applied sciences

MDPI

Article

Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning

Senait Gebremichael Tesfagergish [1], Jurgita Kapočiūtė-Dzikienė [2] and Robertas Damaševičius [1,2,*]

[1] Department of Software Engineering, Kaunas University of Technology, 51368 Kaunas, Lithuania
[2] Department of Applied Informatics, Vytautas Magnus University, 44404 Kaunas, Lithuania
* Correspondence: robertas.damasevicius@ktu.lt

Abstract: We live in a digitized era where our daily life depends on using online resources. Businesses consider the opinions of their customers, while people rely on the reviews/comments of other users before buying specific products or services. These reviews/comments are usually provided in the non-normative natural language within different contexts and domains (in social media, forums, news, blogs, etc.). Sentiment classification plays an important role in analyzing such texts collected from users by assigning positive, negative, and sometimes neutral sentiment values to each of them. Moreover, these texts typically contain many expressed or hidden emotions (such as happiness, sadness, etc.) that could contribute significantly to identifying sentiments. We address the emotion detection problem as part of the sentiment analysis task and propose a two-stage emotion detection methodology. The first stage is the unsupervised zero-shot learning model based on a sentence transformer returning the probabilities for subsets of 34 emotions (anger, sadness, disgust, fear, joy, happiness, admiration, affection, anguish, caution, confusion, desire, disappointment, attraction, envy, excitement, grief, hope, horror, joy, love, loneliness, pleasure, fear, generosity, rage, relief, satisfaction, sorrow, wonder, sympathy, shame, terror, and panic). The output of the zero-shot model is used as an input for the second stage, which trains the machine learning classifier on the sentiment labels in a supervised manner using ensemble learning. The proposed hybrid semi-supervised method achieves the highest accuracy of 87.3% on the English SemEval 2017 dataset.

Keywords: sentiment analysis; emotion detection; sentence transformers; zero-shot model; ensemble learning; natural language processing

Citation: Tesfagergish, S.G.; Kapočiūtė-Dzikienė, J.; Damaševičius, R. Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning. *Appl. Sci.* **2022**, *12*, 8662. https://doi.org/10.3390/app12178662

Academic Editor: Valentino Santucci

Received: 31 July 2022
Accepted: 27 August 2022
Published: 29 August 2022

1. Introduction

For many years, humans have had to adjust their communication style to be 'understood' by computers, but communication in natural language has recently become a new trend. Huge amounts of texts available online are in the unstructured/unannotated form and therefore do not have much value. Such noisy data can be converted into useful information only after proper processing. However, manual processing is a cumbersome and time-consuming process. In contrast, the automatic techniques can help save manual labor, get the result faster, filter through huge amounts of unnecessary data to find appropriate material, and deliver the machine output in the desired format [1]. Natural language processing (NLP) tackles language technology problems by employing Artificial Intelligence (AI) methods for intelligent human–machine interaction. The AI technologies that use data mining, pattern recognition, and NLP, the computer can mimic the way the human brain works. NLP applications, such as machine translation systems, web search engines, natural language assistants, and opinion analysis, are resolving societal problems [2].

Today, the mood (sentiments, emotions) of texts is as important as their content [3]. Sentiment and emotion detection plays a crucial role in analyzing social moods [4,5]. Explosive social media growth enables users to share their opinions more and more and leave feedback online; this, in turn, makes Sentiment analysis become a powerful NLP

tool able to analyze these texts automatically, helping companies/service providers to respond quickly and adequately [6]. However, it can also be misused for spreading disinformation and hate speech, which can be detected automatically using sentiment analysis methods [7–10]. Moreover, sentiment analysis even allows tracking of trends in real time by monitoring the popularity of products/services and even political candidates; and forecasting voting results [11,12].

Here we focus on automatic sentiment analysis as one of the popular NLP problems that have found many areas of applications by analyzing customers' product reviews, social media, survey unstructured responses, etc. The goal of sentiment analysis is to analyze, generalize, and predict whether the text contains subjectivity and expresses the sentiment, apart from which sentiment is the dominant [13,14]. Most sentiment analysis studies focus on assigning positive, negative, and sometimes neutral sentiment values to the given text. A less studied direction in sentiment analysis is to move the aim from analyzing sentiment toward a specific item to the internal mood of the text itself [15]. Zero-shot cross-lingual experiments present the evaluation of monolingual models applied to another language [16–19]. Most models function well with commonly used languages such as English; however, applying these algorithms straight to low-quality corpora frequently yields disappointing results [20]. Cross-lingual sentiment analysis aims to employ high-quality and rich resources in English to improve the classification performance of resource-scarce languages [21]. These methods address the problem of training separate models for each language, but despite that, they lack mechanisms able to modify traditional training methods (classifiers) for the Sentiment analysis task. To our current knowledge, emotion detection and sentiment classification are two different tasks that are often solved independently [22]. However, in this paper, we assume that the sentiment analysis problem can be tackled more effectively if we rely on emotion detection first.

Many accurate pre-trained models are already available for resource-rich languages (see Table 1, including for the sentiment analysis tasks); therefore, machine translation of less-resourced languages into rich-resourced has also been investigated. Improvements in statistical or neural machine translation systems eliminate the need to create separate monolingual Sentiment analysis models for separate languages [23–32]. For example, in [33], neural machine translation is used to convert the multilingual data set into English which is later classified as the English language model. If the source and target languages, computational expenses can be minimized by applying word-to-word translation [34]. Nevertheless, the importance of machine translation cannot be overestimated: the machine translation accuracy of non-normative language texts is not always reliable and can introduce even more noise.

Table 1. Summary of pre-trained NLP models.

Method	Benefit	Solution	Complexity
BERT	Pre-trained model in more than 100 languages and it can be tuned by adding one output layer	Question answering, Abstract summarization	Learns contextual relation between words in a sentence/text
XLM-R	Trained in around 100 different languages	Cross-lingual transfer tasks	It does not require lang tensor to understand which language is used, and should be able to determine the correct language from input ids
ELMo	Improves functions across vast NLP tasks	Answer questions, Textual entailment, Sentiment Analysis	Pre-trained on a huge text-corpus and learned functions from deep bi-directional models (biLM)
XLNet	Unlike BERT it does not need to undergo pre-train fine tuning.	Sentiment Analysis, Question answering, Text classification	Large bidirectional transformer with improved training methodology in terms of large amount data and more computational power to achieve better than BERT prediction
Zero-Shot classification	No training data needed	Text Classification	It classifies objects to a different label that the classifier has not been trained on.

We present the following contributions to the research field.

- The zero-shot model detects emotions first, and later they are used to assign positive, negative, and neutral sentiments. Such a method gradually decreases the dimensionality starting from the high-dimensional sentence transformer input (i.e., vectorized text) mapped into probability values of different emotions; probability values are further mapped into the sentiment labels.
- The second-stage input does not require complicated feature extraction or sophisticated machine learning methods able to catch sentiments directly from the text, which, in turn, speeds up the whole sentiment analysis process.
- The performance of the proposed method is evaluated on three benchmark datasets (IMDB, Sentiment140, and SemEval-2017) and using multiple classifiers, including machine learning, neural network, and ensemble learning.
- The proposed emotion-sentiment detection model requires fewer training data compared to traditional Sentiment analysis detection.

This paper is divided into five more sections. In Section 2, we present the related work of existing solutions. The methods used in this experiment are described in Section 3. In Section 4, we present our experiment results and discuss the results obtained in Section 5. Section 6 summarizes, concludes our work, and provides our thoughts on possible future research directions.

2. Related Work

Sentiment analysis is among the principal tasks of NLP that strives to predict opinion polarity. It often predicts the sentiment as belonging to one of the three categories (negative, neutral and positive) that can be used in many areas such as customer product review [35], political forecasting [11], telehealth services [36], finance [37], etc. [38].

According to Medhat et al. [39], we describe the taxonomy of sentiment analysis techniques and divide it into two main paradigms: rule-/lexicon-based and machine learning. Lexicon-based methods [40] rely on the assumption that the overall sentiment depends on the words that explicitly express these sentiments. Words (adjectives, adverbs, sometimes verbs, and nouns) that define different sentiments are searched in the text and counted: the overall sentiment of the text depends on the majority. In machine learning, the sentiment analysis task is typically formulated as the text classification task and, therefore, can be solved with a whole spectrum of methods for this purpose: traditional machine learning methods (e.g., Support Vector Machine (SVM)), deep learning methods (e.g., Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN)), or innovative sentence transformer models (e.g., Bidirectional Encoder Representation from Transformer (BERT)).

