HEEM, a Complex Model for Mining Emotions in Historical Text

Janneke M. van der Zwaan*, Inger Leemans†, Erika Kuijpers†, and Isa Maks‡
*Netherlands eScience Center, Amsterdam, The Netherlands
Email: j.vanderzwaan@esciencecenter.nl
† Faculty of Humanities, Department of Art and Culture, History, and Antiquity
VU University, Amsterdam, The Netherlands
Email: {i.b.leemans, erika.kuijpers}@vu.nl
‡ Faculty of Humanities, Computational Lexicology and Terminology Lab
VU University, Amsterdam, The Netherlands
Email: e.maks@vu.nl

***Abstract*—Recently, emotions and their history have become a focus point for research in different academic fields. Traditional sentiment analysis approaches generally try to fit relatively simple emotion models (e.g., positive/negative emotion) to contemporary data. However, this is not sufficient for Digital Humanities scholars who are interested in research questions about changes in emotional expressions over time. Answering these questions requires more complex, historically accurate emotion models applied to historical data. The Historic Embodied Emotion Model (HEEM) was developed to study the relationship between body parts and emotional expressions in 17th and 18th century texts. This paper presents the HEEM emotion model and associated dataset from a technical perspective, and examines the performance of a multi-label text classification approach for predicting HEEM labels and labels from two simpler models (i.e., HEEM Emotion Clusters and the Positive/Negative model). The results show that labels in the complex model can be predicted with micro-averaged $F_1 = 0.45$, and macro-averaged $F_1 = 0.24$. Labels with fewer samples (< 40) are not predicted. Overall performance on the simpler emotion models is significantly better, but for individual labels the effect is mixed. We demonstrate that a multi-label text classification approach to learning complex emotion models on historical data is feasible.**

I. Introduction

As different academic fields have taken an affective turn over the last years, emotions and their history have become a focus point for research [1]. At the same time, sentiment analysis and opinion mining have become important research areas, both within and outside academia. So far, however, most techniques are aimed at fitting relatively simple emotion models (positive/negative emotion, or limited sets of at most of six or seven 'basic' emotions (e.g., 'anger', 'disgust/contempt', 'fear', 'interest', 'joy', 'love', 'sadness', and 'surprise')). In addition, these simple models are almost exclusively applied to contemporary, web-based texts.

This poses two problems for Digital Humanities researchers interested in investigating how emotional expressions evolve over time: 1) simple emotion models fail to deal with the complexity and the historical contingency of emotions and their expression (instead, researchers seem to presume that emotional expressions are stable across time); and 2) historical (literary) text differs significantly from contemporary, web-based text, e.g., with respect to spelling.

To address the first problem, we developed the Historic Embodied Emotion Model (HEEM). As we are interested in the relationship between emotional expressions and body parts, HEEM is focused on the bodily enactment of emotions. In addition to 38 historically accurate emotion labels, the model contains four concept types that refer to the embodiment of emotions (i.e., Emotion, Bodily response, Body part, and Emotional action). The 38 emotion labels allow for a rich and fine-grained classification of emotional expressions. HEEM was applied to a corpus of 29 17th and 18th century Dutch theater texts, yielding a dataset consisting of $27,993$ sentences manually annotated with one or more HEEM labels.

This paper presents HEEM and the HEEM dataset from a technical perspective, and examines the performance of standard text classification algorithms on HEEM. Given that related work uses much simpler emotion models, it is not clear to what extent a text classification approach is feasible for predicting labels from a complex model on historical text. Because HEEM prescribes that emotional expressions should be annotated with at least a concept type and an emotion label, we opted for a multi-label text classification approach. A multi-label classification task consist of assigning a (possibly empty) subset of predetermined labels to instances (e.g., sentences). We follow the same approach as Buitinck et al. [2] and train both separate classifiers for each label (Binary Relevance) and an ensemble classifier that takes into account correlations between labels. We also address the second problem of non-standardized spelling in historical text by experimenting with spelling normalization. Finally, we examine the performance of multi-label text classification on two simplifications of HEEM.

This paper is organized as follows. Section II describes related work on sentiment and emotion mining. Section III presents HEEM and the HEEM dataset. We describe our method in section IV. Empirical results follow in section V. Section VI wraps up the paper with a discussion and conclusion.

II. Related Work

HEEM and the HEEM dataset have been developed as part of a methodology to trace changes over time in the way people use their bodies to express emotions. The task of extracting emotional expressions and related concepts, such as body parts from written text amounts to a complex form of sentiment analysis. Much of the previous work on sentiment analysis and opinion mining involves valence classification, i.e., assigning labels 'positive' or 'negative' to written text. A broad introduction to existing sentiment analysis and opinion mining techniques is provided by Pang and Lee [3].

Valence can be seen as a very simple emotion model. Text classification tasks using these kinds of simple models vary in the classes they seek to predict. Several studies on classifying documents as either 'positive' or 'negative' have been carried out; for example on movie reviews [4], [5], [6], customer feedback data [7], and product reviews [8]. Aman and Szpakowicz [9] use a slightly different approach; their work involves classifying sentences from blog posts as either 'emotional' or 'non-emotional'. Alm et al. [10] use three mutually exclusive classes to classify sentences in fairy tales in categories 'positive emotion', 'negative emotion', and 'neutral' (no emotion). Wilson et al. [11] use a two-step approach to classifying sentences from news articles as 'positive', 'negative', 'both', or 'neutral'. In the first step, sentences are classified as 'neutral' or 'polar'. In the second step, polar sentences receive their final classification.

