Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Appraisal Theories for Emotion Classification in Text

Jan Hofmann

Bachelor Thesis

Prüfer:	Prof. Dr. Sebastian Padó
Prüfer und Betreuer:	Dr. Roman Klinger

Beginn der Arbeit:	23.10.2019
Ende der Arbeit:	23.04.2020

Erklärung (Statement of Authorship)
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.[1]

(Jan Hofmann)

[1] Non-binding translation for convenience: This text is the result of my own work, and any material from published or unpublished work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completeley nor partially been published before. The submitted electronic version is identical to this print version.

Abstract

Over the last years the automated classification of emotions from text has become an interesting topic in natural language processing with many applications. Theories from psychological studies on emotions have been widely utilized to support the task of the automated assignment of emotions to textual content. Most commonly used theories are the *fundamental emotions theory* like proposed by Paul Ekman and the *dimensional model of affect* proposed by Albert Mehrabian and James Russell. However, these theories ignore other psychological theories, namely the *coginitive appraisal theories*, which explain emotions as a response to an individual interpretation of a given situation. Such appraisal theories have only been minorly used in the attempt to improve performance of emotion classification. In addition, there are no datasets annotated with appraisal dimensions. This work filled this gap by annotating a dataset with appraisal dimensions. Further, this work conducted several experiments in which classification models utilized these appraisal annotations. Although this work was not able to show a clear improvement in a real-world setting, the results show that appraisal dimensions have the potential to improve the performance of classifiers, which predict emotions from text.

Abstract in German

Die automatische Zuordnung von Emotionen zu Texten wurde in den letzten Jahren zu einer interessanten Thematik in der maschinellen Verarbeitung von natürlicher Sprache. Theorien aus psychologischen Studien zu Emotionen wie die *Basisemotionstheorie* von Paul Ekman und die dimensionale Theorie von Albert Mehrabian und James Russell werden in der automatischen Klassifikation von Emotionen sehr häufig verwendet. Diese Theorien ignorieren jedoch andere psychologische Theorien zu Emotionen, die Bewertungstheorien (engl. appraisal theories), welche Emotionen über die Interpretation von Situationen beschreiben. Solche Theorien wurden bisher nur sehr wenig für die automatische Klassifikation von Emotionen genutzt. Zudem gibt es noch keinen Datensatz, welcher mit Annotationen basierend auf dieser Art von Emotionstheorie versehen ist. Diese Arbeit füllt diese Lücke, indem ein Datensatz mit solchen Bewertungen (engl. appraisals) annotiert wurde. Zudem wurden in dieser Arbeit verschiedene Experimente durchgeführt, bei denen Emotionsklassifikatoren diese Annotationen nutzen. Obwohl diese Arbeit keine deutliche Verbesserung in einem praxisnahen Szenario belegen konnte, zeigten die Resultate, dass solche Bewertungstheorien das Potenzial haben die automatische Klassifikation von Emotionen in Texten zu verbessern.

Acknowledgments

I would like to recognize the invaluable assistance that my supervisor Dr. Roman Klinger and Enrica Troiano provided during this study. Without your knowledge, support, motivation and patience this work would not have been possible.

Finally, I wish to thank my parents and friends for their support and encouragement throughout my study.

Contents

Page

List of Figures

List of Tables

1 Introduction

Emotions play an important role in the life of every human being. They influence the human behavior in many ways (Dolan, 2002) and have been widely and extensively explored by psychologists for a long time (Darwin, 1872; Cannon, 1927; James, 1884). A variety of models and theories have been developed in order to understand the origin and the function of human emotions (Cannon, 1927; James, 1884; Ekman, 1992; Plutchik, 2001; Russell, 1980; Lazarus, 1991). Emotions can be expressed and shared in different ways like verbally and visually, while communicating, or in form of written texts for example in books or a letter.

Due to the increasing popularity of social-media platforms and the World Wide Web in general and the therewith constantly increasing amounts of data, emotions also became an interesting research field for computer scientists. Over the last decade the detection or classification of emotions has become a hot topic in natural language processing and is a key part of affective computing. In affective computing the goal is to develop systems, which can detect or recognize human emotions and include this predicted emotional states into their decision-making process and responses (Picard, 2003).

The automated classification of emotions or sentiments has many applications like social-media analysis (Roberts et al., 2012), opinion mining (Pang and Lee, 2008) or the automated detection of mental health problems (Calvo et al., 2017). Organizations for example can use emotion classification for harvesting insightful informations about the feelings of their customers about their products. The classification of a single emotion from text is essentially multi-class classification task, but is often harder to solve compared to traditional classification tasks due to the complexity and subjectivity of emotions. Therefore, emotional theories from psychological studies are used to improve the performance of the automated classification of emotions.

Emotion theories attempt to reveal all different emotions and how they can be distinguished from each other. For this reason such theories are important elements for the automated classification of emotions. These theories can roughly be divided in three different types: The basic (or fundamental) emotions theories (Ekman, 1999; Plutchik, 2001), the dimensional emotion models of affect (Russell, 1980; Posner et al., 2005) and the cognitive appraisal emotion models (Scherer, 1982; Smith and Ellsworth, 1985; Oatley and Johnson-laird, 1987; Frijda et al., 1989).

Especially for finding annotation schemata, basic emotional theories have already been widely used in the automatic classification of emotions (Strapparava and Mihalcea, 2007a; Alm et al., 2005; Schuff et al., 2017). They provide theoretical foundations on which emotions should be considered as possible classes in a classification scenario and datasets in the field of emotion classification are commonly annotated using a set of these fundamental emotions. Moving away from discrete categories, dimensional models of affect (Russell, 1980) have also become a popular choice as a representation for emotions in the recent years. These dimensional approaches state that all emotions can be described using two (sometimes three) dimensions, namely valence and arousal (and dominance). Several datasets have been annotated using such dimensional models of affect (Buechel and Hahn, 2017a; Preoţiuc-Pietro et al., 2016).

The third category of emotion theories, the cognitive appraisal theories, argue that an emotion is the result of an interpretation of a given event or situation (Moors et al., 2013; Imada and Ellsworth, 2011), This implies that different interpretation invoke different emotions. Such theories have only been minorly utilized for the automated classification of emotions. A work in sense of the automatic classification of emotions from text using appraisal theories was proposed by Balahur et al. (2011). They argued that emotion predictions based on the words in a textual content only is not sufficient for emotion prediction, since emotions are often described or expressed indirectly. According to appraisal theories they stated that emotion classification should make use of the possible interpretations of an event described through the text. For this reason they created a knowledge base, which stores information about the components of situations and their appraisal based emotional response.

However, to date there is no dataset explicitly annotated using an appraisal dimensional approach. This bachelor thesis aims to fill this gap by creating a dataset containing reliable annotations of appraisal dimensions on basis of textual event descriptions.

This work aims to answer the following questions using the created appraisal annotations: Firstly, how well can appraisal dimensions be predicted from text and how well do emotion predictions based on appraisal dimensions perform in comparison to state-of-the-art methods, which predict emotions only on basis of text? Further, this work aims to answer the question if appraisal dimensions can contribute useful information to a classification system in a joint learning scenario in order to improve the performance compared to state-of-the-art methods?

This work aims to answers these questions by conducting several experiments including a real-world pipeline setup in which appraisals are predicted from text followed by emotion predictions based on the predicted appraisals and a multi-task learning framework, which is learning to predict appraisals and emotions jointly.

2 Theoretical Background

2.1 Text Classification

The automatic classification of texts or documents into categories (or classes) has many applications like spam detection (Crawford et al., 2015), abusive language detection (Nobata et al., 2016) or the automated tagging of textual content (Salminen et al., 2019). In general, text classification is a task in which a label or multiple labels are automatically assigned to sentences, documents or words. The algorithm (or function) solving this problem is called a classifier and is associating one or more *classes* (or *labels*) from a finite set of predefined possible classes to a given input. Formally a classifier can be described as follows:

Let $\mathbb{X}$ be a document space and $\mathbb{C}$ a set of classes.
Then a function $f : \mathbb{X} \to \mathbb{C}$ is called a classifier.

This means the classifier or function f is mapping documents in $\mathbb{X}$ to classes in $\mathbb{C}$. Text classification can be divided in *supervised* and *unsupervised learning*. *Unsupervised* techniques use data without predefined labels in order to learn some inherent structure of the data. This is often used for document clustering if creating labeled training documents is not possible or to difficult (Cambero, 2016).

In *supervised* learning on the other hand the function, which is mapping inputs to a class is learned using labeled data. In sense of textual classification this means the algorithm is using textual inputs already assigned with a label, or multiple labels, in order to learn to predict labels for unseen data correctly (Kotsiantis, 2007).

This means that the goal of supervised text classification is to learn a function, which is finding the correct classes for given documents, i.e:

Let d be a document labeled with class c.
Given the document $d \in \mathbb{X}$, determine f such $c = f(d)$.

This function is found by a process called learning. However, most likely a function, which is always finding the correct class can not be found. The goal is to optimize this function, such it performs best on data, which was not seen during the training process. Further a *binary classification* problem is a problem of finding the correct class in a set of two possible classes. A *multiclass classification* problem on the other hand is a problem of finding the correct class in a finite set of more than two possible classes. Like in the binary classification there is also just one of the possible classes correct. In contrast to this in a *multilabel* or *any-of* classification problem the task is to find the set of correct classes in a finite set of possible classes.

Common elements of supervised text classification, as stated by Mirończuk and Protasiewicz (2018) and how they are related are shown in Figure 1. Such elements are, for instance, the annotation of data and the transformation of the data into a format, which is understandable by a classifier. The following sections will give a brief overview of elements, which are relevant in this work.

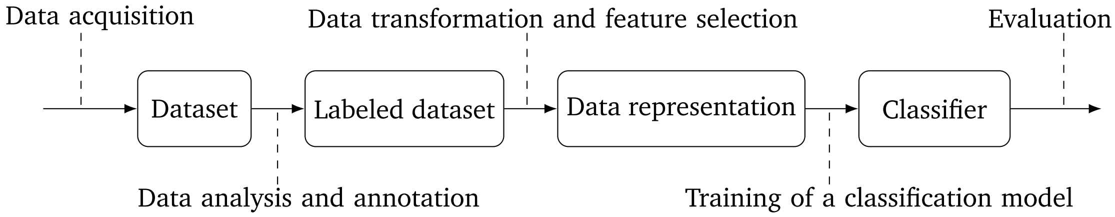

Data acquisition
Data transformation and feature selection
Evaluation
Dataset
Labeled dataset
Data representation
Classifier
Data analysis and annotation
Training of a classification model

Figure 1: Process of text classification. Drawn after Mirończuk and Protasiewicz (2018)

2.1.1 On Annotating a Dataset

In general the annotation of a dataset is the task of manually labeling the instances of the dataset with a set of defined labels. In sense of text classification this is labeling text instances, i.e. sentences or documents, with one or more possible classes. A common method to create annotated datasets is via an *expert annotation* (Bostan and Klinger, 2018). In this method the annotators are experienced in the domain, they are confronted with in the annotation task. Another approach for the acquisition of annotated data *crowdsourcing*, which is a collaborative approach (Sabou et al., 2014), usually distributing the task through the internet. Various platforms on the internet for *crowdsourcing* exist, which enable researchers to collect data (like annotations) from plenty of paid workers or volunteers.

If multiple annotators are tasked to label the same instances the annotation- or inter-annotator- reliability can be measured, which can give insights about the annotation quality or the difficulty of the annotation task. Commonly for this purpose the *Cohens's Kappa* (Cohen's κ) statistic is used (McHugh, 2012; Cohen, 1960). In contrast to a simple percentage agreement calculation, Cohen's κ also considers the probability of an agreement. It is defined as follows:

Let $p(A)$ be the observed agreement and $p(E)$ the expected agreement by probability. Then

$$\kappa = \frac{p(A) - p(E)}{1 - p(E)}$$

However, the κ-score should only be used as a reference point, since it can greatly vary between tasks of different difficulties and there is no theoretical foundation on how to interpret this score.

If the annotation of a dataset (or corpus) is done by multiple annotators different techniques can be used in order to aggregate the individual annotations. Such a technique is for example performing a *majority vote*, which means accepting the label for an instance, which was chosen by the majority of the annotators.

2.1.2 On Preprocessing

Various techniques can be applied in order to prepare the textual input for a classifier before classification. Such techniques are called *preprocessing* and include for example *case folding* (Reducing all letters to lower case), removing *stop words* (Removing extremely common words) or *stemming* (Reducing words to their root form) (Lovins, 1968). The sentence *"I like eating cake VERY MUCH"* could for example result in *"like eat cake much"* after removing some extremely common words and applying *stemming* and *case folding*.

2.1.3 Features in Text Classification

In text classification features are attributes of documents, or a sentence given to a classifier. Based on those attributes the classifier is learning to predict and predicting classes. In the *bag-of-words (unigram) language model* (BOW), for example, a document or sentence is represented by a multiset of the words in it. Features derived by this representation are for example the words and number of occurrences of the words in a document. Note that in this representation spatial information i.e. the order of the words does not remain.

An alternative to this is, for example, to use a *bigram language model*. In this model documents or sentences are also represented by a multiset but in contrast to the unigram model, in which an entry consist of a single word, an entry consists of a word and the word occuring next to it in the document. With this model some of the spatial information remains. A *bigram language model* can be generalized to a *n-gram* language model in which entries are sequences of words of length n.

Another popular representation of documents (or sentences) are *word embeddings*, like for example GloVe (Pennington et al., 2014). A word embedding maps words or phrases to vectors i.e numerical representations. The goal is to provide similar representations for similar words or words which appear in similar contexts. Mapping words like this help learning algorithms to achieve better performance (Mikolov et al., 2013) and are very popular in neural network based classification models (Klinger et al., 2018).

The features in a dataset might be different rich in information for a classifier. Some features might only marginally contribute to a prediction decision, while others grant a classifier more information to make a prediction possible. A common method for calculating this “information richness” of different features is to use the *mutual information* (MI) between the features and the classes in a dataset. With this, features can be ranked according to the information they are providing. This can, for example, help to reduce the amount of features by discarding those which do not contribute any or just minorly contribute information. An information theoretic version of *mutual information*, which provides a measure of how much information the presence or absence of a feature contributes over all classes, can be calculated using the following formula (Xu et al., 2007; Métais et al., 2011):

Let f denote the presence a feature, while $\overline{f}$ is denoting the absence of feature. Further let $\mathbb{C}$ be a set of classes of a classification problem.

$$\begin{aligned}\mathrm{I}(\{f,\overline{f}\},\mathbb{C}) &= \sum_{c\in\mathbb{C}}\sum_{x\in\{f,\overline{f}\}} p(x,c)\log_2\frac{p(x,c)}{p(x)p(c)}\\ &= \sum_{c\in\mathbb{C}} p(f,c)\frac{p(f,c)}{p(f)p(c)} + \sum_{c\in\mathbb{C}} p(\overline{f},c)\log_2\frac{p(\overline{f},c)}{p(\overline{f})p(c)}\end{aligned}$$

Here $p(f)$ is the probability of observing feature f in the dataset, while $p(c)$ denotes the probability of observing class c. $p(f,c)$ represents the *joint probability* of observing feature f together with class c. These probabilities can be estimated by counting the number of observations and normalizing these counts by the size of the dataset (Church and Hanks, 1989).

However, certain features might be related to certain classes. For example if a certain word often or only occurs in combination with a specific class the word is strongly associated with this class. This associativity can be measured using, for example, the pointwise mutual information (Church and Hanks, 1989).

