With

regards to the proactivity labels, expert linguists could annotate tweets with this category because certain linguistic features can clearly represent the difference between a proactive and non-proactive tweet. For instance, a proactive tweet will often contain modal verbs such as "moeten/[must]" or future tense such as "het zal/[it will]"; whereas a non-proactive tweets will distinguishable use past or present tense. Although this was stated in the current annotation guidelines, the Kappa score suggested that it was difficult for non-linguist annotators to make the distinction. Lastly, a way to reduce the limitation of undistributed labels is gathering many more annotations that are also balanced; for instance, 1000 annotated tweets for *proactive* and 1000 annotated tweets for *non-proactive* .

6.3.2 The Automatic Classifiers

The second sub-research question (b) is: **What machine learning methods are best to automatically classify these tweets?**. In this study, two methods inspired by the literature study were implemented: A baseline SVM and a CNN-BiLSTM. Both methods have been shown to provide state-of-the-art results in emotion and sentiment classification tasks (Ge et al., 2019; Liu et al., 2020; Alswaidan and Menai, 2020). Precision and recall scores (P=0.63, R=0.60) and the confusion matrices show that the SVM classifier outperformed the CNN-BiLSTM.

The SVM confusion matrix in figure 6.3 shows that labels that performed with least confusion were the *positive* and *negative* labels, followed by the *neutral* label. This shows that the SVM is able to identify positive and negative emotions better than classifying tweets as neutral. The error analysis conducted over these 3 labels indicate that errors appeared in tweets with both implicit and explicit emotions. Interestingly, one of the errors identified may have been a result of the system failing to identify sarcasm. The error analysis indicated that there is space for improvement in this system with polarity labels, for example, by implementing features that can identify positive and negative emotions as well as features that can identify sarcasm. Another proposed solution is to implement word embeddings, so that the SVM can capture semantic relationships.

Regarding the machine learning methods, there are three possible explanations as to why the SVM performed better than the CNN-BiLSTM. Firstly, there was sufficient training data for the SVM classifier and possibly insufficient training data for the CNN-BiLSTM. The literature suggested that the SVM is known to work well with little data (Alswaidan and Menai, 2020). Secondly, the SVM is a discriminatory model that works well with high-dimensional data. For the case of polarity, it was able to discriminate/classify most of the *positive* and *negative* labels correctly. On the other hand, it was arguably effective in this case because the distribution of *positive, negative* and *neutral* labels were fairly balanced. For instance, the emotion categories has higher dimensions (more labels), however, their distribution is very imbalanced so the SVM may not have had enough features learn from those labels that were less often present in the training data. The last possible reason, is the system's appropriate input, which were the preprocessed tweets vectorized using the tf-idf measure. This method has shown to work better than other conventional methods such as Bag-of-Words, because the tf-idf measure considers the importance of a word by measuring how often it appears in a text (Islam et al., 2017; Stecanella, 2019).

With regards to the CNN-BiLSTM, the literature has shown with strong evidence that it provides state-of-the-art results in emotion classification tasks. The CNN cap-

tures the local features and the BiLSTM considers the sequences of words with their global features. A combination of both can capture the most important features needed for the algorithm to learn well (Liu et al., 2020; Alswaidan and Menai, 2020). However, this was not the case for this study possibly because there is not enough training data for this classifier as a deep-learning neural network needs a lot of training data to generalise well and not overfit. In the recent work with excellent results by Liu et al. (2020), they used "massive text data" , while in another paper by Ge et al. (2019), as part of the SemEval-2019 Task 3, at least 15000 records of emotion classes were provided for training (EmoContext, 2019). With the use of the 300d pre-trained Twitter word embeddings (Nieuwenhuisje, 2018b), and a large annotated corpus of Dutch tweets, there is hope that the CNN-BiLSTM will possibly improve its performance in future work.

Concerning the word embeddings, although they are trained on a large corpus of social media politics and other social media domains, a point for criticism is that they were trained on a corpus from 2017 (Nieuwenhuisje, 2018a). 2017 is relatively recent, but it may not be recent enough to consider topics discussed in the scope of the current data. For instance, a lot of the tweets are about the COVID19 outbreak or other topics that are only spoken in early 2020. For example, words such as 'COVID19' or 'corona' are not present in the word embeddings; whereas the word 'virus' is present, but not in the context of a major outbreak and lockdown. A way of solving this is to re-train the word embeddings on social media data that is more recent.

