data

MDPI

Article

Stance Classification of Social Media Texts for Under-Resourced Scenarios in Social Sciences †

Victoria Yantseva [1,*] and Kostiantyn Kucher [2,3]

[1] Infolab, Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden
[2] Department of Computer Science and Media Technology, Linnaeus University, 351 95 Växjö, Sweden
[3] Department of Science and Technology, Linköping University, 602 33 Norrköping, Sweden
* Correspondence: victoria.yantseva@it.uu.se

† This paper is an extended version of our paper published in the Proceedings of the Swedish Workshop on Data Science (SweDS '21), Växjö, Sweden, December 2–3, 2021. *©2021 IEEE. Reprinted, with permission, from 2021 Swedish Workshop on Data Science (SweDS). V. Yantseva and K. Kucher. Machine Learning for Social Sciences: Stance Classification of User Messages on a Migrant-Critical Discussion Forum. doi:10.1109/SweDS53855.2021.9637718.*

Abstract: In this work, we explore the performance of supervised stance classification methods for social media texts in under-resourced languages and using limited amounts of labeled data. In particular, we focus specifically on the possibilities and limitations of the application of classic machine learning versus deep learning in social sciences. To achieve this goal, we use a training dataset of 5.7K messages posted on Flashback Forum, a Swedish discussion platform, further supplemented with the previously published ABSAbank-Imm annotated dataset, and evaluate the performance of various model parameters and configurations to achieve the best training results given the character of the data. Our experiments indicate that classic machine learning models achieve results that are on par or even outperform those of neural networks and, thus, could be given priority when considering machine learning approaches for similar knowledge domains, tasks, and data. At the same time, the modern pre-trained language models provide useful and convenient pipelines for obtaining vectorized data representations that can be combined with classic machine learning algorithms. We discuss the implications of their use in such scenarios and outline the directions for further research.

Keywords: text mining; machine learning; deep learning; neural networks; stance classification; computational social science; social media; supervised learning; sentiment classification; Swedish language data

Citation: Yantseva, V.; Kucher, K. Stance Classification of Social Media Texts for Under-Resourced Scenarios in Social Sciences. *Data* **2022**, *7*, 159. https://doi.org/10.3390/data7110159

Academic Editor: Kassim S. Mwitondi

Received: 15 July 2022
Accepted: 14 September 2022
Published: 13 November 2022

1. Introduction

Sentiment analysis and opinion mining [1] as text classification tasks have been an object of growing interest in sociology and social sciences due to their applicability in a wide range of analytical tasks, not to mention a growing researchers' interest in social media that provide rich data on users' behavior on the Internet. In social sciences and media studies, for instance, sentiment analysis and opinion mining tasks have been covered in the studies of political orientations and party support [2,3], public attitudes, and opinions on socially relevant issues [4–6], or extremism in online contexts [7]. Likewise, supervised classification methods have been used in social sciences for such topics as the detection of social media users' political orientations [8], identification of violent versus peaceful forms of protest [9], and measurement of cultural change [10]. Thus, the recent developments in computer science and machine learning present growing opportunities to perform this kind of task and allow improving the analytical toolbox traditionally used in social sciences.

In this study, we use a manually labeled dataset originating from Flashback Forum, a Swedish-language online platform that connects users discussing various topics, and, in particular, Swedish immigration and integration policies. The studies of users' behavior

and discussion content on the forum have previously been documented by [11–13] and pointed at oftentimes racist and migrant-critical perspectives on the immigration topic used by its users [13]. To our knowledge, very few previous works [14] used supervised classification of users' sentiments or stances on this forum, although such an analysis would complement the existing knowledge on anti-immigrant media platforms and activism in Sweden, and especially given social scientists' growing interest in this area. A difficulty lies in the absence of sufficient labeled data, as well as in its imbalanced character. Given these constraints, this work provides a contribution by applying machine learning (ML) methodology to real-world textual data in Swedish, as well as providing some guidance for social researchers interested in using machine learning on the pathways to achieve the best model performance given the constraints of real-world social media data in under-resourced languages, for example, for the issues of analysis of hate speech [15,16] or other types of abusive language [17,18]. Finally, the data labeled with the help of the model can be used in studies on users' polarisation, attitudes to immigration, and communication patterns on social media.

Summing up the above-mentioned, our goal with this paper is to demonstrate and evaluate our approach to supervised stance classification on the Flashback data using sparse and imbalanced manually labeled training data from social media in Swedish, which in itself can be described as an under-resourced language in comparison with, for instance, English. We compare the performance of several classic machine learning algorithms, artificial neural network architectures, and pre-trained modern language models and discuss possible solutions that allow achieving the highest classification performance given the character of the training data. Finally, we discuss concerns and opportunities for the use of supervised learning methods in sociology and social sciences, in general. Thus, the results of this work could be relevant for social scientists working on right-wing activism and discourses in social media, as well as for computer scientists working with social media data.

Please note: this article is an extended version of our workshop paper [19] [1]. This version includes further experiments (the third and fourth groups), qualitative evaluation of the results, further discussion, and additional materials (figures and tables).

2. Related Work

2.1. Data-Related Concerns for Text Mining Tasks

Advances in research and industrial applications of computational linguistics over the past decades have led to a variety of approaches and algorithms developed for the tasks of retrieval, processing, and generation of text data. One of the important developments in this regard is the involvement of statistical machine learning (ML) approaches [20,21] in contrast (or in addition) to traditional natural language processing (NLP) approaches, which often rely on custom manually engineered processing steps, rules, and features. The ML paradigm, instead, typically focuses on reducing the given problem to one of the typical ML tasks, such as classification or clustering, which can be achieved by transforming the input text data into a suitable representation, for instance, numerical vectors. Research on this topic has led to the invention of *distributional* and eventually *distributed* representations [22] such as *word embeddings* [23,24] and even sentences and document embeddings [25]. Artificial neural networks and deep networks, in particular, have been actively used for learning such representations and/or eventually for the respective "downstream" tasks such as text classification [26–29].

However, a typical supervised ML approach has a critical requirement of labeled training data of sufficient volume and quality. Deep learning approaches, in particular, can benefit from large training datasets to generalize well and adequately process unseen data when deployed. Collection, curation, annotation, and quality control of such labeled data are all time- and resource-intensive tasks. The proposed strategies for these issues include among others, crowdsourcing [30], active learning [31], distant/weak supervision [32,33], and transfer learning [34] involving existing pre-trained models, e.g., the state-of-the-art

BERT models [35]. All these strategies have their advantages and disadvantages, and we discuss some of them in the following in the context of our more specific task.

2.2. Sentiment and Stance Analysis

In the past decade, the evolution of technology and computational methods for the analysis of large-scale textual data has led to the development of a range of supervised and unsupervised approaches to text classification [36]. The latter, most often represented by lexicon-based approaches such as *VADER* [37] or *TextBlob* [38], have been praised for their relative simplicity and fast speed. However, due to the rather moderate classification accuracy, the preference is often given to supervised and ML-based approaches that can be characterized by higher training time and complexity, but also higher performance [39]. The popularity of supervised approaches has also been stimulated by the introduction of deep learning methods that have been shown to achieve almost error-free classification results [40].

The classification task of interest for the present work is to detect and categorize *subjectivity* in the given text data, e.g., a sentence, paragraph, or complete document. Although a variety of related and overlapping problems have been studied in traditional and computational linguistics in various terms, we can specifically mention *sentiment analysis or classification* [1,41], which is generally defined as the task of categorizing input text data as *negative*, *neutral*, or *positive* with regard to its emotional tone (*polarity* or *valence*). There exist many variations in the scope, expected output, and proposed computational approaches for this task. One task closely related to sentiment analysis is *stance analysis/classification*, which is often defined as the problem of deciding whether a person is in *favor* or *against* a given target (topic, entity, etc.) of interest [42]. The definition and operationalization of the concept of stance can be much broader, as it can involve further aspects of (inter-)subjectivity beyond *agreement/disagreement* and sentiment/emotions, for instance, *uncertainty* or *rudeness* [43,44]; however, we do not follow this broader definition in this work and focus on stance as sentiment/attitude expressed towards a specific topic.

Regarding the general quality and precision of existing supervised and unsupervised tools for sentiment analysis, quite naturally, supervised machine learning approaches have been found to outperform lexicon-based algorithms [45]. Among traditional machine learning methods, support vector machine (SVM) classifiers have often been found to perform best in text classification tasks [46]. More recent deep learning methods have been found to outperform traditional machine learning methods [47,48]. For instance, Lai et al. [49] have demonstrated impressive results for various text classification tasks, including sentiment classification, by combining recurrent (RNN) and convolutional neural network (CNN) architectures, and Chen and Ku [50] have applied a CNN for stance classification of social media data with up to 84% accuracy on online debate forum data in English. In the work of Karakus et al. [40], the use of pre-trained word embeddings together with CNN and long-short-term memory (LSTM) layers in a deep learning model was reported to achieve accuracy as high as 98%. Nevertheless, as demonstrated by Joulin et al. with *fastText* [51], simpler linear models can compete with deep learning approaches if they are used with the right features. Furthermore, deep learning models may not be available in all scenarios or could pose risks related to biases and miscalibration, for instance [52–54].