For many years different languages were treated separately by training monolingual models from scratch for separate tasks and separate languages [41,42]. Recently, many pre-trained models provided by open-source NLP libraries, such as BERT and NTLK, were introduced to minimize the efforts and resources required to learn general knowledge about the language and its structure (i.e., existing words, their meanings, and similarities). These transformer models are typically trained on very large monolingual or multilingual unannotated corpora (i.e., on pure texts) in a self-supervised manner and therefore are not adjusted for specific NLP problems [43]. With the help of transfer learning, the previously acquired general knowledge in the pre-trained word- or sentence-transformer models can be augmented and fine-tuned to tackle the specific NLP problems (including the sentiment analysis task). Once the model is already 'familiar' with the language, it is much easier to adapt it to a specific NLP task: that is, typically, fewer training data are needed. Moreover, some multilingual transformer models are trained on the parallel corpora and tuned for similar tasks in the way they can cross barriers between languages. Cross-lingual methods have recently received more NLP community attention, thus demonstrating promising results when fine-tuning augmented transformer layers on different languages than they are later tested on (a good example is a group of the cross-lingual language model (XLM) transformers) [44].

Pre-trained transformer models can be used for sentiment analysis tasks in very different manners [45,46]: as text classifiers (by adding additional layers connecting the output of the transformer model with the sentiment labels); for the evaluation (by calculating distances between the unseen texts and texts of which the sentiments are already known); as zero-shot models that can evaluate the relatedness of some word (category, narrative, etc.) with the text. Zero-shot models can act as advanced dictionary-based methods that seek emotion or sentiment words both explicitly and implicitly and return probabilities determining how much these words are related to the text. The zero-shot models do not require the training data, but they are not directly adjusted for the sentiment analysis tasks and, therefore, may need additional mechanisms to go their limitations. In Table 2, we summarize the sentiment analysis methods that are the most influential in solving our problem.

Table 2. Summary of analyzed related works. The papers are compared according to the dataset, methods and results achieved.

Paper	Dataset	Methods	Results
Choi et al. [16]	STS benchmark (STSb), Korean (KorSTS), SemEval-2017 Spanish and SemEval-2017 Arabic	SLM RoBERTa (SLM-R) that extend semantic textual similarity (STS), Machine reading comprehension (MRC), Sentiment analysis, and Alignment of sentence embeddings under various cross-lingual settings.	86.38% when the Korean language-tuned model is evaluated using the English dataset.
Pelicon et al. [15]	Slovene: SentiNews dataset and Croatian dataset	Multilingual BERT model for 3-class sentiment classification	The Slovene language-trained model achieved the precision of 59.00 $\pm$ 1.62 and F1-score of 52.41 $\pm$ 2.58, when evaluated on the Croatian language dataset.
Phan et al. [17]	6 languages in Restaurant Domain: English, Russian, Dutch, Spanish, Turkish and French (SemEval 2016-Task 5)	Two main sub-tasks of aspect-based sentiment analysis task are aspect category detection, and opinion target expression using mBERT and XLM-R models	78.94% using the XLM-R English-trained model on the Dutch dataset.
Priban et al. [19]	Movie review dataset (CSFD) Facebook dataset (FB) and Product review dataset (Mallcz)	A binary classification task using BERT-based models (eight models, five of them are multilingual). In the cross-lingual experiment, they tested the ability of four multilingual models to transfer knowledge between English and Czech sentiment classification	91.61 $\pm$ 0.06% when trained on English and tested on Czech, and 93.98 $\pm$ 0.10% when trained on Czech and tested on English
Kumar et al. [18]	SemEval 2017 dataset Task 4 (3-class: Positive, Negative and Neutral) and two Hindi movie and Product reviews	Fine-tuned XLM-RoBERTa model	Cross lingual contextual word embedding and zero-shot transfer learning in projection prediction from resource-rich English to resource-poor Hindi language achieved 60.93% accuracy.
Liang et al. [13]	9 emotion labels: sadness, joy, anger, disgust, fear, surprise, shame, guilt and love.	Unsupervised lexicon-based learning. Top-K based: selects most representative words and designs a distance weighted word vector method to calculate similarity. Weight-based: gives more weight to emotional words and lower weight to noisy words	F1-score is 14.20 (Top-k based), and 16.30 (weigh-based)
Jebbara et al. [47]	SemEval 2016 Task-5. 5 languages: Dutch, English, Russian, Spanish and Turkish	Multilayer CNN for the sequence tagging model. Trained in one language and tested in another language that shares a common vector space.	F1-score for the zero-shot cross-language (from English to Spanish) learning from a single source to a target is 0.5
Sitaula et al. [48]	NepCOV19Tweets (3-class: positive, neutral and negative)	Ensemble CNN of three CNN models $CNN_{ft} = CNN_{fastText(X)}$ $CNN_{ds} = CNN_{domainSpecific(X)}$ $CNN_{da} = CNN_{domainAgnostic(X)}$	The ensemble of the three CNN models achieves the highest accuracy of 68.7%

3. Methodology

3.1. Outline

The proposed two-stage method combines unsupervised and supervised machine learning paradigms in one pipeline (Figure 1). The core of the first stage is the pre-trained zero-shot model, which is applied to (1) the emotion labels (see Table 3) and (2) the inputted text vectorized with the sentence transformer. The output of the zero-shot model is a list of emotion labels mapped to their probabilities for the input text. This output

becomes an input into the second stage (Figure 2): emotion probabilities are converted into a one-hot encoding format and then fed into the sentiment classifier trained to detect positive/negative/neutral (three-class classification scenario) or positive/negative (binary classification scenario) sentiments (see Section 3). For classification, we have used supervised machine learning methods, including neural networks and ensemble learning.

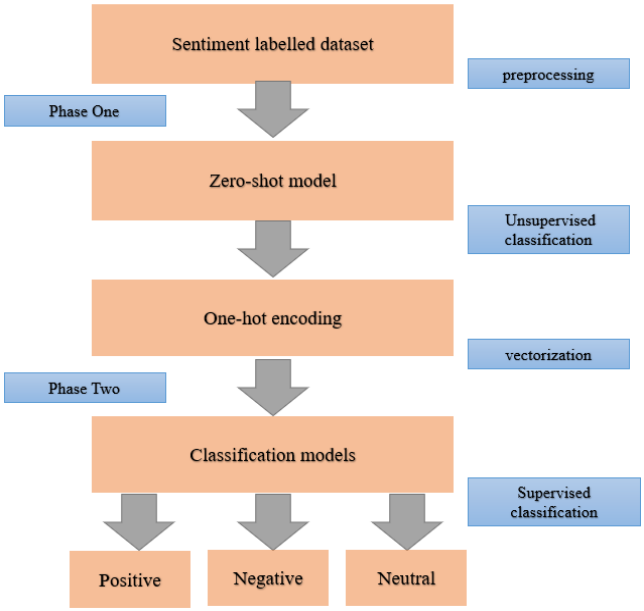

Figure 1. Workflow of the proposed semi-supervised method for sentiment classification. It includes two stages (phases), which combine unsupervised classification for recognizing emotions in a text, and supervised classification for detecting sentiments from emotions.

Table 3. Set of emotions used for zero-shot classification.

Emotion Sets	Emotions
First Set	Anger, sadness, disgust, fear, joy, happiness
Second Set	Admiration, affection, anguish, caution, confusion, desire, disappointment, attraction, envy, excitement
Third Set	Grief, hope, horror, joy, love, loneliness, pleasure, fear, generosity, pleasure
Fourth Set	Rage, relief, sadness, satisfaction, sorrow, wonder, sympathy, shame, terror, panic

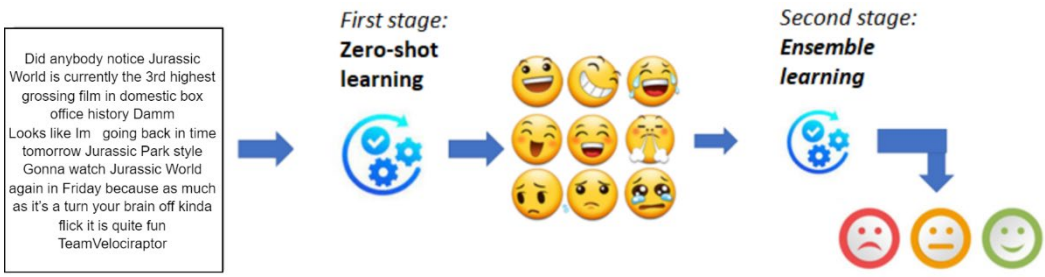

Figure 2. Two stages of the proposed semi-supervised method. The first stage uses unsupervised zero-shot learning by sentence transformers to obtain emotion probabilities. The second stage uses supervised ensemble learning to learn sentiments from emotion probabilities.