As an answer to the second research question, the study showed that the SVM is a more powerful classifier for this task with little training data and when the labels are well distributed and clearly distinguishable, in this case, the polarity labels. This was not the case for the emotion and proactivity labels. The results implied that categorising proactivity and emotions is complex to humans because the aspects that define these two classes are not always explicit or present in the tweets. Furthermore, it is not possible to firmly argue that the SVM works better than the CNN-BiLSTM because its performance for this type of data is unknown when more annotated training data is available.

Chapter 7

Conclusion

The research aimed at building an automatic classifier that could detect emotion and purpose categories in Dutch tweets written by members of the Dutch parliament.

As a starting point, an agreement study was carried out which consisted of identifying emotions and purpose categories that are relevant and representative of the data. This was challenging because the tweets were not personal and contained subtle emotions. This implies that it was difficult to identify the specific emotion or purpose. In the end, three categories of labels were used in six experiments: emotion, proactivity and polarity. The emotion category consists of 6 labels: *admiration, optimism, disappointment, scepticism, no emotion* and *other*. The meaning of the emotion labels can be distinguished in terms of polarity (positive, negative or neutral) and time (present/past or future). The binary proactivity category is a type of purpose category, which distinguishes tweets in terms of *proactive* (desire for change in the future) and *non-proactive* (no desire for change). Lastly, the polarity labels consisted of the emotion categories merged into *positive, negative* and *neutral* emotions. The agreement in the annotated training data showed that the polarity category had the highest Kappa score, followed by the emotion category and lastly, the proactivity category. The agreement study indicated that it was difficult for the annotators to distinguish the emotion and proactivity labels, but were better at distinguishing polarity. This suggested that some of the more complex emotions were not present in the tweets. The training data was annotated 3 times for adjudication purposes and creating the gold labels.

The best performing system is the baseline SVM with the polarity labels. The SVM and the CNN-BiLSTM performed poorly in the other experiments. The possible reasons for the reasonable performance of the SVM include: the presence of polarity in the tweets; the good distribution of labels in the training and test data; and the literature stipulated that the SVM is known to work well with little training data. Furthermore, it was discussed that the SVM system with the polarity labels can improve its performance with the implementation of feature engineering and/or word embeddings. These additions would feed the system with useful features and semantic information that may help identify better the polarity in the tweets. This would also make the classifier more applicable for Red Data.

The challenges suggest that to build an automatic classifier that can detect emotion and purpose in these tweets requires more work on the annotation model for both the SVM and CNN-BiLSTM. It also requires more labelled data for training in the case of the CNN-BiLSTM. The limitations of the annotation model indicated that investigating other emotions that may be present in the tweets and producing more

detailed guidelines for the annotators is a feasible solution. For the proactivity labels, it would be interesting to see whether asking expert linguists to annotate this category could result in better quality labels. Lastly, if the gathering a large volume of annotated tweets of the kind occurs in the future, re-training the present CNN-BiLSTM is also a possibility, as the literature has shown that it is a state-of-the-art architecture for this classification task.

Appendix A

Final Annotation Guidelines

The purpose of these annotations is to train an automatic machine learning classification system. The goal is for the system to learn from these annotations and automatically classify the emotions and purpose of tweets expressed by the Dutch parliament.

GENERAL INSTRUCTIONS

- Read carefully these annotation guidelines.
- The file to annotate contains 4 columns in this order: 'Content', 'Emotion', 'Proactivity' and '(Optional) Comments'.
- The 1st column contains the content of the tweets.
- In the 2nd column, you will be selecting one of the 'emotion' labels. You cannot write any other label here.
- In the 3rd column, you will be selecting one of the 'proactivity' labels. You cannot write any other label here.
- In the 4th column, you may optionally make comments and add your own other labels.
- How to label: You can click on the right arrow and select the label you wish to add as displayed in this screenshot below:

