How to Estimate Continuous Sentiments From Texts Using Binary Training Data

Sandra Wankmüller
Ludwig-Maximilians-Universität
Munich, Germany
https://orcid.org/0000-0002-4003-1704
sandra.wankmueller@gsi.lmu.de

Christian Heumann
Ludwig-Maximilians-Universität
Munich, Germany
chris@stat.uni-muenchen.de

Abstract

Although sentiment is conceptualized as a continuous variable, most text-based sentiment analyses categorize texts into discrete sentiment categories. Compared to discrete categorizations, continuous sentiment estimates provide much more detailed information which can be used for more fine-grained analyses by researchers and practitioners alike. Yet, existing approaches that estimate continuous sentiments either require detailed knowledge about context and compositionality effects or require granular training labels, that are created in resource intensive annotation processes. Thus, existing approaches are too costly to be applied for each potentially interesting application. To overcome this problem, this work introduces CBMM (standing for *c*lassifier-based *b*eta *m*ixed *m*odeling procedure). CBMM aggregates the predicted probabilities of an ensemble of binary classifiers via a beta mixed model and thereby generates continuous, real-valued output based on mere binary training input. CBMM is evaluated on the Stanford Sentiment Treebank (SST) (Socher et al., 2013), the V-reg data set (Mohammad et al., 2018), and data from the 2008 American National Election Studies (ANES) (The American National Election Studies, 2015). The results show that CBMM produces continuous sentiment estimates that are acceptably close to the truth and not far from what could be obtained if highly fine-grained training data were available.

1 Introduction

In natural language processing and computer science, the term *sentiment* typically refers to a loosely defined, broad umbrella concept: Feeling, emotion, judgement, evaluation, and opinion all fall under the term sentiment or are used synonymously with it (Pang and Lee, 2008; Liu, 2015). Interestingly, the broad notion of sentiment is very well captured by the psychological concept of an attitude (Liu, 2015). In psychology, scholars agree that an attitude is a summary evaluation of an entity (Banaji and Heiphetz, 2010; Albarracin et al., 2019). An attitude is the aggregated evaluative response resulting from a multitude of different (and potentially conflicting) information bases relating to the attitude entity (Fabringar et al., 2019). When putting the definition of an attitude as an evaluative summary into mathematical terms, an attitude is a unidimensional, continuous variable ranging from highly negative to highly positive (Cacioppo et al., 1997). This notion that attitudes are continuous is also mirrored in the sentiment analysis literature in which sentiments are devised to vary in their levels of intensity (Liu, 2015).

Despite this conceptualization, in an overwhelming majority of studies textual sentiment expressions are measured as instances of discrete classes. Sentiment analysis often implies a binary or multi-class classification task in which texts are assigned into two or three classes, thereby distinguishing positive from negative sentiments and sometimes a third neutral category (e.g. Pang et al., 2002; Turney, 2002; Maas et al., 2011). Other studies pursue ordinal sentiment classification (e.g. Pang and Lee, 2005; Thelwall et al., 2010; Socher et al., 2013; Kim, 2014; Zhang et al., 2015; Cheang et al., 2020). Here, texts fall into one out of several discrete and ordered categories.

If researchers would generate continuous—rather than discrete—sentiment estimates, this would not only align the theoretical conceptualization of sentiment with the way it is measured but also would provide much more detailed information that in turn can be used by researchers and practitioners for more fine-grained analyses and more fine-tuned responses.

For example, in the plot on the right hand side in Figure 1, the distribution of the binarized sentiment values of the tweets in the V-reg data set (Mo-

hammad et al., 2018) is shown. If researchers and practitioners would operate only on this discrete sentiment categorization, the shape of the underlying continuous sentiment distribution would be unknown. In fact, all distributions shown on the left hand side in Figure 1 produce the plot on the right hand side in Figure 1 if the sentiment values are binarized in such way that tweets with a sentiment value of ≥ 0.5 are assigned to the positive class and otherwise are assigned to the negative class. Imagine that a team of researchers would be interested in the sentiments expressed toward a policy issue and they would only know the binarized sentiment values on the right hand side in Figure 1. The researchers would not be able to conclude whether the expressions toward the policy issue are polarized into a supporting and an opposing side, whether a large share of sentiment expressions is positioned in the neutral middle, or whether the sentiments are evenly spread out. Knowing the continuous sentiment values, however, would allow them to differentiate between these scenarios.

As will be elaborated in Section 2, existing approaches that estimate continuous sentiment values for texts rely on (1) the availability of a comprehensive, context-matching sentiment lexicon and the researcher's knowledge regarding how to accurately model compositionality effects, or (2) highly costly processes to create fine-grained training data.