3.2. Emotions and Sentiments

The emotion models that define the categorization process are a crucial factor to consider for systems that recognize emotions. Although there are various ideas on how to portray emotions, two stand out as the most popular in the field of NLP: the Ekman's

fundamental emotions [49] and the Plutchik's wheel of emotions [50]. Six fundamental emotions are included in the Ekman model: surprise, sadness, happiness, fear, disgust, and anger. Four opposing pairs of axes make up the Plutchik's model, which uses a multidimensional representation method to characterize emotions as points along these axes (dimensions). The axis and intensity are what determine the emotions under this approach. These axis pairings include surprise-anticipation, trust-disgust, anger-fear, and joy-sadness. Other emotions can be produced from these emotions as a combination of other emotions and their intensities, as shown in Figure 3, which is an extraction of the Plutchik model. These axes and intensity are marked with colors in the concentric rings. Most studies on emotion detection only consider a limited selection of these feelings. In this paper, we have subdivided the entire set of emotions into four subsets, as outlined in Table 3. We use four sets of emotions, where each set consists of several taken from the emotions' wheel of emotions (Figure 3): anger, sadness, disgust, fear, joy, happiness, admiration, affection, anguish, caution, confusion, desire, disappointment, attraction, envy, excitement, grief, hope horror, joy, love loneliness, pleasure, fear, generosity, rage, relief, satisfaction, sorrow, wonder, sympathy, shame, terror and panic. Emotions are nonexclusive in the Plutchik's model as they are composable; there are also some correlations between them.

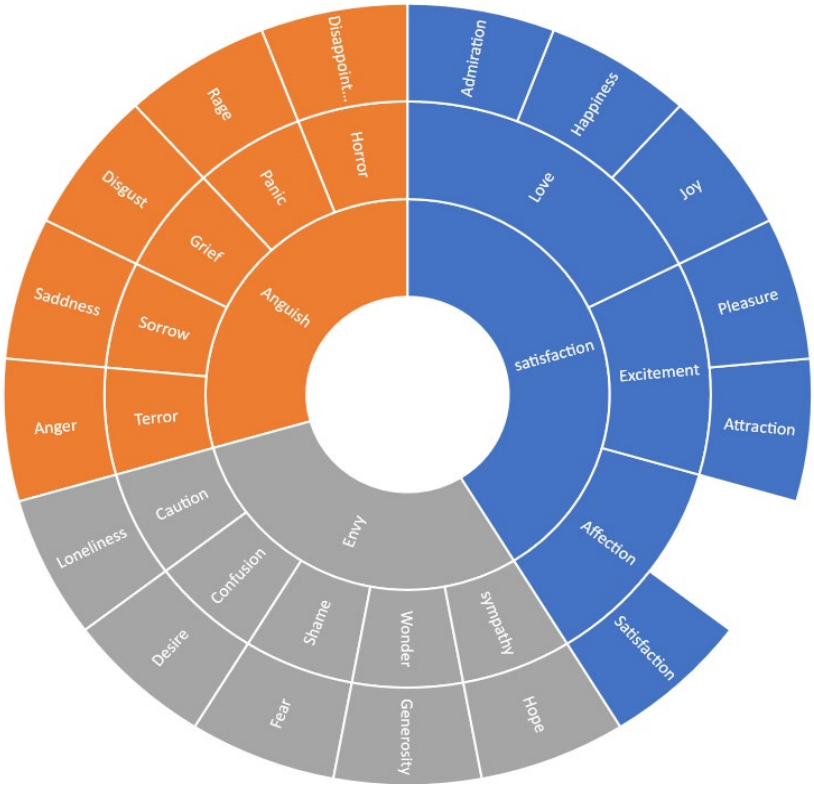

Figure 3. A representation of the wheel of emotions (the Plutchik's model).

3.3. Vectorization

Most machine learning methods perform mathematical calculations during training or testing and, therefore, cannot be directly applied to pure texts. Thus, vectorization becomes a crucial state in any NLP task. In our experiments, we have used the one-hot encoding vectorization technique. It is a discrete token representation method in which the length is equal to the size of the vocabulary. Each token is represented with a unique vector having all zero values except for one value equal to 1. This type of representation was used in the output of categorical data.

3.4. First Stage Zero-Shot Classifiers (Sentence Transformers)

The proposed method has two stages (see its schematic representation in Figure 3). In the first stage, the emotion detection problem is tackled (see Figure 4). The core of this stage is the zero-shot classifier that does not require any training. The idea is to transfer the knowledge it already has to a new task. Zero-shot learning involves training a classifier on a set of labels and then testing it in new data having different labels that the classifier has not been trained on. Classical zero-shot learning needs the provision of a descriptor for an unknown class for a model in order to predict that class without being trained on known representatives of it [51]. This machine learning method is based on a pre-trained model that can observe classes that were not observed during training and has a predictor of which class the input text belongs to. The zero-shot model returns probabilities for the given emotions and thus determines their relations to the input text.

Figure 4. Pipeline of the first stage classification.

In our experiments, we have tested four zero-shot transformer models as follows:

- The *bart-large-mnli* model [52] is a zero-shot sequence classifier proposed in [53]. The model was trained on tweets, emotional occurrences, fairy tales, and artificial sentences. It has nine emotions (anger, disgust, fear, guilt, joy, love, sadness, shame, surprise), as well as the "none" class (if no emotion applies). The approach offers the sequence to be categorized as the multi-genre natural language inference (MNLI) and creates a hypothesis from each possible label. Then, label probabilities are created from the entailment and contradiction probabilities.
- The *Fb-improved-zeroshot* model [54] is a zero-shot model for German and English academic searchlog classification created by ETH Zürich students and based on [53]. The *bart-large-mnli* model was used to train and then fine-tune this model.
- The COVID-Twitter-BERT (CT-BERT), a transformer-based model, is the foundation of the *covid-twitter-bert-v2-mnli* model [55], which was pre-trained on a corpus of Twitter conversations about COVID-19 [56]. CT-BERT was designed to work with the COVID-19 content, particularly from social media. The emotion toward vaccines is captured by the model. The dataset comprises three classes: positive (towards vaccinations), negative, and neutral/others.
- The *bart-large-mnli-yahoo-answers* model [57] refined the *bart-large-mnli* model on Yahoo Answers subject categorization. The model may be used to forecast whether the topic label can be assigned to a certain sequence.

3.5. Second Stage Machine Learning and Ensemble Learning Classifiers

The second stage uses the output of the first stage by transforming it into a one-hot encoding format. These feature vectors were then fed into the classifier in a supervised manner learning to predict positive, negative, and neutral sentiment labels. In our experiments, we used two types of classifiers.

3.5.1. Single-Model Machine Learning Classifiers

Traditional machine learning (as implemented in [58,59]) and deep learning [60–62] classifiers have already been applied to the sentiment analysis problem. Recently, deep

learning methods were combined with ensemble learning [63]. However, the main innovation of our study is that we classify the output of the zero-shot model rather than the vectorized text directly. Due to this reason, we cannot use a whole spectrum of deep learning models such as CNN, LSTM, etc. In our experiments, we have used and evaluated these classifiers described below:

- Feed-forward neural network (FFNN) is suitable for solving tasks as it can learn relationships between independent features. In addition, it is a simple and fast network learning how to adapt the weights of connections between units until the correct output is produced. In this paper, we have used this architecture because of its simplicity of feature selection. The architecture of the model we used in our experiment has one layer of 64 neurons, Rectified Linear Unit (ReLU) activation function in the hidden layer, and sigmoid activation function in the output layer. During training, we used accuracy metrics and Adam optimizer with binary cross-entropy loss.
- Linear regression (LR) is an algorithm used when you want to know how strong the relationship between two variables is and the value of a dependent variable at a certain value of the independent variables. The parameters of this classifier are set to their default values.
- K-nearest neighborhood (KNN). In KNN, similar class-type objects exist in closer proximity. KNN can be used for multiclass classification, and it is useful when the size of the labeled data is smaller. In our case, due to the small amount of data used for this experiment, we chose to test this method. The parameters of these classifiers were set into their default values.
- Support Vector Machine (SVM) is a supervised learning method that is used for classification, regression, and outlier detection. Default values were used in the parameters of this classifier.
- Naive Bayes (NB) predicts the probability of different classes based on several attributes. We use this algorithm because it is mostly used for text classification and multiple classes. We choose this classifier because it does not require much training data. We used the default values of its parameters in our experiment.
- Classifier and Regression Tree (CART). It is a decision tree algorithm used for the classification task. CART can capture non-linear relationships within the dataset, and there is no need for standardization of data when using this model. We used the default values for the parameters of this classifier.