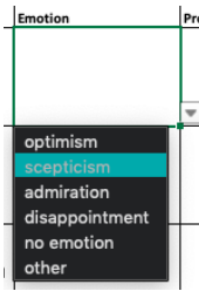

Figure A.1: Screenshot showing how to select the label on Microsoft Excel

- When selecting the label, don't think about who the author is, but think about how the tweet is being expressed. To make this matter easier, the usernames/authors of the tweets have been filtered out.
- The label should be the first thing that pops in your mind.
- When labelling, try to not let your world knowledge influence your decision. For example, if a politician's name is present, do not let your opinion about that politician influence your labelling.
- If there is only a little bit of emotion, please label an emotion. For example, try to avoid using 'no emotion' all the time. 'No emotion' is strictly for tweets that only provide information/facts.

HOW TO ANNOTATE THE 'EMOTION' LABEL ?

For each tweet, you should label 1 of the 'emotion' labels below. If you really think the tweet has more than one of these labels, please write first the 'emotion' that represents the tweet the most and write the second one in the 'Emotion 2' column. However, it is ideal that you select one of the provided labels that represent that tweet the most.

Emotion can be defined as experiences that occur biologically and psychologically in situations that are significant to the individual (Ph.D. Joseph E LeDoux, 2019). Twitter is a social media platform and an environment where emotions are commonly expressed through language.

Dutch political tweets may contain identifiable emotion(s). These guidelines are designed based on trial annotation rounds and thorough research on similar projects. It was decided that Dutch political tweets commonly display subtle forms of emotion. The trials may suggest that the tweets can be classified into: Optimism, Scepticism, Admiration, Disappointment and no emotions. An extra label of 'other' is included in case the annotator strongly thinks none of the provided emotion labels represent that tweet.

To distinguish each emotion, you can think about the 'time' and 'polarity' that represents each label and the tweet. In this case, 'Time' means the space and moment in time the tweet is referring to, such as a past, present or future situation. The past refers to something that has already occurred. The present refers to something that is currently happening. The future refers to something that did not happen yet. E.g. 'Yesterday this happened' refers to the past. Polarity means whether a tweet is positive, neutral or negative e.g. 'This is a sad situation' has a more negative tone. By focusing both on time and polarity, the tweet may be 'easier' to classify. Time and polarity are discussed for every label listed below.

Always ask yourself: is the tweet talking about the past/present or future? And is the tweet positive, negative or neutral?

EMOTION LABELS:

positive + future (1) **OPTIMISIM:** The tweet shows trust, belief and is positive about the future. A type of "everything will be ok" mentality. It is not worried about the future. It will tend to use the grammatical future tense in verbs. E.g. 'wij

zullen.'. Other examples: "Together we will fight coronavirus!", "Things are ok right now", "Don't worry".

positive + present/past

(2) **ADMIRATION:** The tweet shows positivity, proudness, pride or joy of something that happens in the past or present. It will tend to use grammatical present or past tense in verbs. E.g. 'het is..', 'het was...' Other examples: "Nice initiative" or "We are proud of the healthcare workers".

negative + future

(3)**SCEPTICISM:** As opposed to 'optimism', the tweet is negative; it shows disbelief, not trusting and pessimistic about the future. Some tweets will be explicit and display scepticism by showing concern, fear or worry. It will tend to use the grammatical future tense in verbs E.g. 'het zal'. Other examples: "It is too late, the problem will not be solved".

negative + present/past

(4)**DISAPPOINTMENT:** As opposed to admiration, the tweet shows negativity about a past or present situation. This label can represent both sadness and anger. For example, when expectations were not met or they are critical of someone or something currently happening. It will tend to use present or past tense in verbs. E.g. 'het was/is'. Other examples: "Unfortunately, the lockdown happened too late" or "Our country has been let down", "I don't like this!".

neutral

(5) **NO EMOTION:** The tweet is neutral and only gives information. There is no emotion or attitude shown. E.g. "Tonight, this show will be broadcasted" or "A conflict is happening between these 2 countries" In this case, it does not matter whether a tweet is in past/present/future tense. Ask yourself, would the NOS newspaper write this tweet? If not, then there is probably some form of emotion.

Other

(6) **OTHER:** If you think none of the emotions above represent a particular tweet label 'other'. If possible, mention your own label(s) in either English or Dutch in the 'comments' column.