There is also work that uses more complex emotion models, i.e. models with more emotion labels. For example, Yang et al. [12] distinguish four emotion categories, i.e., 'happy', 'joy', 'sad', and 'angry', and train classifiers to predict emotions expressed in blog posts. Danisman and Alpkocak [13] perform experiments with classifiers trained on sentences and snippets containg descriptions of emotional events and test performance on a separate dataset consisting of news headlines. The emotion model consists of categories 'anger', 'disgust', 'fear', 'sad', and 'joy'. The best classifier obtains an overall F_1 score of 0.32. Strapparava and Mihalcea [14] classify newspaper headlines and blog posts into six 'basic' emotions, namely 'anger', 'disgust', 'fear', 'joy', 'sadness', and 'surprise' in blog posts. Buitinck et al. [2] observe that emotion categories are not necessarily mutually exclusive, and propose a multi-label text classification approach to assign (a possibly empty) subset of emotion labels to sentences in movie reviews. Their emotion model consists of seven emotions; 'anger', 'disgust/contempt', 'fear', 'interest', 'joy', 'love', 'sadness', and 'surprise'. This method performs well; the authors report a micro-averaged F_1 score of 0.46.

Our work differs from the related work in two ways. First, compared to other work, we have a very complex emotion model. HEEM consists of 42 non-mutually exclusive labels divided into two layers (concept types and emotion labels). Second, previous work deals mostly with contemporary, web-based data, whereas our corpus consists of historical texts from a relatively long time period (2 centuries).

III. Description of the Data

To investigate the relationship between emotional expressions and body parts, we chose to focus on Dutch language theater texts from the 17th and 18th century (the early modern period). Theater texts are excellent sources for emotion research since staging the passions is one of the main goals of this genre. Characters tend to be very explicit in indicating their feelings, while stage directions and speaking turns provide information about the bodily enactment of emotions.

In this section, we present the Historic Emotion Model (HEEM) and the associated dataset. HEEM differs from 'modern' emotion models in two ways; 1) the model is focused on the bodily enactment of emotions, and 2) allows for a rich and fine-grained classification of emotions in historically accurate classes. Five annotators applied HEEM to a corpus of 29 texts. The resulting dataset consists of 27,993 sentences of which 13.55% is emotional. The corpus does not suffer from 'short-termism' in the sense that it is historical and that it spans a large time period (1600–1800). We also describe a method for spelling normalization, as spelling in the 17th and 18th century was not yet standardized, and this might pose a challenge to conventional text classification algorithms. Finally, we report the results of an inter-annotator agreement study.

A. The Historic Embodied Emotion Model

Because we are interested in historical interpretations of emotional expressions, and want to be able to track changes in emotional expressions over time, we chose to create a new emotion model, instead of using an existing one. The most important differences between our model and existing models are 1) a focus on the role of the human body in experiencing and expressing emotions (embodiment), and 2) the number of labels available to classify emotions. HEEM contains 38 emotion labels, whereas most modern emotion models consist of fewer labels; for example, Strapparava and Mihalcea use a model that contains six emotions [14], and Buitinck et al. distinguish seven emotions [2].

The HEEM annotations consist of two levels: concept types and emotion labels. In the model, an expression of emotion is characterized by one or more concept types and one or two emotion labels. In the context of this work, emotion is defined as a strong feeling deriving from one's circumstances or relationships with others involving cognitive appraisal, bodily symptoms, (a readiness for) action, motorical expression (for instance in face or voice) and subjective awareness (cf. Scherer's definition of emotions [15]).

HEEM concept types are listed in table I. The concept type *Emotion* is used to mark expressions of emotions in the text. *Body part*s are tagged only, when they are used to express emotions. A *Bodily process* refers to reactions of the body to an emotion, while an *Emotional action* is an action triggered by an emotion. The main difference between these concept types is the extent to which the person expressing the emotion is in control of her actions: a *Bodily process* cannot be controlled (e.g., to tremble of fear), while an *Emotional*

Table I
HEEM CONCEPT TYPES

Concept type	Explanation	Example
Emotion	See definition in the text	to loathe, to be in love, to be sad
Body part	A body part involved in an emotional action or expression (internal or external)	eyes, hand, blood, mind
Bodily process	Uncontrollable reactions of the body to emotions	to cry, to sigh, to blush, to tremble
Emotional action	Controllable human action triggered by an emotion	to embrace, to scorn, to scold

action is a purposeful action (e.g., to embrace someone out of happiness)[1].

In addition to one or more concept types, expressions of emotions are further specified by one or two emotion labels. The model contains 38 emotion labels that are relevant for the interpretation of emotions in 17th and 18th century texts (see table II). The list includes the label *Other* to specify cases where one of the other labels does not apply, but an emotion is expressed nevertheless. The list was composed by domain experts based on our definition of emotion and a preliminary exploration of a subset of corpus texts.

The fine-grained emotion classification can be mapped on simpler models. We propose two: one based on historically accurate emotion clusters in HEEM and a binary one of the positive and negative emotions. Table III shows the division of HEEM labels in 12 HEEM historically accurate emotion clusters.

Table IV lists the HEEM labels divided in positive and negative emotions. Again, the division was chosen to be historically accurate. These historical interpretations do not always coincide with contemporary ones. For example, *Aquiescence* is a positive emotion in the Positive/Negative model, whereas nowadays it is often seen as a negative emotion. In addition, some emotions can be either positive or negative depending on the context (e.g., *Pride*), these emotions have been assigned to either the positive or negative category, based on whether the emotion is in essence positive or negative for the person experiencing it (e.g., *Pride* was chosen to be a positive emotion).