In relation to the previous statement, the choice of particular methods (such as for instance, traditional machine learning versus deep learning) has been found to play only a partial role in the final classification accuracy, while the nature of the data has been found to be a more determinant factor for model performance [55] (similar conclusions about the importance of data over model were reached, for example, by the prior work on the related task of hate speech detection [56]). Although the classification accuracy can reach 95% in state-of-the-art work, real-world classification problems, especially in cases with sparse annotated textual data, present certain methodological challenges and, thus, do not allow achieving the same performance. In particular, classification problems with three or more

classes, imbalanced datasets, and data coming from social networks have been named as factors that negatively influence the classifiers' performance [55]. Data sparsity and short document length have also been named as challenges for the successful sentiment analysis of social media data [57].

Another challenge related to sentiment classification tasks is the size and quality of annotated data. In particular, supervised methods for text classification require thousands or, preferably, tens of thousands of annotated documents to build classifiers that can successfully distinguish between the document categories. Outsourcing the annotation task to annotators on various platforms such as Amazon Mechanical Turk has been a popular solution; however, it has been shown that non-expert annotators cannot provide the same quality as professionals [30]. Another constraint in relation to the annotation task is the ambiguity of the social media messages, which causes substantial disagreement between human annotators and makes labeling even more problematic [58].

Indeed, while the recent results for sentiment analysis of English texts are impressive, they are not perfect and they also tend to rely on certain assumptions, such as for instance, text genres. The results from SemEval-2017, for example, demonstrate the difficulty of reaching high performance with real-world data, especially in the case of the newer and smaller datasets used [59]. In another work from 2017, that is concerned with texts in German, the benchmark solutions showed F1 scores between 0.44 and 0.65 for a three-class sentiment classification problem for the test data [60]. The application of supervised learning to languages beyond English, namely Greek (F1 score 0.77–0.80 for three-class data) [61], Russian (F1 score of 0.72 for five-class classification) [62], and Czech (F1 score of 0.69 for three-class data) [63], has also shown similar results. The importance of creating language resources and models for other languages—or developing multilingual models with comparable performance—is, thus, evident (for instance, Schmidt and Wiegand discuss this as an important challenge for the related task of hate speech detection [16]). The existing sentiment classification approaches for Swedish texts are discussed next.

2.3. Sentiment Analysis Approaches for Swedish Text Data

Sentiment classification in under-resourced languages presents further methodological challenges, often due to the limited availability of unsupervised classification tools and the absence of labeled training data or large text corpora, which makes users give preference to the approaches that can be created from scratch.

For the Swedish language, in particular, a review of language resources and natural language processing tools was carried out by Elenius et al. [64] some time ago. Some particular resources and studies for Swedish text data include the *Talbanken05* Swedish treebank with phrase structure and dependency annotations by Nivre et al. [65]; dependency parsing for Swedish text by Øvrelid and Nivre [66]; *Stagger*, the part-of-speech tagger for Swedish by Östling [67]; named entity recognition in short text messages in Swedish by Ek et al. [68] or Swedish clinical texts by Skeppstedt et al. [69]. To address data availability and multilingual support issues (including Nordic languages such as Swedish and Finnish), Lundberg et al. [70] have even foregone the involvement of text data as input for their classification of bots on Twitter.

With respect to resources and models directly applicable to sentiment and stance analysis of Swedish texts, a few lexicons, such as *SenSALDO* [71], or more complex algorithms, such as VADER [37] and the more recent BERT [35,72] model (as well as the version of *Sentence-BERT* [73,74] for Swedish [75]) have become available in recent years. The application of unsupervised sentiment analysis methods to Swedish textual data has been demonstrated, for instance, by several social scientists [3,76]. Furthermore, the prior work on the *ABSAbank-Imm* annotated corpus and the respective analyses by Rouces et al. [14,77,78] are very relevant and useful for our work.

Although there are considerably fewer studies on classification performance for texts in Swedish, the existing evidence suggests that existing models tend to achieve accuracy scores in the range of 70–80%. For instance, in a study of hate detection with universal

language models, an application of the *LMFit* model allowed achieving classification performance of 79% on flashback data [79]. Borg and Boldt [80] used an SVM classifier in conjunction with the VADER sentiment lexicon to perform the classification of e-mails in Swedish. Their results demonstrate accuracy scores in the range of 82–85% [80]. The highest classification results (accuracy 0.96) have been reported by Fernquist et al. [81], who used ML for the identification of tweets posted by bots. However, it seems that Swedish social research, in general, and sociology, in particular, have not yet fully exploited the possibilities of NLP and ML for the application of text data in Swedish. Thus, in this work, we aim to address both issues of the training data quality and the performance of models built with the help of such data, and provide a reflection on the use of ML for the Swedish social media data.

3. Methods

3.1. Datasets and Tasks

An overview of the data sources, annotation process, and resulting labeled data for our work is provided in Figure 1.

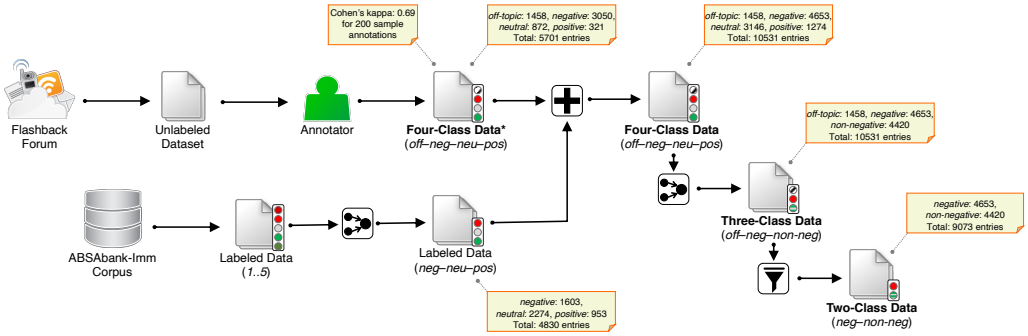

Figure 1. An overview of the data used in this work. The initial dataset consists of forum data labeled by one annotator (one of the authors of this work) at the document/message level with one of four class labels (i.e., a single label for a document potentially consisting of several paragraphs and sentences)—see "Four-Class Data*" in the figure. To increase the amount of training and test data for our stance classification models, the data was augmented with a transformed subset of the ABSAbank-Imm corpus labeled at the paragraph level (i.e., a single label for a paragraph from some message, which might potentially consist of several sentences). Afterward, we transformed the four-class dataset into three classes by merging a pair of categories, and finally filtering another category, arriving at a two-class dataset. *©2021 IEEE. Reprinted, with permission, from 2021 Swedish Workshop on Data Science (SweDS). V. Yantseva and K. Kucher. Machine Learning for Social Sciences: Stance Classification of User Messages on a Migrant-Critical Discussion Forum. doi:10.1109/SweDS53855.2021.9637718.*

The data used in this study come from Flashback Forum, a Swedish online discussion platform, and, in particular, from its "Integration and Immigration" section dedicated to the discussion of various aspects of Swedish immigration policies. A total of 1.3 M messages dated 2012–2019 were collected in January 2020 using the *rvest* package in the R programming language [82], and a random sample of 5701 messages were manually labeled by one of the authors. The documents were initially coded as belonging to one of the four categories: *off-topic*, *negative*, *neutral*, and *positive*. Intra-rater reliability of labeling was evaluated using a test dataset of 200 messages (Cohen's kappa = 0.69) in a fashion similar to inter-rater reliability testing [83].

In this work, our aim was to perform a *stance* rather than *sentiment* classification of user messages. Previous research has suggested that sentiment represents the general tonality of the messages and emotions expressed [42]. Stance, on the other hand, reflects the attitude of an author towards a specific target and its approval or disapproval [42]. Since most discussions on the forum are concerned with immigration as the main topic, our goal was to perform stance classification of the forum messages in such a way that

would allow distinguishing between users' attitudes to immigration and corresponding policies in Sweden rather than identifying general tonality of the messages, as it is usually done in sentiment analysis. Therefore, the messages that were not directly related to the immigration issue were coded as off-topic.

To compensate for data imbalance in our dataset (since positive messages comprise only 5% of the data), we have also tested our models for three-class data (*off-topic*/*negative*/*non-negative*, with the latter category being formed by merging *neutral* and *positive*). Finally, we evaluated the models' performance for the two-class (*negative*/*non-negative*; *off-topic* items ignored) classification problem. Since previous research has demonstrated that Flashback can be described as a right-wing communication platform, we have been primarily interested in distinguishing between those who hold migrant-critical versus non-critical views, which underpinned our decision to merge neutral and positive messages into one category.