Let f be a feature and c be a class, then

$$\mathrm{PMI}(f,c) = \log_2\frac{p(f,c)}{p(f)p(c)}$$

Note that there is a fundamental difference between *mutual information* as formulated above and PMI: *Mutual information* includes the information gain if a term is absent, while PMI ignores this information and only includes information gain if a term is present (Yang and Pedersen, 1997). Further, the PMI measure can easily be normalized into values between $[-1, +1]$ (Bouma, 2009):

$$\mathrm{NPMI}(f, c) = \frac{\mathrm{PMI}}{-\log_2 p(f, c)}$$

Here $\mathrm{NPMI}(f, c) = 1$ means the feature and the class only occur together, while $\mathrm{NPMI}(f, c) = 0$ means they are independently distributed. A NPMI value of -1 on the other hand means that the feature and the class do never occur in combination in the dataset (Bouma, 2009).

2.1.4 Performance Metrics

In order to be able to rate different classifiers along their performance some metrics need to be defined. Generally as a result of a classification task four different outcomes can occur (Note that not all are mutually exclusive):

- A *true positive (TP)* prediction, which describes a scenario in which the predicted class corresponds to the correct class.
- A *false positive (FP)* prediction, which describes a scenario in which the classifier is predicting a class, which does not correspond to the correct class.
- A *true negative (TN)* prediction occurs, which describes a scenario in which the classifier is predicting a class to be not the correct class, which indeed is not the correct class.
- A *false negative (FN)* prediction occurs, which describes a scenario in which the classifier is predicting a class to be not the correct class although it is the correct class.

In order to determine the performance of a classifier the metrics *precision (P)* , *recall (R)* and a combined measure of *recall* and *precision,* called *F-score* are widely used. The metrics *precision* and *recall* are calculated using the amount of *true positives*, *true negatives*, *false positives* and *false negatives* observed in the evaluation process of a classifier and are formulated as follows:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F_\beta = (1 + \beta^2) * \frac{P * R}{(P * \beta^2) + R}$$

With different values for β, the importance of precision against recall can be adjusted. The most frequently used is $\beta = 1$, which gives equal weight to *recall* and *precision*. This is called the balanced *F-score* or F$_1$-score and is the harmonic mean between *precision* and *recall.*

$$F_1 = 2 * \frac{P * R}{P + R}$$

Further, the metric *accuracy* (Acc) is occasional used for performance evaluation:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

With this metrics, the performance of the classifer on a single class can be evaluated. In order to calculate the performance scores over all classes *micro-* and *macro-averaging* is used. In *micro-averaging* the number of *true positives*, *true negatives*, *false positives* and *false negatives* for each class are aggregated and then used to compute the performance metrics.

Formally:

Let

$$P_\mu = \frac{\sum_{c \in \mathbb{C}} TP_c}{\sum_{c \in \mathbb{C}} TP_c + \sum_{c \in \mathbb{C}} FP_c} \quad \textit{and} \quad R_\mu = \frac{\sum_{c \in \mathbb{C}} TP_c}{\sum_{c \in \mathbb{C}} TP_c + \sum_{c \in \mathbb{C}} FN_c}$$

then

$$\textit{Micro-averaged } F_1 = 2 * \frac{P_\mu * R_\mu}{P_\mu + R_\mu}$$

Note that in a multiclass classification setting $\sum_{c \in \mathbb{C}} FN = \sum_{c \in \mathbb{C}} FP$ following that the micro-averaged *precision*, *recall*, *F-score* and *accuracy* are equal.

In *macro-averaging* on the other hand the metrics are separately calculated for every class and then simply averaged, i.e.

Let F_{1c} be the F_1-score of class c. Then

$$\textit{Macro-averaged } F_1 = \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} F_{1c}$$

An advantage of macro averaging is that it will give equal weight to all classes even if the classes are not balanced in the dataset. However, in this work both, *micro-* and *macro- averaging* are stated in the experiments.

2.1.5 Evaluation of a Classification Algorithm

When evaluating a classifier on a specific dataset, this dataset is typically divided into three subsets. A *training set*, *validation set* and a *test set*. The *training set* is used for training the classifier i.e. optimizing the parameters of the classifier.

The *validation set* is a separate set on which the classifier can be tested during training and is also used for optimizing meta- or hyperparameters like the number of neurons in a artificial neural network layer. Note that a meta- or hyperparameter in machine learning is a variable usually manually selected before training a classifier. In contrast to this model-parameters are selected by the training process (Probst, 2019). The *test set* is used for the final evaluation of the model, containing only data, the classifier has never seen.

A more sophisticated method for evaluating a classifier is the *k-fold cross validation*. The idea of this method is to split the dataset into k equal sized parts (folds). After this $k-1$ folds are used as training set and one fold as test set. This training and testing is repeated k times such that every fold is used as the test set once. Further a *repeated k-fold cross validation* can be done by repeatedly doing a *k-fold cross validation*, in which the data is split into folds differently in every repetition.

2.1.6 Maximum Entropy Classifier

In text classification various models exist like, for example, *Naive Bayes*, *Support Vector Machines* or the *Maximum Entropy classifier* (MaxEnt). The MaxEnt classifier is a probabilistic, linear classifier and due to its simplicity and easy to reproduce results, commonly used in text classification. In addiation the MaxEnt classifier is often en par with artificial neural models (Schuff et al., 2017). In text classification maximum entropy modeling is using an iterative optimization process in order to estimate the *conditional probabilities* from labeled training data (Nigam et al., 1999).

In sense of text classification the *conditional probability* can be defined as follows:

Given class $c \in \mathbb{C}$ and document $d \in \mathbb{X}$.
Let $p(c \mid d)$ be the conditional probability of observing the class c given document d.

The maximum entropy model has the following parametric form:

Let $c \in \mathbb{C}$ be a class and $d \in \mathbb{X}$ a document.

$$\begin{aligned} p_\lambda(c \mid d) &= \frac{\exp \sum_i \lambda_i f_i(d,c)}{\sum_{c'} \exp \sum_i \lambda_i f_i(d,c')} \\ &= \frac{1}{Z(d)} \exp \sum\nolimits_i \lambda_i f_i(d,c), \end{aligned}$$

$$\text{with } Z(d) = \sum\nolimits_{c'} \exp \sum\nolimits_i \lambda_i f_i(d,c'),$$

in which c is a class variable, d is the input data, λ_i are parameters (or weights) to be learned and $Z(d)$ is a normalization factor. Note that here the function f is a so called *feature function* (Berger et al., 1996). This binary valued function is used for extracting *features*, denoted as $f_i(x,c)$, from a given input and class.

The following will give a definition on how such features can be derived from a bag-of-words model containing the words $w_1, ..., w_i$:

Let $c \in \mathbb{C}$ be a class and $d \in \mathbb{X}$ a document labeled with class $c' \in \mathbb{C}$.

$$f_i(d,c) = \begin{cases} 1, & \text{if } d \text{ contains } w_i \text{ and c} = c' \\ 0, & \text{otherwise.} \end{cases}$$

Learning the parameters λ_i for the features $f_i(x,c)$ is a crucial part of the MaxEnt classifier. A parameter λ_i is a real-valued weight associated with feature f_i (Berger, 1997) and measures the "importance" of a feature. Different methods exist on how these parameters can be learned (or optimized) like, for example, *improved iterative scaling* (Berger, 1997) or more generally numerical optimization techniques like stochastic gradient descent with its variants like the *Adam* optimization algorithm (Kingma and Ba, 2014).

However, all those optimization algorithms try to minimize the negative conditional log likelihood of the training data given to a model, i.e.

Given a set of classes $\mathbb{C}$ and a set of training documents $\mathbb{D}$,

$$\min_\lambda - \log p_\lambda(\mathbb{C} \mid \mathbb{D}) = \sum_{(c,d)\in(\mathbb{C},\mathbb{D})} -\log p_\lambda(c \mid d)$$

Finally, a document (or sentence) d can be mapped (or classified) to a class $c \in \mathbb{C}$ with the maximum entropy classifier using the following equation:

$$\text{prediction}(d) = \arg\max_{c\in\mathbb{C}} p_\lambda(c \mid d)$$

2.1.7 Artificial Neural Networks

Popular classifiers for text classification are also neural network based models. The basis of an *artificial neural network* (ANN) is a collection of so called *neurons* or

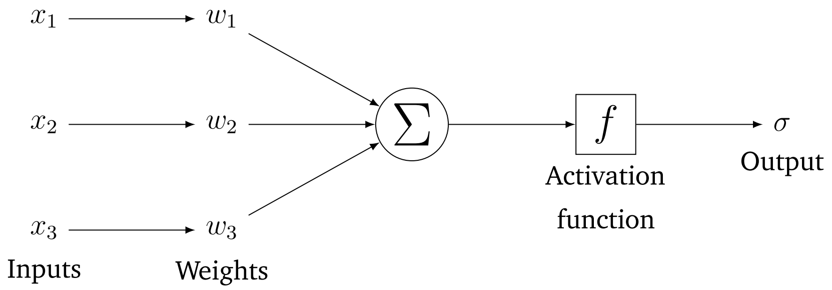


Figure 2: Schematic model of an artificial neuron with three inputs, weights, an activation function, an output and arrows representing the information flow.

units. If a *neuron* receives an input signal it processes the signal and transfers the result to other connected neurons or an output function. These connections are often called edges (Shiruru, 2016). With this concept an *artificial neural network* tries to imitate parts of the human brain. Figure 2 shows a simplified schematic model of this apporach with input signals, weights, an artificial neuron, an activation function and an output. Note that the arrows represent information going to and out of the artificial neuron.

Activation functions define the output at a given input and weights. A commonly used activation function for hidden layer neurons is, for instance, the ReLU function (Nair and Hinton, 2010), while common activation functions for an output layer are, for example, the sigmoid or the softmax function. They are defined as follows:

$$\text{sigmoid}(x_i) = \frac{1}{1+e^{-x_i}} \quad \text{softmax}(x_i) = \frac{\exp\ x_i}{\sum_n \exp\ x_n} \quad \text{reLU}(x_i) = \text{max}(0, x_i)$$

Formally the output of a artificial neuron given an input, weights and the activation function can be defined as follows:

Let $x_1, ..., x_i$ be the inputs of a neuron, $w_1, ..., w_i$ the weights of the edges connected to the neuron and f the activation function.
Then the output of a neuron is defined by

$$\sigma = f(\sum_i x_i w_i)$$

However, an artificial neural network does not consist of only one neuron. Multiple artificial neurons connected to the same inputs are a so called layer. Usually an artificial neural network consists of an input layer, one or multiple hidden layers and an output layer. Note that hidden layers are the layers of neurons between the input and the output layer. A simple *feedforward* ANN with one input layer, containing four neurons, one hidden layer with five neurons and an output layer with two neurons is shown in Figure 3. Note that in a *feedforward* ANN the information is only transferred in one direction from layer to layer and there are no cycles (Montana and Davis, 1989). In this figure the information transfer from layer to layer is represented by the arrows.

There exist different architectures (or types) of neural networks such as *Convolutional Neural Networks* (CNNs), which is also a feedforward network, or Recurrent Neural Networks (RNNs), which is not a feedforward network. Both types of neural networks are popular choices for text classification (Kowsari et al., 2019).

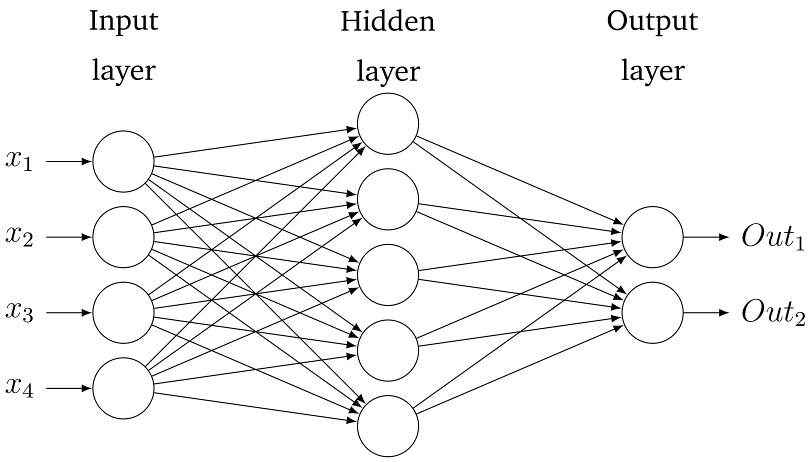


Figure 3: Schematic model of an artificial neural network with input-, hidden-, output-layer and arrows indicating information transfers.

Kim (2014) proposed a popular CNN architecture for different sentence classification tasks, which produced state-of-the-art results in various tasks. He also showed that unsupervised pretrained *word embeddings*, like GloVe (Pennington et al., 2014), for example, are an important ingredient for deep learning in text classification.

A CNN uses a mathematical operation on two functions called convolution. In case of a CNN one function is represented by the input to a convolutional layer and the other function is represented by a so called filter kernel. In addition, convolutional networks usually use a technique called pooling. A pooling layer is used to compute the a value within a small neighborhood of each convolutional data output (Ranzato et al., 2007).

2.1.8 Multi-task Learning

Multi-task learning is an approach in machine learning in which multiple task are learned simultaneously. The idea of this learning approach is to use information which is shared between the tasks and prefer solutions which solve both tasks over solutions which only solve specific tasks (Ruder, 2017). This can improve the generalization of a classifier (Caruana, 1997), i.e. can help a classifier to perform better on unseen data compared to a traditional single-task learning approach if the learned tasks are related. The idea of multi-task learning can be compared to the way humans are learning and solving tasks. When we are trying to solve new tasks they often apply knowledge obtained by learning related tasks (Ruder, 2017; Yu Zhang, 2018).

In sense of neural networks this usually means that tasks, which are learned simultaneously are sharing one or more hidden layers.

2.2 Emotions

Emotions, their origin and their purpose have been widely and extensively explored by scientists from various fields. A variety of theories have been developed in order to understand the origin and the function of human emotions. Some theories, for instance, state that emotions can be seen as an evolutional result developed by our ancestors (Darwin, 1872).

Theories like the James-Lange-Theory, for instance, state that emotions appear as responses to physiological reactions (James, 1884). However other theories like the Cannon-Bard-Theory state that emotions and physiological changes are independent and can appear at the same time (Cannon, 1927).

Theories on emotions also attempt to reveal all different emotions and how they can be distinguished from each other. For this reason such theories are important elements for the automated classification of emotions. They can roughly be divided in three different types of theories: The basic (or fundamental) emotions theories (Ekman, 1999; Plutchik, 2001), the dimensional emotion theories (Russell, 1980; Posner et al., 2005) and the cognitive appraisal theories (Scherer, 1982; Smith and Ellsworth, 1985; Lazarus, 1991).

2.2.1 Basic Emotions Theories

After studying the relation between emotions and culture Ekman (1992) identified the six emotions *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* as so called basic emotions. The term basic refers to the conclusion of this study that these six emotions are distinguishable by every human being in facial expressions independently from their origin and culture. This challenged earlier views that the interpretation of facial expressions are learned and concluded that those six emotions are universal and innate. Plutchik (2001) agreed with the idea of having basic emotions and stated that all other emotions, so called complex emotions are derived by mixing these basic ones.

However, in contrast to Ekman's set of six emotions, Plutchik identified the eight basic emotions *anger*, *anticipation disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*. He proposed an emotion model known as *Plutchik's wheel of emotions* (Shown in Figure 4). In this model, complex emotions like *love* or *optimism* are retrieved by mixing basic emotions. The emotion *love*, for instance, is described as a mix of the two basic emotions *joy* and *trust*.