	Past/Present	***Future***
Positive	admiration	optimism
Negative	disappointment	scepticism
Neutral	No emotion	

Figure A.2: A table to distinguish between the polarity and the time/tense of the labels.

HOW TO ANNOTATE THE 'PROACTIVITY' LABEL ?
For each tweet, you should label whether it is 'proactive' or 'non-proactive'.

Proactivity is a type of purpose. Purpose is to have an intention; a reason for which something is being done or something exists. It does not need to be emotional. Tweets have a reason as to why they are being posted and this is supported by a paper on 'The Politics of Twitter' by Duncombe (2019). Duncombe suggests that political tweets represent emotions but also provoke emotions, in other words, they are purposeful (Duncombe, 2019).

A possibly common type of purpose in political tweets is whether there is proactivity or not, for example, whether someone insists in making a change or not. Ask yourself: is the tweet talking about the past/present or future? Does the tweet show the desire to make a change?

PROACTIVITY LABELS:

(1) **PROACTIVE:** The tweeter wants to take action before a future situation occurs. The tweet talks about something that will happen in the future. The person posting the tweet wants to do something and wants a situation to be changed. It usually uses future grammatical tense. E.g. 'wij moeten...' Other examples, E.g. 'we will financially support them' or 'we need to fix this economic recession' or 'Tonight this will be broadcasted'.

(2) **NON-PROACTIVE**: As opposed to proactive, the tweet will not engage in desired action or change. Or, the tweeter does not explicitly say that a change is needed. It is describing something of the present or the past. It uses grammatical past/present tense E.g. 'Het is fijn', 'Dit is raar', 'Ik heb dat gedaan'. Other examples: E.g. 'We are proud of the healthcare workers';'This is happening over there'.

Past/present + no desire to change	***Future + wants a change***
Non-proactive	proactive

Figure A.3: A table to distinguish between the time/tense of the labels.

Appendix B

Confusion Matrices of the 5 trial annotations rounds

Table B.1: (Round 1) Confusion Matrix of Annotators A and B

	A					
B	**admiration**	**disappointment**	**no emotion**	**optimistic**	**sceptical**	**Grand Total**
admiration	2		1	1		4
disappointment		1			2	3
no emotion		1	2	1	1	5
optimistic	2	1	1	1		5
proactive	2		3	3		8
sceptical				1	1	2
surprise		1	1			2
Grand Total	**6**	**4**	**8**	**7**	**4**	**29**

Table B.2: (Round 1) Confusion Matrix of Annotators A and C

	C						
A	**admiration**	**disappointment**	**no emotion**	**optimistic**	**proactive**	**sceptical**	**Grand Total**
admiration	3	1		1	1		6
disappointment		1		1		2	4
no emotion			3	2	2		8
optimistic	2		1	2	2		7
sceptical	1	1		1		1	4
Grand Total	**7**	**3**	**4**	**7**	**5**	**3**	**29**

Table B.3: (Round 1) Confusion Matrix of Annotators B and C

	C						
B	**admiration**	**disappointment**	**no emotion**	**optimistic**	**proactive**	**sceptical**	**Grand Total**
admiration	2	1		1			4
disappointment	1	1				1	3
no emotion		1	3	1			5
optimistic	3			2			5
proactive				3	5		8
sceptical	**1**					**1**	**2**
surprise			1			1	2
Grand Total	**7**	**3**	**4**	**7**	**5**	**3**	**29**

Table B.4: (Round 2) Confusion Matrix of Emotion Labels for Annotators A and B

	A						
B	**admiration**	**anger**	**disappointment**	**no emotion**	**sceptical**	**surprise**	**Grand Total**
admiration			1	3			4
disappointment		1	1	1			3
no emotion				4	1		5
optimistic	2			3		2	7
sadness				1			1
sceptical			1	2	1		4
Grand Total	**2**	**1**	**3**	**14**	**2**	**2**	**24**

Table B.5: (Round 2) Confusion Matrix of Emotion Labels for Annotators A and C

	A						
C	admiration	anger	disappointment	no emotion	sceptical	surprise	Grand Total
admiration	1						1
anger				1			1
disappointment			1				1
no emotion			1	10			11
optimistic				2		2	4
other: irritated					1		1
sceptical	**1**	**1**	**1**	**1**	**1**		**5**
Grand Total	**2**	**1**	**3**	**14**	**2**	**2**	**24**