B. Dataset

The dataset consists of 29 Dutch theater texts (1600–1800) that were manually annotated with the concepts and emotion labels from HEEM (see section III-A). The 29 texts were selected from a larger corpus of ~220 theater texts that are publicly available in Nederlab[2]. Texts were selected to ensure coverage of different genres (tragedy, comedy, and farce), and different time periods (Renaissance, Classicism, and Enlightenment). Table V shows the division of the texts

[1]The distinction between controllable and uncontrollable actions is problematic. Our inter-annotator agreement study shows that our annotators do not always agree on what is a *Bodily process* and what is an *Emotional action* (see Section III-C). However, we decided to keep the distinction between *Bodily process* and *Emotional action*.

[2]http://www.nederlab.nl/

Table II
ABSOLUTE LABEL FREQUENCIES

Label	All	Test sets
Emotion	2978	277–325
Bodily process	905	77–102
Body part	703	61–80
Emotional action	458	36–54
Love	733	67–80
Sadness	710	63–80
Fear	621	55–74
Joy	460	36–66
Anger	386	33–43
Despair	229	16–30
Vindictiveness	228	13–30
Desire	203	13–25
Hatred	173	11–25
Hope	121	6–18
Compassion	112	6–16
Remorse	84	4–14
Worry	83	4–20
Happiness	80	5–13
Other	79	4–14
Shame	70	3–10
Heavy-heartedness	62	2–9
Honor	60	2–9
Disgust	59	2–10
Loyalty	56	1–12
Wonder	49	3–9
Spitefulness	45	0–7
Moved	44	1–9
Annoyance	40	1–7
Envy	37	2–5
Acquiescence	30	0–6
Benevolence	30	1–6
Pride	29	1–6
Suspicion	26	0–6
Offended	25	0–5
Dedication	23	0–5
Unhappiness	20	0–4
Disappointment	18	0–4
Greed	16	0–4
Trust	15	0–3
Awe	14	0–3
Relief	12	0–2
Loss	9	0–2

Table III
HEEM EMOTION CLUSTERS

Label	Freq.	HEEM labels
Sadness	834	Sadness, Heavy-heartedness, Disappointment, Loss, Remorse
Love	785	Love, Loyalty, Dedication
Anger	778	Anger, Hatred, Annoyance, Spitefulness, Offended, Vindictiveness
Fear	718	Fear, Worry, Suspicion, Awe
Joy	697	Joy, Happiness, Hope, Relief, Wonder
Desire	249	Desire, Greed, Envy
Despair	249	Despair, Unhappiness
Disgust	129	Disgust, Shame
PosSentiments	117	Moved, Acquiescence, Trust, Benevolence
Compassion	112	Compassion
PrideHonor	88	Pride, Honor
Other	79	Other

Table IV
THE POSITIVE/NEGATIVE MODEL

Label	Freq.	HEEM labels
Positive	2380	Love, Joy, Desire, Hope, Compassion, Happiness, Honor, Loyalty, Wonder, Moved, Acquiescence, Benevolence, Pride, Dedication, Trust, Awe, Relief
Negative	1805	Sadness, Fear, Anger, Despair, Vindictiveness, Hatred, Remorse, Worry, Shame, Heavy-heartedness, Disgust, Spitefulness, Annoyance, Envy, Suspicion, Offended, Unhappiness, Disappointment, Greed, Loss
Other	79	Other

Table V
TEXTS IN THE HEEM CORPUS

	Renaissance (1600–1669)	Classicism (1670–1749)	Enlightenment (1750–1800)
Tragedy	5	2	6
Comedy	1	8	4
Farce	1	2	0
Total	7	12	10

over the genres and time periods.

The texts were annotated in KAFAnnotator[3] by a group of five experienced readers of 17th and 18th century texts. All texts were annotated by a single annotator. Inter-annotator agreement was assessed on two separate texts. The results of this study can be found in Section III-C. After annotation, the texts were split into sentences, yielding $27,993$ sentences of which 13.55% was annotated with at least one HEEM label (concept type and/or emotion label)[4]. Table II shows the absolute label frequencies of the different labels. The mean label cardinality per sentence [16] is 0.362 and the mean label density is 0.009. For the experiments, classifier performance is assessed by using 10-fold cross validation (see section IV). The number of positive samples per label in the test sets is also reported in table II.

C. Inter-Annotator Agreement

To check the consistency and reliability of the annotations, an inter-annotator study was carried out. For this study, five annotators independently annotated two additional texts consisting of approximately 1100 sentences in total. Agreement was calculated using the metric *agr* as proposed by [17] and [18]. This metric calculates pair-wise agreement by first measuring precision of annotator A's annotations on B's annotations, and then measuring precision of annotator B's annotations on annotator A's annotations. Agreement between 2 annotators is the mean of the 2 scores and overall agreement is the mean of all pair-wise scores. Moreover, we considered annotations to match if there was an overlap of at least one word.

Table VI reports the overall agreement scores for the 4 concept types (*Emotion*, *Body part*, *Bodily process* and *Emotional action*. The results show that concept types *Emotion* and *Body part* are reliably identifiable, with overall scores of 0.73. *Bodily process*s and *Emotional action*s, however, have low agreement scores which do not allow for definite conclusions.