Another way to compensate for data imbalance was to supplement it with a paragraph-level annotated set of 852 documents (4872 paragraphs) from the same forum provided as part of the ABSAbank-Imm corpus [77]. Since that dataset included labels on the scale between 1 (*very negative*) and 5 (*very positive*), they were recoded for them to correspond to the three categories in our main dataset (1–2: *negative*; 3: *neutral*; 4–5: *positive*). SMOTE resampling [84] was another alternative for class imbalance compensation.

3.2. Processing Pipeline

An overview of the feature engineering and text classification approaches used in our work, resulting in four groups of experiments, is provided in Figure 2.

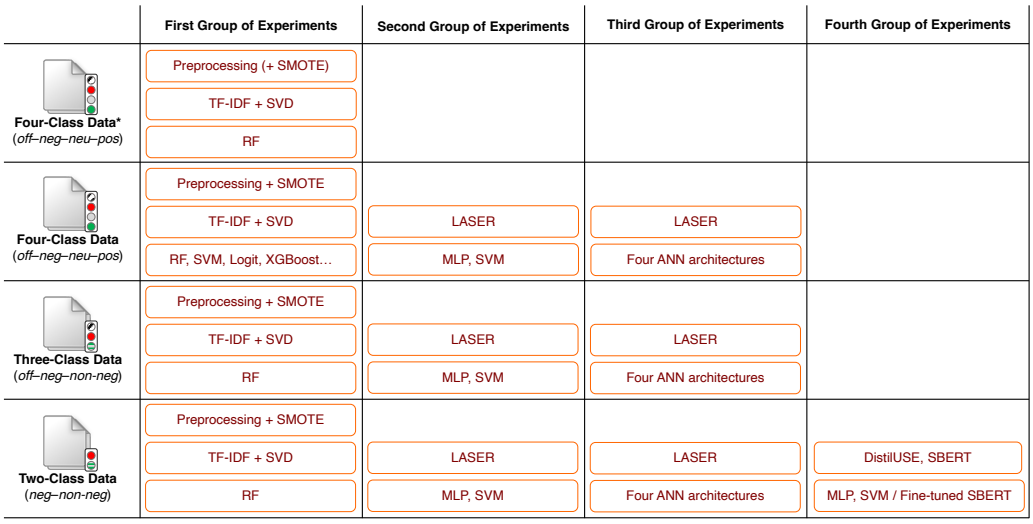

Figure 2. An overview of the feature engineering and text classification approaches used in this work (see the description of the data sets in Figure 1). Our first group of experiments included custom preprocessing, data augmentation, and feature engineering techniques combined with classic machine learning algorithms for text classification. The second group of experiments relied on the embeddings produced by the LASER model as features combined with several classic classification algorithms. The third group of experiments replaced the classification algorithms with several artificial neural network architectures while still relying on LASER embeddings as input features. Finally, the fourth group of experiments included other modern models for producing embedding features (combined with several classification algorithms) as well as a fine-tuned end-to-end modern model.

Our first group of experiments involved traditional machine learning methods for stance classification. The documents were pre-processed using a standard pre-processing pipeline: we removed URLs, numbers, punctuation, and stop words from the dataset; stemmed the tokens and added bigrams; applied *tf-idf* weighting and performed singular value decomposition (SVD) to create a dense vector representation of the corpus [20,21].

The dataset was randomly split into training (70%) and test sets (30%). Subsequently, we evaluated a range of traditional algorithms used in the classification tasks, in particular, random forest (*RF*), support vector machines (*SVM*), logistic regression (*Logit*), extreme gradient boosting (*XGBoost*) and adaptive boosting (*AdaBoost*). All analyses were performed with the help of several Python packages, in particular, *NLTK* (stemming) [85], *imbalanced-learn* (SMOTE resampling) [84], *scikit-learn* (machine learning) [86], and *xgboost* (the XGBoost algorithm) [87].

For the second group of experiments, we used a different pipeline with the features produced by Facebook's Language-Agnostic SEntence Representations (*LASER*) model [88]. LASER uses a neural network architecture to produce a distributed representation of the sentences or documents in the corpus. It is a language-agnostic model that includes support for more than 90 languages, including Swedish, which underpinned our choice of this model. Several algorithms from *scikit-learn* were used to train and test the respective models, with a focus on SVM and multilayer perceptrons (*MLP*) with several hidden layers. This choice of algorithms was motivated by their feasibility for relatively small text datasets, roughly similar computational time and resource requirements, and practicality of the classification pipelines, with the latter being a definite strength when it comes to the applicability of machine learning in the social sciences.

In the third group of experiments, we tested several neural network architectures:

- a network with three dense layers;
- a network with two long short-term memory layers (*LSTM*) [29] layers and one dense layer;
- a network with one 1D convolution layer (*Conv1D*) [29] layer and one dense layer; and
- a network with one LSTM, one Conv1D, and one dense layer.

The same features produced by the LASER model were used in this step for model training to check whether the utilization of artificial neural network architectures would help to beat the results of the best-performing ML model for classification purposes. Network training was performed in *TensorFlow* [89]. As in the previous case, the dataset was split into training (70%) and test (30%) data.

In the fourth group of experiments, we considered other modern language models available for Swedish text data and focused on the two-class problem, considering the results from previous experiments. Here, we relied on the multilingual DistilUSE [90] model (namely, *sentence-transformers/distiluse-base-multilingual-cased-v2*) and the Swedish Sentence-BERT [75] model (namely, *KBLab/sentence-bert-swedish-cased*) to produce embedding vectors (length 512 and 768, respectively). These vectors were used together with the SVM and MLP classification algorithms based on their performance in the second group of our experiments. We have also conducted *fine-tuning* [28,29] of the pretrained Swedish Sentence-BERT model [75] for our downstream stance classification task using the *PyTorch* [91] and Huggingface *Transformers* [92] libraries. The dataset was split into training (70%) and test (30%) data for these experiments in the same way as above.

Finally, we supplemented our experiments with a qualitative evaluation of the model output. As an example, we took labels for ABSAbank-Imm corpus [77] produced by the MLP algorithm with LASER features for two-class data and evaluated cases with high-class probability and label coherence with ground truth, as well as cases with mixed class membership and incoherent labels.

4. Results

4.1. Training Classic Machine Learning Models

In the first and second sets of experiments, we manipulated a range of model configurations to identify those achieving the best classification performance, in particular: the number of classes; classification algorithm; type of feature-vector representation; and methods to compensate for class imbalance. We report macro-averaged results as they favor minority classes for imbalanced data classification [93,94], but also list the F1 score [95] results for the majority class for reference in Table 1.

Table 1. Model Parameters and Results—First and Second Groups of Experiments.

Classification Model	Accuracy, Training (CV)	F1, Training (CV)	Accuracy, Test Data	Precision, Test Data	Recall, Test Data	F1, Test Data	F1, Test Data, *Negative*
Four classes*, SVD, no resampling, RF	0.60	0.31	0.61	0.75	0.33	0.31	**0.74**
Four classes*, SVD, SMOTE resampling, RF	**0.86**	**0.86**	**0.59**	**0.42**	**0.37**	**0.38**	0.73
Four classes, SVD, SMOTE resampling, ADABoost	0.55	0.55	0.42	0.41	0.43	0.40	0.51
Four classes, SVD, SMOTE resampling, XGBoost	0.85	0.85	0.51	0.46	0.44	0.44	0.63
Four classes, SVD, SMOTE resampling, Logit	0.50	0.51	0.50	0.46	0.49	0.46	0.61
Four classes, SVD, SMOTE resampling, SVM	0.74	0.74	0.51	0.47	0.49	0.47	0.62
Four classes, SVD, SMOTE resampling, RF	**0.86**	**0.86**	0.54	0.49	0.46	0.47	0.64
Four classes, LASER, MLP	0.59	0.52	**0.61**	**0.59**	0.53	0.54	**0.70**
Four classes, LASER, SVM	0.57	0.55	0.58	0.55	**0.59**	**0.56**	0.65
Three classes, SVD, SMOTE resampling, RF	**0.70**	**0.69**	0.61	0.59	0.57	0.57	0.63
Three classes, LASER, MLP	0.66	0.64	**0.67**	**0.68**	0.64	**0.66**	**0.68**
Three classes, LASER, SVM	0.66	0.65	**0.67**	0.65	**0.70**	**0.66**	0.67
Two classes, SVD, SMOTE resampling, RF	0.70	0.70	0.67	0.68	0.66	0.66	0.72
Two classes, LASER, MLP	0.72	0.72	0.71	0.71	0.71	0.71	**0.73**
Two classes, LASER, SVM	**0.73**	**0.73**	**0.72**	**0.72**	**0.72**	**0.72**	**0.73**

Note: The baseline models listed in the first two rows were trained with a smaller dataset annotated by one of the authors, and other models were trained with larger datasets merged with ABSAbank-Imm paragraph-level annotated data (see Figure 1). The models were trained with 70% of the labeled data for the respective problems (four classes: *off-topic*/*negative*/*neutral*/*positive*; three classes: *off-topic*/*negative*/*non-negative*; two classes: *negative*/*non-negative*). The macro averaged accuracy and F1 score results for the training data are based on average outcomes of a 5-fold cross-validation. Macro averaged results are listed for precision, recall, and F1 score values for the test data (the held-out 30% of the labeled data for the respective problems). The last column provides the F1 score results for the majority class (*negative*) on the test data. The best scores for each column for each respective problem and data are marked in bold.