Table B.6: (Round 2) Confusion Matrix of Emotion Labels for Annotators B and C

	C							
B	admiration	anger	disappointment	no emotion	optimistic	other: irritated	sceptical	Grand Total
admiration				3	1			4
disappointment			1				2	3
no emotion		1		3		1		5
optimistic	1			2	3		1	7
sadness				1				1
sceptical				2			2	4
Grand Total	**1**	**1**	**1**	**11**	**4**	**1**	**5**	**24**

Table B.7: (Round 2) Confusion Matrix of Purpose Labels for Annotators A and B

	A									
B	**agree**	**commemorate**	**criticize**	**criticize**	**entertain**	**inform**	**proactive**	**question**	**ridicule**	**Grand Total**
agree	1					1				2
commemorate		2			2		1			5
criticize			1			2				3
entertain				1	1	1		1		4
inform						2				2
proactive	1					2	2			5
question			1					1		2
ridicule									1	1
Grand Total	**2**	**2**	**2**	**1**	**3**	**8**	**3**	**2**	**1**	**24**

Table B.8: (Round 2) Confusion Matrix of Purpose Labels for Annotators A and C

	A									
C	**agree**	**commemorate**	**criticize**	**criticize**	**entertain**	**inform**	**proactive**	**question**	**ridicule**	**Grand Total**
commemorate	1	1								2
criticize			1				1			2
inform					2	4				6
no purpose	1	1		1	1	1	1			6
proactive						3	1			4
question			1					1		2
ridicule								1	1	2
Grand Total	**2**	**2**	**2**	**1**	**3**	**8**	**3**	**2**	**1**	**24**

Table B.9: (Round 2) Confusion Matrixof Purpose Labels for Annotators B and C

	C							
B	**commemorate**	**criticize**	**inform**	**no purpose**	**proactive**	**question**	**ridicule**	**Grand Total**
agree	1				1			2
commemorate	1		1	3				5
criticize		1	2					3
entertain			1	2		1		4
inform			2					2
proactive		1		1	3			5
question						1	1	2
ridicule							**1**	**1**
Grand Total	**2**	**2**	**6**	**6**	**4**	**2**	**2**	**24**

Table B.10: (Round 3) Confusion Matrix of Emotion Labels for Annotators A and B

	A							
B	**admiration**	**disappointment**	**irritation**	**no emotion**	**optimism**	**other**	**scepticism**	**Grand Total**
admiration	3	1			1			5
disappointment		2						2
irritation		1	2				1	4
no emotion	1	3	4	9	5	1	2	25
optimism	2	2			3			7
other		1	1					2
scepticism		2					1	3
Grand Total	**6**	**12**	**7**	**9**	**9**	**1**	**4**	**48**

Table B.11: (Round 3) Confusion Matrix of Emotion Labels for Annotators A and C

	A							
C	**admiration**	**disappointment**	**irritation**	**no emotion**	**optimism**	**other**	**scepticism**	**Grand Total**
admiration	3	1	2	1	6			13
disappointment	2	1		3			1	7
irritation			5				1	6
no emotion		1		4	1	1	1	8
optimism	1	1			1			3
scepticism		8		1	1		1	11
Grand Total	**6**	**12**	**7**	**9**	**9**	**1**	**4**	**48**

Table B.12: (Round 3) Confusion Matrix of Emotion Labels for Annotators B and C

	C						
B	**admiration**	**disappointment**	**irritation**	**no emotion**	**optimism**	**scepticism**	**Grand Total**
admiration	4	1					5
disappointment		1				1	2
irritation			3			1	4
no emotion	7	4	2	6		6	25
optimism	2			1	3	1	7
other			1			1	2
scepticism		1		1		1	3
Grand Total	**13**	**7**	**6**	**8**	**3**	**11**	**48**