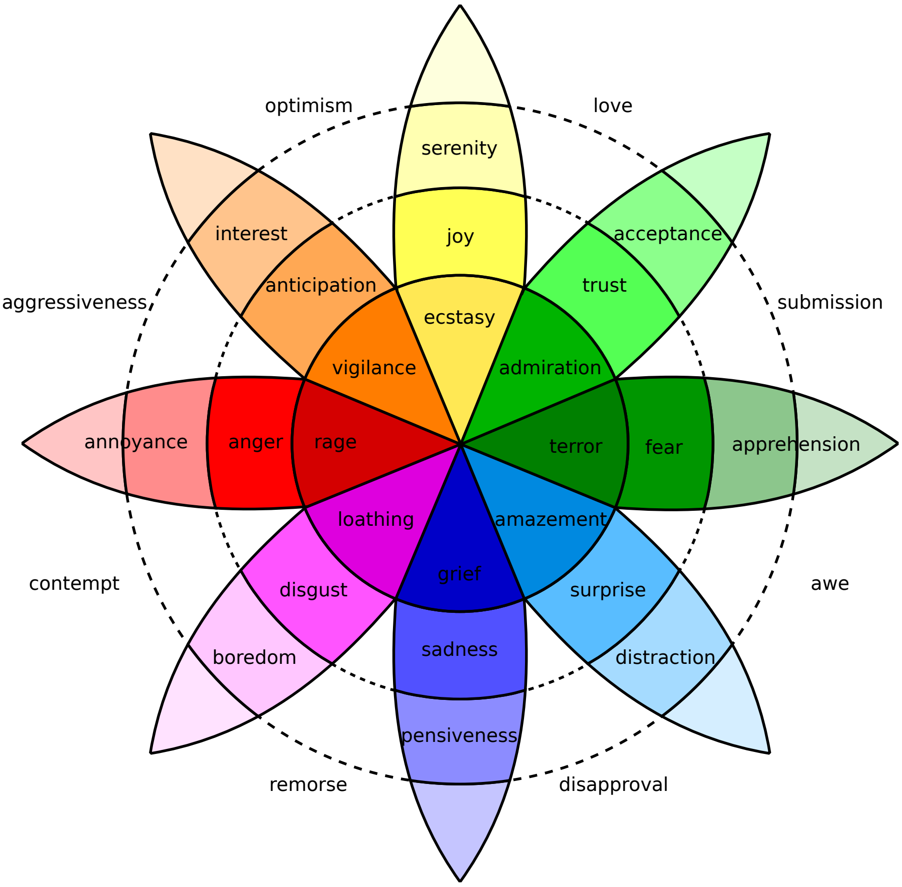


Figure 4: Plutchik's wheel of emotions.

2.2.2 Dimensional Emotion Theories

Other theorists try to distinguish emotions using dimensional models. A popular model is the circumplex model of affect developed by Russell (1980), which states that emotions are distributed in a circle around a two-dimensional space containing a *valence* dimension (the degree of pleasantness) and an *arousal* dimension (the degree of activation). In this model emotions are described using only these two dimensions i.e. the amout of *valence* and *arousal* (See Figure 5). The *valence* dimension is used to describe how unpleasant or pleasant an emotional experience is. A high level of *valence* is, for example, associated with *happy*, *relaxed* or *cheerfull*. On the other hand a low level of *valence* is associated with *anger*, *sadness* or *disgust*. The *arousal* – or activation – dimension is used to describe how intense this emotional experience is felt. A high level of *arousal* is, for example, associated with *surprise*, while a low level of *arousal* is associated with *boredom*. A high level of *arousal* combined with low level of *valence* is, for example, indicating *stress* or *fear*.

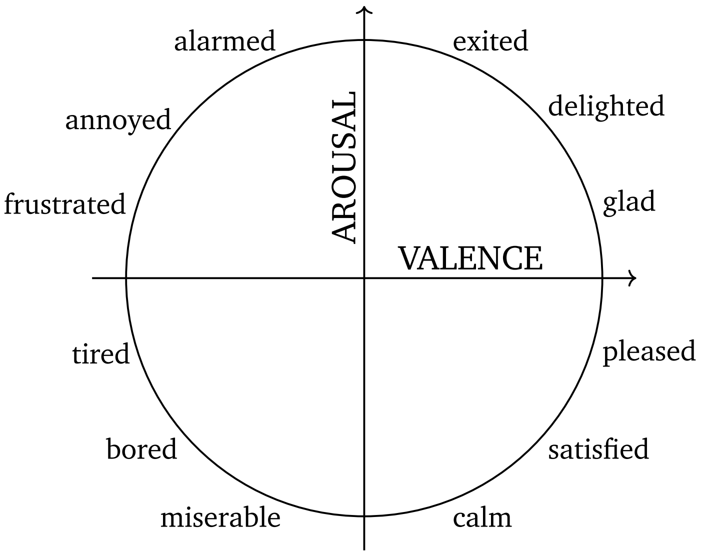


Figure 5: Circumplex model of emotions: The horizontal axis is representing the valence dimension, the vertical axis is representing the arousal dimension (drawn after Russell (1980); Kim and Klinger (2019))

An extended version of the circumplex model is the Valence-Arousal-Dominance model (VAD- or PAD-model) (Mehrabian, 1996; Russell and Mehrabian, 1977), which suggests an additional dimension of dominance. The dominance dimension is used to describe the degree of control one feels to have over the situation that causes an emotion (Kim and Klinger, 2019).

2.2.3 Cognitive Appraisal Theories

Cognitive appraisal theories argue that an emotion is the result of an interpretation of given events or situations (Moors et al., 2013; Imada and Ellsworth, 2011). An appraisal as defined by Ellsworth and Scherer (2009) is a cognitive process of evaluating stimuli or events, which then result in a specific emotional experience. Like dimensional theories the appraisal theories for emotions are componential. Researchers state that there are different cognitive appraisal dimensions (or components), which are evaluated separately while reacting to a stimulus or a situation (Roseman, 1984; Smith and Ellsworth, 1985). Afterwards, accordingly to this evaluations the emotion is determined. Appraisal theories try to specify different appraisal dimensions, which are most important for differentiating between emotions. After specifying the different appraisal dimensions theorists try to find patterns in the evaluation of this dimensions, which are linked to specific emotions. Such patterns can afterwards be used for predicting an emotion.

A popular appraisal theory and the theory this study refers to has been proposed by Smith and Ellsworth (1985). This theory differentiates an emotional experience using the six dimensions *pleasantness, self-other responsibility/control, certainty, attentional activity, anticipated effort* and *situational control*. These dimensions are a result of independant research by different theorists of cognitive appraisal theories (Roseman, 1984; Scherer, 1982). Smith and Ellsworth (1985) proposed an experiment for the rating a set of emotions along those dimensions. In this experiment they asked participants about their experiences feeling a specific emotion using questionnaire systematically designed to rate specific emotional experiences along the proposed appraisal dimensions. These ratings showed that each emotion is associated with a different pattern of appraisal between the six dimensions. They demonstrated the use of this model in an analysis, in which 15 different emotions were correctly predicted over 40% of the time using the cognitive appraisal dimension pattern (Smith and Ellsworth, 1985).

The following will give a brief overview of these six appraisal dimensions according to the results and observations mentioned by Smith and Ellsworth (1985).

The *attentional activity* dimension describes the level of attention in an emotional experience. If we are confronted with a stimulus or a situation our first reaction is either to attend to it, ignore it or avoid it (Scherer, 1982). *Frustration* for example is an emotion indicated with a increased level of *attention*, while *boredom* or *disgust* is indicated with low levels of *attention*.

The *certainty* dimension describes how well understood or predictable the consequences of a situation are. *Surprise* and *fear* are emotional experiences, which are described with very low levels of certainty. If one fears something he is not able to predict what is going to be the outcome of a situation. *Anger* and *guilt* on the other hand are described as experiences with high levels of *certainty*. If we feel guilty we know we did something wrong i.e. we are certain about that.

The *anticipated effort* dimension describes how much effort one feels a specific situation will require. According to Smith and Ellsworth (1985) this dimension is mainly used to differentiate pleasant emotions. *Challenge*, for instance, is associated with a large amount of anticipated effort, while *happiness* or *pride* are associated with very low levels of anticipated effort.

The *pleasantness* dimension describes how pleasant or unpleasant an emotional experience is and is very similar to the valence dimension in the Valence-Arousal-Model. This dimension is mainly used to differentiate between pleasant and unpleasant emotional experiences.

The *responsibility/control* dimension considers the level of control one feels to have over a situation (comparable to the dominance dimension in the PAD-Model) and the level of responsibility one feels to have for bringing about the situation. High levels of responsibility/control are associated with increased *self-responsibility/-control*, while low levels of this dimension is associated with increased *other-responsibility/-control*. This dimension can be used, for example, for differentiating between the emotions guilt and anger. If we feel guilty, we feel responsible for bringing about the situation. If we feel angry we blame other for being responsible for the situation.

The *situational control* dimension is used to describe to what extent a situation is controlled by circumstances versus to what extent the situation is controlled by a human (Smith and Ellsworth, 1985). This is important because while evaluating responsibility and control of a situation, people not only distinguish between self- and other- responsibility/control but also to what extend the situation is caused by circumstances beyond anyone's control (Smith and Ellsworth, 1985). This dimension is, for example, used for distinguishing *sadness* from *guilt* or *anger*. If we feel sad, we feel like a victim of circumstances, while if we experience *anger* or *guilt*, we think the situation was caused by someone (someone else or our self).

Their main findings on appraisal dimension evaluations regarding different emotions are shown in Table 1. Note that in this table a high level of *unpleasant* is indicating an increased level of *unpleasantness*, while a high level of *uncertainty* is indicating increased uncertainty. Further a high level of *responsibility/control* is indicating increased *self-responsibility/contol*, while a low level of *responsibility/control* is indicating a increased level of *other-responsibility/contol*, which means that in the emotion causing situation is evaluated as being controled by someone else or someone else is respobsible for the situation. A high level of *situational control* indicates that the situation is controlled by circumstance, which means that the situation is judged as being not controllable by anyone. A a low level of *situational control* indicates that the situation is evaluated as being controlled or controllable by a human. Further, a high level of *attention* or *effort* indicate increased *attentional activity* or *anticipated effort* respectively.

These findings show patterns, which can be used to differentiate between different emotions. The emotion *happiness*, for example, is described as an emotion with a high level of attentional activity and a low level of *uncertainty* and *anticipated effort*. Also, the emotion *happiness* is described with a moderate level of the *responsibility/control* appraisal, which translates to *increased self-responsibility/control*. The emotional

Emotion	Unpleasant	Resp./Control	Uncertainty	Attention	Effort	Sit. Control
Happiness	−1.46	0.09	−0.46	0.15	−0.33	−0.21
Sadness	0.87	−0.36	0.00	−0.21	−0.14	1.15
Anger	0.85	−0.94	−0.29	0.12	0.53	−0.96
Boredom	0.34	−0.19	−0.35	−1.27	−1.19	0.12
Challenge	−0.37	0.44	−0.01	0.52	1.19	−0.20
Hope	−0.50	0.15	0.46	0.31	−0.18	0.35
Fear	0.44	−0.17	0.73	0.03	0.63	0.59
Interest	−1.05	−0.13	−0.07	0.70	−0.07	−0.63
Contempt	0.89	−0.50	−0.12	0.08	−0.07	−0.63
Disgust	0.38	−0.50	−0.39	−0.96	0.06	−0.19
Frustration	0.88	−0.37	−0.08	0.60	0.48	0.22
Surprise	−1.35	−0.94	0.73	0.40	−0.66	0.15
Pride	−1.25	0.81	−0.32	0.02	−0.31	−0.46
Shame	0.73	1.31	0.21	−0.11	0.07	−0.07
Guilt	0.60	1.31	−0.15	−0.36	0.00	−0.29

Table 1: The locations of emotions along appraisal dimensions according to Smith and Ellsworth (1985), Table 6. Emotions considered in this study are marked as boldface. Resp./Control: Responsibility/Control; Sit. Control: Situational Control

experiences *surprise* and *happiness*, for example, can be distinguished using the *uncertainty* and the *responsibility/control* dimensions. In contrast to happiness, *surprise* is described as an emotion with a high level of *uncertainty* and a low level of *responsibility/control*. This low level of *responsibility/control* translates into an increased level of *other-responsibility/control*.

The *unplesant* emotions *disgust* and *guilt* on the other hand can be distinguished using the *anticipated effort* and the *responsibility/control* dimensions. In the table *guilt* is described as an emotional state with a high level of *self-responsibility/control* and a high level of *anticipated effort*. *Disgust* on the other hand is described as an emotional state with a low level of *anticipated effort* and a high level of *other-responsibility/control*.

Further, *anger* is described as an *unpleasant* emotion with increased *other-responsibility/control*, a low level of *uncertainty* and *situational control* and high level of *attention* and *anticipated effort*. With this insights the emotion *anger* can be distinguished from *disgust*, or *guilt* which both are, in contrast to *anger*, indicated with a low level of *attention*.

2.3 Emotion Classification

The automated classification or identification of emotion based on text is a task in which documents, sentences or words are associated with emotions. From a natural language processing perspective the automated classification of emotions based on text is usually *multiclass* – though sometimes also a *multilabel* – classification problem in which one or multiple emotions are predicted for a given textual input. In contrast to a verbal conversation, in which emotions can also be judged and predicted using gestures and facial expressions, the automated emotion classification based on text can only use the given words for predictions.

For this reason even the manual annotation of data with a finite set of possible emotions can be a challenging task as seen in the work presented by Schuff et al. (2017), in which the inter annotator agreement for the emotion *fear*, for instance, only ranged from $\kappa = 0.08$ to $\kappa = 0.25$ between the annotators. The emotional standpoint of an annotator does also have an impact on the annotation quality (Buechel and Hahn, 2017b). The interpreted emotion from the view of the writer of a sentence can, for example, differ from the view of the reader of a sentence. In addition emotion predictions based on a sentence can heavily vary between different persons, contexts or even by the age and lifelong experience. The sentence

“I felt ... when my mom offered me curry”

(from enISEAR dataset (Troiano et al., 2019)), for instance, can be interpreted with a variety of different emotions. People enjoying eating curry are likely to associate the sentence with the emotion *joy*, while another person could associate *disgust*.

Most approaches in emotion classification are based on emotion theories statet in psychology studies. The set of possible emotions, for instance, often follows fundamental emotion theories like proposed by Ekman (1992) or Plutchik (1980). For training an emotion classifier various datasets from different domains (or genres) exist. An early work, for example, is the dataset *Tales* (Alm et al., 2005), in which sentences of fairy tales are annotated with Ekmans’s six basic emotions. Other domains are, for example, news headlines in the dataset *AffectiveText* (Strapparava and Mihalcea, 2007b), blogs in the dataset *Blogs* (Aman and Szpakowicz, 2007) or microblogs like in the datasets *SSEC* or *TEC* (Schuff et al., 2017; Mohammad, 2012). Other dataset, like *ISEAR* (Scherer and Wallbott, 1997) or *enISEAR* (or *deISEAR*) (Troiano et al., 2019) use descriptions of event, which caused specific emotions as the basis for classification. Further, some datasets like *EmoBank* (Buechel and Hahn, 2017a), for instance are annotated with Valence-Arousal-Dominance (VAD) (Russell and Mehrabian, 1977) meta representations for emotions.