Sentiment analysis thus would benefit from a technique that generates continuous sentiment predictions for texts and is less demanding concerning the required information or resources. To meet this need, this work explores in how far the here proposed *c*lassifier-based *b*eta *m*ixed *m*odeling approach (CBMM) can produce valid continuous (i.e. real-valued) sentiment estimates on the basis of mere binary training data. The method comprises three steps. First, for each training set document a binary class label indicating whether the document is closer to the negative or the positive extreme of the sentiment variable has to be created or acquired. Second, an ensemble of J classifiers is trained on the binary class labels to produce for each of N test set documents J predicted probabilities to belong to the positive class. Third, a beta mixed model with N document random intercepts and J classifier random intercepts is estimated on the predicted probabilities. The N document random intercepts are the documents' continuous sentiment estimates.

In the following section, existing approaches

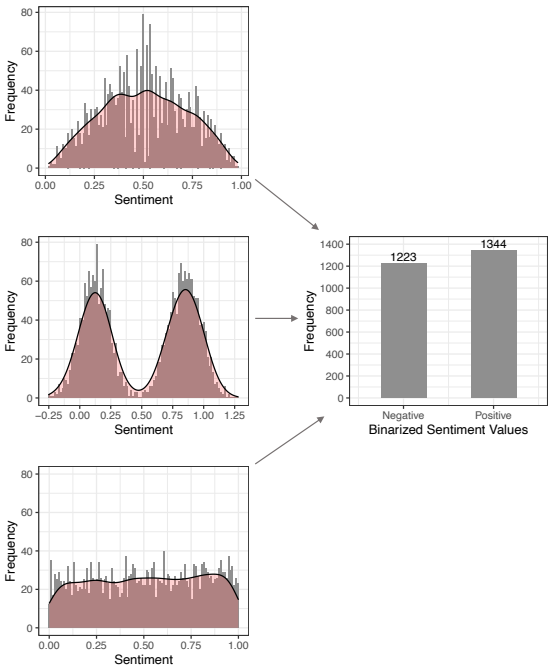

Figure 1: Continuous and Discrete Sentiment Distributions. Right plot: Binarized sentiment values of the tweets in the V-reg data set (Mohammad et al., 2018). Left plots: Histograms and kernel density estimates for three continuous distributions of sentiments that produce the plot on the right hand side if the continuous sentiment values are binarized such that tweets with values of ≥ 0.5 are assigned to the positive class and otherwise are assigned to the negative class. The unimodal distribution at the top is the true distribution of sentiment values but the other two distributions would generate the same binary separation of tweets into positive and negative.

that generate continuous sentiments are reviewed (Section 2). Then, CBMM is introduced in detail (Section 3) before it is evaluated on the basis of the Stanford Sentiment Treebank (SST) (Socher et al., 2013), the V-reg data set (Mohammad et al., 2018), and data from the 2008 American National Election Studies (ANES) (The American National Election Studies, 2015) (Section 4). A concluding discussion follows in Section 5.

2 Related Work

This work is concerned with the estimation of continuous values for texts in applications in which the underlying, unidimensional, continuous variable (e.g. sentiment) is well defined and the researcher seeks to position the texts along exactly this variable. Hence, this work does not consider unsupervised approaches (e.g. Slapin and Proksch,

2008) and only considers methods in which information on the definition of the underlying variable explicitly enters the estimation of the texts' values. Among these methods, one can distinguish two major approaches: lexicon-based procedures and regression models that operate on fine-grained training data.[1]

2.1 Lexicon-Based Approaches

An ideal sentiment lexicon covers all features in the corpus of an application and precisely assigns each feature to the sentiment value the feature has in the thematic context of the application (Grimmer and Stewart, 2013; Gatti et al., 2016). A major difficulty of lexicon-based approaches, however, is that even such an ideal sentiment lexicon will not guarantee highly accurate sentiment estimates. The reason is that sentiment builds up through complex compositional effects (Socher et al., 2013). These compositional effects either can be modeled via human-created rules or can be learned by supervised machine learning algorithms. Approaches that try to model compositionality via human-created rules range from simple formulas (e.g. Paltoglou et al., 2013; Gatti et al., 2016) to elaborate procedures (e.g. Moilanen and Pulman, 2007; Thet et al., 2010). Human-coded compositionality rules, however, tend to be outperformed by supervised machine learning algorithms (compare e.g. Gatti et al., 2016, Table 12 and Socher et al., 2013, Table 1). In the latter case, sentiment lexicons serve the purpose of creating the feature inputs to regression approaches—which are discussed next.

2.2 Regression Approaches

The second major set of approaches that generate real-valued sentiment estimates makes use of highly granular training data (e.g. as in the SST data set where each text is assigned to one out of 25 distinct values (Socher et al., 2013)). In these approaches, the fine-grained annotations are treated as if they were continuous and a regression model is applied.[2] Typically, the mean squared error (MSE) between the true granular labels and the real-valued predictions from the regression model is minimized. Regression approaches have shown to be able to generate continuous sentiment predictions that are quite close to the true fine-grained labels (Mohammad et al., 2018; Wang et al., 2018). Yet, the prerequisite for implementing such an approach is that fine-grained labels for the training data are available. Generating such granular annotations, however, is difficult and costly: Categorizing a training text into few ordinal categories is arguably a more easy task than assigning a text into one out of a large number of ordered values or even rating a text on a real-valued scale. As the number of distinct values increases, the number of inter- and intra-rater disagreements is likely to increase (Krippendorff, 2004). Hence, to produce reliable text annotations, it is advantageous to have each document rated several times by independent raters. The independent ratings then can be aggregated by taking the median or the mean of the ratings to obtain the final value (see e.g. Kiritchenko and Mohammad, 2017). The larger the number of raters for a document, the more reliable the final value assigned to the document. For this reason, generating reliable fine-grained labels for training documents via rating scale annotations requires a resource intensive annotation process.