Table B.13: (Round 3) Confusion Matrix of Purpose Labels for Annotators A and B

Count of A	**A**						
B	**ask**	**commemorate**	**criticize**	**inform**	**proactive**	**unknown purpose**	**Grand Total**
ask			1				1
commemorate	1	5	1	1			8
criticize		1	4	1	1		7
inform	1	1	5	9	4		20
other						1	1
proactive			2	4	2		8
unknown purpose			**1**	**1**		**1**	**3**
Grand Total	**2**	**7**	**14**	**16**	**7**	**2**	**48**

Table B.14: (Round 3) Confusion Matrix of Purpose Labels for Annotators A and C

	C					
A	**ask**	**commemorate**	**criticize**	**inform**	**proactive**	**Grand Total**
ask	2					2
commemorate	1	4	1	1		7
criticize	2	1	6	5		14
inform		3	1	9	3	16
proactive	1			4	2	7
unknown purpose				1	1	2
Grand Total	**6**	**8**	**8**	**20**	**6**	**48**

Table B.15: (Round 3) Confusion Matrix of Purpose Labels for Annotators B and C

	C					
B	**ask**	**commemorate**	**criticize**	**inform**	**proactive**	**Grand Total**
ask	1					1
commemorate	1	6		1		8
criticize	2		2	3		7
inform	1	2	4	11	2	20
other					1	1
proactive	1		1	3	3	8
unknown purpose			**1**	**2**		**3**
Grand Total	**6**	**8**	**8**	**20**	**6**	**48**

Table B.16: (Round 4) Confusion Matrixof Emotion Labels for Annotators A and B

	A						
B	**admiration**	**disappointment**	**no emotion**	**optimism**	**other**	**scepticism**	**Grand Total**
admiration	2	2	1				5
disappointment		10	1				11
no emotion	1	4	8	4		1	18
optimism	1	2		3		1	7
scepticism	1	5			1		7
Grand Total	**5**	**23**	**10**	**7**	**1**	**2**	**48**

Table B.17: (Round 4) Confusion Matrix of Emotion Labels for Annotators A and C

	C						
A	**admiration**	**disappointment**	**no emotion**	**optimism**	**scepticism**	**Grand Total**	**Grand Total**
admiration	2		2	1		5	5
disappointment	1	7	9	2	4	23	11
no emotion		1	9			10	18
optimism	1		4	2		7	7
other			1			1	7
scepticism	**1**		**1**			**2**	**48**

Table B.18: (Round 4) Confusion Matrix of Emotion Labels for Annotators B and C

	B						
C	**admiration**	**disappointment**	**no emotion**	**optimism**	**scepticism**	**Grand Total**	**Grand Total**
admiration	2		2	1		5	5
disappointment		5		1	2	8	11
no emotion	2	5	15	3	1	26	18
optimism	1		1	2	1	5	7
scepticism		1			3	4	7
Grand Total	**5**	**11**	**18**	**7**	**7**	**48**	**48**

Table B.19: (Round 4) Confusion Matrix of Proactivity Labels for Annotators A and B

	A		
B	**inactive**	**proactive**	**Grand Total**
inactive	18	11	29
proactive	4	15	19
Grand Total	**22**	**26**	**48**

Table B.20: (Round 4) Confusion Matrix of Proactivity Labels for Annotators A and C

	C		
A	**inactive**	**proactive**	**Grand Total**
inactive	20	2	22
proactive	8	18	26
Grand Total	**28**	**20**	**48**

Table B.21: (Round 4) Confusion Matrix of Proactivity Labels for Annotators B and C

	B		
C	**inactive**	**proactive**	**Grand Total**
inactive	21	7	28
proactive	8	12	20
Grand Total	**29**	**19**	**48**

Table B.22: (Round 5) Confusion Matrix of Emotion Labels for Annotators A and B

	A					
B	**admiration**	**disappointment**	**no emotion**	**optimism**	**other**	**Grand Total**
admiration	7		3	1		11
disappointment	2	10		1	3	16
no emotion			**8**	1	2	11
optimism				2		2
scepticism		5	1	1	2	9
Grand Total	**9**	**15**	**12**	**6**	**7**	**49**

Table B.23: (Round 5) Confusion Matrix of Emotion Labels for Annotators A and C

	A					
C	**admiration**	**disappointment**	**no emotion**	**optimism**	**other**	**Grand Total**
admiration	6	1	1	2	1	11
disappointment		10	1	1	2	14
no emotion	**1**	**1**	**10**	1	2	15
optimism	2			2		4
scepticism		3			2	5
Grand Total	**9**	**15**	**12**	**6**	**7**	**49**