3.5.2. Ensemble Learning Classifiers

Ensemble learning methods use multiple combined machine learning classifiers (instead of a single classifier) to achieve better predictive performance. Each of these methods is trained to solve the same problem, but their results are combined. In our experiments, we have used the following ensemble learning methods:

- Adaptive Boosting (AdaBoost) classifier re-assigns weights to each data sample, i.e., higher weights are assigned to wrongly classified data. AdaBoost is less likely to overfit because input parameters are not optimized jointly.
- AdaBoost regressor is a meta-estimator that, first, fits a regressor on the original dataset, and then it fits subsequent copies of the regressor while the weights of the instances are changed in accordance with the error of the most recent prediction.
- Bagging classifier is used to lower a variance within the noisy dataset. A bagging classifier fits base classifiers on randomly selected subsets of the dataset and then combines their predictions (by averaging or by voting) to get a prediction.
- Bagging regressor is a meta-estimator that fits base regressors to individual random subsets of the dataset and then combines each prediction to get the final prediction. By adding randomization to the process of building a black-box estimator (such as a decision tree), a meta-estimator lowers the variance of the estimator.
- Extremely Randomized Trees (ExtraTress) classifier is similar to Random Forest but has two key differences: it samples without replacement; in this case, bootstrap is

equal to False by default, and nodes are split based on random splits rather than best splits. The advantage of this estimator is its low variance.
- Histogram Gradient Boosting (HistGradientBoost) classifier buckets continuous feature values into discrete bins, and then it uses these bins to generate feature histograms during training. The histogram-based algorithm is very efficient in both memory consumption and training speed.
- Stacking classifier stacks several machine learning classifiers such as Random Forest Classifier, KNN, decision tree, SVM, NB, and Support Vector Regression.

3.6. Evaluation and Statistical Analysis of Performance

The tested methods were evaluated with the commonly used accuracy, precision, recall, and F-score metrics. With the null hypothesis that the medians of the two variables differ, we used the Wilcoxon rank-sum test with the null hypothesis indicator H and the significance level p-value to determine whether the performance differences between sentence transformers (used as a baseline) and the suggested method were statistically significant. We have used the Friedman test and the post hoc Nemenyi test to examine the effectiveness of various machine learning techniques. The Friedman test is a strong non-parametric statistical ranking test that does not require the assumption of normality. It has been used in various studies in the past to evaluate the effectiveness of machine learning techniques. All pairwise algorithm comparisons were performed using the non-parametric Nemenyi test, with a 0.05 significance level. The critical distance (CD) diagram [64] is used to represent the outcomes (mean rankings of compared methods).

4. Experiments and Results

4.1. Settings

Sentiment analysis is a text classification task, where given written text as an input, positive, neutral, or negative class is returned as the output. Here we perform the binary (2-class, positive and negative) and 3-class (positive, negative, and neutral) sentiment classification.

Our method was implemented using Tensorflow and Keras libraries with python programming language. Our experiments were executed with the datasets described in Section 4.2 and using the methods described in Sections 3.4 and 3.5. The results of the experiments are presented in Tables 4–9.

4.2. Datasets

We have used the following sentiment datasets (for detailed statistics, see Figure 5):

- IMDB [65] is the English dataset that has 50K movie reviews (with ~300 words per review on average) annotated with positive or negative labels. This dataset contains only highly polarized reviews (with a score of $\leq$4 of 10 for negative and $\geq$ 7 of 10 for positive). It is highly researched, with more than 1000 research papers using it. The task analyzed in this paper differs from the traditional text classification, and it does not require a large, annotated dataset. Therefore, we have randomly selected 5000 samples of positive and negative classes to create a new dataset used for our experiments.
- Sentiment140 [66] is an English dataset has 1.6 million tweets extracted using the Twitter API and annotated with two classes (positive and negative). For our experiments, we randomly selected a subset of 5000 texts for each class.
- SemEval-2017 [67] is the dataset (the English version) that was first presented in the scientific SemEval competitions. It has three classes (positive, neutral, and negative), but it is imbalanced. For binary classification and comparison, we have omitted the neutral class and tested with the positive and negative classes only.

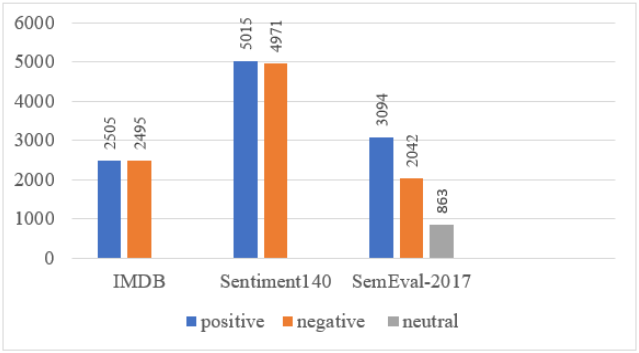

Figure 5. The distribution of texts among positive/negative/neutral sentiment categories in the IMDB, Sentiment 140, and SemEval-2017 datasets.

The Sentiment140 and SemEval-2017 datasets are retrieved from the Twitter social network, and they contain symbols of emojis and weblinks that were filtered out in the data pre-processing step.

4.3. Results

The results of experiments on zero-shot classification (first stage) are summarized in Table 4. We compared four zero-shot models (i.e., *bart-large-mnli*, *Fb_improved_zeroshot*, *covid-twitter-bert-v2-mnli*, and *bart-large-mnli-yahoo-answers*). The models were employed for zero-shot classification via a pipeline in the Hugging face's transformers package. The determined most accurate zero-shot model (i.e., *bart-large-mnli*), which gives the best performance of 0.747 on the Sentiment140 dataset with the single-model machine learning classifiers, was later used in our further experiments.

Table 4. The impact of zero-shot models on the accuracy of machine learning classifiers for the binary sentiment classification with the Sentiment140 dataset. The best result is shown in bold.

Method	*bart-large-mnli-yahoo-answers*	*bart-large-mnli*	*covid-twitter-bert-v2-mnli*	*Fb_improved_zeroshot*
Linear regression	0.727	0.740	0.670	0.693
KNN	0.650	**0.747**	0.740	0.663
CART	0.700	0.730	0.677	0.693
SVM	0.727	0.733	0.670	0.693
Naïve Bayes	0.513	0.723	0.680	0.523

Table 5 represents the accuracies of single model machine learning and ensemble classifiers with different sets of emotions (from Table 3) on the SemEval-2017 dataset using three-class classification. The best overall accuracy was achieved by the stacking classifier on the first set of emotions (0.627).

Table 5. Accuracy of classifiers on the SemEval-2017 dataset using three-class classification with different sets of emotions. The best result is shown in bold.

Classification Methodology	Method	First Set	Second Set	Third Set	Fourth Set
Single-model machine learning	FFNN	0.338	0.433	0.484	0.458
	Linear regression	0.611	0.546	0.575	0.516
	KNN	0.577	0.501	0.541	0.484
	SVM	0.611	0.546	0.575	0.516
	Naïve Bayes	0.555	0.538	0.575	0.520
	CART	0.611	0.544	0.574	0.516
Ensemble learning	AdaBoost Classifier	0.611	0.551	0.578	0.519
	AdaBoost regressor	0.292	0.256	0357	0.219
	Bagging classifier	0.611	0.551	0.578	0.519
	Bagging regressor	0.263	0.266	0.270	0.207
	ExtraTrees classifier	0.611	0.551	0.578	0.519
	HistGradientBoost classifier	0.611	0.551	0.578	0.519
	Stacking classifier	**0.627**	0.544	0.578	0.509

Table 6 shows the accuracy of single-model machine learning and ensemble classifiers on the SemEval-2017 dataset (of two-class classification without the neutral class) with different sets of emotions. The best overall accuracy was also achieved by the stacking classifier on the third set of emotions (0.873).

Table 6. Accuracy of classifiers on the SemEval-2017 dataset (of two-class classification without considering the neutral class) with different sets of emotions. The best result is shown in bold.