Table VI
INTER-ANNOTATOR AGREEMENT ON HEEM CONCEPTS

Concept	Agreement
Emotion	0.73
Body part	0.73
Bodily process	0.47
Emotional action	0.30
Emotion labels (38 classes)	0.69

The remaining agreement scores refer to the emotions labels identified in the text. The 38 HEEM classes can be identified and classified ($agr = 0.69$). However, this number might be overestimate the actual agreement as the diversity of emotions in the test set is rather low; 75% of the labels assigned come from 7 classes (i.e., *Fear*, *Sadness*, *Dedication*, *Joy*, *Hope*, *Love*, *Hatred*, and *Despair*).

IV. METHOD

The problem of predicting concept types and emotion labels from HEEM is treated as a multi-label text classification task. We tested two algorithms for multi-label classification: Binary Relevance (BR) and Random *k*-Labelsets (RAkEL) [16]. BR solves the multi-label problem by training a binary classifier for each label in the label set, while RAkEL takes into account correlation between labels by reducing it to multi-class learning. Linear Support Vector Machines (SVM) using standard bag-of-words features with tf-idf weighting and stop word removal were trained for both classification algorithms. The classifiers were implemented in scikit-learn [19], [20].

A. Classification Algorithms

Multi-label classification concerns the task of assigning to samples a (possibly empty) subset of labels $Y \in L$. Generally, two approaches to multi-label classification exists: 1) problem transformation methods, and 2) algorithm adaptation methods [21]. Problem transformation methods reduce multi-label classification into one or more single-label classification problems, while algorithm adaptation methods that extend existing learning algorithms to handle multi-label data directly. To predict labels from HEEM, we tested two problem transformation methods.

1) Binary Relevance: This method learns $|L|$ binary classifiers $H_l : X \rightarrow \{l, \neg l\}$, one for each label in L. Each classifier learns to distinguish one label from the rest. The classification of a new sample x consists of the union of the labels that are predicted by the $|L|$ classifiers:

$$H(x) = \bigcup_{l \in L} \{l\} : H_l(x) = l \quad (1)$$

A disadvantage of the Binary Relevance method is that correlations between labels are not taken into account.

[3]http://www.cltl.nl/results/software/kafnafannotator/

[4]https://github.com/NLeSC/embem-ml-dataset

2) RAkEL [16]: This method is a variant of the label power set (LP) method. Label power set is a problem transformation approach to multi-label classification that learns a single-label classifier $H : X \rightarrow P(L)$, where $P(L)$ is the power set of L [21]. This method takes into account correlations between labels, but quickly becomes computationally very expensive to train, and suffers from having to train classifiers for label sets with only a few positive samples [22]. RAkEL mitigates these problems by training an ensemble H of m classifiers, each trained on a subset of labels R of a fixed size k, randomly chosen from the set L^k without replacement, where L^k is the set of all distinct k-labelsets of L [16], [22]. In our experiments, we take $k = 3$ and $m = 2|L|$ as advised by Tsoumakas et al. [22].

Each classifier h_i outputs binary predictions $h_i(x, l_j)$ for $l_j \in L$ in the corresponding k-labelset R_i. New samples x are classified by calculating the mean of the predictions for all labels $l_j \in L$ by each classifier $h_i \in H$. A label is assigned to x if the mean prediction > 0.5.

B. Evaluation Measures

Classifier performance is measured by F_1 score. F_1 is the harmonic mean of precision and recall. Given the number of true positives (tp), true negative (tn), false positives (fp), and false negatives (fn), F_1 is defined as:

$$F_1 = \frac{2 \cdot tp}{2 \cdot (tp + fp + fn)} \tag{2}$$

In addition to F_1 scores for the individual labels, overall classifier performance is reported by micro-averaged and macro-averaged F_1. Micro-averaging and macro-averaging [23] are ways to aggregate performance scores over multiple labels. Let tp_l, tn_l, fp_l, and fn_l be the true positives, true negatives, false positives, and false negatives of label l, and $M = |L|$. Micro-avaraged F_1 and macro-averaged F_1 are calculated:

$$\text{Micro } F_1 = F_1(\sum_{l=1}^{M} tp_l, \sum_{l=1}^{M} tn_l, \sum_{l=1}^{M} fp_l, \sum_{l=1}^{M} fn_l) \tag{3}$$

$$\text{Macro } F_1 = \frac{1}{M} \sum_{l=1}^{M} F_1(tp_l, tn_l, fp_l, fn_l) \tag{4}$$

In micro-averaging, predictions are weighted by sample, while for macro-averaging predictions are weighted by label. Therefore, micro-averaged F_1 is dominated by labels with many positive samples, and is really a measure of performance on the 'large' labels. Macro-averaged F_1, on the other hand, gives a sense of classifier performance on 'small' labels (i.e., labels with few positive samples) [24]. Because in the HEEM dataset, there are both labels with many samples, and labels with few samples (see table II), we report both micro-averaged F_1 and macro-averaged F_1.

C. Spelling Normalization

A problem with 17th and 18th century Dutch is that a standardized spelling did not yet exist. This means that words were written using different spellings (e.g., 'wraeck', 'wraecke', 'wrake', 'wraake' and 'wraak' for *wraak* (revenge)). Sometimes, different spellings of a word are used within the same text. Spelling variation is a problem, because most text classification approaches –including the one used in this paper– use bag-of-words features to predict whether a text belongs to a class, and rely on spelling to determine whether words are indicative of a class. Currently, there are no NLP tools available for 17th or 18th century Dutch, making it impossible to extract reliable information on word stems, lemmas, or part-of-speech (POS) tags.