The best-worst scaling (BWS) method in which coders have to identify the most positive and the most negative document among tuples of documents (typically 4-tuples), alleviates the problems of inter- and intra-rater inconsistencies (Kiritchenko and Mohammad, 2017). Yet, in order for the rankings among document tuples to generate valid real-valued ratings via the counting procedure implemented in BWS, it is essential that each document occurs in many different tuples such that each document is compared to many different other documents. This implies that a substantive number of unique tuples have to be annotated—which, in turn, demands respective human coding resources.

An alternative to the labeling of texts by human coders is the usage of already available information (e.g. if product reviews additionally come with numerical star ratings). The problem here, however, is that such information—if available at all—often comes in the form of discrete variables with only few distinct values (e.g. 5-star rating systems).

[1] Techniques for estimating continuous document positions on an a priori defined unidimensional latent variable also have been developed in political science. These methods either are at their core lexicon-based approaches (Watanabe, 2021) or require continuous values for the training documents (Laver et al., 2003)—and thus have the same shortcomings as either lexicon-based or regression approaches.

[2] Note that here, in correspondence with machine learning terminology, regression refers to statistical models and algorithms that model a real-valued response variable—which typically is assumed to follow a normal distribution.

To conclude, it is difficult and resource intensive to create or acquire fine-grained training data that is so detailed that it can be treated as if it were continuous. Not each team of researchers or practitioners will have the resources to create detailed training annotations and thus regression models cannot be applied to each substantive application of interest. Hence, the question that this work addresses is: Can one generate continuous sentiments with fewer costs in a setting where inter- and intra-rater inconsistencies are likely to be small? For example based on a simple binary coding of the training data?

3 Procedure

In the following, the three steps of the proposed CBMM procedure—(1) generating binary class labels, (2) training and applying an ensemble of classifiers, as well as (3) estimating a beta mixed model—are explicated. CBMM assumes that the documents to be analyzed are positioned on a latent, unidimensional, continuous sentiment variable. The aim is to estimate the test set documents' real-valued sentiment positions. The test set documents are indexed as $i \in \{1 \dots N\}$ and their sentiment positions are denoted as $\boldsymbol{\theta} = [\theta_1 \dots \theta_i \dots \theta_N]^\top$.

3.1 Generating Binary Class Labels

The CBMM procedure starts by generating binary class labels for the training set documents, e.g. via human coding. The coders classify the training documents into two classes such that the binary class label of each training set document indicates whether the document is closer to the negative (0) or the positive (1) extreme of the sentiment variable. Alternatively to human coding, binarized external information (such as star ratings associated with texts) can be used as class label indicators.

3.2 Training and Applying an Ensemble of Classifiers

In the second step, an ensemble of classification algorithms, indexed as $j \in \{1 \dots J\}$, is trained on the binary training data. The classifiers in the ensemble may differ regarding the type of algorithm, hyperparameter settings, or merely the seed values initializing the optimization process. After training, each classifier produces predictions for the N documents in the test set and each classifier's predicted probabilities for the test set documents to belong to the positive class are extracted. Thus, for each document i, a predicted probability to belong to class 1 is obtained from each classifier j, such that there are J predicted probabilities for each document: $\hat{\boldsymbol{y}}_i = [\hat{y}_{i1} \dots \hat{y}_{ij} \dots \hat{y}_{iJ}]$; whereby $\hat{y}_{ij}$ is classifier j's predicted probability for document i to belong to class 1.

3.3 Estimating a Beta Mixed Model

In step three, the aim is to infer the unobserved documents' continuous values on the latent sentiment variable from the observed predicted probabilities that have been generated by the set of classifiers. The approach taken here is similar to item response theory (IRT) in which unobserved subjects' values on a latent variable of interest (e.g. intelligence) are inferred from the observed subjects' responses to a set of question items (Hambleton et al., 1991). Central to IRT is the assumption that a subject's value on the latent variable of interest *affects* the subject's responses to the set of question items (Hambleton et al., 1991). For example, a subject's level of intelligence is postulated to influence his/her answers in an intelligence test. In correspondence with this assumption, the consistent mathematical element across all types of IRT models is that the observed subjects' responses are regressed on the unobserved subjects' latent levels of ability.

Here, there are documents rather than subjects and classifiers rather than question items. Yet, the aim is the same: to infer unobserved latent positions from what is observed. As in IRT, the idea here is that a document's value on the latent sentiment variable *affects* the predicted probabilities the document obtains from the classifiers. For example, a document with a highly positive sentiment is assumed to get rather high predicted probabilities from the classifiers. Consequently, the predicted probabilities are regressed on the documents' latent sentiment positions.