Classification Methodology	Method	First Set	Second Set	Third Set	Fourth Set
Single-model machine learning	FFNN	0.82	0.826	0.873	0.776
	Linear regression	0.845	0.801	0.863	0.790
	KNN	0.830	0.782	0.823	0.639
	SVM	0.845	0.801	0.863	0.790
	Naïve Bayes	0.845	0.801	0.854	0.790
	CART	0.845	0.801	0.863	0.790
Ensemble learning	AdaBoost classifier	0.844	0.800	0.863	0.790
	AdaBoost regressor	0.519	0.404	0.506	0.315
	Bagging classifier	0.844	0.800	0.863	0.790
	Bagging regressor	0.460	0.318	0.519	0.284
	ExtraTrees classifier	0.844	0.800	0.863	0.790
	HistGradientBoost classifier	0.844	0.800	0.863	0.790
	Stacking classifier	0.819	0.826	**0.873**	0.776

Table 7 compares the accuracy of single-model machine learning and ensemble classifiers on three analyzed datasets. The best overall accuracy was matched by the stacking classifier and FFNN on the SemEval-2017 dataset (without using the neutral class) (0.873).

Table 7. Accuracy of classifiers on three benchmark (IMDB, Sentiment140 and SemEval-2017) datasets. The best result is shown in bold.

Classification Methodology	Method	IMDB	Sentiment140	SemEval-2017 (*w/o* Neutral Class)
Single-model machine learning	FFNN	0.773	0.728	0.873
	Linear regression	0.767	0.715	0.863
	KNN	0.760	0.655	0.823
	SVM	0.767	0.715	0.863
	Naïve Bayes	0.766	0.715	0.854
	CART	0.767	0.715	0.863
Ensemble learning	AdaBoost classifier	0.767	0.714	0.863
	AdaBoost regressor	0.423	0.177	0.506
	Bagging classifier	0.767	0.714	0.863
	Bagging regressor	0.332	0.047	0.519
	ExtraTrees classifier	0.767	0.714	0.863
	HistGradientBoost classifier	0.767	0.714	0.863
	Stacking classifier	0.772	0.728	**0.873**

The experiment with single-model and ensemble learning methods shows the superiority of ensemble methods (see Tables 5 and 6). It is explainable: they combine the knowledge from several classifiers. The highest accuracy for both the binary and 3-class classification problems was achieved with the ensemble learning type methods 0.873 and 0.627, respectively, using the SemEval-2017 dataset.

The confusion matrix for the three-class classification case is presented in Figure 6. Note most common misclassifications occur between the "adjacent" classes, i.e., between neutral and negative sentiments and between neutral and positive sentiments.

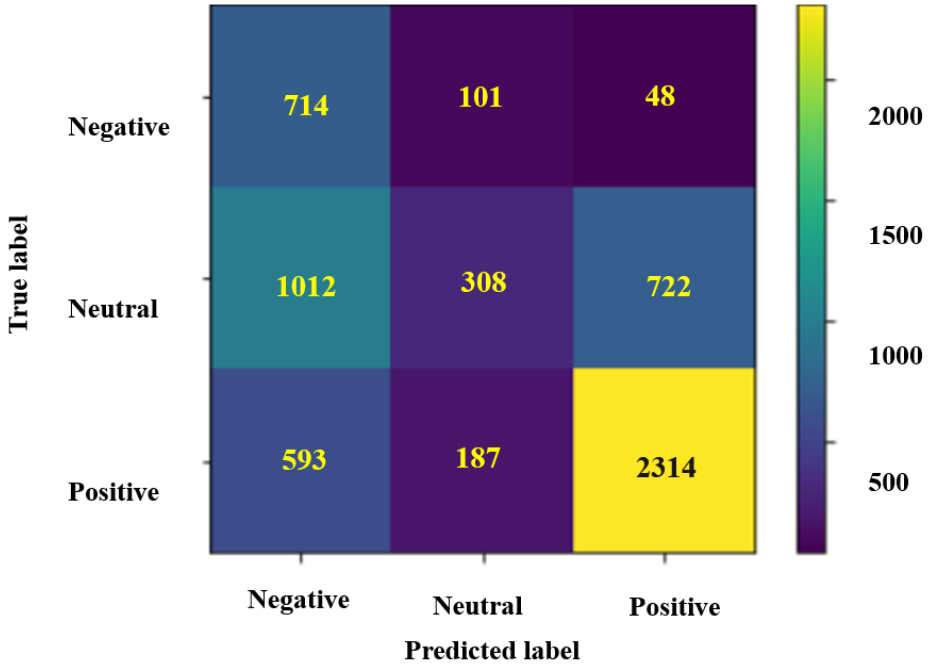

Figure 6. Confusion matrix of 3-class (negative, neutral and positive) classification.

Table 9 shows some of the examples of misclassifications. Note that many misclassifications may have occurred due to mislabeling of the original text in the dataset.

Table 8. Performance result comparison for binary and 3-class classification.

	Precision	Recall	F1-Score	Accuracy
Binary classification	0.863	0.908	0.884	0.873
3-class classification	0.562	0.627	0.554	0.627

Table 9. Misclassified instances and their probability score for the binary classification of the SemEval-2017 dataset and ensemble learning method.

Text	Score	Labels	Predicted	True Class
Did anybody notice Jurassic World is currently the 3rd highest grossing film in domestic box office history Damm	0.082	Fear	Positive	Negative
Looks like Im going back in time tomorrow Jurassic Park style	0.3746	Fear	Positive	Negative
Gonna watch Jurassic World again in Friday because as much as it's a turn your brain off kinda flick it is quite fun TeamVelociraptor	0.9961	Pleasure	Positive	Positive
Justin is lost in the 1st minute No experience	0.7968	Horror	Negative	Negative

4.4. Ablation Study

To compare the result of the traditional sentiment analysis classification task and our proposed method, we perform an experiment on the SemEval-2017 (without neutral class) dataset using sentence transformer and single-model machine learning classifiers.

We have analyzed different sizes of training datasets (from 100 to 1000 samples, see the vertical axis in Figure 7) with the fixed-size testing set using 500 instances. The result shows that our proposed method can achieve almost the same and even better in most cases than sentence transformers with only a small dataset required for training.

Figure 7. Accuracy vs. number of training instances for sentence transformer + machine learning classifiers and our proposed method.

4.5. Statistical Analysis

We have analyzed the results statistically to compare our approach with the result achieved using sentence transformers (Figure 8). We used the ranking-based non-parametric Wilcoxon test. The improvement in accuracy was statistically significant for decision Tree ($p < 0.001$), FFNN ($p < 0.001$), KNN ($p < 0.01$), and Random Forest ($p < 0.001$) classifiers, however, there was no significant difference for Log regression and Naïve Bayers classifiers. The results of the Wilcoxon test show that the performance of the sentence transformers and the proposed two-stage semi-supervised methodology are statistically different.