In order to try to reduce the problem of spelling variation, we made a spelling-normalized version of the HEEM dataset. Because (historical) spelling normalization is not the main focus of our work, the procedure was kept as simple as possible. We automatically created a substitution list that maps historical terms to modern terms. The mapping was created by looking up historical spelling variants of words in the BWNT[5] in the INL LexiconService[6]. The BWNT is a glossary of $111,796$ common Dutch words for which the LexiconService provided a total of $103,062$ spelling variants in the time period from 1600 to 1800. Subsequently, the mapping from modern to historical terms was reversed. By far most historical terms map unambiguously to a single modern variant (96.7%). For historical terms that map to multiple modern variants, a single variant was selected (i.e., the first variant sorted in alphabetical order).

Applying the historical-to-modern mapping to the HEEM dataset resulted in 59.2% of the words being replaced by a modern variant. Inspection of the mapping shows that noise is introduced by this preprocessing step (e.g., some verbs are replaced by nouns, and different types of pronouns are merged into the same term). The spelling normalization was completed by removing accents from accented characters. The effect of spelling normalization on classification is discussed in section V-B.

V. Results

Four multi-label text classification experiments were performed on the HEEM dataset. First, we investigate the effect of taking into account correlations between labels by training classifiers using the Binary Relevance method and compare the results to classifiers trained using RAkEL. The second experiment explores the effect of spelling normalization. The third and fourth experiment concern simplifications of HEEM, i.e., HEEM Basic, and the Positive/Negative model. Classifier performance reported for experiments is obtained by perform-

[5]Bronbestand Woordenlijst Nederlandse Taal 2005 as provided by Nederlandse Taalunie (http://tst.inl.nl/producten/GB05/?db_select=GB05_001; in Dutch)

[6]http://sk.taalbanknederlands.inl.nl/LexiconService/, http://www.succeed-project.eu/wiki/index.php/Lexicon_Service

ing 10-fold cross validation. F_1 scores were compared using Welch's two-sided t-test.

A. HEEM

Table VII lists the results for BR vs. RAkEL on the HEEM labels. The results are summarized in the last column of the table. Upward pointing arrows (↗) indicate statistically significant improvements, while downward pointing arrows (↘) indicate statistical significant decreases. Overall performance (micro and macro-averaged F_1) is reported in the last two rows of the table. The results show RAkEL is superior over BR for 18 individual labels. In particular, the 'middle' classes (i.e., labels for which the number of samples is > 40 and < 200) seem to benefit from taking into account correlations between labels. There is no statistically significant difference between micro-averaged F_1. Macro-averaged F_1 for RAkEL is significantly higher than for BR.

The results for the individual labels are consistent with the agreement scores reported in section III-C; concept types *Emotion* and *Body part* have relatively high *agr* (> 0.7) and F_1 (> 0.4), while the scores for *Bodily process* and *Emotional action* are lower ($agr < 0.5$; $F_1 < 0.4$). Also, especially larger emotion labels (i.e., labels that occur > 200 times) have both high agreement ($agr > 0.69$), and, generally, high F_1 scores ($(0.4 < F_1 < 0.2)$). Interestingly, many labels have an F_1 score of 0. This means that the label is not predicted at all. For BR, 19 labels have an F_1 score of 0. For RAkEL, the performance of 14 of these labels does not differ significantly from 0. All these labels occur < 40 times in the data, and the minimum number of test samples is 0. This explains why performance is so low. These labels also did not occur in the texts annotated for the agreement study. So, again, the text classification results are consistent with the agreement study.

Overall, the results show that predicting labels from a complex emotion model is feasible. Performance is comparable to related work with simpler emotion models (see section II). In addition, we show that taking into account correlations between labels is useful for HEEM.

B. Spelling Normalization

Table VIII lists the results for RAkEL on the spelling normalized data. Compared to the results on the original data, both micro and macro-averaged F_1 scores increase significantly. This increase is entirely due to increases in performance of dominant classes, i.e., labels that have > 150 samples (with the exception of *Disgust*). There are no labels for which performance goes from zero to non-zero. The results for BR vs. spelling normalized are comparable (significant improvements in F_1 for a limited set of labels, again mainly the dominant classes).

We conclude that spelling normalization helps to improve performance for dominant labels. For the other experiments, we only report results on the spelling normalized data. For both experiments, the results on the original data are comparable, but significantly worse on the overall F_1 scores, and F_1 scores of a number of dominant classes.