In doing so, it has to be accounted for that the predicted probabilities are grouped in a crossed non-nested design: In step 2, for each of the N documents, J predicted probabilities (one from each classifier) are produced such that there are $N \times J$ predicted probabilities. These predicted probabilities cannot be assumed to be independent. The J predicted probabilities for one document are likely to be correlated because they are repeated measurements on the same document. Additionally, the N predicted probabilities produced by one classifier also are generated by a common source. They come

from the same classifier that might systematically differ from the others, e.g. produce systematically lower predicted probabilities.

Moreover, the data generating process is such that the documents are drawn from a larger population of documents. The population distribution of the probability to belong to class 1 might inform the probabilities obtained by individual documents. Similarly, the classifiers are sampled from a population of classifiers with a population distribution in the generated predicted probabilities that may inform an individual classifier's predicted probabilities. To account for this data generating process, a mixed model with N document random intercepts and J classifier random intercepts seems the adequate model of choice. (On mixed models see for example Fahrmeir et al. (2013, chapter 7)).

As the predicted probabilities, $\hat{y}_{ij}$, are in the unit interval [0,1], it is assumed that the $\hat{y}_{ij}$ are beta distributed. Following the parameterization of the beta density employed by Ferrari and Cribari-Neto (2004) the beta mixed model is:

$$\hat{y}_{ij} \sim B(\mu_{ij}, \phi) \tag{1}$$

$$g(\mu_{ij}) = \beta_0 + \theta_i + \gamma_j \tag{2}$$

$$\theta_i \sim N(0, \tau_\theta^2) \tag{3}$$

$$\gamma_j \sim N(0, \tau_\gamma^2) \tag{4}$$

In the model described here, $\hat{y}_{ij}$ (the probability for document i to belong to class 1 as predicted by classifier j) is assumed to be drawn from a beta distribution with conditional mean μ_{ij}. μ_{ij} assumes values in the range (0,1) and $\phi > 0$ is a precision parameter (Cribari-Neto and Zeileis, 2010). μ_{ij} is determined by the fixed global population intercept β_0, the document-specific deviation θ_i from this population intercept, and the classifier-specific deviation γ_j from the population intercept. As the documents are assumed to be sampled from a larger population, the document-specific θ_i are modeled to be drawn from a shared distribution (see equation 3).[3] The same is true for the classifier-specific γ_j. To ensure that the results from the linear predictor in equation 2 are kept between 0 and 1, the logit link is chosen as the link function $g(\cdot)$.[4]

Note that in the beta distribution $Var(\hat{y}_{ij}) = \mu_{ij}(1-\mu_{ij})/(1+\phi)$ (Cribari-Neto and Zeileis, 2010). This means that the variance of $\hat{y}_{ij}$ not only depends on precision parameter ϕ but also depends on μ_{ij}, which implies that the model naturally exhibits heteroscedasticity (Cribari-Neto and Zeileis, 2010). In the given data structure, documents that express very positive (or very negative) sentiments are likely to be easy cases for the classifiers and it is likely that all classifiers will predict very high (or very low) values. Documents that express less extreme sentiments, in contrast, are likely to be more difficult cases and the classifiers are likely to differ more in their predicted probabilities. This is, predicted probabilities are likely to exhibit a higher variance for documents positioned in the middle of the sentiment value range. To additionally account for this effect, the beta mixed model described in equations 1 to 4 can be extended with a dispersion formula describing the precision parameter ϕ as a function of document-specific fixed effects:[5]

$$h(\phi_i) = \delta_i \tag{5}$$

To keep $\phi_i > 0$, $h(\cdot)$ here is the log link (Brooks et al., 2017).[6] In the following, CBMM is implemented with and without the dispersion formula in equation 5. The variant of CBMM that includes equation 5 is denoted CBMMd.

With or without a dispersion formula, the θ_i describe the document-specific deviations from the fixed population mean β_0. Hence, the θ_i —in linear relation to β_0—position the documents on the real line and thus are taken as the CBMM and CBMMd estimates for the continuous sentiment values.

4 Applications

4.1 Data

The effectiveness of CBMM in generating continuous sentiments using binary training data is evaluated on the basis of four data sets:

The Stanford Sentiment Treebank (SST) (Socher et al., 2013) contains sentiment labels for 11,855 sentences [train: 9,645; test: 2,210] taken from movie reviews. Each of the sentences was assigned one out of 25 sentiment score values ranging from highly negative (0) to highly positive (1) by three independent human annotators.

[3] Note that the usually employed assumption is that the random effects are independent and identically distributed according to a normal distribution (Fahrmeir et al., 2013).

[4] Thus, equation 2 is $log(\mu_{ij}/(1-\mu_{ij})) = \beta_0 + \theta_i + \gamma_j$.