Emotion classification can, similar to text classification in general, be divided in rule-based algorithms and machine learning approaches (Bostan and Klinger, 2018). Rule-based classification are usually based on lexical resources like emotionally charged words and use rules to identify an affiliation of a given input to an emotion. Supervised feature-based machine learning systems on the other hand use pre-labeled datasets in order to learn to classify emotions from text. Machine learning models used in emotions classification are the same as used in text classification. Such models are, for instance, Naive Bayes, Support Vector Machines (SVM) or Maximum Entropy models.

However, state-of-the-art models for the automated classification of emotions are, like in text classification, neural network based models and linear models like MaxEnts and SVMs are mostly used in order to retrieve a baseline performance for comparison (Zhang et al., 2018b; Schuff et al., 2017; Klinger et al., 2018). Popular artificial neural network based models are for example convolutional neural networks (CNNs) or recurrent neural networks (RNNs), like the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or BiLSTMs (Bidirectional Long Short-Term Memory) (Schuster and Paliwal, 1997).

In the shared task of predicting emotions from text *IEST: WASSA-2018*, all top performing submissions used neural model in combination with embeddings (Klinger et al., 2018). In the three submissions, which were showing the best results all models used BiLSTMs, or BiLSTMs in combination with other neural model (Rozental et al., 2018; Balazs et al., 2018; Chronopoulou et al., 2018). Further, Schuff et al. (2017) use neural models (CNN, LSTM, BiLSTM) and compare them to linear classifiers, namely SVM and MaxEnt. In their study on the dataset SSEC the BiLSTM shows the best results.

Multi-task learning has also been shown to be effective in the automated classification of emotions as shown by Akhtar et al. (2019). They proposed a multi-task learning framework that jointly performs sentiment and emotion analysis on different inputs like text, acoustic and visual frames from a video. This approach led to performance improvements for both sentiment analysis and emotion recognition. Bostan and Klinger (2018) showed, transferring models trained on dataset to another dataset can lead to performance drops. Tafreshi and Diab (2018) proposed a multi-task learning framework, in which a classifier is learning to predict emotions using datasets of different domains. With this work they showed that this approach helps countering such performance drops in the cross-corpus evaluation of models.

3 Corpus Creation and Analysis

3.1 Datasets

In this bachelor thesis the set of appraisal dimension from the findings presented by Smith and Ellsworth (1985) were annotated on the recently published *enISEAR* dataset (Troiano et al., 2019). This dataset contains 1001 instances in form of event descriptions written in the English language, which are single labeled with the seven emotions *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*. The labels are well-balanced with 143 event descriptions for each emotional label.

For comparison the proposed models were also tested on the additional datasets *ISEAR* (International Survey On Emotion Antecedents And Reactions (Scherer and Wallbott, 1997)), *TEC* (Twitter Emotion Corpus (Mohammad, 2012)) and SSEC (Schuff et al., 2017; Mohammad et al., 2016a).

Like the *enISEAR*, the *ISEAR* dataset provides English descriptions focused on emotional events labeled with the same seven emotions *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*. The labels are almost balanced with about 1096 descriptions for each emotion label. The *ISEAR* and the *enISEAR* dataset were created using a *self reporting* process, in which participants created event descriptions for given emotions from a writers perspective.

The *TEC* dataset contains 21051 collected Twitter-Posts (tweets) annotated with the six basic emotions proposed by Ekman (*anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*). This dataset was created by collecting microblogs from a social media platform, namely

Emotion	enISEAR	ISEAR	TEC	SSEC
Anger	143 (14.3%)	1,096 (14.3%)	1,555 (7.4%)	2,902 (59.6%)
Anticipation	—	—	—	2,700 (55.5%)
Disgust	143 (14.3%)	1,096 (14.3%)	761 (3.6%)	2,183 (44.8%)
Fear	143 (14.3%)	1,095 (14.3%)	2,816 (13.4%)	1,840 (37.8%)
Guilt	143 (14.3%)	1,093 (14.3%)	—	—
Joy	143 (14.3%)	1,094 (14.3%)	8,240 (39.1%)	2,067 (42.5%)
Sadness	143 (14.3%)	1,096 (14.3%)	3,830 (18.2%)	2,644 (54.3%)
Shame	143 (14.3%)	1,096 (14.3%)	—	—
Surprise	—	—	3,849 (18.3%)	1,108 (22.8%)
Trust	—	—	—	1,713 (35.2%)
# Instances	1,001	7,666	21,051	4,868

Table 2: Comparision of the datasets used in this work, showing labels used in the specific dataset and their distribution. In addition the total amount of instances in the datasets is shown.

Twitter, in which users used hashtags to notify others about the emotion associated with the written messages (tweets). These hashtags were then used as class labels and removed from the instances to form a classification problem. This can be seen as a crowdsourcing-like annotation in which most annotators only annotated one instance. Note that they collected 21051 instances from 19059 authors (Mohammad, 2012). In contrast to *enISEAR* and *ISEAR*, this dataset is not balanced across the different classes with 39.1% of the instances labeled with the emotion *joy* and only 3.6% of the instances labeled with the emotion *disgust*.

In addition, the *SSEC* dataset is used for baseline model validation. This dataset is in contrast to the other datasets labeled with multiple labels per instance. The data in this corpus is labeled with the emotions *anger, anticipation, disgust*, *fear, joy*, *sadness*, *surprise* and *trust*. The dataset is the result of a post-annotation with emotion labels of the SemEval 2016 Stance Data corpus (Mohammad et al., 2016b). Further the *SSEC* dataset is divided, in contrast to the other datasets, in a train set and test set and contains a total of 4.868 instances. Like in the *TEC* dataset, the amount of instances labeled with specific emotions is not balanced in the *SSEC* dataset. *Anger*, for example, is annotated on 2902 instances (59.6%), while *surprise* is only annotated on only 1108 instances (22.8%).

Table 2 shows a comparison between the different datasets. The table shows the amount of instances, which are labeled with specific emotions and their relative counts notmialized by the total amount of instances. Note that the total amout of instances for the *enISEAR*, *ISEAR* and *TEC* dataset is equal to the sum of instances labeled with the specific emotions. This is not true for the *SSEC* dataset, since instances in *SSEC* are labeled with multiple emotions.

3.2 Annotation Guideline

In order to guide the annotators through the process of annotating the instances of the *enISEAR* corpus an annotation guideline was created. In this guideline the different appraisal dimensions were formulated as sentences for an easier understanding. Discussions with the annotators led to the understanding that they find it hard to distinguish the appraisals control and situational control. For this reason the two dimensions *situational control* and *responsibity/control* were split into three separately evaluated dimensions. *Responsibilty* was used as a label only for juding the *responsibilty* of the author the of a sentence for bringing about the situations, while the label *control* was used to evaluate *self-* vs *other- control*, i.e. is the writer of the sentence in control of the situation or someone else. Further another label called *circumstance* was introduced, which was used to evaluate human vs. non-human *control*

In contrast to the continuous scales used by Smith and Ellsworth (1985) for rating the appraisal dimensions, a binary annotation setting was used in this work. Early experiments with more than two possible values for the evaluation of a single appraisal dimensions were producing too much dissimilarities between the annotators and a worse agreement compared to a binary setting.

For the final annotation the annotators were instructed to read the event descriptions from the *enISEAR* dataset and answer the following questions [2].

Most probably, at the time when the event happened, the writer...

- ...wanted to devote further attention to the event. (*Attention*)
- ...was certain about what was happening. (*Certainty*)
- ...had to expend mental or physical effort to deal with the situation. (*Effort*)
- ...found that the event was pleasant. (*Pleasantness*)
- ...was responsible for the situation. (*Responsibility*)
- ...found that he/she was in control of the situation. (*Control*)
- ...found that the event could not have been changed or influenced by anyone. (*Circumstance*)

3.3 Annotation Procedure and Analysis

The annotation of the event descriptions from the *enISEAR* dataset was done by three annotators between the age 25 and 30. One annotator is a female Ph.D. student of computational linguistics. The other two annotators are a male graduate students of software engineering. The annotators were trained on the task using a total of four training iterations. Every iteration consisted of 15-20 hand-picked samples from the *ISEAR* dataset (Scherer and Wallbott, 1997). After every iteration, differences in the annotation were discussed in face-to-face meetings and the annotation guideline was refined in order to clear ambiguities.

The first iteration started only with two annotators and a Cohen's κ score of 0.62 was observed. All other training iterations were performed by all three annotators. In the second iteration pairwise Cohen's scores of 0.83, 0.15 and 0.15 were observed, showing misconceptions of the annotator who joined in this iteration. In the third and fourth iteration, scores of 0.71, 0.73, 0.67 and 0.72, 0.64, 0.64 were observed. Except for

[2] For the original guideline presented to the annotators see the Appendix.

iteration two the annotators did not have access to the emotion labels associated with the instances. Giving the annotators access to the emotion label led to a substantial improvement in the agreement (from κ=0.62 to κ=0.83). In the final annotation each instance from the enISEAR dataset was annotated by all three annotators. However, in order to evaluate the annotators performance similar to the automatic model, the annotators were not given access to the emotional label assigned with the instances. The left side of Table 3 is showing the pairwise inter-annotator scores of the final annotation. The scores show that the difficulty of annotating appraisals varies between the dimensions. However, the annotations between the annotators on specific appraisal dimensions show a similar agreement. The dimension *pleasantness* (∅.89) shows the highest agreement, followed by *responsibility* (∅.63) and *control* (∅.58). The lowest agreement scores are observed for *attention* and *certainty* (both ∅.31). Across all annotators and appraisal dimensions an average agreement score of κ=0.58 was observed. Overall these result reveal that rating appraisal dimensions on given event descriptions is a challenging task.

The final post-annotated dataset was created using the majority vote between the individual annotators. This majority vote was then compared to the individual annotators. The right side of Table 3 shows the pairwise agreement between each annotator and the majority vote. The highest agreement is again observed on the dimensions *pleasantness* (∅.94), followed by *responsibility* (∅.81) and *control* (∅.78). Again *certainty* (∅.62), *attention* (∅.64) show the lowest agreement. With an average agreement score persistently above κ=0.62 and an average score of κ=0.76 across all annotators and dimensions an acceptable agreement between annotators and the majority vote was observed.

	Cohen's κ							
	between annotators				annotator–majority			
Appraisal Dimension	A1/A2	A1/A3	A2/A3	∅	A1	A2	A3	∅
Attentional Activity	.28	.24	.41	.31	.50	.76	.66	.64
Certainty	.41	.23	.29	.31	.62	.77	.46	.62
Anticipated Effort	.38	.33	.26	.32	.69	.67	.62	.66
Pleasantness	.89	.88	.90	.89	.93	.96	.94	.94
Responsibility	.68	.57	.63	.63	.80	.88	.76	.81
Control	.65	.56	.52	.58	.84	.81	.70	.78
Circumstance	.52	.32	.28	.37	.80	.69	.49	.66
Average	.59	.48	.52	.53	.77	.82	.70	.76

Table 3: Cohen's κ between all annotator pairs (left), between each annotator and the majority vote (right) and average scores.

3.4 Post-annotation Analysis

Table 4 shows the co-occurrence counts and normalized pointwise mutual information (PMI) across emotion and appraisal dimension pairs. In addition the table shows the mutual information of the different appraisal dimensions across all emotions. The most frequently annotated appraisal reveals to be *certainty*, followed by *attention* and *anticipated effort* with about 76%, 67% and 40% of the instances annotated respectively. The most rarely annotated appraisal is *pleasantness*, followed by *control* and *circumstance* with about 15%, 23% and 24% of the instances.

Instances labeled with the emotion *anger* show a strong association with the dimensions *attention* and *certainty* and are moderately associated with *anticipated effort*. *Disgust* on the other hand is mostly annotated with *certainty* and moderately with *attention* and *anticipated effort*. *Fear* is very often annotated with *attention* and *anticipated effort* and moderately with *circumstance*. The emotion *guilt* is mostly associated with the dimensions *certainty* and *responsibility*, while also showing a moderate association to *control*. Non surprisingly *joy* is almost always annotated with the *pleasantness* dimension. Further, *joy* is annotated often in combination with *attention* and *certainty* and shows the most annotations with the appraisal *attention* compared to other emotions. *Sadness* shows the most co-occurrences with the appraisal *circumstance* and also a lot of co-occurrences with *attention* and *certainty*. The emotion *shame* shows, similarly to *guilt*, many co-occurrences with *certainty*, *responsibility* and moderate association with *control*.

Surprisingly two instances labeled with *disgust*, four with *fear*, one with *sadness* and

	Appraisal Dimension														
Emotion	Attention		Certainty		Effort		Pleasant		Respons.		Control		Circum.		Total
Anger	129	+0.14	119	+0.04	60	+0.02	0	−1.00	9	−0.38	1	−0.50	5	−0.36	323
Disgust	67	−0.13	134	+0.10	40	−0.11	2	−0.38	14	−0.31	11	−0.24	24	−0.09	292
Fear	129	+0.14	13	−0.49	121	+0.35	4	−0.30	43	−0.07	18	−0.15	66	+0.24	394
Guilt	55	−0.19	132	+0.10	36	−0.14	0	−1.00	133	+0.45	88	+0.41	11	−0.25	455
Joy	139	+0.19	140	+0.13	4	−0.48	141	+0.97	65	+0.07	41	+0.07	25	−0.09	555
Sadness	122	+0.11	112	+0.01	88	+0.18	1	−0.44	7	−0.41	2	−0.45	97	+0.45	429
Shame	32	−0.32	111	+0.01	51	−0.04	1	−0.44	106	+0.30	67	+0.27	12	−0.24	380
Total	673	(0.28)	761	(0.30)	400	(0.21)	149	(0.47)	377	(0.36)	228	(0.21)	240	(0.19)	

Table 4: Instance counts and pointwise mutual information across emotions and appraisal annotations. In addition the mutual information for the appraisal dimensions across all emotions is shown on the bottom in brackets.

one with *shame*, were annotated with the appraisal dimension *pleasantness* by the majority vote. These instances are shown in the upper part of Table 5. For *disgust* such an instance was, for example, the text ”I felt ... when I had bean soup”, which can be interpreted in several ways depending on one’s personal background. Someone, who likes bean soup would associate this sentence with *joy* and therfore as *pleasant*. While others could associate *disgust* and *unpleasantness*. In addition two instances labeled with the emotion *joy* were not annotated with the appraisal *pleasantness* (shown in the lower part of Table 5). All sentences in Table 5 show such multiple ways of interpretation. Most likely, these dissimilarities emerged from the differences in the annotations setups. While in the annotation of the enISEAR datasets subjects recalled own experiences, the annotators in this study interpreted the situations from a readers perspective not knowing preferences of the writer of the sentence.

The results of the appraisal annotations are also interesting in comparison to the findings by Smith and Ellsworth (1985) (See Table 1). Most of the results of the annotations in this work are consistent with their findings. In their findings the emotion *anger*, for instance, is indicated with the lowest level of *responsibility* and *situational control*, which is also observed in this study. The emotion *fear* shows a strongest association to the dimension *anticipated effort*, a strong association with *situational control* and association to *certainty*, which is similar to the results in this study. In addition, the emotion *sadness* shows the highest level of *situational control* in the findings of Smith and Ellsworth (1985) which is also consistent with the results of the annotation of the enISEAR dataset. *Shame* and *guilt* are strongly associated with *responsibility*, which corresponds to blaming the self (Tracy and Robins, 2006), while *shame* is less associated with *certainty* than *guilt*. The results of the annotations in

Emotion	P	Text
Disgust	1	I felt ... when my mom offered me curry.
Disgust	1	I felt ... when I had bean soup.
Fear	1	I felt ... when I first flew on a plane.
Fear	1	I felt ... when I cycled down a mountain in Scotland.
Fear	1	I felt ... when I was abseiling down a cliff-face.
Fear	1	I felt ... when I rode a rollercoaster at a theme park.
Sadness	1	I felt ... today when I thought of an anniversary that today is for me. I relived different moments and I know I will for the next few days.
Shame	1	I felt ... that a member of staff was being nice to me when I was not able to afford what they showed me and which suited me.
Joy	0	I felt ... when Will Young won Pop Idol because he was nicer than Gareth Gates.
Joy	0	I felt ... when I knew that I was going back to Florida a year earlier than I thought I would.