Reducing the number of classes. We tested a range of solutions with four (*off-topic*, *negative*, *neutral*, and *positive*), three (*off-topic*, *negative*, and *non-negative*) and two (*negative* and *non-negative*) classes. As follows from Table 1, reducing the number of classes appears to be an effective strategy to achieve a higher model performance for this task. A two-class problem compared to a four-class problem, for instance, allowed boosting model performance by almost 20% (F1 macro score of 0.47 versus 0.66, respectively). The downside of this approach, of course, is a need to train two separate classifiers: one to distinguish between off-topic and on-topic messages, and another one to identify their stances.

Compensating for class imbalance. To mitigate the effects of class imbalance, we applied SMOTE resampling to upsample observations in minority classes, which helped to raise the F1 macro score from 0.31 to 0.38 and, as expected, was more effective than traditional upsampling of minority classes. Adding data with paragraph labels from the ABSAbank-Imm corpus also resulted in considerable improvements (F1 macro score of 0.47 vs. 0.38 with the same RF algorithm).

Choosing the best classification algorithm. RF and SVM appeared to be the most effective alternatives among the first group of models (F1 macro scores of 0.47 for both algorithms and a four-class classification problem). RF seems to provide better precision (0.49 versus 0.47), while SVM is better at achieving greater recall (0.49 versus 0.46). RF and SVM were followed by Logit (F1 macro 0.46), XGBoost (F1 macro 0.44) and, finally, ADABoost (F1 macro 0.40). Further, we performed a grid search to find optimal RF classifier hyper-parameters; however, using optimal hyper-parameters did not affect the final outcome. For the second group of models that used LASER embeddings, MLP allowed us to increase the macro-F1 score to 0.54 compared to a score of 0.47 for an RF classifier. However, the SVM classifier provided even better performance for four-class and two-class problems (F1 macro scores of 0.56 and 0.72, respectively), while being tied with MLP for the three-class problem (F1 macro score of 0.66).

Manipulating model features. For the first group of our models, we applied traditional text preprocessing and singular value decomposition to obtain feature vectors that were used for subsequent training (50 dimensions; we also tested other dimensionalities; however, this did not improve the performance). In the second stage, instead of preprocessing and SVD, we used vectors (1024 dimensions) obtained with the help of a pre-trained LASER model. The two-class SVM model with LASER embeddings yielded an F1 macro score of 0.72, which was the best solution tested in this stage of our experiments. Table 2 provides further details on the best models.

Table 2. Detailed Test Data Results for the Best-Performing Models in the Second Group of Experiments.

Model/Class	Precision	Recall	F1	Support
LASER, SVM, *off-topic*	0.52	0.80	0.63	438
LASER, SVM, *negative*	0.73	0.58	0.65	1396
LASER, SVM, *neutral*	0.55	0.55	0.55	944
LASER, SVM, *positive*	0.39	0.45	0.42	382
LASER, MLP, *off-topic*	0.70	0.56	0.62	438
LASER, MLP, *negative*	0.64	0.73	0.68	1396
LASER, MLP, *non-negative*	0.69	0.64	0.66	1326
LASER, SVM, *off-topic*	0.53	0.79	0.63	438
LASER, SVM, *negative*	0.71	0.63	0.67	1396
LASER, SVM, *non-negative*	0.70	0.67	0.68	1326
LASER, SVM, *negative*	0.73	0.72	0.73	1396
LASER, SVM, *non-negative*	0.71	0.72	0.71	1326

Note: The results are listed for the test data (the held-out 30% of the labeled data for the respective problems). The last column provides details about the number of test data items with the respective label for each case.

4.2. Training Neural Network Models

In our third set of experiments, we tested the performance of several neural network architectures (dense layers only, LSTM and dense layers, Conv1D and dense layers, as well as a combination of the three) [29] on four-, three-, and two-class training data. Just as in the second group of experiments, we used 1024 features generated by the LASER model to check whether neural networks would be able to beat the results of the best performing SVM model (cf. Table 1). In this step, we also tried to manipulate a range of model parameters, such as the number of neurons in each of the layers, the activation function used, and dropout rates. However, none of these measures helped to significantly improve final classification results.

The main takeaway from our tests is that the reduction of class number by one allows boosting the final model's accuracy on test data by roughly 6%, while the choice of actual layer types seems to play a secondary role in model performance. As presented in Table 3, the best-performing model is the one with a combination of dense 1D and 1D convolution layers, which typically provides the highest accuracy and the lowest loss. However, it performs only marginally better than networks with other layer combinations (1–2% in model accuracy in test data). This network is followed by one with all three types of layers (dense, Conv1D, and LSTM). Quite expectedly, a network with three dense layers is the least successful in our task—while the final accuracy is not that much different from other networks, it seems to overfit the data and provides the highest loss on the test data.

Table 3. Model Parameters and Results—Third Group of Experiments.

Classification Model	Accuracy, Training	Loss, Training	Accuracy, Test Data	Loss, Test Data
Four classes, LASER, three dense layers	**0.85**	**0.40**	**0.58**	1.60
Four classes, LASER, LSTM and dense layers	0.74	0.67	**0.58**	1.14
Four classes, LASER, Conv1D and dense layers	0.67	0.81	**0.58**	**0.99**
Four classes, LASER, Covd1D, LSTM, and dense layers	0.70	0.75	**0.58**	1.12
Three classes, LASER, three dense layers	**0.90**	**0.24**	0.64	1.48
Three classes, LASER, LSTM and dense layers	0.80	0.48	0.65	0.98
Three classes, LASER, Conv1D and dense layers	0.76	0.55	**0.66**	0.76
Three classes, LASER, Conv1D, LSTM, and dense layers	0.69	0.67	**0.66**	**0.72**
Two classes, LASER, three dense layers	**0.90**	**0.23**	0.70	1.02
Two classes, LASER, LSTM and dense layers	0.82	0.37	0.70	0.69
Two classes, LASER, Conv1D and dense layers	0.80	0.43	**0.72**	**0.57**
Two classes, LASER, Conv1D, LSTM, and dense layers	0.82	0.40	0.71	0.63

Note: The models were trained with 70% of the labeled data for the respective problems (four classes: *off-topic*/*negative*/*neutral*/*positive*; three classes: *off-topic*/*negative*/*non-negative*; two classes: *negative*/*non-negative*). The accuracy and loss results are listed for the training data as well as the test data (the held-out 30% of the labeled data for the respective problems). The best scores for each column for each respective problem and data are marked in bold.

The most important observation, however, is that neural networks used as classifiers do not perform better than the classic machine learning models for our type and scale of data—rather, both approaches were able to classify slightly more than two-thirds of messages correctly in the test data. The most obvious explanation for this conclusion is, of course, the size of training data available: the results for a scenario with massive training data, e.g., large collections of texts in English, could probably have been more favorable for deep neural network models when training classifiers from scratch.

4.3. Making Use of Modern Language Models

Given the results from the previous groups of experiments above, we considered alternative approaches to search for better stance classification results, at least for the two-class problem. While the 1024-dimensional LASER embedding vectors were used as input features in the second and third groups of experiments, we modified this part of our data processing pipeline and computed 512-dimensional DistilUSE [90] vectors as well as 768-dimensional Sentence-BERT [73–75] vectors for our input texts in Swedish. We combined these two feature sets with the SVM and MLP algorithms (which demonstrated the best results in the second group of our experiments, cf. Table 1), resulting in four models.

As demonstrated in Table 4, these models perform similarly or even better than the best models described above, with the SVM classifier using Sentence-BERT embeddings

achieving the value of 0.76 in several performance metrics on the test data (as well as the F1 score of 0.77 for the majority *negative* class). Given the availability of the respective Sentence-BERT model for Swedish text data [75], such a pipeline provides a viable option for this and further application scenarios.

Table 4. Model Parameters and Results—Fourth Group of Experiments.