Table B.24: (Round 5) Confusion Matrix of Emotion Labels for Annotators B and C

	A					
B	admiration	disappointment	no emotion	optimism	other	Grand Total
admiration	7		3	1		11
disappointment	2	10		1	3	16
no emotion			8	1	2	11
optimism				2		2
scepticism		5	1	1	2	9
Grand Total	**9**	**15**	**12**	**6**	**7**	**49**

Table B.25: (Round 5) Confusion Matrix of Proactivity Labels for Annotators A and B

	A		
B	**non-proactive**	**proactive**	**Grand Total**
non-proactive	35	3	38
proactive	3	8	11
Grand Total	**38**	**11**	**49**

Table B.26: (Round 5) Confusion Matrix of Proactivity Labels for Annotators A and C

	A		
C	**non-proactive**	**proactive**	**Grand Total**
non-proactive	31	3	34
proactive	7	8	15
Grand Total	**38**	**11**	**49**

Table B.27: Confusion Matrix of 'Proactivity' Labels for Annotators B and C in Final Trial Round 5

	A		
B	non-proactive	proactive	Grand Total
non-proactive	35	3	38
proactive	3	8	11
Grand Total	38	11	49

Bibliography

N. Alswaidan and M. Menai. A survey of state-of-the-art approaches for emotion recognition in text. In *Knowledge and Information Systems*. Springer Link, march 2020. URL https://link.springer.com/article/10.1007/s10115-020-01449-0.

L. F. Barrett, Z. Khan, J. Dy, and D. Brooks. Nature of emotion categories: Comment on cowen and keltner, Feb 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6014873/.

L.-A.-M. Bostan and R. Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1179.

J. Brownlee. What is deep learning?, Dec 2019. URL https://machinelearningmastery.com/what-is-deep-learning/.

J. Brownlee. How to implement baseline machine learning algorithms from scratch with python, May 2020. URL https://machinelearningmastery.com/implement-baseline-machine-learning-algorithms-scratch-python/.

Cambridge. Purpose: meaning in the cambridge english dictionary, 2020. URL https://dictionary.cambridge.org/dictionary/english/purpose.

S. Dandge. saitejdandge/sentimental$_a nalysis_l stm_c onv1d, Feb2019. URL$.

M. Donaldson. Plitchik's wheel of emotions–2017 update, 2017. URL https://www.6seconds.org/2017/04/27/plutchiks-model-of-emotions/.

C. Duncombe. The Politics of Twitter: Emotions and the Power of Social Media. *International Political Sociology*, 13(4):409–429, 08 2019. ISSN 1749-5679. 10.1093/ips/olz013. URL https://doi.org/10.1093/ips/olz013.

P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. 10.1080/02699939208411068.

EmoContext. Emocontext, 2019. URL https://www.humanizing-ai.com/emocontext.html.

S. Ge, T. Qi, C. Wu, and Y. Huang. THU_NGN at SemEval-2019 task 3: Dialog emotion classification using attentional LSTM-CNN. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 340–344, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. 10.18653/v1/S19-2059. URL https://www.aclweb.org/anthology/S19-2059.

Y. Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, Nov 2016. 10.1613/jair.4992.

A. Gonfalonieri. How to build a data set for your machine learning project, Feb 2019. URL https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-p

Y. He, L.-C. Yu, K. R. Lai, and W. Liu. YZU-NLP at EmoInt-2017: Determining emotion intensity using a bi-directional LSTM-CNN model. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 238–242, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 10.18653/v1/W17-5233. URL https://www.aclweb.org/anthology/W17-5233.

M. S. Islam, F. E. M. Jubayer, and S. I. Ahmed. A support vector machine mixed with tf-idf algorithm to categorize bengali document. *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2017. 10.1109/ecace.2017.7912904. URL https://ieeexplore.ieee.org/document/7912904.

Kaggle. Lstm with word2vec embeddings, Apr 2017. URL https://www.kaggle.com/lystdo/lstm-with-word2vec-embeddings.