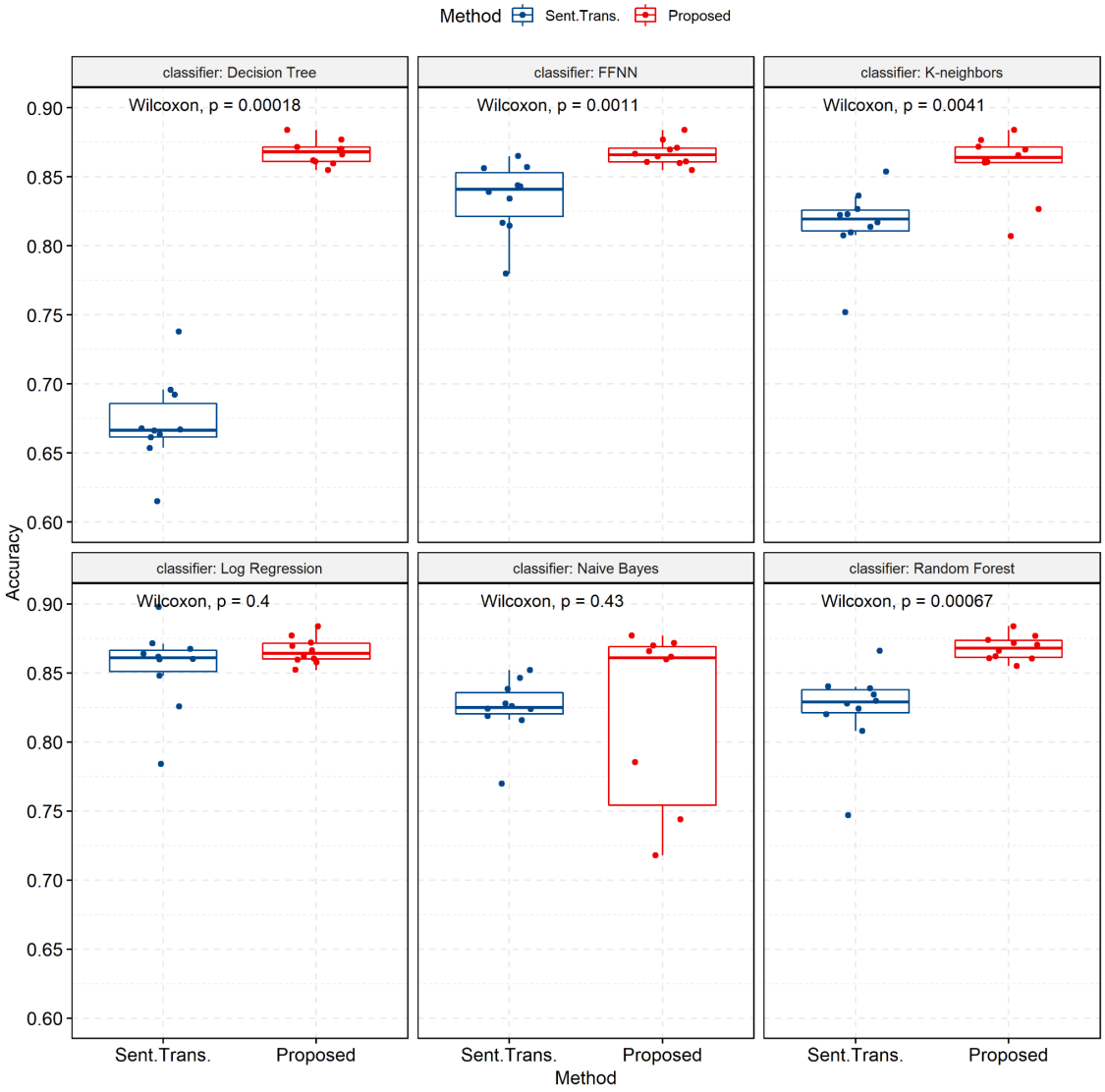

Figure 8. Results of statistical significance testing using non-parametric Wilcoxon test. Boxplots show an accuracy of classification using sentence transformers (Sent. Trans.) and the proposed method.

Figure 9 shows the critical distance diagram from the post hoc Nemenyi test for the two-class and three-class classification scenarios. The best performance across four emotion subsets was demonstrated by FFNN (the mean rank is 1.33) and Histogram Gradient Boosting classifier (the mean rank is 2.88), although the performance of other machine learning classifiers (excluding Bagging regressor and AdaBoost regressor) was not significantly different (within a critical distance of 10.534 for the 2-class classification scenario, and within a critical distance of 9.123 for the 3-class classification scenario).

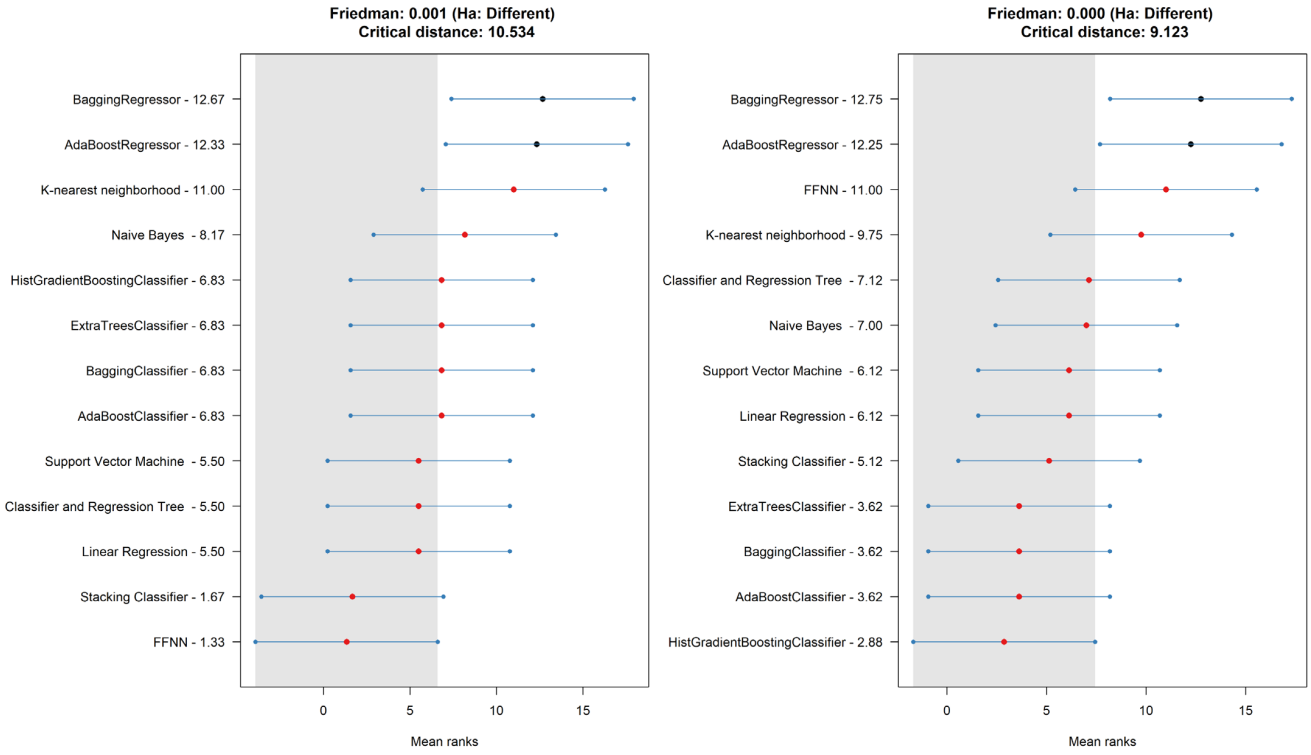

Figure 9. Critical Distance diagram from post hoc Nemenyi test: (**a**) 2-class classification scenario, and (**b**) 3-class classification scenario showing mean ranks of the methods. The mean ranks falling within a grey box marking the critical distance are not statistically different. Red dot marks the mean rank. Blue dots mark the confidence interval.

5. Discussion

The previous studies (see a discussion in Section 2) have demonstrated that multilingual pre-trained transformer models can be adjusted for the sentiment analysis problem. These multilingual transformer models already store semantics about the languages they support, thus decreasing the need for very large, supervised training data. However, text sentiments (positive, negative, neutral) often depend not only on the text content but also on different emotions (joy, sadness, anger, etc.) that are often mixed and ambiguous.

In this study, we assume that sentiment labels are easier to determine if we already know exactly what emotion the text represents. Due to this reason, we are solving the two-staged sentiment analysis problem by detecting emotions in the first stage and, based on it, detecting the exact sentiments. Our experimental investigation proves that such a methodology is effective. When detecting emotions, we rely on the zero-shot classification method that does not require any training, but it can return the probabilities of emotions for the input text. These probabilities represent the strength/impact of the detected emotions in the text. Later, we map these emotion probabilities into hot encoding vectors by strengthening the impact of dominated emotion. During the second stage, we train machine learning classifiers (single-model or ensemble) with the training data using one-hot encodings as feature vectors. Thus, our method to solve the sentiment analysis problem is very different from the typical solutions (see a review of methods described in [2,4,5]), relying on the textual content directly. However, by relying on the semantics kept in the zero-shot method and its ability to determine emotions, we reduce the need for larger training data (see Figure 7), which is important for resource-poor languages [68].

The proposed method can be further investigated and potentially improved by:

1. Applying a classification threshold. We have performed an error analysis of the misclassified instances, and most of them received the lowest probability score for certain

emotions (see Table 9). Correctly classified emotions have the highest probability score when classified using the zero-shot model. Therefore, setting a certain threshold for emotions can increase the accuracy of the model, then emotions with a lower score than the threshold might potentially be in the neutral class. In our classification, the highest misclassified class was the neutral class (see Figure 6), which can be confused either with a positive or a negative class.