Classification Model	Accuracy, Training (CV)	F1, Training (CV)	Accuracy, Test Data	Precision, Test Data	Recall, Test Data	F1, Test Data	F1, Test Data, *Negative*
Two classes, DistilUSE, MLP	0.71	0.71	0.70	0.70	0.70	0.70	0.71
Two classes, DistilUSE, SVM	0.73	0.73	0.71	0.71	0.71	0.71	0.73
Two classes, Sentence-BERT, MLP	0.73	0.73	0.72	0.72	0.72	0.72	0.72
Two classes, Sentence-BERT, SVM	**0.75**	**0.75**	**0.76**	**0.76**	**0.76**	**0.76**	**0.77**
Two classes, Sentence-BERT, fine-tuning	-	-	**0.76**	**0.76**	**0.76**	**0.76**	0.76

Note: The models were trained with 70% of the labeled data for the respective problems (two classes: *negative*/*non-negative*). The macro averaged accuracy and F1 score results for the training data are based on average outcomes of a 5-fold cross-validation (results not provided for the fine-tuned Sentence-BERT model due to the extensive training time necessary). Macro averaged results are listed for precision, recall, and F1 score values for the test data (the held-out 30% of the labeled data for the respective problems). The last column provides the F1 score results for the majority class (*negative*) on the test data. The best scores for each column for each respective problem and data are marked in bold.

Finally, we considered an *end-to-end* approach that would see the pre-trained Swedish Sentence-BERT model *fine-tuned* [28,29] for our two-class stance classification task, i.e., the model would take the text messages as input and produce classification results for the respective classes at the output layer of the underlying neural network. We ran the fine-tuning process for 3 epochs and evaluated the model with the test data, resulting in the performance metric values approximately the same as the Sentence-BERT + SVM combination discussed above. Further experiments with different hyperparameters (including the learning rate, etc.) could potentially lead to even better results; however, there is another concern that we should mention: the experiments in this group were run on a physical machine (Apple Studio M1 Max, 8 + 2 CPU cores, 24 GPU cores, 32 GB of unified memory, macOS Monterey), and the time necessary for Sentence-BERT fine-tuning was significantly longer than the classic ML model training (198 minutes compared to approximately 25 s for the SVM, provided the pre-computed 768-dimensional vectors). Particular time and memory concerns might be resolved in the future with better optimization of the ML/DL libraries for particular software and hardware platforms (as well as cloud-based solutions); however, such practicalities (beyond the model performance on its own) are still part of the considerations in applied scenarios, including the workflows of social science researchers.

4.4. Qualitative Assessment of the Classification Results

In this work, we are not only concerned with identifying the best-achieving model for our task, but also with the subsequent possibility to qualitatively assess the results and label assignment. The reason is that social scientists are often interested in identifying patterns and the reasons for particular label assignments, rather than simply using the results as they are.

Please note: we have not included any particular message excerpts below to protect users' privacy and integrity.

To assess the results of the label assignment qualitatively, we used the best performing MLP model for two-class data (*negative*/*non-negative*; cf. Table 1) to generate labels for the annotated ABSAbank-Imm corpus [77] (paragraph level annotations, recoded ground truth labels for *negative*/*neutral*/*positive* classes, see Section 3.1). The MLP model was applied here instead of SVM, since it directly provides us with probability estimates for the respective classes, i.e., it provides estimates of classification decision confidence for particular text instances. Our close reading of several documents allows suggesting that the model captures well the overall difference between negative and non-negative documents or sentences. We conclude that the model outputs coherent labels for the paragraphs that include some typical terms associated with a negative ("rapes", "criminality") or positive ("solidarity", "respect", "human rights") stance on the topic, or the paragraphs that include a specific lexicon used by some members of the Flashback audience (e.g., "luck seekers", "immigrants" ["lycksökare" and "blattar" in Swedish]). When it comes to the cases where labels become confused, some texts do indeed have ambiguous meanings, which makes it challenging to classify those texts. In other cases, shorter texts seem to be more likely to be classified as non-negative. Finally, one more case is when misclassification of some documents as negative might be associated with the presence of some typical terms associated with a negative stance to the topic (i.e., "homophobia", "poverty", "genetics", etc.), despite the non-negative stance of the respective document's author.

It should also be mentioned that other limitations to the final model quality arise from the potential errors that can occur in the process of manual document labeling, text meaning ambiguities, and subsequent inconsistencies in evaluators' labels [83,96]. Another constraint that might affect the quality of the final classification is the size of the documents, since longer documents could potentially confuse the final labeling, even in cases where the majority of sentence labels were classified correctly. These observations might be used in future work to mitigate the drawbacks of the final model. For instance, the predicted label for a paragraph or a document/post could take the results of classification into account at the same and underlying hierarchical levels into account (e.g., a paragraph could be classified as a single input to the model, but also as a collection of sentences, and the predicted labels could all be considered).

5. Discussion

In this section, we discuss the limitations and concerns related to this work. The main issue in relation to our classification task has been data imbalance and sparsity, which seemed to have the highest effect on the models' performance. Reducing the number of classes has led to a significant improvement of the results in this work, which is in line with previous research that has suggested that two-class classification problems tend to produce significantly better results compared to three-class problems for sentiment analysis [55] and hate speech detection [15] tasks. However, this solution might require training additional classifiers or filtering the input data by topic, depending on the application. Another way to compensate for class imbalance has been to augment the dataset with additional paragraph-level annotations. If available, this also appears to be an effective strategy in cases where human resources for labeling data are limited, as it requires minimal additional work while helping to fill in the observations in the minority classes.

Moreover, in line with earlier findings, SVM turned out to be one of the best performing algorithms for our task [46]. The MLP algorithm has also shown very good results in the second and fourth groups of experiments (see Tables 1 and 4). However, contrary to our expectations and existing evidence [48], none of the neural network models with more advanced architectures has been able to beat the results of the SVM algorithm, achieving the same performance at best in the third group of experiments (see Table 3). Although neural networks have been shown to be able to provide extremely high results in sentiment classification (e.g., see the work by Karakus et al. [40]), their application to real-world data demonstrates that classic machine learning still represents a competitive approach that can actually outperform these (more advanced) models in certain tasks, especially with

limited training resources. The fine-tuned Sentence-BERT model has allowed us to reach some of the top results, but, as discussed in Section 4.3, this can also come at the cost of considerable time and effort required compared to the scenario of using more traditional classification algorithms with the features computed by the more modern models.

On top of this, we notice that we were unable to achieve accuracy (and several other metrics) higher than 76% in the test data even with the help of neural networks. Previous attempts to apply more traditional ML or modern DL-based approaches to corpora in various languages, such as Swedish [79], German [60], or even English [50], have also been consistent with this result. Therefore, it seems that the precisions between 70% and 80% seem to serve as a rather common threshold for given types of task irrespective of the methodology used, providing evidence of the limits of machine learning capabilities. This observation also confirms the earlier statement that the nature of the data, rather than the classification algorithm used, plays a key role in the classification results [55], at least in practical domain application scenarios. Although recent very large language models have shown very high results for sentiment classification, at least for texts in English, the availability and application of these models may also pose certain challenges, as mentioned previously [52–54].

Therefore, in the social science domain, classic machine learning or deep learning approaches still need to compete with more conventional methods, such as qualitative analysis (close reading) of a smaller subset of documents, which represents a challenge for the integration of computational methods into social sciences. Thus, considering the fact that social sciences frequently deal with complex and ambiguous real-world data (rather than benchmark datasets used to build many of the deep learning models achieving very high accuracy), more simple classification approaches seem to be the way to go in many cases when automated classification of novel text corpora is needed (although there is still the issue of transforming the texts into a representation suitable for classification).

We also note that one of the key limitations of the use of DL approaches is that they require larger training datasets, which is not always possible when dealing with novel real-world corpora (such as data from a specific platform), since labeling in most cases needs to be performed from scratch. However, various transfer learning strategies [34] could be a potential solution in some scenarios. Another concern relates to the additional knowledge and methodological skills required for social scientists to build a model with a neural network architecture. In contrast to deep learning, classic approaches require less time to grasp the key ideas and principles behind their work, which is important given that far from all social scientists possess advanced skills in statistics and programming. However, given the fast-paced development of deep learning methods, it is reasonable to suggest that more user-friendly GUI-based tools for deep learning would be introduced, which would make them more accessible to this group of researchers. For instance, the computational linguistics community dealing with NLP and other text mining tasks has recently considered tools for the development and testing of models for various NLP tasks [97], and even more advanced interactive approaches that rely on the body of knowledge in (text) visualization and visual (text) analytics are possible [98–102]. Additionally, the modern pre-trained language models provide useful and convenient pipelines for obtaining vectorized data representations that can be used as inputs for the classic classification algorithms (or other downstream applications), as demonstrated by most of our experiments and results.

Despite the above-mentioned, the final model accuracy or other classification performance metric [95] is not the only important factor when it comes to the application of ML or DL in the social science domain. Since social sciences pay great attention to the interpretation and explanation of the phenomena they study, the model *interpretability* [103,104] may be a more important factor than accuracy [105,106]. However, since neural networks function as black boxes, a preference could be given to other, sometimes less sophisticated approaches that could explain label assignment, although they could provide lower precision [105]. Although testing such approaches specifically with a focus on interpretability or explainability was not a task for this work, future research could also

test the application of these approaches to our dataset and compare the performance of "black box" and interpretable models in the social sciences. Furthermore, the application of interactive visual interfaces for constructing and investigating models and their results also provides further opportunities, as mentioned above.