Y. Katariya. tf.keras.activations.softmax : Tensorflow core v2.2.0, Jan 2020. URL https://www.tensorflow.org/api_docs/python/tf/keras/activations/softmax.

Keras. Keras documentation: Bidirectional lstm on imdb, May 2020. URL https://keras.io/examples/nlp/bidirectional_lstm_imdb/.

D. Kirange and R. R. Deshmukh. Emotion classification of news headlines using svm, May 2012. URL https://www.researchgate.net/profile/Ratnadeep_Deshmukh/publication/262791078_Emotion_Classification_of_News_Headlines_Using_SVM/links/00463538dae0691c15000000.pdf.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, 1977. 10.2307/2529310.

Z.-X. Liu, D.-G. Zhang, G.-Z. Luo, M. Lian, and B. Liu. A new method of emotional analysis based on cnn–bilstm hybrid neural network. *Cluster Computing*, Mar 2020. 10.1007/s10586-020-03055-9.

M. L. McHugh. Interrater reliability: the kappa statistic, 2012. URL https://www.ncbi.nlm.nih.gov/pubmed/23092060/.

S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Inf. Process. Manage.*, 51(4):480–499, July 2015. ISSN 0306-4573. 10.1016/j.ipm.2014.09.003. URL https://doi.org/10.1016/j.ipm.2014.09.003.

S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

L. Mollema, I. A. Harmsen, E. Broekhuizen, R. Clijnk, H. D. Melker, T. Paulussen, G. Kok, R. Ruiter, and E. Das. Disease detection or public opinion reflection? content analysis of tweets, other social media, and online newspapers during the measles outbreak in the netherlands in 2013. *Journal of Medical Internet Research*, 17(5), 2015. 10.2196/jmir.3863. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4468573/.

P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. SemEval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June 2016. Association for Computational Linguistics. 10.18653/v1/S16-1001. URL https://www.aclweb.org/anthology/S16-1001.

A. Nieuwenhuisje. Open-source dutch word embeddings, Jul 2018a. URL https://www.linkedin.com/pulse/open-source-dutch-word-embeddings-alexander-nieuwenhuijse/.

A. Nieuwenhuisje. dutch-word-embeddings. Jul 2018b. URL https://github.com/coosto/dutch-word-embeddings.

C. Olah. Colah's blog, Aug 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs

S. Patel. Chapter 2 : Svm (support vector machine) - theory, May 2017. URL https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72.

R. Plutchik. A general psychoevolutionary theory of emotion. *Theories of Emotion*, page 3–33, 1980. 10.1016/b978-0-12-558701-3.50007-7.

M. Polignano, P. Basile, M. D. Gemmis, and G. Semeraro. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP19 Adjunct*, Jun 2019. 10.1145/3314183.3324983.

A. U. Rehman, A. K. Malik, B. Raza, and W. Ali. A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis. pages 26597–26613, Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan. Springer.

S. Rosenthal, N. Farra, and P. Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017. Association for Computational Linguistics.

E. T. K. Sang, H. Kruitbosch, M. Broersma, and M. E. D. Valle. Determining the function of political tweets. *2017 IEEE 13th International Conference on e-Science (e-Science)*, 2017. 10.1109/escience.2017.60.

S. Scott and S. Matwin. Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems*, 1998.

A. Seyeditabari, N. Tabari, and W. Zadrozny. Emotion detection in text: a review. *ArXiv*, abs/1806.00674, 2018.

K. P. Shung. Accuracy, precision, recall or f1?, Apr 2020. URL https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

S. Singha. How to remove stopwords in python: Stemming and lemmatization, Jun 2020. URL https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-norma

B. Stecanella. What is tf-idf?, Dec 2019. URL https://monkeylearn.com/blog/what-is-tf-idf/.

C. Strapparava and R. Mihalcea. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S07-1013.

S. Wang and C. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, pages 90–94, Dec. 2012. ISBN 9781937284251. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 ; Conference date: 08-07-2012 Through 14-07-2012.

W. Ying, R. Xiang, and Q. Lu. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 10.18653/v1/D19-5541. URL https://www.aclweb.org/anthology/D19-5541.

C. Zaiontz. Cohen's kappa, 2018. URL http://www.real-statistics.com/reliability/interrater-reliability/cohens-kappa/.