2. Skipping one-hot vectorization. The current method transforms the outputs of the zero-shot method into one-hot encoding vectors used as features in the supervised training. We may expect possible improvement if, instead of determining one dominant emotion, we provide the whole spectrum of their influence (i.e., returned probabilities). Then the supervised machine learning model can be trained on the real values instead of binary (i.e., one-hot encoded) vectors.
3. Choosing more specific emotions. In our experiments, we tested four sets of emotions. The third set achieved the best result compared to all other tested sets (see Table 6). Using a larger set of emotions and a different split of emotions into subsets may allow for improving the result.

6. Conclusions

In this paper, we have addressed the binary (positive, negative) and three-class (positive, negative, neutral) sentiment analysis problem for the English language with three datasets used for evaluation. Our proposed method is completely different from how such tasks are usually solved. We formulate our sentiment analysis problem as a two-stage classification problem: the first stage determines emotions, and based on it, the second stage determines sentiments. The core of the first stage is the zero-shot transformer model, which does not require any training, and can extract probabilities of emotions for the given text. The second stage takes the zero-shot classification results, converts them into the one-hot encoding vector (used as features), and trains the supervised machine learning classifier.

In our experiments, we have investigated a large variety of different machine learning methods, i.e., traditional machine learning, deep learning, single-model, and ensemble methods. The best accuracy equal to 0.87 and 0.63 for the binary and three-class classification problems was achieved with the set of 10 and 6 emotions, respectively. We have determined that the best zero-shot model is *bart-large-mnli*, and the best classifier is based on ensemble learning (a stacking classifier of Random Forest, KNN, decision tree, SVM, Naïve Bayes, and Support Vector Regression). Compared with previous research in [13], our proposed method achieved an improvement of 44%. The performance of our method is stable (differences are insignificant), even having small training datasets.

Our proposed method reduces the effort of training the vectorizers to map the text into a real vector space and the need for a large training dataset. Due to its simplified structure, under-researched languages can benefit from our research findings. Most importantly, our research validates the application of emotion detection can help to detect the sentiment of a given text.

In the future, we will consider testing (1) all possible emotions; (2) domain-dependent ones. Theoretically, different emotions in different contexts and domains may lead to different sentiments. It would be interesting to test this idea experimentally.

Author Contributions: Conceptualization, J.K.-D. and R.D.; methodology, J.K.-D. and R.D.; software, S.G.T.; validation, S.G.T., J.K.-D. and R.D.; formal analysis, S.G.T., J.K.-D. and R.D.; investigation, S.G.T., J.K.-D. and R.D.; data curation, S.G.T.; writing—original draft preparation, S.G.T.; writing—review and editing, J.K.-D. and R.D.; visualization, S.G.T. and R.D.; supervision, J.K.-D. and R.D.; funding acquisition, J.K.-D. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The IMDB dataset is available at https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (accessed on 28 July 2022). The Sentiment140 dataset is available at https://www.kaggle.com/datasets/kazanova/sentiment140 (accessed on 28 July 2022). The English SemEval 2017 dataset is available at https://github.com/cbaziotis/datastories-semeval2017-task4 (accessed on 28 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sagnika, S.; Pattanaik, A.; Mishra, B.S.P.; Meher, S.K. A review on multi-lingual sentiment analysis by machine learning methods. *J. Eng. Sci. Technol. Rev.* **2020**, *13*, 154–166. [CrossRef]
2. Acheampong, F.A.; Nunoo-Mensah, H.; Chen, W. Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artif. Intell. Rev.* **2021**, *54*, 5789–5829. [CrossRef]
3. Kanclerz, K.; Milkowski, P.; Kocon, J. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Comput. Sci.* **2020**, *176*, 128–137. [CrossRef]
4. Mutanov, G.; Karyukin, V.; Mamykova, Z. Multi-class sentiment analysis of social media data with machine learning algorithms. *Comput. Mater. Contin.* **2021**, *69*, 913–930. [CrossRef]
5. Krishnan, H.; Elayidom, M.S.; Santhanakrishnan, T. A comprehensive survey on sentiment analysis in twitter data. *Int. J. Distrib. Syst. Technol.* **2022**, *13*, 52. [CrossRef]
6. Kilimci, Z.H.; Omurca, S.I. Extended feature spaces based classifier ensembles for sentiment analysis of short texts. *Inf. Technol. Control.* **2018**, *47*, 457–470. [CrossRef]
7. Alonso, M.A.; Vilares, D.; Gómez-Rodríguez, C.; Vilares, J. Sentiment analysis for fake news detection. *Electronics* **2021**, *10*, 1348. [CrossRef]
8. Aldjanabi, W.; Dahou, A.; Al-Qaness, M.A.A.; Elaziz, M.A.; Helmi, A.M.; Damaševičius, R. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. *Informatics* **2021**, *8*, 69. [CrossRef]
9. Karayiğit, H.; Akdagli, A.; Aci, Ç.İ. Homophobic and hate speech detection using multilingual-BERT model on turkish social media. *Inf. Technol. Control.* **2022**, *51*, 356–375. [CrossRef]
10. Tesfagergish, S.G.; Damaševičius, R.; Kapočiūtė-Dzikienė, J. Deep fake recognition in tweets using text augmentation, word embeddings and deep learning. In Proceedings of the 21st International Conference on Computational Science and Its Applications, ICCSA 2021, Cagliari, Italy, 13–16 September 2021; Part VI. pp. 523–553. [CrossRef]
11. Anstead, N.; O'Loughlin, B. Social media analysis and public opinion: The 2010 UK general election. *J. Comput. Mediat. Commun.* **2015**, *20*, 204–220. [CrossRef]
12. Lampert, J.; Lampert, C.H. Overcoming rare-language discrimination in multi-lingual sentiment analysis. In Proceedings of the 2021 IEEE International Conference on Big Data, Big Data 2021, Orlando, FL, USA, 15–18 December 2021; pp. 5185–5192.
13. Liang, M.; Zhou, J.; Sun, Y.; He, L. Working with few samples: Methods that help analyze social attitude and personal emotion. In Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021, Dalian, China, 5–7 May 2021; pp. 1135–1140. [CrossRef]
14. Nazir, A.; Rao, Y.; Wu, L.; Sun, L. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 845–863. [CrossRef]
15. Pelicon, A.; Pranjić, M.; Miljković, D.; Škrlj, B.; Pollak, S. Zero-shot learning for cross-lingual news sentiment classification. *Appl. Sci.* **2020**, *10*, 5993. [CrossRef]
16. Choi, H.; Kim, J.; Joe, S.; Min, S.; Gwon, Y. Analyzing zero-shot cross-lingual transfer in supervised NLP tasks. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2020; pp. 9608–9613. [CrossRef]
17. Phan, K.T.; Ngoc Hao, D.; Thin, D.V.; Luu-Thuy Nguyen, N. Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In Proceedings of the 2021 International Conference on Multimedia Analysis and Pattern Recognition, MAPR, Hanoi, Vietnam, 15–16 October 2021. [CrossRef]
18. Kumar, A.; Albuquerque, V.H.C. Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor indian language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2021**, *20*, 1–13. [CrossRef]
19. Pribán, P.; Steinberger, J. Are the multilingual models better? Improving czech sentiment with transformers. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP, Online, 1–3 September 2021; pp. 1138–1149. [CrossRef]
20. Musa, I.H.; Xu, K.; Liu, F.; Zamit, I.; Abro, W.A.; Qi, G. A cross-lingual sentiment topic model evolution over time. *Intell. Data Anal.* **2020**, *24*, 253–266. [CrossRef]
21. Wang, D.; Jing, B.; Lu, C.; Wu, J.; Liu, G.; Du, C.; Zhuang, F. Coarse alignment of topic and sentiment: A unified model for cross-lingual sentiment classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 736–747. [CrossRef]
22. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef]
23. Al-Saffar, A.; Awang, S.; Tao, H.; Omar, N.; Al-Saiagh, W.; Al-bared, M. Malay sentiment analysis based on combined classification approaches and senti-lexicon algorithm. *PLoS ONE* **2018**, *13*, e0194852. [CrossRef]