[5] Note that the document-specific δ_i are fixed effects that are not modeled to be sampled from a shared population distribution. The reason is that current software implementations of mixed models that use maximum likelihood estimation only allow for inserting fixed effects but no random effects in the dispersion model formula (Brooks et al., 2017).

[6] Thus, equation 5 here is $log(\phi_i) = \delta_i$.

The V-reg data set from the SemEval-2018 Task 1 on "Affect in Tweets" (Mohammad et al., 2018) contains 2,567 tweets [train: 1,630; test: 937] that are likely to be rich in emotion. The tweets' real-valued valence scores are in the range (0,1) and were generated via BWS, whereby each 4-tuple was ranked by four independent coders.

Furthermore, *two data sets from the 2008 American National Election Studies (ANES)* (The American National Election Studies, 2015) are used. The feeling thermometer question, in which participants have to rate on an integer scale ranging from 0 to 100 in how far they feel favorable and warm vs. unfavorable and cold toward parties, is posed regularly in ANES surveys. In the 2008 pre-election survey, participants were additionally asked in open-ended questions to specify what they specifically like and dislike about the Democratic and the Republican Party.[7] Here, the aim is to generate continuous estimates of the sentiments expressed in the answers based on the binarized feeling thermometer scores. For the Democrats there are 1,646 answers [train: 1,097; test: 549]. This data set is named ANES-D. For the Republicans there are 1,523 answers [train: 1,015; test: 508] that make up data set ANES-R. For comparison with the other applications, the true scores from ANES are rescaled by min-max normalization from range [0,100] to [0,1].

To create binary training labels for the CBMM procedure, in all training data sets the fine-grained sentiment values are dichotomized such that the class label for a document is 1 if its score is $\geq$ 0.5 and is 0 otherwise. CBMM's continuous sentiment estimates for the test set documents then are compared to the original fine-grained values. Note that these four data sets are selected for evaluation precisely because they provide fine-grained sentiment scores against which the CBMM estimates can be compared to. In each of the four data sets, the detailed training annotations are the result of resourceful coding processes or—in the case of ANES—lucky coincidences. For example, around 50,000 annotations were made for the V-reg data set that comprises 2,567 tweets (Mohammad et al., 2018). Such resources or coincidences, however, are unlikely to be available for each potentially interesting research question. Thus, whilst these data sets are selected because they come with fine-grained labels that can be used for evaluating CBMM, the settings in which CBMM will be especially valuable are those in which external information that may serve as a granular training input is unavailable and the available amounts of resources are not sufficient for a granular coding of texts.

4.2 Generating Continuous Sentiment Estimates via CBMM

Step 2 of the CBMM procedure consists in training an ensemble of classifiers on the binary training data to then obtain predicted probabilities for the test set documents. Here, for all four applications, a set of 10 pretrained language representation models with the RoBERTa architecture (Liu et al., 2019) are fine-tuned to the binary classification task. The 10 models within one ensemble merely differ regarding their seed value that initializes the optimization process and governs batch allocation.[8] As the seed values are randomly generated, this neatly fits with the assumption encoded in the specified mixed models that classifiers are randomly sampled from a larger population of classifiers. As a Transformer-based model for transfer learning, RoBERTa is likely to yield relatively high prediction performances in text-based supervised learning tasks also if—as is the case for the selected applications—training data sets are small.

In step 3 of CBMM, two different beta mixed models as presented in equations 1 to 5—one model with and the other without a dispersion formula—are estimated. In each mixed model, the estimate for θ_i is taken as the sentiment value predicted for document i.

Steps 1 and 3 of the CBMM procedure are conducted in R (R Core Team, 2020). The beta mixed models are estimated with the R package glmmTMB (Brooks et al., 2017). In step 2, fine-tuning is conducted in Python 3 (Van Rossum and Drake, 2009) making use of PyTorch (Paszke et al., 2019). Pretrained RoBERTa models are accessed via the open-source library provided by HuggingFace's Transformers (Wolf et al., 2020). The source code to replicate the findings is available at https://doi.org/10.6084/m9.figshare.14381825.v1.

[7] The survey contains one question asking what the participant likes and a separate question asking what the participant dislikes about a party. For each respondent, the answers to these two questions are concatenated into a single answer.

[8] The 10 models applied for one application also have the same hyperparameter settings. In all four applications, a grid search across sets of different values for the batch size, the learning rate and the number of epochs is conducted via a 5-fold cross-validation. The hyperparameter setting that exhibits the lowest mean loss across the validation folds and does not suffer from too strong overfitting is selected.

4.3 Evaluation

4.3.1 Comparisons to Other Methods

The sentiment estimates from CBMM and CBMMd are compared to the following methods.

Mean of Predicted Probabilities [Pred-Prob-Mean]. For each document, this procedure simply takes the mean of the predicted probabilities across the ensemble of classifiers: $\hat{\theta}_i = \frac{1}{J}\sum_{j=1}^{J}\hat{y}_{ij}$.