Table VII
F_1 SCORES FOR BINARY RELEVANCE (BR) AND RAKEL

	BR	RAkEL	
Emotion	0.543 ± 0.025	0.542 ± 0.023	
Bodily process	0.215 ± 0.050	0.311 ± 0.037‡	↗
Body part	0.309 ± 0.049	0.419 ± 0.047‡	↗
Emotional action	0.028 ± 0.031	0.129 ± 0.036‡	↗
Love	0.484 ± 0.060	0.536 ± 0.038†	↗
Sadness	0.441 ± 0.051	0.473 ± 0.045	
Fear	0.493 ± 0.079	0.549 ± 0.046	
Joy	0.439 ± 0.100	0.432 ± 0.058	
Anger	0.288 ± 0.084	0.323 ± 0.062	
Despair	0.208 ± 0.105	0.277 ± 0.051	
Vindictiveness	0.193 ± 0.103	0.385 ± 0.073‡	↗
Desire	0.025 ± 0.052	0.172 ± 0.062‡	↗
Hatred	0.588 ± 0.064	0.572 ± 0.090	
Hope	0.386 ± 0.125	0.520 ± 0.111†	↗
Compassion	0.100 ± 0.106	0.251 ± 0.140†	↗
Remorse	0.059 ± 0.135	0.238 ± 0.145†	↗
Worry	0.000 ± 0.000	0.068 ± 0.080†	↗
Happiness	0.152 ± 0.193	0.323 ± 0.142†	↗
Other	0.000 ± 0.000	0.031 ± 0.048	
Shame	0.029 ± 0.086	0.248 ± 0.186†	↗
Heavy-heartedness	0.000 ± 0.000	0.072 ± 0.079†	↗
Honor	0.125 ± 0.157	0.250 ± 0.157	
Disgust	0.033 ± 0.100	0.077 ± 0.068	
Loyalty	0.000 ± 0.000	0.148 ± 0.132†	↗
Wonder	0.018 ± 0.055	0.245 ± 0.210†	↗
Spitefulness	0.108 ± 0.135	0.296 ± 0.189†	↗
Moved	0.000 ± 0.000	0.190 ± 0.157†	↗
Annoyance	0.000 ± 0.000	0.029 ± 0.086	
Envy	0.117 ± 0.236	0.250 ± 0.258	
Acquiescence	0.000 ± 0.000	0.000 ± 0.000	
Benevolence	0.000 ± 0.000	0.000 ± 0.000	
Pride	0.000 ± 0.000	0.127 ± 0.220	
Suspicion	0.000 ± 0.000	0.230 ± 0.201†	↗
Offended	0.000 ± 0.000	0.077 ± 0.132	
Dedication	0.000 ± 0.000	0.000 ± 0.000	
Unhappiness	0.000 ± 0.000	0.000 ± 0.000	
Disappointment	0.000 ± 0.000	0.000 ± 0.000	
Greed	0.000 ± 0.000	0.000 ± 0.000	
Trust	0.000 ± 0.000	0.017 ± 0.050	
Awe	0.000 ± 0.000	0.000 ± 0.000	
Relief	0.000 ± 0.000	0.042 ± 0.085	
Loss	0.000 ± 0.000	0.090 ± 0.181	
Micro-averaged F_1	0.388 ± 0.019	0.404 ± 0.017	
Macro-averaged F_1	0.128 ± 0.017	0.213 ± 0.011‡	↗

† statistically significant at $p < 0.05$
‡ statistically significant at $p < 0.001$

C. HEEM Emotion Clusters

Table IX reports results for BR vs. RAkEL on the HEEM Emotion Clusters (spelling normalized data). Compared to BR, RAkEL significantly improves performance of four emotion labels. The results for the four concept types are very similar to the results obtained in the original experiment (significant performance improvement for 3 labels).

In table X performance for the dominant HEEM label in the merged HEEM Basic labels is compared to performance for the new subdivision in HEEM Basic emotion labels. Performance for four labels decreases significantly, one significantly increases, and the others do not change. Performance for the four concept types is the same. With regard to the overall

Table VIII
F_1 SCORES FOR RAKEL AND SPELLING NORMALIZED RAKEL

	RAkEL	sp. norm. RAkEL	
Emotion	0.542 ± 0.023	0.578 ± 0.027†	↗
Bodily process	0.311 ± 0.037	0.370 ± 0.053†	↗
Body part	0.419 ± 0.047	0.422 ± 0.041	
Emotional action	0.129 ± 0.036	0.168 ± 0.040†	↗
Love	0.536 ± 0.038	0.572 ± 0.047	
Sadness	0.473 ± 0.045	0.562 ± 0.032‡	↗
Fear	0.549 ± 0.046	0.615 ± 0.047†	↗
Joy	0.432 ± 0.058	0.568 ± 0.054‡	↗
Anger	0.323 ± 0.062	0.357 ± 0.069	
Despair	0.277 ± 0.051	0.345 ± 0.114	
Vindictiveness	0.385 ± 0.073	0.453 ± 0.069	
Desire	0.172 ± 0.062	0.242 ± 0.054†	↗
Hatred	0.572 ± 0.090	0.624 ± 0.079	
Hope	0.520 ± 0.111	0.486 ± 0.108	
Compassion	0.251 ± 0.140	0.301 ± 0.131	
Remorse	0.238 ± 0.145	0.284 ± 0.124	
Worry	0.068 ± 0.080	0.087 ± 0.118	
Happiness	0.323 ± 0.142	0.317 ± 0.147	
Other	0.031 ± 0.048	0.022 ± 0.048	
Shame	0.248 ± 0.186	0.383 ± 0.104	
Heavy-heartedness	0.072 ± 0.079	0.091 ± 0.111	
Honor	0.250 ± 0.157	0.255 ± 0.228	
Disgust	0.077 ± 0.068	0.211 ± 0.104†	↗
Loyalty	0.148 ± 0.132	0.163 ± 0.069	
Wonder	0.245 ± 0.210	0.262 ± 0.166	
Spitefulness	0.296 ± 0.189	0.336 ± 0.161	
Moved	0.190 ± 0.157	0.244 ± 0.150	
Annoyance	0.029 ± 0.086	0.025 ± 0.051	
Envy	0.250 ± 0.258	0.257 ± 0.215	
Acquiescence	0.000 ± 0.000	0.025 ± 0.075	
Benevolence	0.000 ± 0.000	0.000 ± 0.000	
Pride	0.127 ± 0.220	0.144 ± 0.165	
Suspicion	0.230 ± 0.201	0.243 ± 0.239	
Offended	0.077 ± 0.132	0.035 ± 0.070	
Dedication	0.000 ± 0.000	0.000 ± 0.000	
Unhappiness	0.000 ± 0.000	0.058 ± 0.118	
Disappointment	0.000 ± 0.000	0.000 ± 0.000	
Greed	0.000 ± 0.000	0.000 ± 0.000	
Trust	0.017 ± 0.050	0.038 ± 0.077	
Awe	0.000 ± 0.000	0.000 ± 0.000	
Relief	0.042 ± 0.085	0.000 ± 0.000	
Loss	0.090 ± 0.181	0.067 ± 0.200	
Micro-averaged F_1	0.404 ± 0.017	0.449 ± 0.024‡	↗
Macro-averaged F_1	0.213 ± 0.011	0.243 ± 0.009‡	↗