Further application of our proposed methodology for stance classification on Flashback may provide several important insights to social research of right-wing online activism in the Swedish context. For instance, Åkerlund reports that influential right-wing Swedish Twitter users tend to use surprisingly more neutral rather than negative language (potentially, as a consciously chosen discourse strategy), based on analyses of metadata and sentiment in tweets with VADER [3]. Similarly, cross-platform analysis of sentiments using a lexicon-based approach demonstrates that the tonality of messages depends on the particular frames of the immigration issue used in them [6]. However, while both works deploy unsupervised sentiment analytic tools, and given a limited number of studies available on sentiments and stances of users on Swedish radical right resources, the application of the models proposed in this work to Swedish forum data with similar considerations in mind is an opportunity for future work and can be successfully combined with other methods, such as topic modeling, network analysis, or qualitative analysis of the social media discourse on right-wing platforms. Before conclusions about the underlying phenomena are drawn based on particular data and model, we should note, though, that potential biases (texts collected from a particular source and time range), data quality (less than perfect annotator agreement, which could be expected for such complex and controversial discussion topics), and ML/DL training process deficiencies (lack of global optimal solution guarantees) present potential threats to the validity of this and similar studies, and, thus, further research efforts are welcome.

6. Conclusions

In this paper, we have demonstrated the application of machine learning approaches to Swedish social media data and evaluated their effectiveness on the stance classification task based on a sparse and unbalanced manually annotated dataset, with the final SVM model (using Swedish Sentence-BERT text embeddings as input features) and the fine-tuned Swedish Sentence-BERT model achieving F1 macro of 0.76 on the test data for a two-class classification problem (*negative*/*non-negative* stance for immigration discussions). Despite promising results so far, we see possibilities to extend our analysis, such as, for instance, to involve further feature computation approaches and pre-trained models, such as pre-trained word vectors from fastText [107], which were applied for named entity recognition tasks in Swedish by Adewumi et al. [108]. As mentioned above, another way to extend the analysis could be to test further applicability of "black box" versus interpretable machine learning approaches [105,106] and to compare their performance on the tasks relevant for social science research. Extension of the training dataset (including inter-annotator agreement checks), replication of our approach with different datasets (primarily in Swedish), analysis of particular classification errors with respect to labeled data and human judgment [96,109], and application of our approach for related tasks such as detection of hate speech [16,94] and other types of abusive language [18] are also interesting and important research opportunities. In addition to purely computational approaches, we plan to integrate our models with visual analytic approaches, which have been proposed for various analyses of text, sentiment, and social media [98–101], including the machine learning explainability, interpretability, and trustworthiness concerns in particular [102].

Finally, we aim to put our model to use with regard to previously unseen data from the same forum to classify forum users' stances. The predicted results will be used further for the analysis of communication patterns on this online platform [6]. Given the growing interest of social scientists in right-wing activism and the use of online media by radical groups, as well as the development of the field of computational social science, our work could be relevant for professionals in political science, sociology, and media studies working with social media data in Sweden and other countries.

Author Contributions: Conceptualization, V.Y. and K.K.; methodology, V.Y. and K.K.; software, V.Y. and K.K.; validation, V.Y. and K.K.; formal analysis, V.Y. and K.K.; investigation, V.Y. and K.K.; data curation, V.Y.; writing—original draft preparation, V.Y. and K.K.; writing—review and editing, V.Y. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partly funded by EU CEF grant number 2394203 (NORDIS-NORdic observatory for digital media and information DISorder). The authors would also like to thank Linnaeus University and Linköping University for supporting the work on this publication.

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

Note

[1] *©2021 IEEE. Reprinted, with permission, from 2021 Swedish Workshop on Data Science (SweDS). V. Yantseva and K. Kucher. Machine Learning for Social Sciences: Stance Classification of User Messages on a Migrant-Critical Discussion Forum. doi:10.1109/SweDS53855.2021.9637718*

References

1. Pang, B.; Lee, L. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135, https://doi.org/10.1561/1500000011.
2. Ceron, A.; Curini, L.; Iacus, S.M.; Porro, G. Every Tweet Counts? How Sentiment Analysis of Social Media can Improve our Knowledge of Citizens' Political Preferences With an Application to Italy and France. *New Media Soc.* **2014**, *16*, 340–358, https://doi.org/10.1177/1461444813480466.
3. Åkerlund, M. The Importance of Influential Users in (Re)Producing Swedish Far-right Discourse on Twitter. *Eur. J. Commun.* **2020**, *35*, 613–628, https://doi.org/10.1177/0267323120940909.
4. Loureiro, M.L.; Alló, M. Sensing Climate Change and Energy Issues: Sentiment and Emotion Analysis with Social Media in the U.K. and Spain. *Energy Policy* **2020**, *143*, 111490, https://doi.org/10.1016/j.enpol.2020.111490.
5. Pope, D.; Griffith, J. An Analysis of Online Twitter Sentiment Surrounding the European Refugee Crisis. In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 16, Porto, Portugal, 9–11 November 2016; SciTePress: Setúbal, Portugal, 2016; pp. 299–306, https://doi.org/10.5220/0006051902990306.
6. Yantseva, V. Migration Discourse in Sweden: Frames and Sentiments in Mainstream and Social Media. *Soc. Media + Soc.* **2020**, *6*, https://doi.org/10.1177/2056305120981059.
7. Macnair, L.; Frank, R. The Mediums and the Messages: Exploring the Language of Islamic State Media Through Sentiment Analysis. *Crit. Stud. Terror.* **2018**, *11*, 438–457, https://doi.org/10.1080/17539153.2018.1447226.
8. Colleoni, E.; Rozza, A.; Arvidsson, A. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *J. Commun.* **2014**, *64*, 317–332, https://doi.org/10.1111/jcom.12084.
9. Anastasopoulos, L.J.; Williams, J.R. A Scalable Machine Learning Approach for Measuring Violent and Peaceful Forms of Political Protest Participation with Social Media Data. *PLoS ONE* **2019**, *14*, e0212834, https://doi.org/10.1371/journal.pone.0212834.
10. Sheetal, A.; Savani, K. A Machine Learning Model of Cultural Change: Role of Prosociality, Political Attitudes, and Protestant Work Ethic. *Am. Psychol.* **2021**, *76*, 997–1012, https://doi.org/10.1037/amp0000868.
11. Blomberg, H.; Stier, J. Flashback as a Rhetorical Online Battleground: Debating the (Dis)guise of the Nordic Resistance Movement. *Soc. Media + Soc.* **2019**, *5*, https://doi.org/10.1177/2056305118823336.
12. Törnberg, A.; Törnberg, P. Muslims in Social Media Discourse: Combining Topic Modeling and Critical Discourse Analysis. *Discourse, Context & Media* **2016**, *13 Pt B*, 132–142, https://doi.org/10.1016/j.dcm.2016.04.003.
13. Malmqvist, K. Satire, Racist Humour and the Power of (Un)Laughter: On the Restrained Nature Of Swedish Online Racist Discourse Targeting EU-Migrants Begging for Money. *Discourse Soc.* **2015**, *26*, 733–753, https://doi.org/10.1177/0957926515611792.
14. Rouces, J.; Borin, L.; Tahmasebi, N. Creating an Annotated Corpus for Aspect-Based Sentiment Analysis in Swedish. In Proceedings of the Conference of the Association of Digital Humanities in the Nordic Countries. CEUR Workshop Proceedings, DHN '20, Riga, Latvia, 21–23 October 2020; pp. 318–324.
15. Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; Tesconi, M. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In Proceedings of the First Italian Conference on Cybersecurity. CEUR Workshop Proceedings, ITASEC '17, Venice, Italy, 17–20 January 2017; Volume 1816, pp. 86–95.
16. Schmidt, A.; Wiegand, M. A Survey on Hate Speech Detection using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP '17, Valencia, Spain, 3 April 2017; ACL: Stroudsburg, PA, USA, 2017; pp. 1–10, https://doi.org/10.18653/v1/W17-1101.
17. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. *Proc. Int. AAAI Conf. Web Soc. Media* **2017**, *11*, 512–515, https://doi.org/10.1609/icwsm.v11i1.14955.
18. Waseem, Z.; Davidson, T.; Warmsley, D.; Weber, I. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In Proceedings of the First Workshop on Abusive Language Online, ALW '17, Vancouver, BC, Canada, 4 August 2017; ACL: Stroudsburg, PA, USA, 2017; pp. 78–84, https://doi.org/10.18653/v1/W17-3012.