Lexicon-Based Approaches. Two lexicons are made use of. First, the SST provides for each textual feature in the SST corpus a fine-grained human annotated sentiment value that indicates the feature's sentiment in the context of movie reviews. Hence, the SST constitutes an all-encompassing and perfectly tailored lexicon for the SST application and is employed as a lexicon here. Second, the SentiWords lexicon (Gatti et al., 2016), that is based on SentiWordNet (Esuli and Sebastiani, 2006) and contains prior polarity sentiment values for around 155,287 English lemmas, is used. For the SST and the SentiWords lexicons, the sentiment value estimates are generated by computing the arithmetic mean of a document's matched features' values. The procedures here are named SST-Mean and SentiWords-Mean.

Regression approaches, that make use of the true fine-grained sentiment values rather than the binary training data, are also applied. Note that the evaluation results for the regression-based procedures signify the levels of performance that can be achieved *if* one is in the ideal situation and possesses fine-grained training annotations. Hence, the regression approaches constitute a reference point against which the other approaches' performances can be related to.

Here, in all four applications, J = 10 RoBERTa regression models are trained on the training set and then make real-valued predictions for the documents in the test set such that there are J = 10 predictions for each test set document: $\hat{\boldsymbol{z}}_i = [\hat{z}_{i1} \dots \hat{z}_{ij} \dots \hat{z}_{iJ}]$; whereby $\hat{z}_{ij}$ is the real-valued prediction of regression model j for document i. To have a fair comparison to CBMM, the same procedures for aggregating the predicted values are explored. Thus, there are three different aggregation methods. First, the mean of the 10 models' predictions is taken such that the sentiment estimate is: $\hat{\theta}_i = \frac{1}{J}\sum_{j=1}^{J}\hat{z}_{ij}$ [Regr-Mean]. Second and third, a mixed model with and without a dispersion formula is estimated on the basis of the $\hat{z}_{ij}$. The estimates for the θ_i are extracted as the continuous sentiment predictions. Yet, to account for the data generating process of the $\hat{z}_{ij}$, a linear mixed model (LMM)—instead of a beta mixed model—is estimated:

$$\hat{z}_{ij} \sim N(\mu_{ij}, \sigma^2) \tag{6}$$

$$\mu_{ij} = \beta_0 + \theta_i + \gamma_j \tag{7}$$

$$\theta_i \sim N(0, \tau_\theta^2) \tag{8}$$

$$\gamma_j \sim N(0, \tau_\gamma^2) \tag{9}$$

This approach is named Regr-LMM. The LMM with a dispersion formula, Regr-LMMd, additionally has: $h(\sigma_i^2) = \delta_i$; with $h(\cdot)$ being the log link.

4.3.2 Evaluation Metrics

The generated continuous sentiment estimates are evaluated by comparing them to the original granular sentiment labels. Three evaluation metrics are used: the mean absolute error (MAE), the Pearson correlation coefficient r, and Spearman's rank correlation coefficient ρ. The evaluation metrics are selected such that there is a measure of the average absolute distance (MAE) as well as a measure of the linear correlation (r) between the original true sentiment values and the estimated values. Note that Spearman's ρ assesses the correlation between the ranks of the true and the ranks of the estimated values and thus evaluates in how far the order of documents from negative to positive sentiment as produced by the evaluated approaches reflects the order of documents according to the true scores.

4.4 Results

Table 1 presents the evaluation results across all applied data sets. Figure 2 visualizes distributions of the true and estimated sentiment values for the SST data. Across the four employed data sets (each with a different shape of the to be approximated distribution of the true sentiment values) the performance levels vary for all approaches. Yet, the main result remains consistent: the continuous sentiment estimates generated by CBMM correlate similarly with the truth and get only slightly less closer to the truth as the predictions generated by regression approaches that operate on fine-grained training data. At times, CBMM estimates even slightly outperform regression predictions. Hence, researchers that seek to get continuous sentiment estimates but do not have the resources to produce highly detailed training annotations can apply CBMM on binary training labels and thereby obtain estimated continuous sentiments whose performance is likely

	SST			V-reg			ANES-D			ANES-R		
	MAE	r	ρ	MAE	r	ρ	MAE	r	ρ	MAE	r	ρ
SST-Mean	0.190	0.554	0.574	0.171	0.437	0.487	0.242	−0.013	−0.033	0.252	0.059	0.058
SentiWords-Mean	0.201	0.428	0.429	0.177	0.429	0.475	0.254	0.067	0.079	0.289	−0.009	0.005
Regr-Mean	0.099	0.892	0.876	0.090	0.871	0.869	0.195	0.655	0.653	0.191	0.618	0.627
Regr-LMM	0.099	0.892	0.876	0.090	0.871	0.869	0.195	0.655	0.653	0.191	0.618	0.627
Regr-LMMd	0.099	0.892	0.876	0.090	0.872	0.870	0.195	0.655	0.653	0.192	0.618	0.627
Pred-Prob-Mean	0.216	0.859	0.856	0.198	0.804	0.844	0.202	0.646	0.649	0.218	0.624	0.613
CBMM	0.161	0.874	0.856	0.164	0.819	0.842	0.191	0.667	0.648	0.207	0.621	0.613
CBMMd	0.137	0.877	0.856	0.133	0.835	0.844	0.200	0.668	0.649	0.205	0.620	0.612

Table 1: Evaluation Results. For the SST, V-reg, ANES-D, and ANES-R test data sets, the mean absolute error (MAE), the Pearson correlation coefficient r, and Spearman's rank correlation coefficient ρ between the true and the estimated sentiment values are presented. The shading of the cells is a linear function of the approaches' level of performance. The darker the shading, the higher the performance. For computing the MAE, the predicted sentiment values are rescaled via min-max normalization to the range of the true sentiment values.