Table 5: Instances labeled with a negative emotion and the positive appraisal pleasantness (top) and instances labeled with joy and not the appraisal pleasantness (bottom). P: Pleasantness annotation

this study also reflect this finding. Further the emotion *joy*, or *happiness* in Table 1, is associated as very pleasant, the highest amount of *certainty* and the least amount of *anticipated effort*, which is also observed in this study.

There are also some difference between the findings of Smith and Ellsworth (1985) compared to the results of this study. For instance, *anger* is associated with a high level of *anticipated effort* in their study and a low level in this study. Further, is *sadness* associated with a low level of *anticipated effort*, while in this study a moderate level is observed. Presumably, these differences arose from the different annotations setups. While their subjects recalled personal events, which were then rated along the appraisal dimensions, the annotators in this study evaluated the instances only from a reader's perspective.

However, the analysis of the appraisal annotation reveals that the distribution of appraisal dimensions differs across the different emotions. Similarly to the findings by Smith and Ellsworth (1985), different patterns in the evaluation of the appraisal dimensions for specific emotions can be observed. Figure 6 tries to visualize such patterns for the different emotions. In this figure every appraisal dimensions is represented by an axis, in which one line equals to 20 annotations. This method of visualizing is creating polygonal shapes for every emotions. The figure shows that these shapes differ between the different emotions, except for *guilt* and *shame*, which appear similar.

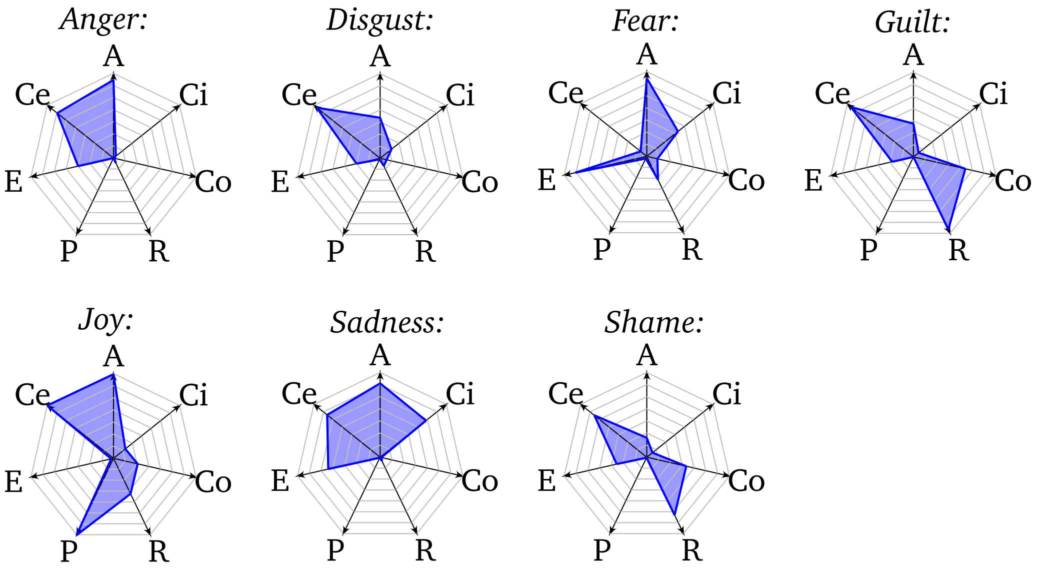

Anger: *Disgust:* *Fear:* *Guilt:* *Joy:* *Sadness:* *Shame:*

Figure 6: Emotion labels visualized using the amount of appraisal annotations. A: Attention, Ce: Certainty, E: Effort, P: Pleasantness, R: Responsibility, Co: Control, Ci: Circumstance. One line is equal to 20 annotations.

4 Methods for Appraisal based Emotion Prediction

4.1 Task T→E: Baseline System

The experiment in this study is divided in four tasks. In the first task (*Task T→E*) a traditional system, which predicts *emotions from text* (Schematically shown in Figure 7) was created in order to receive a baseline performance on the enISEAR dataset.

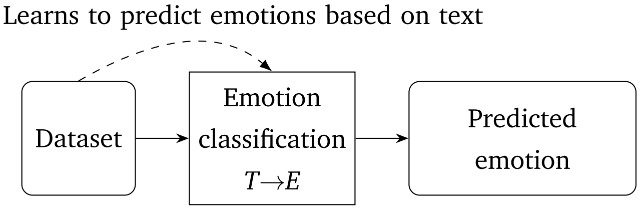


Figure 7: Visual representation of the baseline system, which is predicting emotions on basis of text.

For this baseline performance two different configurations were evaluated. A *Maximum Entropy* (MaxEnt) classifier (Berger et al., 1996) with unigram bag-of-words (BOW) features and L2 regulation, which corresponds to a neural network with one hidden layer (shallow neural network) with softmax activation and categorical cross-entropy loss. The motivation to use a MaxEnt are the easy to reproduce results MaxEnt models provide. Further MaxEnt models have shown to be able to perform often almost en par on emotion classification tasks in comparison to neural approaches (Schuff et al., 2017).

In addition a *Convolutional Neural Network* (CNN), following Kim (2014), was created. The hyperparameter configuration of this CNN was the result of a *grid search* (See Table 6) evaluating different paramaters according to their performance and training complexity on the enISEAR dataset. In this *grid search* different embedding dimensions,

Hyper-parameter	Configurations	Selected
Embedding dimension	50, 100, 300	300
Number of filters	64, 128, 256, 512	128
Kernel sizes	2, 3, 4, 5	2, 3, 4
Dropout rate	0, 0.1, 0.2, 0.5	0.5
Batch size	16, 32, 64, 128	32

Table 6: Parameter configurations evaluated with a grid search and the selected configuration for the CNN baseline system.

convolutional filter sizes, convolutional kernel sizes, dropout rates and batch sizes were testet. The final CNN uses a 300-dimensional embedding layer with a pretrained GloVe (Glove840B) embeddings (Pennington et al., 2014) [3] and convolutional filter sizes of 2, 3, 4. The convolutional layers are followed by a max-pooling layer of lenght 2. After the pooling layer a dropout (Srivastava et al., 2014) of 0.5 followed by a fully connected layer is used. In addition the *Rectified Linear Unit* (ReLU) (Nair and Hinton, 2010) function is used as an activation function for fully connected and convolutional layers.

4.2 Task T→A, A→E: Pipeline System

The second task is the first step of a real-world pipeline setup, which first predicts appraisal dimensions from text and afterwards predicts an emotion based on the predicted appraisal. This task is schematically visualized through the first three steps shown in Figure 8. In this task a classifier was created, which learns to predict appraisal dimension based on text, also referred as *Task T→A*. For this task the classifier is using the annotated appraisals on the enISEAR dataset created during this study. The model configurations (MaxEnt and CNN) were kept the same for this task except for the use of the sigmoid activation function and binary cross-entropy loss instead of the softmax activation function and categorical cross-entropy loss. This is necessary because the task of predicting appraisals is a multilabel classification task instead of the multiclass classification task like it was in the first task.

The third task, also refered as *Task A→E*, is the second step in this pipeline setup.

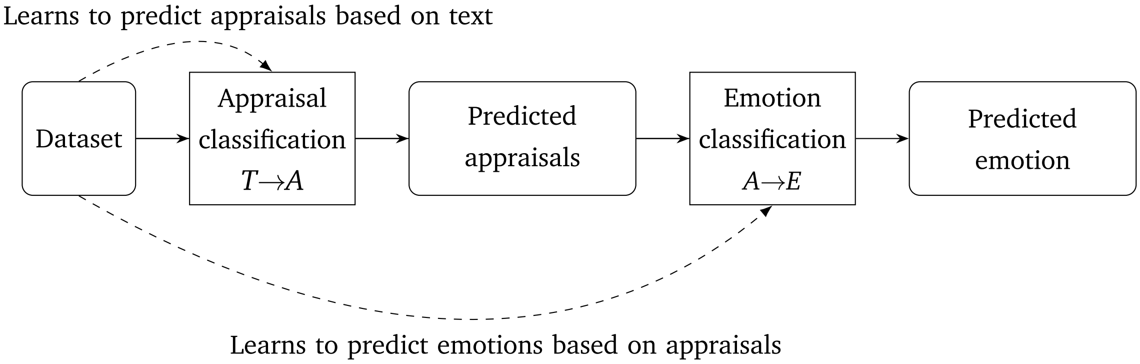

Learns to predict appraisals based on text
Dataset
Appraisal classification T→A
Predicted appraisals
Emotion classification A→E
Predicted emotion
Learns to predict emotions based on appraisals

Figure 8: Visual representation of the pipeline system, which is predicting appraisals in the first step and predicting emotions based on the predicted appraisals in the second step.

[3] https://nlp.stanford.edu/projects/glove/

This task is schematically visualized through the last three steps shown in Figure 8. In this task the created classifier, is learning to predict emotions from appraisal dimensions based on the annoteted appraisals on the enISEAR dataset.

For this task the MaxEnt model was kept the same. The CNN on the other hand was changed to a simple shallow neural network with two hidden (fully connected) layers, since the features for this model are only seven boolean variables (the appraisal dimensions annotated on the enISEAR dataset). In addtion the ReLU functions was used as activation function for the hidden layers and a dropout of 0.5 was applied after each hidden layer.

4.3 Task T→A/E: Multi-task System

The fourth task, also referred as *T→A/E* consist of a CNN system with a similar structure to the first task (T→E). This time the convolutional layer is shared between the task of predicting *emotions from text* and predicting *appraisals from text*, which corresponds to a multi-task system (See Figure 9). The model uses two output layers, one for emotion prediction with softmax activation and one for appraisal predictions with a sigmoid activation.

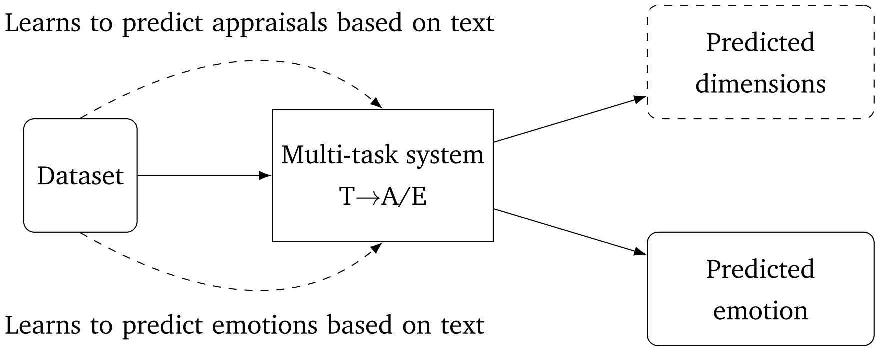

Learns to predict appraisals based on text
Predicted dimensions
Dataset
Multi-task system T→A/E
Predicted emotion
Learns to predict emotions based on text

Figure 9: Visual representation of the multi-task system, which is learning to predict emotions and appraisals jointly.

4.4 Ensemble System

Finally, in addition to the previously presented configurations another experiment was conducted in which an oracle is selecting a prediction of either the baseline configuration (Task T→E) or the pipeline configuration (Task T→A, A→E). In this experiment it is assumed that the oracle is capable of always selecting the correct prediction if one of the two different configurations provided a correct prediction for an emotion classification task.

5 Results

In the experiments all models were trained using ADAM (Kingma and Ba, 2014) and validated using repeated 10-fold cross-validation (CV) with a total of 10 repetitions. In every cross-validation repetition the dataset was randomly shuffled. In order to restrict further randomness between the evaluation of the different experiments these 10 randomly arranged instances of the dataset were kept the same across all experiments. Further, no preprocessing other than case folding was done in the different datasets.

5.1 Baseline System (Task T→E)

The first experiment started with the evaluation of the baseline model configurations (Task T→E). This experiment consists of a classifier predicting emotions only on basis of the input texts and was evaluated using the two model configurations for a MaxEnt classifier and a CNN as described previously.

The results of the model evaluation performed on the *enISEAR* dataset is shown in Table 9 in the columns labeled with T→E. With an average-micro and average-macro F_1 score of .46 and .60 for the MaxEnt and the CNN model respectively the results show that the CNN model is outperforming the MaxEnt model substantially. In addition, the CNN model also exceeds the performance of the MaxEnt model if single emotions are compared. The emotion *anger* shows the biggest performance gap

	ISEAR						TEC						SSEC					
	MaxEnt			CNN			MaxEnt			CNN			MaxEnt			CNN		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	49	48	48	52	55	54	48	31	38	50	34	41	78	70	74	78	77	78
Anticipation		—			—			—			—		72	61	66	72	63	67
Disgust	59	59	59	64	62	63	45	22	30	47	26	33	64	54	58	67	60	63
Fear	70	69	69	72	73	72	67	54	60	66	57	61	59	40	48	64	38	47
Guilt	50	50	50	55	50	52		—			—			—			—	
Joy	72	76	74	73	79	76	64	80	71	66	80	72	55	58	56	59	60	59
Sadness	64	62	63	66	63	65	49	47	48	51	49	50	63	67	65	66	65	66
Shame	47	49	48	53	51	52		—			—			—			—	
Surprise		—			—		57	50	53	57	51	54	49	21	29	46	16	23
Trust		—			—			—			—		61	43	50	59	51	55
Macro ∅	59	59	59	62	62	62	55	47	50	56	50	52	63	52	56	64	54	57
Micro ∅			59			62			59			61	65	55	60	67	58	62

Table 7: Performance comparison of the Text-to-Emotion baseline (T→E) model configurations on the datasets *enISEAR*, *ISEAR* and *TEC*.

between the two models with an F_1 score of .34 for the MaxEnt model and .52 for the CNN model. The lowest performance gap shows the emotion *guilt* with scores of .44 and .36 respectively. Further, both models show an optimal balance between precision and recall scores.

In addition, the baseline systems were evaluated on the datasets *ISEAR*, *TEC* and *SSEC*. Note that in contrast to the other dataset *SSEC* consist of a training set and a test set, which were used to evaluate the models. Further, this evaluation using the test set and training set on *SSEC* was repeated ten times and then averaged. The result of this evaluations are shown in Table 7. These tests of the baseline configurations provide consistent results to other studies with a MaxEnt configuration like the work presented by Bostan and Klinger (2018). While they report a micro F_1 score of .64 on *ISEAR*, and .56 on *TEC* for a MaxEnt model, this study produced micro F_1 scores of .59 for both datasets. Note that Bostan and Klinger (2018) excluded the emotions *shame* and *guilt* in the *ISEAR* dataset, which explains differences in the micro F_1 score.

Further the results for predicting emotions from text with the CNN model are comparable to the baseline results observed by Zhang et al. (2018a) on *ISEAR* and *TEC*. They report an accuracy score of .64 for *ISEAR* and .62 for *TEC*, while this study reports micro F_1 scores of .62 and .61, respectively. Note that accuracy and micro F_1 score are equal in a multi-class classification setting.