19. Yantseva, V.; Kucher, K. Machine Learning for Social Sciences: Stance Classification of User Messages on a Migrant-Critical Discussion Forum. In Proceedings of the Swedish Workshop on Data Science, SweDS '21, Växjö, Sweden, 2–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; https://doi.org/10.1109/SweDS53855.2021.9637718.
20. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
21. Aggarwal, C.C. *Machine Learning for Text*; Springer: Berlin/Heidelberg, Germany, 2018.
22. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828, https://doi.org/10.1109/TPAMI.2013.50.
23. Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
24. Almeida, F.; Xexéo, G. Word Embeddings: A Survey. *arXiv* **2019**, arXiv:1901.09069.
25. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the International Conference on Machine Learning, PMLR, ICML '14, Beijing, China, 21–26 June 2014; pp. 1188–1196.
26. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
27. Liu, J.; Chang, W.C.; Wu, Y.; Yang, Y. Deep Learning for Extreme Multi-Label Text Classification. In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Tokyo, Japan, 7–11 August 2017; ACM: Singapore, 2017; pp. 115–124, https://doi.org/10.1145/3077136.3080834.
28. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained Models for Natural Language Processing: A Survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897, https://doi.org/10.1007/s11431-020-1647-3.
29. Otter, D.W.; Medina, J.R.; Kalita, J.K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 604–624, https://doi.org/10.1109/TNNLS.2020.2979670.
30. Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, Honolulu, HI, USA, 25–27 October 2008; ACL: Stroudsburg, PA, USA, 2008; pp. 254–263.
31. Settles, B. Active Learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2012**, *6*, 1–114, https://doi.org/10.2200/S00429ED1V01Y201207AIM018.
32. Mohammad, S.; Turney, P. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET '10, Los Angeles, CA, USA, 5 June 2010; ACL: Stroudsburg, PA, USA, 2010; pp. 26–34.
33. Hamilton, W.L.; Clark, K.; Leskovec, J.; Jurafsky, D. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '16, Austin, TX, USA, 1–5 November 2016; ACL: Stroudsburg, PA, USA, 2016; pp. 595–605, https://doi.org/10.18653/v1/D16-1057.
34. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76, https://doi.org/10.1109/JPROC.2020.3004555.
35. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), NAACL-HLT '19, Minneapolis, MN, USA, 2–7 June 2019; ACL: Stroudsburg, PA, USA, 2019; pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.
36. Hemmatian, F.; Sohrabi, M.K. A Survey on Classification Techniques for Opinion Mining and Sentiment Analysis. *Artif. Intell. Rev.* **2019**, *52*, 1495–1545, https://doi.org/10.1007/s10462-017-9599-6.
37. Hutto, C.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the International AAAI Conference on Weblogs and Social Media, ICWSM '14, Ann Arbor, MI, USA, 27–29 May 2014; AAAI: Menlo Park, CA, USA, 2014.
38. Loria, S. TextBlob: Simplified Text Processing. 2013. Available online: https://github.com/sloria/TextBlob (accessed on 15 July 2022).
39. Jindal, K.; Aron, R. A Systematic Study of Sentiment Analysis for Social Media Data. *Mater. Today Proc.* **2021**, https://doi.org/10.1016/j.matpr.2021.01.048.
40. Ay Karakuš, B.; Talo, M.; Hallaç, İ.R.; Aydin, G. Evaluating Deep Learning Models for Sentiment Classification. *Concurr. Comput.* **2018**, *30*, e4783, https://doi.org/10.1002/cpe.4783.
41. Mohammad, S.M. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In *Emotion Measurement*; Woodhead Publishing: Thorston, UK, 2016; pp. 201–237, https://doi.org/10.1016/B978-0-08-100508-8.00009-6.
42. Mohammad, S.M.; Sobhani, P.; Kiritchenko, S. Stance and Sentiment in Tweets. *ACM Trans. Internet Technol.* **2017**, *17*, 1–23, https://doi.org/10.1145/3003433.
43. Skeppstedt, M.; Simaki, V.; Paradis, C.; Kerren, A. Detection of Stance and Sentiment Modifiers in Political Blogs. In *Speech and Computer*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10458, pp. 302–311, https://doi.org/10.1007/978-3-319-66429-3_29.
44. Simaki, V.; Paradis, C.; Skeppstedt, M.; Sahlgren, M.; Kucher, K.; Kerren, A. Annotating Speaker Stance in Discourse: The Brexit Blog Corpus. *Corpus Linguist. Linguist. Theory* **2020**, *16*, 215–248, https://doi.org/10.1515/cllt-2016-0060.
45. Eisenstein, J. Unsupervised Learning for Lexicon-Based Classification. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 3188–3194.
46. Raza, H.; Faizan, M.; Hamza, A.; Mushtaq, A.; Akhtar, N. Scientific Text Sentiment Analysis Using Machine Learning Techniques. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 157–165, https://doi.org/10.14569/IJACSA.2019.0101222.

47. Abd El-Jawad, M.H.; Hodhod, R.; Omar, Y.M. Sentiment Analysis of Social Media Networks Using Machine Learning. In Proceedings of the International Computer Engineering Conference, ICENCO '18, Cairo, Egypt, 29–30 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 174–176, https://doi.org/ICENCO.2018.8636124.
48. Zhang, L.; Wang, S.; Liu, B. Deep Learning for Sentiment Analysis: A Survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, https://doi.org/10.1002/widm.1253.
49. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent Convolutional Neural Networks for Text Classification. In *Proc. AAAI Conf. Artif. Intell.* **2015**, *29*, 2267–2273, https://doi.org/10.1609/aaai.v29i1.9513.
50. Chen, W.F.; Ku, L.W. UTCNN: A Deep Learning Model of Stance Classification on Social Media Text. In Proceedings of the International Conference on Computational Linguistics—Technical Papers, COLING '16, Osaka, Japan, 11–16 December 2016; ACL: Stroudsburg, PA, USA, 2016; pp. 1635–1645.
51. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, EACL '17, Valencia, Spain, 3–7 April 2017; ACL: Stroudsburg, PA, USA, 2017, pp. 427–431.
52. Bender, E.M.; Koller, A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '20, Online, 5–10 July 2020; ACL: Stroudsburg, PA, USA, 2020; pp. 5185–5198, https://doi.org/10.18653/v1/2020.acl-main.463.
53. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Online, 3–10 March 2021; ACM: New York, NY, USA, 2021; pp. 610–623, https://doi.org/10.1145/3442188.3445922.
54. Desai, S.; Durrett, G. Calibration of Pre-trained Transformers. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '20, Online, 16–20 November 2020; ACL: Stroudsburg, PA, USA, 2020; pp. 295–302, https://doi.org/10.18653/v1/2020.emnlp-main.21.
55. Heitmann, M.; Siebert, C.; Hartmann, J.; Schamp, C. More Than a Feeling: Benchmarks for Sentiment Analysis Accuracy. *Soc. Sci. Res. Netw.* **2020**, *in press*, https://doi.org/10.2139/ssrn.3489963.
56. Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; Asokan, N. All You Need is "Love": Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, AISec '18, Toronto, ON, Canada, 19 October 2018; ACL: Stroudsburg, PA, USA, 2018; pp. 2–12, https://doi.org/10.1145/3270101.3270103.
57. Giachanou, A.; Crestani, F. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv.* **2016**, *49*, 1–41, https://doi.org/10.1145/2938640.
58. Santos, J.; Bernardini, F.; Paes, A. Measuring the Degree of Divergence when Labeling Tweets in the Electoral Scenario. In Proceedings of the Brazilian Workshop on Social Network Analysis and Mining, BraSNAM '21, Florianópolis, SC, Brazil, 18–22 July 2021; SBC: Porto Alegre, RS, Brazil, 2021; pp. 127–138, https://doi.org/10.5753/brasnam.2021.16131.
59. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the International Workshop on Semantic Evaluation, SemEval-2017, Vancouver, BC, Canada, 3–4 August 2017; ACL: Stroudsburg, PA, USA, 2017; pp. 502–518, https://doi.org/10.18653/v1/S17-2088.
60. Cieliebak, M.; Deriu, J.M.; Egger, D.; Uzdilli, F. A Twitter Corpus and Benchmark Resources for German Sentiment Analysis. In Proceedings of the International Workshop on Natural Language Processing for Social Media, SocialNLP '17, Valencia, Spain, 3 April 2017; ACL: Stroudsburg, PA, USA, 2017; pp. 45–51, https://doi.org/10.18653/v1/W17-1106.
61. Alexandridis, G.; Varlamis, I.; Korovesis, K.; Caridakis, G.; Tsantilas, P. A Survey on Sentiment Analysis and Opinion Mining in Greek Social Media. *Information* **2021**, *12*, 331, https://doi.org/10.3390/info12080331.
62. Rogers, A.; Romanov, A.; Rumshisky, A.; Volkova, S.; Gronas, M.; Gribov, A. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. In Proceedings of the International Conference on Computational Linguistics, COLING '18, Santa Fe, NM, USA, 20–26 August 2018; ACL: Stroudsburg, PA, USA, 2018; pp. 755–763.
63. Habernal, I.; Ptáček, T.; Steinberger, J. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA '13, Atlanta, GA, USA, 14 June 2013; ACL: Stroudsburg, PA, USA, 2013; pp. 65–74.
64. Elenius, K.; Forsbom, E.; Megyesi, B. Language Resources and Tools for Swedish: A Survey. In Proceedings of the International Conference on Language Resources and Evaluation, LREC '08, Marrakech, Morocco, 28–30 May 2008; ELRA: Paris, France, 2008.
65. Nivre, J.; Nilsson, J.; Hall, J. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In Proceedings of the International Conference on Language Resources and Evaluation, LREC '06, Genoa, Italy, 22–28 May 2006; ELRA: Paris, France, 2006.
66. Øvrelid, L.; Nivre, J. When Word Order and Part-of-speech Tags are not Enough—Swedish Dependency Parsing with Rich Linguistic Features. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '07, Borovets, Bulgaria, 27–29 September 2007.
67. Östling, R. Stagger: An Open-Source Part of Speech Tagger for Swedish. *North Eur. J. Lang. Technol.* **2013**, *3*, 1–18, https://doi.org/10.3384/nejlt.2000-1533.1331.
68. Ek, T.; Kirkegaard, C.; Jonsson, H.; Nugues, P. Named Entity Recognition for Short Text Messages. *Procedia-Soc. Behav. Sci.* **2011**, *27*, 178–187, https://doi.org/10.1016/j.sbspro.2011.10.596.