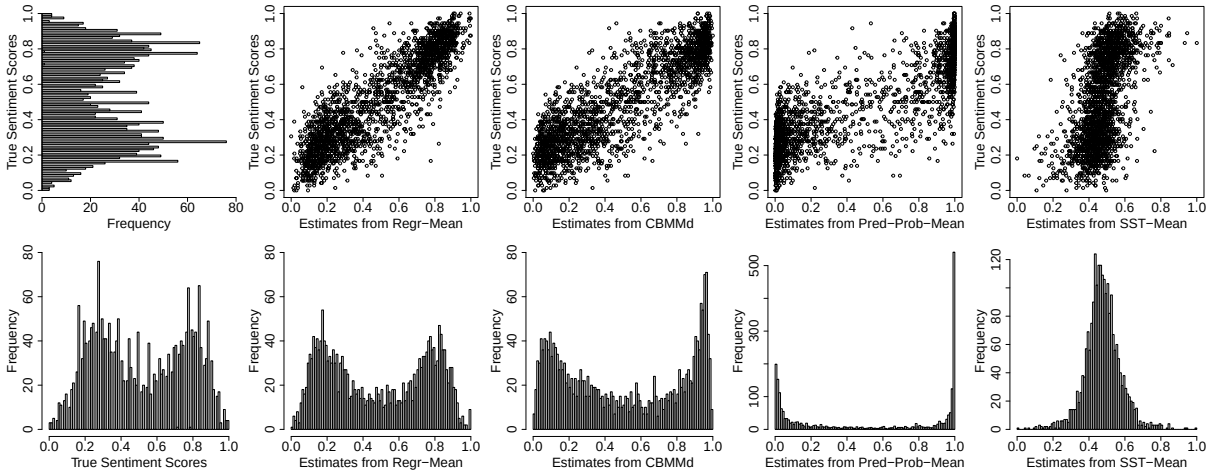

Figure 2: True and Estimated Sentiment Values for the SST Data. First column: Histograms of the true sentiment scores. Remaining columns, top row: Estimates from Regr-Mean, CBMMd, Pred-Prob-Mean, and SST-Mean plotted against the true sentiment values. Remaining columns, bottom row: Histograms of the estimates from Regr-Mean, CBMMd, Pred-Prob-Mean, and SST-Mean.

to be only slightly lower compared to predictions from regression models. Beside this main finding, the following aspects are revealed:

Lexicon-based approaches do not perform very well. The predicted sentiments are centered in the middle of the sentiment value range and changes in a document's sentiment are not strongly reflected in changes in the sentiment values predicted by the lexicons. (As an example see the most right column of Figure 2.) Consequently, the lexicon generated sentiment estimates exhibit relatively low levels of correlation with the true sentiment values. Especially the case of the SST lexicon for the SST data shows that it is not sufficient to have a lexicon that has a coverage of 100% and is perfectly tailored to the context it is applied to. In order to get valid sentiment estimates, one requires an aggregation procedure that accounts for the complex building up of sentiment in texts.

Regression Approaches. The continuous sentiment predictions generated by regression approaches tend to have the smallest distances to and the highest correlations with the true sentiment scores. Hence, the results demonstrate that *if* one has detailed training annotations available that can be treated as if they were continuous, regression approaches constitute an effective way to bring sentiment estimates as close as possible to the true sentiment values.

Across applications, the estimates obtained from Regr-Mean, Regr-LMM, and Regr-LMMd are highly similar. The reason is that the variance for the document-specific intercepts, τ_θ^2, is high relative to the error variance σ^2, and the classifier-

specific variance τ^2_γ.[9] Thus, the LMM estimator is close to a fully unpooled solution in which a separate model for each document is estimated (Fahrmeir et al., 2013, p. 355-356). The sentiment predictions from Regr-LMM are therefore highly correlated with Regr-Mean that computes a separate mean for each document. Furthermore, adding a dispersion formula does not strongly affect the predictions from Regr-LMM.

Pred-Prob-Mean leads to acceptable results. Yet, the estimates from Pred-Prob-Mean still strongly mirror the binary coding structure (see the fourth column of Figure 2). Moreover, MAE tends to decrease and r tends to increase further if the predicted probabilities are aggregated via beta mixed models in CBMM.