The results on the SSEC corpus, are comparable to the performance observed by Schuff et al. (2017). While they report a micro-average F_1 score of .58 for a MaxEnt configuration, in this study a micro-average F_1 score of .60 was observed. For a CNN configuration they report a micro-average F_1 score of .60, while this study observed .62. The performance of the baseline classifiers on SSEC from this work on the different emotions is also very similar to what was observed by Schuff et al. (2017). In the MaxEnt model and the emotion *Anger*, for example, the same F_1 score of .74 was observed.

5.2 Pipeline System (Task T→A/A→E)

5.2.1 Appraisal Prediction (Task T→A)

The second task (Task T→A), which is the first of two steps in the pipeline system, is predicting *appraisal dimensions from text*. The result of the models predicting appraisals on the *enISEAR* dataset are shown in Table 8. The results reveal that the CNN model outperforms the MaxEnt model in this task as well, with micro-average F_1

scores of .75 and .70 respectively. For the appraisal dimensions *attention*, *certainty* and *anticipated effort* the models perform similar, while *pleasantness*, *responsibility* and *control* show major performance drops in the MaxEnt configuration. The biggest loss in performance shows the appraisal *pleasantness* with an F_1 score of .58 in the MaxEnt configuration and an F_1 score of .70 in the CNN configuration.

Noticeable is also the decreased precision score on the appraisals *pleasantness*, *responsibility* and *control* in the MaxEnt configuration. Overall the precision is much better in the CNN model, with a macro-average precision score of .73 compared to the score of .64 in the MaxEnt model. Further, recall is slightly better in the CNN model with a macro-average recall score of .68 compared to the score of .64 in the MaxEnt configuration.

A difficulty in this task was the unbalanced training data, i.e. the amount of instances annotated with certain appraisals varied. The appraisals *attention* and *certainty*, which were annotated very frequently show the best and also a similar performance in both classifier configurations. The appraisal *attention* shows F_1 scores of .80 and .82 for the MaxEnt and the CNN model respectively, while the appraisal *certainty* shows scores of .84 (MaxEnt) and .85 (CNN). In addition, in the CNN configuration, these two appraisals are the only ones, in which a higher recall than precision was observed. Overall, the results show that the classifier perform worse on appraisals annotated annotated less frequently, like *control* or *circumstance*, than on appraisals which were annotated more often.

However, in the CNN configuration this is not true for the appraisal *pleasantness*, which was annotated in the least amount of instances comparing all appraisal dimensions. The model performs far better on the appraisal *pleasantness* in comparison to other

	Task T→A					
	MaxEnt			CNN		
Appraisal Dimension	P	R	F_1	P	R	F_1
Attention	80	79	80	81	84	82
Certainty	85	84	84	84	86	85
Effort	60	69	65	68	68	68
Pleasantness	62	54	58	79	63	70
Responsibility	58	62	60	74	68	71
Control	47	44	46	63	49	55
Circumstance	57	57	57	65	58	61
Macro ∅	64	64	64	73	68	70
Micro ∅	70	71	70	77	74	75

Table 8: Classifier performance on predicting appraisal dimensions.

less frequent annotated appraisals. These results reveal that classification of the appraisal *pleasantness* is therefore somehow “easier” for a classification model to solve in comparison to other less frequent annotated appraisals like *control* or *circumstance*.

In addition, the results reveal that the CNN configuration is able to cope with this imbalance better, since it shows a better performance on the less frequent annotated appraisals compared to the MaxEnt configuration.

5.2.2 Emotion Prediction based on Appraisals (Task A→E)

The second step of the pipelined is a classifier learning to predict *emotions from appraisals* (Task A→E). The results of this task are shown in Table 9 in the column A→E (Gold). Note that this classifier is learning to predict appraisals using the annotated appraisal dimensions and not the predicted appraisals of the previous Task T→A. Experiments showed that learning to predict emotions from predicted appraisals performs worse than learning to predict appraisals using the annotated appraisal dimensions.

The results of the second step of the pipelined classifier (Task A→E) show that for this task the MaxEnt model is almost en par with the shallow artificial neural network model. The MaxEnt model achieves a micro F_1 score of .64, while the neural model achieves a micro F_1 score of .66. The biggest differences in performance is observed the emotions *sadness* and *fear*, in which the ANN performs 8 and 7 percentage points

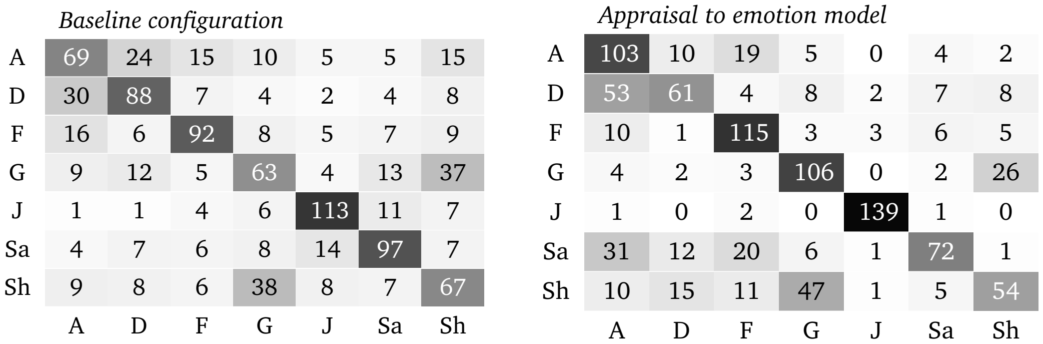
	MaxEnt									CNN			CNN, ANN			ANN		
	T→E			T→A,A→E			A→E (Gold)			T→E			T→A, A→E			A→E (Gold)		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	38	31	34	25	49	33	49	71	58	51	52	52	34	62	44	55	71	62
Disgust	48	48	48	37	28	32	59	42	49	65	63	64	59	34	43	53	48	51
Fear	60	59	59	50	51	51	63	81	71	69	71	70	55	55	55	79	78	78
Guilt	34	38	36	30	36	33	59	71	65	47	42	44	38	50	43	57	70	63
Joy	61	60	60	69	50	58	95	98	97	74	80	77	77	69	72	94	98	96
Sadness	53	55	54	58	41	48	71	48	58	69	67	68	58	40	47	69	63	66
Shame	27	28	28	33	20	25	58	36	45	44	45	45	36	24	29	56	35	43
Macro ∅	46	46	46	39	43	41	65	64	63	60	60	60	51	48	48	66	66	65
Micro ∅			46			39			64			60			48			66

Table 9: Comparison of the Text-to-Emotion baseline (T→E) with the performance of first prediction appraisal followed by emotion analysis (T→A,A→E) and the prediction of emotions learned on the annotated dimensions (A→E)

better than the MaxEnt model. Overall, in this setting a clear improvement is shown compared to the emotion classification only on basis of text (T→E). Comparing the micro-average F_1 scores between these tasks, the *appraisal to emotion* configuration outperforms the *text to emotion* configuration by 18 percentage points in the MaxEnt model and by 6 percentage points in the neural model.

In the neural configuration the biggest improvement is observed on the emotions *joy*, with an increased micro-average F_1 score from .77 to .96. The emotion *guilt* shows the second biggest increase of performance, with a F_1 score of .44 in the *text to emotion* configuration compared to a F_1 score of .63 in the *appraisal to emotion* configuration. However, the performance on predicting the emotions *disgust*, *sadness* and *shame* decreases in the *appraisal to emotion* configuration compared to *text to emotion* the configuration. The biggest performance drop is observed on the emotion *disgust* with a loss of 13 percentage points in the F_1 score

In Figure 10 the confusion matrices for the neural network configurations on the *text to emotion* task (left) the *appraisal to emotion* task (right). This confusion matrices represent one run of the 10-fold cross validation. Note that the folds where the same in both tasks. Predicted emotions are shown in the rows, while the actual labeled emotions are shown in the columns. The *appraisal to emotion* configuration confuses the emotion *disgust* with *anger* much more often than the *text to emotion* configuration. Most likely, this is a result of the similar representation of the two emotions. Both are very often annotated with the appraisal *certainty* and moderately with *anticipated effort*. The only difference is that *anger* is very often annotated with *attention* and *disgust* only moderately.



Baseline configuration

	A	D	F	G	J	Sa	Sh
A	69	24	15	10	5	5	15
D	30	88	7	4	2	4	8
F	16	6	92	8	5	7	9
G	9	12	5	63	4	13	37
J	1	1	4	6	113	11	7
Sa	4	7	6	8	14	97	7
Sh	9	8	6	38	8	7	67

Appraisal to emotion model

	A	D	F	G	J	Sa	Sh
A	103	10	19	5	0	4	2
D	53	61	4	8	2	7	8
F	10	1	115	3	3	6	5
G	4	2	3	106	0	2	26
J	1	0	2	0	139	1	0
Sa	31	12	20	6	1	72	1
Sh	10	15	11	47	1	5	54

Figure 10: Confusion matrices for the neural *text to emotion* classifier (left) and the neural *appraisal to emotion* classifier (right) of one randomly selected 10-fold cross validation run. Columns: predicted emotions, rows: labeled emotions. A: Anger D: Disgust, F: Fear, G: Guilt, J: Joy, Sa: Sadness, Sh: Shame

The confusion matrices also reveal that in contrast to the *emotion to text* model the *appraisal to emotion* model is, confusing *sadness* with *anger*, *fear* and sometimes also with *disgust*. This is also most likely due to the shared representation. *Sadness* and *anger* are both very often annotated with the appraisals *attention* and *certainty*. The only difference is the *circumstance* appraisal, which is associated with *sadness* and not with *attention*. However, this appraisal was annotated very sparsely. Further, *shame* is more often confused by the appraisal model with the emotion *disgust*, *fear* and *guilt* compared to the *text to emotion* model.

The *text to emotion* model on the other hand is more often confusing the emotion *anger* with *disgust* or *shame* than the appraisal model. This is most likely caused by the different appraisal representation of the emotions *anger* and *disgust*. However, there are also some emotions, which are confused by both models. Especially *shame* is very often confused by with *guilt* by both models.

5.2.3 Pipeline Setting (Task T→A,A→E)

The results for the final pipeline configurations are shown in Table 9 in column T→A,A→E. The MaxEnt pipeline configuration shows only a slight performance drop compared to the MaxEnt baseline model with micro-average F_1 scores of .39 and .46, respectively. While there is almost no loss in performance comparing the emotions *anger*, *joy*, *guilt* and *shame*, the results show that the pipelines model significantly performs worse in predicting the emotions *disgust*, *fear* and *sadness*.

While the MaxEnt comparison only shows a performance decrease in the micro-average F_1 score of 7 percentage points, the CNN/ANN pipeline configuration shows a more substantial performance decrease of 12 percentage points. The CNN baseline model shows a micro F_1 score of .60, while the pipelined configuration with appraisals achieves a micro-average F_1 score of .48. The biggest performance drop is shown by the emotions *sadness* and *disgust* with a drop of 21 percentage points comparing the F_1 scores of these emotions between the neural configurations. The lowest performance drop is observed by the emotion *guilt* with only one percentage point, followed by *joy* with 5 percentage points. Overall, the neural pipeline configuration outperforms the MaxEnt *text to emotion* baseline configuration and the MaxEnt pipeline configuration.

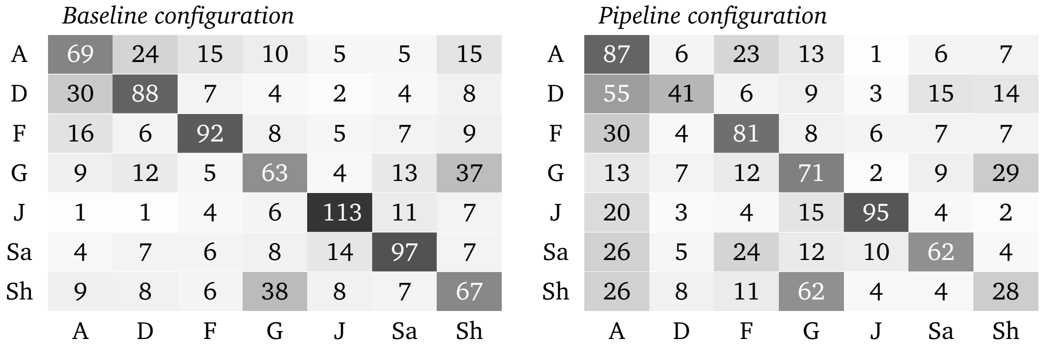
Comparing the neural pipeline results (T→A, A→E) to the neural baseline results (T→E), clear differences in the balance of precision and recall can be observed for some emotions. In the emotions *anger* and *guilt*, for instance, precision and recall scores are almost balanced in the baseline configuration, while the pipeline configuration shows a much higher recall, with a lower precision score. The emotions *disgust*, *joy*, *sadness* and *shame* on the other hand show lower recall scores than precision scores in the pipeline configuration, while the baseline configuration these scores are almost balanced in the baseline configuration.

The confusion matrices for the neural network configurations are shown in Figure 11. On left side the *text to emotion* task is shown while the right side represents the *pipeline* configuration. Again, these confusion matrices represent one run of the 10-fold cross validation and the folds where the same in both tasks. Predicted emotions are shown in the rows, while the actual labeled emotions are shown in the columns.

The matrix for the pipeline configuration reveals that the emotion *anger* is confused with all other emotions frequently. Clearly this confusion is introduced by the *text to appraisal* prediction model and not by the *appraisal to emotion model*. Since the appraisals *attention* and *certainty* are strongly associated with *anger* and show much higher recall scores than the other appraisals, *anger* is predicted more often. This is also reflected in the high recall score of .62, compared to the low precision score of .34 in the emotion *anger* in the pipeline configuration.

Further, comparing the matrix from the *appraisal to emotion* model (Figure 10 on the right) to the confusion matrix of the pipeline configuration reveals that confusions, which arise in the *appraisal to emotion* model are passed to the pipeline configuration. However, there are also some differences between those two matrices. *Disgust*, for example, is more often confused with *sadness* and *shame* by the pipeline configuration.

As an additional experiment the neural pipeline system was tested on the datasets *ISEAR* and *TEC*. Since there are no appraisal annotations for these datasets, the *text to*



Baseline configuration

A	69	24	15	10	5	5	15
D	30	88	7	4	2	4	8
F	16	6	92	8	5	7	9
G	9	12	5	63	4	13	37
J	1	1	4	6	113	11	7
Sa	4	7	6	8	14	97	7
Sh	9	8	6	38	8	7	67
	A	D	F	G	J	Sa	Sh

Pipeline configuration

A	87	6	23	13	1	6	7
D	55	41	6	9	3	15	14
F	30	4	81	8	6	7	7
G	13	7	12	71	2	9	29
J	20	3	4	15	95	4	2
Sa	26	5	24	12	10	62	4
Sh	26	8	11	62	4	4	28
	A	D	F	G	J	Sa	Sh

Figure 11: Confusion matrices for the neural *text to emotion* classifier (left) and the neural *pipeline* configuration (right) of one randomly selected 10-fold cross validation run. Columns: predicted emotions, rows: labeled emotions. A: Anger D: Disgust, F: Fear, G: Guilt, J: Joy, Sa: Sadness, Sh: Shame

appraisal (T→A) and *appraisal to emotion* (A→E) models were trained on the *enISEAR* dataset. For this experiment, the evaluation method was slightly changed to a method in which the *enISEAR* corpus was used a training set and the *ISEAR* and *TEC* datasets were used as test sets only. Note that for the *TEC* dataset, the emotion *surprise* was excluded because it is not present in the *enISEAR* dataset. This experiment was then compared to the *text to emotion* model. For a fair comparison the *text to emotion* (T→E) was also trained on the *enISEAR* dataset. This methodology was repeated 10 times and then averaged. The results are shown in Table 10.