† statistically significant at $p < 0.05$
‡ statistically significant at $p < 0.001$

Table IX
F_1 SCORES FOR BINARY RELEVANCE (BR) AND RAKEL ON HEEM EMOTION CLUSTERS (SPELLING NORMALIZED DATA)

	sp. norm. BR	sp. norm. RAkEL	
Emotion	0.579 ± 0.028	0.585 ± 0.029	
Bodily process	0.295 ± 0.049	0.378 ± 0.035‡	↗
Body part	0.334 ± 0.051	0.428 ± 0.051†	↗
Emotional action	0.043 ± 0.037	0.184 ± 0.043‡	↗
Sadness	0.490 ± 0.039	0.511 ± 0.027	
Love	0.501 ± 0.047	0.532 ± 0.029	
Anger	0.428 ± 0.040	0.473 ± 0.033†	↗
Fear	0.535 ± 0.086	0.550 ± 0.043	
Joy	0.469 ± 0.068	0.507 ± 0.0406	
Desire	0.071 ± 0.058	0.221 ± 0.0567‡	↗
Despair	0.245 ± 0.119	0.290 ± 0.0737	
Disgust	0.152 ± 0.133	0.221 ± 0.040	
PosSentiments	0.042 ± 0.064	0.114 ± 0.049†	↗
Compassion	0.200 ± 0.124	0.227 ± 0.098	
PrideHonor	0.085 ± 0.107	0.161 ± 0.0529	
Other	0.000 ± 0.000	0.033 ± 0.030†	↗
Micro-averaged F_1	0.4501 ± 0.0231	0.4432 ± 0.0185	
Macro-averaged F_1	0.2792 ± 0.0298	0.3384 ± 0.0156	↗

† statistically significant at $p < 0.05$
‡ statistically significant at $p < 0.001$

Table X
F_1 SCORES FOR DOMINANT HEEM LABELS AND HEEM EMOTION CLUSTERS LABELS (RAKEL, AND SPELLING NORMALIZED DATA)

	HEEM	HEEM Emotion Clusters	
Emotion	0.578 ± 0.027	0.585 ± 0.029	
Bodily process	0.370 ± 0.053	0.378 ± 0.035	
Body part	0.422 ± 0.041	0.428 ± 0.051	
Emotional action	0.168 ± 0.040	0.184 ± 0.043	
Sadness (Sadness)	0.562 ± 0.032†	0.511 ± 0.027	↘
Love (Love)	0.572 ± 0.047†	0.532 ± 0.029	↘
Anger (Anger)	0.357 ± 0.069	0.473 ± 0.033‡	↗
Fear (Fear)	0.615 ± 0.047†	0.550 ± 0.043	↘
Joy (Joy)	0.568 ± 0.054†	0.507 ± 0.041	↘
Desire (Desire)	0.242 ± 0.054	0.221 ± 0.057	
Despair (Despair)	0.345 ± 0.114	0.290 ± 0.074	
Disgust (Disgust)	0.211 ± 0.104	0.221 ± 0.040	
Moved (PosSentiments)	0.244 ± 0.150†	0.114 ± 0.049	↘
Compassion	0.301 ± 0.131	0.227 ± 0.098	
Honor (PrideHonor)	0.255 ± 0.228	0.161 ± 0.053	
Other	0.022 ± 0.048	0.033 ± 0.030	
Micro-averaged F_1	0.449 ± 0.024	0.443 ± 0.019	
Macro-averaged F_1	0.243 ± 0.009	0.338 ± 0.016‡	↗

† statistically significant at $p < 0.05$
‡ statistically significant at $p < 0.001$

performance measures, micro-averaged F_1 does not change, and macro-averaged F_1 significantly increases. This can be explained by the absence of labels that are not predicted. Otherwise, it seems that merging labels has no positive effect on performance.

D. The Positive/Negative Model

The results for BR vs. RAkEL on the Positive/Negative model are displayed in table XI. In this experiment, RAkEL does not outperform BR on the emotion labels (the results for the concept types are similar to the performance obtained in the original experiment). Apparently, there are no longer significant correlations between concept types and emotion labels and between emotion labels themselves. Given that there are only two mutually exclusive emotion labels (not counting *Other*, which is a very small class), this is not surprising.

When comparing the overall performance of the Positive/Negative model to HEEM and HEEM Emotion Clusters, we see that both micro-averaged and macro-averaged F_1 for the Positive/Negative model are significantly higher than HEEM and HEEM Emotion Clusters ($p < 0.001$). No statistically significant differences were found for the performance of the HEEM concept types.