24. Balaguer, P.; Teixidó, I.; Vilaplana, J.; Mateo, J.; Rius, J.; Solsona, F. CatSent: A catalan sentiment analysis website. *Multimed. Tools Appl.* **2019**, *78*, 28137–28155. [CrossRef]
25. Smetanin, S. The applications of sentiment analysis for russian language texts: Current challenges and future perspectives. *IEEE Access* **2020**, *8*, 110693–110719. [CrossRef]
26. Osorio Angel, S.; Peña Pérez Negrón, A.; Espinoza-Valdez, A. Systematic literature review of sentiment analysis in the spanish language. *Data Technol. Appl.* **2021**, *55*, 461–479. [CrossRef]
27. Pota, M.; Ventura, M.; Catelli, R.; Esposito, M. An effective bert-based pipeline for twitter sentiment analysis: A case study in italian. *Sensors* **2021**, *21*, 133. [CrossRef] [PubMed]
28. Ranathunga, S.; Liyanage, I.U. Sentiment analysis of sinhala news comments. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2021**, *20*, 1–23. [CrossRef]
29. Kydros, D.; Argyropoulou, M.; Vrana, V. A content and sentiment analysis of greek tweets during the pandemic. *Sustainability* **2021**, *13*, 16150. [CrossRef]
30. Obiedat, R.; Al-Darras, D.; Alzaghoul, E.; Harfoushi, O. Arabic aspect-based sentiment analysis: A systematic literature review. *IEEE Access* **2021**, *9*, 152628–152645. [CrossRef]
31. Aydln, C.R.; Güngör, T. Sentiment analysis in Turkish: Supervised, semi-supervised, and unsupervised techniques. *Nat. Lang. Eng.* **2021**, *27*, 455–483. [CrossRef]
32. Khan, L.; Amjad, A.; Afaq, K.M.; Chang, H. Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. *Appl. Sci.* **2022**, *12*, 2694. [CrossRef]
33. Fujihira, K.; Horibe, N. Multilingual sentiment analysis for web text based on word to word translation. In Proceedings of the 2020 9th International Congress on Advanced Applied Informatics, IIAI-AAI, Kitakyushu, Japan, 1–15 September 2020; pp. 74–79. [CrossRef]
34. Baliyan, A.; Batra, A.; Singh, S.P. Multilingual sentiment analysis using RNN-LSTM and neural machine translation. In Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development, INDIACom, New Delhi, India, 17–19 March 2021; pp. 710–713. [CrossRef]
35. Ji, Z.; Pi, H.; Wei, W.; Xiong, B.; Wozniak, M.; Damasevicius, R. Recommendation based on review texts and social communities: A hybrid model. *IEEE Access* **2019**, *7*, 40416–40427. [CrossRef]
36. Omoregbe, N.A.I.; Ndaman, I.O.; Misra, S.; Abayomi-Alli, O.O.; Damaševičius, R. Text messaging-based medical diagnosis using natural language processing and fuzzy logic. *J. Healthc. Eng.* **2020**, *2020*, 8839524. [CrossRef]
37. Liapis, C.M.; Karanikola, A.; Kotsiantis, S. A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting. *Entropy* **2021**, *23*, 1603. [CrossRef]
38. Agüero-Torales, M.M.; Abreu Salas, J.I.; López-Herrera, A.G. Deep learning and multilingual sentiment analysis on social media data: An overview. *Appl. Soft Comput.* **2021**, *107*, 107373. [CrossRef]
39. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
40. Sattar, K.; Umer, Q.; Vasbieva, D.G.; Chung, S.; Latif, Z.; Lee, C. A multi-layer network for aspect-based cross-lingual sentiment classification. *IEEE Access* **2021**, *9*, 133961–133973. [CrossRef]
41. Kapočiūtė-Dzikienė, J.; Damaševičius, R.; Woźniak, M. Sentiment analysis of Lithuanian texts using traditional and deep learning approaches. *Computers* **2019**, *8*, 4. [CrossRef]
42. Kapočiūtė-Dzikienė, J.; Damaševičius, R.; Woźniak, M. *Sentiment Analysis of Lithuanian Texts using Deep Learning Methods. Proceedings of the ICIST 2018: Information and Software Technologies, Vilnius, Lithuania, 4–6 October 2018*; Communications in Computer and Information Science Book Series; Springer: Cham, Switzerland, 2018; Volume 920. [CrossRef]
43. Sarkar, A.; Reddy, S.; Iyengar, R.S. Zero-shot multilingual sentiment analysis using hierarchical attentive network and BERT. In Proceedings of the NLPIR 2019: 2019 the 3rd International Conference on Natural Language Processing and Information Retrieval, Tokushima, Japan, 28–30 June 2019; ACM International Conference Proceeding Series. pp. 49–56. [CrossRef]
44. Xu, Y.; Cao, H.; Du, W.; Wang, W. A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations. *Data Sci. Eng.* **2022**. [CrossRef]
45. Syed, A.A.; Gaol, F.L.; Matsuo, T. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access* **2021**, *9*, 13248–13265. [CrossRef]
46. Tiwari, D.; Nagpal, B. KEAHT: A Knowledge-Enriched Attention-Based Hybrid Transformer Model for Social Sentiment analysis. *New Gener. Comput.* **2022**, *11*, 1–38. [CrossRef]
47. Jebbara, S.; Cimiano, P. Zero-shot cross-lingual opinion target extraction. In Proceedings of the NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 2486–2495. [CrossRef]
48. Sitaula, C.; Basnet, A.; Maintali, A.; Shahi, T.B. Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets. *Comput. Intell. Neurosci.* **2021**, *2021*, 215884. [CrossRef]
49. Ekman, P. Basic Emotions. In *Handbook of Cognition and Emotion*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 1999; Chapter 3; pp. 45–60. [CrossRef]
50. Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*; Elsevier: Amsterdam, The Netherlands, 1980.

51. Romera-Paredes, B.; Torr, P.H.S. An embarrassingly simple approach to zero-shot learning. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 3, pp. 2142–2151.
52. Facebook/Bart-Large-Mnli Hugging Face. Available online: https://huggingface.co/facebook/bart-large-mnli (accessed on 26 March 2022).
53. Yin, W.; Hay, J.; Roth, D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP/IJCNLP), Hong Kong, China, 3–7 November 2019; Volume 1, pp. 3912–3921. [CrossRef]
54. Oigele/Fb_Improved_Zeroshot Hugging Face. Available online: https://huggingface.co/oigele/Fb_improved_zeroshot (accessed on 26 March 2022).
55. Digitalepidemiologylab/Covid-Twitter-Bert-V2-Mnli Hugging Face. Available online: https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2-mnli (accessed on 26 March 2022).
56. Müller, M.; Salathé, M.; Kummervold, P.E. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv* **2020**, arXiv:2005.07503.
57. Joeddav/Bart-Large-Mnli-Yahoo-Answers Hugging Face. Available online: https://huggingface.co/joeddav/bart-large-mnli-yahoo-answers (accessed on 26 March 2022).
58. Rosenthal, S.; Nakov, P.; Kiritchenko, S.; Mohammad, S.M.; Ritter, A.; Stoyanov, V. SemEval-2015 task 10: Sentiment analysis in Twitter. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), Denver, CO, USA, 4–5 June 2015; pp. 451–463.
59. Robnik-Šikonja, M.; Reba, K.; Mozetic, I. Cross-lingual transfer of sentiment classifiers. *Slovenscina 2.0* **2021**, *9*, 1–25. [CrossRef]
60. Peng, S.; Cao, L.; Zhou, Y.; Ouyang, Z.; Yang, A.; Li, X.; Yu, S. A survey on deep learning for textual emotion analysis in social networks. *Digit. Commun. Netw.* **2021**, in press. [CrossRef]
61. Sharma, T.; Kaur, K. Benchmarking deep learning methods for aspect level sentiment classification. *Appl. Sci.* **2021**, *11*, 542. [CrossRef]
62. Etaiwi, W.; Suleiman, D.; Awajan, A. Deep learning based techniques for sentiment analysis: A survey. *Informatica* **2021**, *45*, 89–95. [CrossRef]
63. Luo, S.; Gu, Y.; Yao, X.; Fan, W. Research on text sentiment analysis based on neural network and ensemble learning. *Rev. Intell. Artif.* **2021**, *35*, 63–70. [CrossRef]
64. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
65. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Christopher Potts, C. Learning Word Vectors for Sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011), Portland, OR, USA, 19–24 June 2011.
66. Go, A.; Bhayani, R.; Huang, L. *Twitter Sentiment Classification Using Distant Supervision*; CS224N Project Report; Stanford: Stanford, CA, USA, 2009.
67. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment analysis in Twitter. *arXiv* **2019**, arXiv:1912.00741.
68. Tesfagergish, S.G.; Kapočiūtė-Dzikienė, J. Part-of-speech tagging via deep neural networks for northern-ethiopic languages. *Inf. Technol. Control.* **2020**, *49*, 482–494. [CrossRef]