First of all, the results reveal that learning to predict emotions based on text on the *enISEAR* dataset and testing on *ISEAR* and *TEC* leads to substantial performance drops in both dataset. The micro-average F_1 score in the *ISEAR* dataset drops from .62 to .41, while the score in the *TEC* dataset drops from .61 to .36. Comparing the results on specific emotion on *ISEAR*, reveals that the same emotions suffer from a performance drop like the comparison on the *enISEAR* between the *text to emotion* and the pipeline configuration. There is just a small performance loss observed in the emotions *anger*, *guilt* and *shame*. Like in all pipeline configurations the results on *ISEAR* and *TEC* show that the recall score for *anger* is way higher than in the baseline configuration.

In the *TEC* dataset the emotions *anger*, *disgust* and *joy* show only a slight performance drop, while *sadness* shows the biggest performance drop of 15 percentage point in the F_1 score. However, the emotion *fear* shows a performance increase of 7 percentage point in the pipeline configuration. Further, *fear* shows a much better recall and precision score.

	ISEAR						TEC					
	T→E			T→A,A→E			T→E			T→A,A→E		
Appraisal Dimension	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	31	33	32	22	47	30	17	33	22	12	55	19
Disgust	53	35	42	35	31	33	13	47	20	14	33	19
Fear	57	45	51	42	45	43	21	18	19	26	25	26
Guilt	32	29	30	23	25	24		—			—	
Joy	57	54	56	62	30	40	68	44	54	75	35	48
Sadness	63	40	49	55	22	32	31	30	30	26	11	15
Shame	25	51	34	25	23	24		—			—	
Macro ∅	45	41	42	38	32	32	30	34	29	31	32	25
Micro ∅			41			32			36			30

Table 10: Neural network pipeline configuration evaluated on *ISEAR* and *TEC* in comparison to a baseline, which consist of training on *enISEAR* and testing on *ISEAR* and *TEC*.

5.3 Multitask and Oracle Ensemble System

Finally, the results of the multitask system (T→A/E) and the ensemble system are presented in Table 11. The results show that the proposed multi-task learning model does not improve emotion predictions compared to the CNN baseline with similar F_1 scores across all emotions. With this result the question if the real-world pipeline setting (T→A, A→E) learns the same things as the baseline setting or if there are some inherent relationships between text, appraisal dimensions and emotions, still remains.

In order to answer this question an ensemble-like system was designed. This system is using an oracle which is selecting the correct result of predictions provided by the baseline system (Task T→E) and the pipeline system (Task T→A, A→E). In this model a prediction was accepted as a true positive if one of the two configurations provided the correct emotion. The result of this configuration is shown in Table 11 in the column "Ensemble". The enseble setup reveals a clear improvement over the baseline model. While the baseline model achieves an micro-average and macro-average F_1 score of .60 the ensemble model achieves an micro-average and macro-average average F_1 score of .70. The most noticeable improvement shows the emotion *anger* with an improvement of 21 percentage points, followed by the emotion *guilt* with an improvement of 18 percentage points. In contrast to this the emotion *joy* improves the least with F_1 scores of .77 and .80 for the baseline and the ensemble model respectively. These results show that the a text-based classifier and the appraisal based pipeline setup behave differently. This means that on some instances an appraisal based model is providing better predictions than the text-based model. This provides evidence that a model, which is informed about appraisal dimensions has the potential to contribute in predicting the correct emotion.

	T→E			T→A/E			Ensemble		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1
Anger	51	52	52	51	52	52	66	81	73
Disgust	65	63	64	64	64	64	78	68	73
Fear	69	71	70	70	68	69	76	77	77
Guilt	47	42	44	45	42	44	60	63	62
Joy	74	80	77	77	77	77	79	80	80
Sadness	69	67	68	68	68	68	74	70	72
Shame	44	45	45	43	43	43	58	51	54
Macro ∅	60	60	60	60	59	59	70	70	70
Micro ∅			60			59			70

Table 11: Comparison of the baseline CNN model (T→E), the multitask model (T→A/E) and the oracle ensemble experiment.

6 Discussion and Analysis

The results of the experiments show that the approach of predicting emotions based on appraisal has a potential to improve performance compared to predictions only on basis of text. However, this work was not able to show a direct improvement in a real-world setting, in which appraisals are predicted from text as a basis for emotion prediction. Table 12 shows example predictions provided by the real-world setting (pipeline configuration). The top part of this table consists of examples in which the pipeline configuration predicts the appraisal dimensions as they were annotated and then predicts the correct emotion based on the predicted appraisals.

The second part of the table shows examples in which the pipeline configuration is predicting the correct appraisals but not the correct emotion. This shows that correct appraisal predictions can lead to wrong emotion predictions if certain appraisal patterns are strongly associated to certain emotions. In the sentence "because I did something silly", annotated with the emotion *shame*, the appraisals *certainty*, *responsibility* and *control* were predicted as they were annotated. Further this combination of appraisals makes sense since if we experience an emotion after doing "something silly", we are certain about that. Also we are responsible and most likely we were in control of the situation. However, this combination of appraisals is more often associated with the emotion *guilt*, while *shame* is more often associated with the absence of the *control* appraisal. Therefore, the classifier is predicting *guilt* instead of *shame*.

Such a scenario is also observed in the sentence "a huge spider just plopped on down on the sofa besides me, staring me out" in which the classifier correctly predicted the appraisals *attention*, *certainty*, *anticipated effort* and *circumstance*. This combination of appraisals is associated with the emotion *sadness* and therefore the *appraisal to emotion* classifier is predicting *sadness* wrongly. The correct emotion for this sentence is *fear*, which is associated with almost the same appraisals with the difference that the appraisal *certainty* is not present in *fear*. On the other hand, according to the annotation guideline used in this work, the positive annotation of *certainty* makes sense in this sentence because the writer of this sentence "presumably was certain about what was happening", i.e the writer was certain that there was a spider. Most likely, this could have been avoided if the guideline for *certainty* would have been more precise and closer to the definition provided by Smith and Ellsworth (1985), because they also include the "predictability" of a situation in their *certainty* dimension. Most likely, this would have resulted in a negative annotation of the dimension *certainty* because the spider popping up was not predictable by the experiencer of the emotion.

Further, with the absence of the appraisal *certainty* the *appraisal to emotion* classifier would most likely predict the correct emotion *fear* for this instance.

The third part of Table 12 shows a subset of instances in which appraisals and emotions were not correctly predicted. The appraisals, which were not predicted as annotated by the *text to appraisal* classifier are shown in bold. In the sentence "when I saw bees coming back to my garden after few years of absence" a correctly predicted *pleasantness* appraisal would lead to a correct the emotion prediction. In the sentence "I feel ... because I can't stand when people lie." the *appraisal to emotion* system

		Appraisal							
	Emotion (G/P)	A	Ce	E	P	R	Co	Ci	Text
Appr+Emo correct	Anger	1	1	0	0	0	0	0	when my neighbour started to throw rubbish in my garden for no reason.
	Disgust	0	1	0	0	0	0	0	to watch someone eat insects on television.
	Fear	1	0	1	0	0	0	1	when our kitten escaped in the late evening and we thought he was lost.
	Guilt	0	1	0	0	1	1	0	when I took something without paying.
	Joy	1	1	0	1	1	0	0	when I found a rare item I had wanted for a long time.
	Sadness	1	1	1	0	0	0	1	when my dog died. He was ill for a while. Still miss him.
	Shame	0	1	0	0	1	0	0	when I remember an embarrassing social faux pas from my teenage years.
Emo incorrect	Anger/Fear	1	0	1	0	0	0	0	when someone drove into my car causing damage and fear to myself – then drove off beforeexchanging insurance details.
	Disgust/Anger	1	1	0	0	0	0	0	when I saw a bird being mistreated when on holiday.
	Fear/Sadness	1	1	1	0	0	0	1	a huge spider just plopped on down on the sofa besides me, staring me out.
	Guilt/Disgust	0	1	0	0	0	0	0	when I watched a documentary that showed footage of farms of pigs and chickens and as a meat eater I felt awful guilt at how they are treated.
	Sadness/Anger	1	1	0	0	0	0	0	when I saw a group of homeless people and it was cold outside.
	Shame/Guilt	0	1	0	0	1	1	0	because I did something silly.
Ap+Emo incorrect	Anger/Shame	**0**	1	0	0	**1**	0	0	I feel ... because I can't stand when people lie.
	Disgust/Anger	**1**	1	0	0	0	**0**	0	when I saw a medical operation on a TV show.
	Fear/Guilt	1	0	**0**	0	**1**	0	**0**	when I was on a flight as I am ... of flying.
	Guilt/Shame	0	1	**1**	0	1	**1**	0	when I lost my sister's necklace that I had borrowed.
	Joy/Anger	1	1	0	**0**	0	0	**0**	when I saw bees coming back to my garden after few years of absence.
	Sadness/Guilt	**1**	1	0	0	**1**	0	**0**	when I watched some of the sad cases of children in need.
	Shame/Guilt	0	1	0	0	1	**1**	0	when I forgot a hairdressers appointment.
Appr incorrect	Anger	**1**	1	0	0	0	0	0	when Liverpool FC lost against Wolves..
	Disgust	**0**	1	0	0	**0**	0	**0**	when I stepped into a pile of dog excrement on the pavement.
	Fear	1	**0**	**1**	0	0	0	0	when a drunk man kept knocking at my door and shouting at me late at night.
	Guilt	1	1	**0**	0	**1**	**1**	0	when I couldn't visit my mum every day, whilst she was in hospital.
	Joy	1	1	0	1	**1**	0	0	when I saw my child perform in the school play.
	Sadness	1	**1**	1	0	0	0	1	when my grandad passed away.
	Shame	0	**1**	**0**	0	1	0	0	when I turned up drunk at a party and made a show of myself..

Table 12: Examples for the prediction of the pipeline setting (T→A, A→E). The first emotion mention is the gold (G), the second is the prediction (P). If the predicted emotions is equal to the gold emotions only one is given. In the third and fourth part appraisals not correctly predicted are shown in bold. A: Attention, Ce: Certainty, E: Effort, P: Pleasantness, R: Responsibility, Co: Control, Ci: Circumstance.

predicted *shame* because the appraisals *certainty* and *responsibility* were predicted. A correct prediction of the appraisal *attention* would most likely result in the correct prediction of the emotion *anger*, since *attention* and *certainty* are strongly associated with *anger*. This shows, that wrongly predicted appraisals can lead to wrong emotion predictions.

However, there are also instances in which the correct emotion is predicted although the wrong appraisals are predicted. Examples of such instances are shown in the fourth part of the table. Interestingly for all these examples, except for the examples labeled with the emotion *joy* and *shame*, is that a correct appraisal prediction would lead to a wrong emotion prediction. The sentence "I felt ... when I stepped into a pile of dog excrement on the pavement", for example, is annotated with the appraisals *attention*, *certainty responsibility* and *circumstance*. However, predicting these appraisals would most likely result in a prediction of the emotion *sadness*, since the combination of *attention*, *certainty* and *circumstance* is strongly associated with *sadness*. The appraisal prediction classifier however is only predicting the appraisal *certainty*, which then leads to the correct emotion prediction *disgust*. Also for the sentence "when a drunk man kept knocking at my door and shouting at me late at night" the appraisal prediction system did predict *certainty* as absent and *anticipated effort* as present. In the annotations for this instance *certainty* was labeled as present and *anticipated effort* as absent. If the *text to appraisal* classifier would have predicted *certainty* as present most likely the emotion label *anger* would be predicted for this sentence.

Particularly interesting are also the instances in which emotion predictions based on appraisals are correct and predictions based on text are wrong. A subset of such instances are shown in Table 13. Emotions in boldface are correct predictions. Note that those predictions from *appraisal to emotion* classifier are based on the annotated appraisal dimension and not on predicted ones.

In this comparison the appraisal based emotion classification seems to recognize the interpretation of a described event and the resulting emotion. The text based classifier on the other hand is predicting emotions only on specific words, which were learned in the training data to indicate specific emotion. In the sentence "when someone overtook my car on a blind bend and nearly caused an accident", for example, the words "car" and "accident" are likely to indicate the emotion *fear* in the *text to emotion* classifier while the appraisal based classifier is interpreting the sentence correctly with the resulting emotion *anger*. Further, in the sentence "when my mom caught me lying" the text based classifier is likely to associate the word "lying" with *anger*, while the *appraisal to emotion* classifier is correctly interpreting the described event and predicting the correct emotion *shame*.

However, the predictions in some instances in which the *text to emotion* classifier fails and the *appraisal to emotion* classifier predicts the correct emotion are reasonable and could arguably be seen as the correct emotion. For the sentence "when I took something without paying" labeled with *guilt* the prediction of the emotion *shame* by the *text to emotion* classifier also makes sense.

The lower part of Table 13 shows example instances in which the *appraisal to emotion*

Pred. Emotion		Appraisal Annotation							
A→E	T→E	A	Ce	E	P	R	Co	Ci	Text
Anger	Disgust	1	1	0	0	0	0	0	when I saw someone mistreating an animal.
Anger	Fear	1	0	0	0	0	0	0	when someone overtook my car on a blind bend and nearly caused an accident.
Disgust	Fear	0	1	1	0	0	0	0	when I was on a ferry in a storm and lots of people were vomiting.
Disgust	Shame	0	1	0	0	0	0	1	because the milk I put in my coffee had lumps in it.
Fear	Shame	1	0	1	0	0	0	1	because I had to have a general anaesthetic for an operation.
Fear	Sadness	1	0	1	0	0	0	0	when my 2 year old broke her leg, and we felt helpless to assist her.
Guilt	Joy	1	1	1	0	1	1	0	for denying to offer my kids what they demanded of me.
Guilt	Anger	0	1	0	0	1	1	0	when I had not done a job for a friend that I had promised to do.
Joy	Shame	1	1	0	1	1	1	0	when I managed to complete a cryptic crossword.
Joy	Disgust	1	1	0	1	1	0	1	when I found a twenty pound note on the ground outside.
Sadness	Fear	1	1	0	0	0	0	1	when it was raining this morning as I been planning to go on a camping trip.
Sadness	Joy	1	1	0	0	0	0	1	when I see the Christmas decorations come down, and know they won't be up again for another year.
Shame	Joy	0	1	1	0	1	0	0	when I failed my ninth year at high school.
Shame	Anger	0	1	0	0	1	0	0	when my mom caught me lying.
Fear	**Anger**	1	0	1	0	0	0	0	because my ex-husband bullied me and my children, and threatened to knock down our house door.
Guilt	**Anger**	1	1	1	0	1	0	0	when as a young hairdresser I was closing up the shop. My boss came in and was annoyed and angry that I had let my girlfriend into the shop while I was closing up. We had an altercation over the incident.
Anger	**Disgust**	1	1	1	0	0	0	0	when I had to clean up after people.
Shame	**Disgust**	0	1	1	0	1	1	0	when I unblocked a drain filled with raw sewage.
Anger	**Fear**	1	0	0	0	0	0	0	when awaiting the email results of an important and very expensive accountancy exam I had taken a few months previously.
Sadness	**Fear**	1	0	0	0	0	0	1	because there was a very loud and unexpected bang behind me.
Shame	**Guilt**	0	1	1	0	1	1	0	because I didn't really want my eldest son to come over for Christmas when I should have left old feelings aside.
Shame	**Guilt**	0	1	0	0	1	0	1	when I couldn't visit a relative because I was ill.
Anger	**Joy**	1	1	0	0	0	0	0	when Will Young won Pop Idol because he was nicer than Gareth Gates.
Sadness	**Joy**	1	1	0	0	0	0	1	when I knew that I was going back to Florida a year earlier than I thought I would.
Fear	**Sadness**	0	0	1	0	0	0	1	I feel ... because I am depressed.
Guilt	**Sadness**	1	1	0	0	1	1	0	when I watched a sad movie.
Fear	**Shame**	1	0	1	0	1	0	0	when I was arrested for stealing.
Guilt	**Shame**	0	1	0	0	1	1	0	when I found money and did not hand it in.