Table XI
F_1 SCORES FOR BINARY RELEVANCE (BR) AND RAKEL ON THE POSITIVE/NEGATIVE MODEL (SPELLING NORMALIZED DATA)

	BR	RAkEL	
Emotion	0.579 ± 0.028	0.603 ± 0.028	
Bodily process	0.295 ± 0.049	$0.376 \pm 0.035^{\dagger}$	↗
Body part	0.333 ± 0.051	$0.428 \pm 0.047^{\ddagger}$	↗
Emotional action	0.043 ± 0.037	$0.204 \pm 0.046^{\ddagger}$	↗
Negative	0.565 ± 0.027	0.578 ± 0.017	
Positive	0.476 ± 0.048	0.511 ± 0.041	
Other	0.000 ± 0.000	$0.023 \pm 0.024^{\dagger}$	↗
Micro-averaged F_1	0.494 ± 0.019	0.492 ± 0.020	
Macro-averaged F_1	0.327 ± 0.013	$0.389 \pm 0.019^{\ddagger}$	↗

† statistically significant at $p < 0.05$
‡ statistically significant at $p < 0.001$

VI. DISCUSSION AND CONCLUSION

This paper presented the Historic Embodied Emotion Model (HEEM) and HEEM dataset from a technical perspective. HEEM and the associated dataset are relevant for typical Digital Humanities research questions that require more complex and historically accurate emotion models, and aim at applying these models to historical data spanning larger time periods (e.g., centuries). HEEM was designed to study the relationship between body parts and emotional expression in the 17th and 18th century texts. The second contribution of this paper is a study of the performance of a multi-label text classification approach for predicting HEEM labels and the labels of two simpler models derived from HEEM. We performed four experiments: 1) the Binary Relevance (BR) method was compared to Random *k*-Labelsets (RAkEL) on the HEEM model, 2) the effect of a simple method for spelling normalization was measured, 3) performance of HEEM was compared to HEEM Emotion Clusters, an emotion model created by merging HEEM labels into 12 classes of historically accurate emotion clusters, and 4) performance of HEEM was compared to the Positive/Negative model consisting of HEEM labels divided into positive and negative emotions.

The results show that HEEM labels can be predicted with micro-averaged $F_1 = 0.45$, and macro-averaged $F_1 = 0.24$ on the spelling normalized data. These results are comparable to performance from related work for simpler emotion models (e.g., Danisman and Alpkocak [13] obtain an F_1 score of 0.32 on a five-way multi-class prediction problem, and Buitinck et al. [2] report a micro-averaged F_1 score of 0.46 for an emotion model with 8 labels). The large difference between micro-averaged and macro-averaged F_1 can be explained by the fact that the classifiers never learn to predict 19-15 labels at all. Labels with many samples (> 200, i.e., the dominant classes), generally have better performance than labels with few samples (< 40). These 'small' labels do not always occur in the test set, which explains why performance is sometimes 0. It seems that a minimum number of samples in both the train and test sets is required to get any predictions at all. One way to solve this problem is by using stratified datasets [25].

In the second experiment, we show that spelling normalization significantly improves both micro and macro-averaged F_1. However, it is mainly the dominant labels that benefit (i.e., *Sadness*, *Joy*, *Fear*, and *Desire*). No additional labels were predicted. As we used a very simple method to normalize spelling that introduced quite some noise, better results can be expected from a more sophisticated method; for example, by taking into account additional linguistic information, such as POS tags. Adding linguistic information to the dataset might also improve classifier performance. However, since no NLP tools exist for 17th and 18th century Dutch, it will be challenging to add useful linguistic features.

The results from the final two experiments show that the performance for simpler models is not necessarily better. Although the overall performance measures significantly improve (with the exception of micro-avaraged F_1 for HEEM Basic Emotion Clusters compared to HEEM), this is not true for performance of individual labels. The large increase in macro-averaged F_1 can be explained by the fact that the simple models do not contain labels that are not predicted. A limitation of the simple models we used is that they were retrofitted, i.e., created based on HEEM, and dominant labels in the HEEM dataset. Overall, we conclude that a multi-label text classification approach to learning complex emotion models on historical data is feasible.

In future work, we intend to explore the effects of dividing the data into sets for the different time periods and/or genres. We expect that performance will increase, given that there is a sufficient number of samples for a label in a dataset. This might also mean that a classifier trained on a subset of the data will predict labels that are currently not predicted, for example, when a specific emotion label only occurs in a specific time period or genre.

REFERENCES

[1] R. Boddice, "The affective turn: historicising the emotions," in *Psychology and history: Interdisciplinary explorations*, C. Tileag and J. Byford, Eds. Cambridge: Cambridge University Press, 2014, pp. 147–156.

[2] L. Buitinck, J. van Amerongen, E. Tan, and M. de Rijke, "Multi-Emotion Detection in User-Generated Reviews," in *ECIR 2015: 37th European Conference on Information Retrieval*, 2015.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002, pp. 79–86.

[5] T. Mullen and N. Collier, "Sentiment Analysis using Support Vector Machines with Diverse Information Sources," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[6] F. Salvetti, S. Lewis, and C. Reichenbach, "Impact of lexical filtering on overall opinion polarity identification," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

[7] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[8] K. Dave, S. Lawrence, and D. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 519–528.

[9] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Text, Speech and Dialogue*. Springer, 2007, pp. 196–205.

[10] C. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 579–586.

[11] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 2005, pp. 347–354.

[12] C. Yang, K. Lin, and H. Chen, "Emotion classification using web blog corpora," in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2007, pp. 275–278.

[13] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model," in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, 2008.

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

[15] K. Scherer, "What are emotions? And how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[16] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Machine learning: ECML 2007*, 2007, pp. 406–417.

[17] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[18] J. Read and J. Carroll, "Annotating expressions of appraisal in English," *Language resources and evaluation*, vol. 467, no. 3, pp. 421–447, 2012.

[19] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquax, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop on Languages for Machine Learning*, 2013.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int J Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[22] G. Tsoumakas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[23] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[24] C. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007.

[25] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, 2011, pp. 145–158.