CBMM produces continuous sentiment estimates that exhibit performance levels that are relatively close to those of the regression-based procedures. When considering the MAE and r, CBMMd tends to slightly outperform CBMM. As the predicted probabilities across all four data sets are characterized by a high degree of heteroskedasticity[10] additionally accounting for heteroskedasticity via the dispersion formula thus tends to further improve the estimates.

Interestingly, across the three approaches based on predicted probabilities (Pred-Prob-Mean, CBMM, CBMMd) Spearman's ρ nearly remains unchanged. This implies that the predicted order of documents on the latent sentiment variable is largely determined by the predicted probabilities from the ensemble of classifiers. Thus, whilst Pred-Prob-Mean, CBMM and CBMMd operate on the same order of documents,[11] it is the aggregation of the predicted probabilities by a beta mixed model—and the accounting for heteroskedasticity—that enables CBMM and CBMMd to alter the distances between the documents' positions on the sentiment variable such that the distribution of true sentiment values can be approximated more closely. (Compare the histograms of the values predicted by CBMMd and Pred-Prob-Mean in Figure 2.)

[9] Yet, across all evaluated data sets, a Restricted Likelihood-Ratio-Test (based on the approximation presented by Scheipl et al. (2008) as implemented in the RLRsim R-package) testing the null hypothesis that $\tau^2_\gamma = 0$, reveals that this null hypothesis can be rejected at a significance level of 0.01.

[10] To assess heteroskedasticity, Breusch-Pagan Tests (Breusch and Pagan, 1979) are conducted. For all applications and tested linear models, the Breusch-Pagan Test suggests that the null hypothesis of homoskedasticity can be rejected at a significance level of 0.01.

[11] Spearman's ρ between the estimates from Pred-Prob-Mean and CBMMd equals 0.999 across all applications.

5 Conclusion

This work introduced CBMM—a *c*lassifier-based *b*eta *m*ixed *m*odeling technique that generates continuous estimates for texts by estimating a beta mixed model based on predicted probabilities from a set of classifiers. CBMM's central contribution is that it produces continuous output based on binary training input, thereby dispensing the requirement of regression approaches to have (possibly prohibitively costly to create) fine-grained training data. Evaluation results demonstrate that CBMM's continuous estimates perform well and are not far from regression predictions.

CBMM here is applied in the context of sentiment analysis. Yet, it can be applied to any context in which the aim is to have continuous predictions but the resources only allow for creating binary training annotations.

References

Dolores Albarracin, Aashna Sunderrajan, Sophie Lohmann, Man pui Sally Chan, and Duo Jiang. 2019. The psychology of attitudes, motivation, and persuasion. In Dolores Albarracin and Blair T. Johnson, editors, *The Handbook of Attitudes*, pages 3–44. Routledge, New York, NY.

Mahzarin R. Banaji and Larisa Heiphetz. 2010. Attitudes. In Susan T. Fiske, Daniel T. Gilbert, and Gardner Lindzey, editors, *Handbook of Social Psychology*, pages 348–388. John Wiley & Sons, New York, NY.

Trevor S. Breusch and Adrian R. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler, and Benjamin M. Bolker. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

John T. Cacioppo, Wendi L. Gardner, and Gary G. Berntson. 1997. Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, 1(1):3–25.

Brian Cheang, Bailey Wei, David Kogan, Howey Qiu, and Masud Ahmed. 2020. Language representation models for fine-grained sentiment classification. *arXiv preprint*. arXiv:2005.13619v1 [cs.CL].

Francisco Cribari-Neto and Achim Zeileis. 2010. Beta regression in R. *Journal of Statistical Software*, 34(2):1–24.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Leandre R. Fabringar, Tara K. MacDonald, and Duane T. Wegener. 2019. The origins and structure of attitudes. In Dolores Albarracin and Blair T. Johnson, editors, *The Handbook of Attitudes*, pages 109–157. Routledge, New York, NY.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. 2013. *Regression*. Springer-Verlag, Berlin.

Silvia Ferrari and Francisco Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2016. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Sage, Newbury Park, California.

Yoon Kim. 2014. 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. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage Publications, Thousand Oaks.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Bing Liu. 2015. *Sentiment Analysis*. Cambridge University Press, New York.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*. arXiv:1907.11692v1 [cs.CL].

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 378–382, Borovets, Bulgaria.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2013. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 4(1):106–115.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, Michigan, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035. Curran Associates, Inc.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Fabian Scheipl, Sonja Greven, and Helmut Küchenhoff. 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52(7):3283–3299.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

The American National Election Studies. 2015. *ANES 2008 Time Series Study*. Inter-University Consortium for Political and Social Research, Ann Arbor, MI. https://electionstudies.org/data-center/2008-time-series-study/.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Tun Thura Thet, Jin-Cheon Na, and Christopher S.G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848.

Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Jin Wang, Bo Peng, and Xuejie Zhang. 2018. Using a stacked residual LSTM model for sentiment intensity prediction. *Neurocomputing*, 322:93–101.

Kohei Watanabe. 2021. Latent Semantic Scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, 15(2):81–102.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint*. arXiv:1910.03771v5 [cs.CL].

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 649–657, Montreal, Canada. MIT Press.