Table 13: Top part: Examples in which the appraisal model (on gold appraisal annotation) predicts the correct emotion and the baseline system does not. Bottom part: Examples in which the appraisal model predicts the not correct emotion and the baseline system does predict the correct emotion. Correct predictions are shown in bold. A: Attention, Ce: Certainty, E: Effort, P: Pleasantness, R: Responsibility, Co: Control, Ci: Circumstance.

classifier fails to predict the correct emotion, while the *text to emotion* classifier predicts the correct emotion. These examples show, that for some instances in which the *appraisal to emotion* classifier fails, the predicted emotion could also arguably be seen as a correct emotion. The sentence “because my ex-husband bullied me and my children, and threatened to knock down our house door” is such an example. It is labeled with *anger* but could also be seen as a situation in which the writer of such a sentence is experiencing the emotion *fear*, like predicted by the *appraisal to emotion* model. In addition, the annotations for this sentence with *attention* and *anticipated effort* make sence, since such a situation requires *attention* and *anticipated effort*. A similar example is the sentence “when I had to clean up after people”, which is labeled with the emotion *disgust*. The *appraisal to emotion* model however is predicting *anger*, which is also a reasonable emotional response to the situation.

The analyisis also shows that *guilt* and *shame* are confused due to the presence or absence of the *control* appraisal. Both emotions are associated with *certainty* and *responsibility*. The *appraisal to emotion* model however, is favouring *guilt* instead of *shame* if the *control* appraisal is present. This makes sence, since if one feels guilty in a situation the person feels like being in control of the situation. In the sentence “when I found money and did not hand it in” the classifier is wrongly predicting *guilt* instead of *shame*, most likely due to the presence of the appraisal *control*. The sentence “when I couldn’t visit a relative because I was ill” on the other hand is wrongly predicted as *shame* instead of *guilt* due to the absence of the appraisal annotation *control*. However, not annotating the appraisal *control* as present makes sense because the writer of the sentence was most likely not in control of “getting ill”.

In the two sentences in which the *appraisal to emotion* model failed to predict the correct emotion *joy*, the absence of an annotation in the *pleasantness* led to the wrong prediction. However, the predictions make also sense. The sentence “when I knew that I was going back to Florida a year earlier than I thought I would”, for instance, could also be seen as a *sad* experience. The lower part of the table also shows instances in which the predictions by the *appraisal to emotion* do not make sense. The sentence “when I watched a sad movie”, which is labeled with *sadness*, was predicted as *guilt*. This sentence is annotated with *attention*, *certainty*, *responsibility* and *control*. Especially the combination of the appraisals *responsibility* and *control* is associated with the emotion *guilt*. Most likely the annotaters annotated *responsibility* and *control* because they think if someone is watching a movie the person is responsible for doing that and also in control of the situation. From such a perspective these annotations make sense. However, a typical *sad* emotional experience is, according to the findings by Smith and Ellsworth (1985), not associated with *responsibility* and *control*.

7 Conclusion and Future Work

In this work, a new approach for the automated classification of emotions in text was presented. At first a corpus was annotated using a cognitive appraisal theory. Afterwards, machine learning models, namely a maximum entropy model, a convolutional neural network and a simple two layer artificial neural network, were used to investigate the hypothesis that giving an emotion prediction model access to cognitive appraisal dimensions is beneficial for the performance of the model.

Under the assumption of perfect appraisal predictions the results of this work show that emotion classification based on appraisals performs better than text-based classification. In addition, the oracle based experiment showed that the a text-based classifier and the appraisal based pipeline setup behave differently. However, this work was not able to show a clear improvement in the automated prediction of emotions in a real-world pipeline, in which appraisals are predicted as basis nor in a multi-task setting. Although the proposed appraisal prediction model achieves a reasonable performance, this shows that the real-world pipeline setting suffers from error propagation.

Nevertheless, this study in emotion classification using appraisal theories raises some interesting research questions: Are there other neural models or multi-task architectures, which benefit more from the information appraisal dimensions can contribute to an automated emotion prediction task or would more annotated data be sufficient in order to improve the performance of the proposed models? Therefore, experiments with larger datasets and other neural classification architectures, like a LSTM or BiLSTM, remain interesting.

Finally, the question remains if a different annotation setup would have changed the results of the appraisal based emotion classification provided in this work. Such a different setup could, for example, be giving the annotators access to the emotion label in the annotation process or a completely automated annotation process without human annotators on basis of the patterns derived by Smith and Ellsworth (1985) or other appraisal theories. Further, a limitation, which was introduced to the classification models by the annotation setup of this work were the binary-valued appraisal annotations. Therefore, it should be investigated if a continuous-valued annotation method would be useful for emotion predictions based on appraisals. Such an approach could not only improve the performance on a classification model predicting emotions from appraisals but also could lead to more fine-grained appraisal predictions compared to the binary setting proposed in this study.

References

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1034. URL `https://www.aclweb.org/anthology/N19-1034`.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/H05-1073`.

Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74628-7.

Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems*, NLDB'11, page 27–39, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642223266.

Jorge Balazs, Edison Marrese-Taylor, and Yutaka Matsuo. IIIDYT at IEST 2018: Implicit emotion classification with deep contextualized word representations. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–56, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6208. URL `https://www.aclweb.org/anthology/W18-6208`.

Adam Berger. The improved iterative scaling algorithm: A gentle introduction, 1997.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1): 39–71, 1996. URL `https://www.aclweb.org/anthology/J96-1002`.

Laura Ana Maria Bostan and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/C18-1179`.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*, 01 2009.

Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April 2017a. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-2092`.

Sven Buechel and Udo Hahn. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain, April 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-0801. URL `https://www.aclweb.org/anthology/W17-0801`.

Rafael A. Calvo, David N. Milne, M Sazzad Hussain, and Helen Christensen. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685, 2017. doi: 10.1017/S1351324916000383.

Angel Cambero. A comparative study of twitter sentiment analysis methods for live applications. Master's thesis, Rochester Institute of Technology, 2016.

Walter Bradford Cannon. The james-lange theory of emotions: a critical examination and an alternative theory. *The American Journal of Psychology*, 39:106–124, 1927. URL `https://doi.org/10.2307/1415404`.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, July 1997. ISSN 0885-6125. doi: 10.1023/A:1007379606734. URL `https://doi.org/10.1023/A:1007379606734`.

Alexandra Chronopoulou, Aikaterini Margatina, Christos Baziotis, and Alexandros Potamianos. NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–64, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6209. URL `https://www.aclweb.org/anthology/W18-6209`.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada, June 1989. Association for Computational Linguistics. doi: 10.3115/981623.981633. URL `https://www.aclweb.org/anthology/P89-1010`.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL `https://doi.org/10.1177/001316446002000104`.

Michael Crawford, Taghi Khoshgoftaar, Joseph Prusa, Aaron Richter, and Hamzah Al-Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2:23, 10 2015. doi: 10.1186/s40537-015-0029-9.

Charles Darwin. *The Expression of the Emotions in Man and Animals*. 1872. The original was published 1898 by Appleton, New York. Reprinted 1965 by the University of Chicago Press, Chicago and London.

R. J. Dolan. Emotion, cognition, and behavior. *Science*, 298(5596):1191–1194, 2002. ISSN 0036-8075. doi: 10.1126/science.1076358. URL `https://science.sciencemag.org/content/298/5596/1191`.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Paul Ekman. Basic emotions. In Tim Dalgleish and Mick J. Power, editors, *Handbook of Cognition and Emotion*. John Wiley & Sons, Sussex, UK, 1999.

Phoebe Ellsworth and Klaus Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, pages 572–595, 01 2009.

Nico H. Frijda, Peter W Kuipers, and Elisabeth ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2):212–228, 08 1989.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Toshie Imada and Phoebe Ellsworth. Proud americans and lucky japanese: Cultural differences in appraisal and corresponding emotion. *Emotion (Washington, D.C.)*, 11:329–45, 04 2011. doi: 10.1037/a0022855.

William James. II.—WHAT IS AN EMOTION ? *Mind*, os-IX(34):188–205, 04 1884. ISSN 0026-4423. doi: 10.1093/mind/os-IX.34.188. URL `https://doi.org/10.1093/mind/os-IX.34.188`.

Evgeny Kim and Roman Klinger. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift fuer Digitale Geisteswissenschaften*, 4, 2019. doi: http://dx.doi.org/10.17175/2019_008.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL `https://www.aclweb.org/anthology/D14-1181`.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. URL `http://arxiv.org/abs/1412.6980`.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium, October 2018. Association for Computational

Linguistics. doi: 10.18653/v1/W18-6206. URL https://www.aclweb.org/anthology/W18-6206.

S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*, page 3–24, NLD, 2007. IOS Press. ISBN 9781586037802.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown, Laura Id, and Barnes. Text classification algorithms: A survey. *Information (Switzerland)*, 10, 04 2019. doi: 10.3390/info10040150.

Richard S. Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *The American psychologist*, 46(8):819–834, 1991.

Julie Beth Lovins. Development of a stemming algorithm. *Mech. Translat. Comp. Linguistics*, 11:22–31, 1968.

Mary McHugh. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82, 10 2012. doi: 10.11613/BM.2012.031.

A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 11:261–292, 1996. URL http://dx.doi.org/10.1007/BF02686918.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

Marcin Mirończuk and Jaroslaw Protasiewicz. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106, 03 2018. doi: 10.1016/j.eswa.2018.03.058.

Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S12-1033.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia, May 2016a. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1623.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016b. Association for Computational Linguistics. doi: 10.18653/v1/S16-1003. URL `https://www.aclweb.org/anthology/S16-1003`.

David J. Montana and Lawrence Davis. Training feedforward neural networks using genetic algorithms. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'89, page 762–767, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

Agnes Moors, Phoebe Ellsworth, Klaus Scherer, and Nico Frijda. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5:119–124, 03 2013. doi: 10.1177/1754073912468165.

Elisabeth Métais, Andrés Montoyo, and Rafael Muñoz. Natural language processing and information systems - 16th international conference on applications of natural language to information systems. 06 2011.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, Haifa, Israel, 2010. Omnipress. URL `http://www.icml2010.org/papers/432.pdf`.

Kamal Nigam, John D. Lafferty, and Andrew McCallum. Using maximum entropy for text classification. 1999.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883062. URL `https://doi.org/10.1145/2872427.2883062`.

Keith Oatley and P. Johnson-laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1:29–50, 03 1987. doi: 10.1080/02699938708408362.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL `http://dx.doi.org/10.1561/1500000011`.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://www.aclweb.org/anthology/D14-1162`.

Rosalind W. Picard. Affective computing: Challenges. *Int. J. Hum.-Comput. Stud.*, 59 (1–2):55–64, July 2003. ISSN 1071-5819. doi: 10.1016/S1071-5819(03)00052-1. URL `https://doi.org/10.1016/S1071-5819(03)00052-1`.

Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.

Robert Plutchik. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. doi: 10.1017/S0954579405050340.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0404. URL `https://www.aclweb.org/anthology/W16-0404`.

Philipp Probst. Hyperparameters, tuning and meta-learning for random forest and other machine learning algorithms. July 2019. URL `http://nbn-resolving.de/urn:nbn:de:bvb:19-245579`.

Marc'Aurelio Ranzato, Fu Jie Huang, Y. Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07*, 2007. ISBN 1424411807. doi: 10.1109/CVPR.2007.383157.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. EmpaTweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3806–3813, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/201_Paper.pdf`.

Ira Roseman. Cognitive determinants of emotion: A structural theory. *Rev. Pers. Soc. Psychol.*, 5, 01 1984.

Alon Rozental, Daniel Fleischer, and Zohar Kelrich. Amobee at IEST 2018: Transfer learning from language models. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–49, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6207. URL `https://www.aclweb.org/anthology/W18-6207`.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098, 2017.

James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 09 1977. doi: 10.1016/0092-6566(77)90037-X.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf.

J. Salminen, Vignesh Yoganathan, J. Corporan, B.J. Jansen, and S.-G. Jung. Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type, 2019. URL http://hdl.handle.net/10454/17058.

Klaus R. Scherer. Emotion as a process: Function, origin and regulation. *Social Science Information*, 48:555–70, 1982.

Klaus R. Scherer and Harald G. Wallbott. The ISEAR questionnaire and codebook. Geneva Emotion Research Group, 1997. URL https://www.unige.ch/cisa/research/materials-and-online-research/research-material/. https://www.unige.ch/cisa/research/materials-and-online-research/research-material/.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5203. URL https://www.aclweb.org/anthology/W17-5203.

M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997. ISSN 1053-587X. doi: 10.1109/78.650093. URL https://doi.org/10.1109/78.650093.

Kuldeep Shiruru. An introduction to artificial neural network. *International Journal of Advance Research and Innovative Ideas in Education*, 1:27–30, 09 2016.

Craig. A. Smith and Phoebe. C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838, 1985.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, page 70–74, USA, 2007a. Association for Computational Linguistics.

Carlo Strapparava and Rada 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 2007b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S07-1013`.

Shabnam Tafreshi and Mona Diab. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1246`.

Jessica L. Tracy and Richard W. Robins. Appraisal antecedents of shame and guilt: Support for a theoretical model. *Personality and social psychology bulletin*, 32(10): 1339–1351, 2006.

Enrica Troiano, Sebastian Padó, and Roman Klinger. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1391. URL `https://www.aclweb.org/anthology/P19-1391`.

Yan Xu, Gareth Jones, Jintao Li, Bin Wang, and Chunming Sun. A study on mutual information-based feature selection for text categorization. *Journal of Computational Information Systems*, 3, 03 2007.

Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, page 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1558604863.

Qiang Yang Yu Zhang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018. doi: https://doi.org/10.1093/nsr/nwx105. URL `http://engine.scichina.com/publisher/ScienceChinaPress/journal/NationalScienceReview/5/1/10.1093/nsr/nwx105`.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4595–4601. AAAI Press, 2018a. ISBN 9780999241127.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4595–4601. AAAI Press, 2018b. ISBN 9780999241127.

Appendix

Annotation Guideline for Appraisal Dimensions [4]

In this task, annotators provide judgements about emotional events using appraisal dimensions.
You will read the description of a real-life situation.
The description reports an experience that occurred in the life of its writer.
Next, you will be asked to judge some properties of such emotion-inducing events (Attention, Certainty, Pleasantness, Responsibility, Anticipated Effort, Situational Control).

The only possible values for each of these dimensions are: 0 (no) and 1 (yes).

Notes:

- Judgments must refer to the time in which the event occured
- A description may contain multiple events: in that case, judgments must be relative only to the event that caused the emotion.

Most probably, at the time when the event happened, the writer ...

* wanted to devote further attention to the event.	**ATTENTION**
* was certain about what was happening.	**CERTAINTY**
* had to expend mental or physical effort to deal with the situation.	**EFFORT**
* found that the event was pleasant.	**PLEASANTNESS**
* was responsible for bringing about the situation.	**RESPONSIBILITY**
* found that the he/she was in control of the situation.	**CONTROL**
* found that the event could not have been changed or influenced by anyone. (Fate)	**CIRCUMSTANCE**

[4] Motivated by Smith and Ellsworth (1985)