69. Skeppstedt, M.; Kvist, M.; Dalianis, H. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In Proceedings of the International Conference on Language Resources and Evaluation, LREC '12, Istanbul, Turkey, 23–25 May 2012; ELRA: Paris, France, 2012; pp. 1250–1257.
70. Lundberg, J.; Nordqvist, J.; Laitinen, M. Towards a Language Independent Twitter Bot Detector. In Proceedings of the Conference of the Association of Digital Humanities in the Nordic Countries. CEUR Workshop Proceedings, DHN '19, Copenhagen, Denmark, 6–8 March 2019; Volume 2364, pp. 308–319.
71. Rouces, J.; Borin, L.; Tahmasebi, N.; Rødven Eide, S. SenSALDO: A Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox. In *Proceedings of the Selected Papers from the CLARIN Annual Conference 2018*; Linköping University Electronic Press: Linköping, Sweden 2019; pp. 177–187.
72. Malmsten, M.; Börjeson, L.; Haffenden, C. Playing with Words at the National Library of Sweden—Making a Swedish BERT. *arXiv* **2020**, arXiv:2007.01658.
73. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), EMNLP-IJCNLP '19, Hong Kong, China, 3–7 November 2019; ACL: Stroudsburg, PA, USA, 2019; pp. 3982–3992, https://doi.org/10.18653/v1/D19-1410.
74. Reimers, N.; Gurevych, I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '20, Online, 16–20 November 2020; ACL: Stroudsburg, PA, USA, 2020; pp. 4512–4525, https://doi.org/10.18653/v1/2020.emnlp-main.365.
75. Rekathati, F. The KBLab Blog: Introducing a Swedish Sentence Transformer. 2022. Available online: https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/ (accessed on 15 July 2022).
76. Heidenreich, T.; Eberl, J.M.; Lind, F.; Boomgaarden, H. Political Migration Discourses on Social Media: A Comparative Perspective on Visibility and Sentiment Across Political Facebook Accounts in Europe. *J. Ethn. Migr. Stud.* **2020**, *46*, 1261–1280, https://doi.org/10.1080/1369183X.2019.1665990.
77. Berdicevskis, A. Svensk ABSAbank-Imm 1.0: An Annotated Swedish Corpus for Aspect-Based Sentiment Analysis (A Version of Absabank). 2021. Available online: https://spraakbanken.gu.se/resurser/absabank-imm (accessed on 15 July 2022).
78. Rouces, J.; Borin, L.; Tahmasebi, N. Tracking Attitudes Towards Immigration in Swedish Media. In Proceedings of the Conference of the Association of Digital Humanities in the Nordic Countries. CEUR Workshop Proceedings, DHN '19, Copenhagen, Denmark, 6–8 March 2019; Volume 2364, pp. 387–393.
79. Fernquist, J.; Lindholm, O.; Kaati, L.; Akrami, N. A Study on the Feasibility to Detect Hate Speech in Swedish. In Proceedings of the IEEE International Conference on Big Data, Big Data '19, Los Angeles, CA, USA, 9–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4724–4729, https://doi.org/10.1109/BigData47090.2019.9005534.
80. Borg, A.; Boldt, M. Using VADER Sentiment and SVM for Predicting Customer Response Sentiment. *Expert Syst. Appl.* **2020**, *162*, 113746, https://doi.org/10.1016/j.eswa.2020.113746.
81. Fernquist, J.; Kaati, L.; Schroeder, R. Political Bots and the Swedish General Election. In Proceedings of the IEEE International Conference on Intelligence and Security Informatics, Miami, FL, USA, 9–11 November 2018; ISI '18, IEEE: Piscataway, NJ, USA, 2018; pp. 124–129, https://doi.org/10.1109/ISI.2018.8587347.
82. Wickham, H. rvest: Easily Harvest (Scrape) Web Pages. 2016. Available online: https://cran.r-project.org/web/packages/rvest (accessed on 15 July 2022).
83. Artstein, R.; Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.* **2008**, *34*, 555–596, https://doi.org/10.1162/coli.07-034-R2.
84. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
85. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly: Sebastopol, CA, USA, 2009.
86. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830, https://doi.org/10.5555/1953048.2078195.
87. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, CA, USA, 13–17 August 2016; ACL: Stroudsburg, PA, USA, 2016; pp. 785–794, https://doi.org/10.1145/2939672.2939785.
88. Schwenk, H.; Douze, M. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In Proceedings of the Workshop on Representation Learning for NLP, RepL4NLP '17, Vancouver, BC, Canada, 3 August 2017; ACL: Stroudsburg, PA, USA, 2017; pp. 157–167, https://doi.org/10.18653/v1/W17-2619.
89. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In Proceedings of the USENIX Symposium on Operating Systems Design and Implementation, OSDI '16, Savannah, GA, USA, 2–4 November 2016; USENIX Association: Berkeley, CA, USA 2016; pp. 265–283.

90. Yang, Y.; Cer, D.; Ahmad, A.; Guo, M.; Law, J.; Constant, N.; Hernandez Abrego, G.; Yuan, S.; Tar, C.; Sung, Y.h.; et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. In Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL '20, Online, 5–10 July 2020; ACL: Stroudsburg, PA, USA, 2020; pp. 87–94, https://doi.org/10.18653/v1/2020.acl-demos.12.
91. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Sydney, Australia, 2019; Volume 32.
92. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP '20, Online, 16–20 November 2020; ACL: Stroudsburg, PA, USA, 2020; pp. 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6.
93. Zheng, Z.; Wu, X.; Srihari, R. Feature Selection for Text Categorization on Imbalanced Data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 80–89, https://doi.org/10.1145/1007730.1007741.
94. MacAvaney, S.; Yao, H.R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate Speech Detection: Challenges and Solutions. *PLoS ONE* **2019**, *14*, e0221152, https://doi.org/10.1371/journal.pone.0221152.
95. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437, https://doi.org/10.1016/j.ipm.2009.03.002.
96. Plank, B.; Hovy, D.; Søgaard, A. Linguistically Debatable or Just Plain Wrong? In Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL '14, Baltimore, MD, USA, 22–27 June 2014; ACL: Stroudsburg, PA, USA, pp. 507–511, https://doi.org/10.3115/v1/P14-2083.
97. Ribeiro, M.T.; Wu, T.; Guestrin, C.; Singh, S. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '20, Online, 5–10 July 2020; ACL: Stroudsburg, PA, USA, 2020; pp. 4902–4912, https://doi.org/10.18653/v1/2020.acl-main.442.
98. Chen, S.; Lin, L.; Yuan, X. Social Media Visual Analytics. *Comput. Graph. Forum* **2017**, *36*, 563–587, https://doi.org/10.1111/cgf.13211.
99. Kucher, K.; Paradis, C.; Kerren, A. The State of the Art in Sentiment Visualization. *Comput. Graph. Forum* **2018**, *37*, 71–96, https://doi.org/10.1111/cgf.13217.
100. Alharbi, M.; Laramee, R.S. SoS TextVis: An Extended Survey of Surveys on Text Visualization. *Computers* **2019**, *8*, 17, https://doi.org/10.3390/computers8010017.
101. Baumer, E.P.S.; Jasim, M.; Sarvghad, A.; Mahyar, N. Of Course It's Political! A Critical Inquiry into Underemphasized Dimensions in Civic Text Visualization. *Comput. Graph. Forum* **2022**, *41*, 1–14, https://doi.org/10.1111/cgf.14518.
102. Chatzimparmpas, A.; Martins, R.M.; Jusufi, I.; Kucher, K.; Rossi, F.; Kerren, A. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Comput. Graph. Forum* **2020**, *39*, 713–756, https://doi.org/10.1111/cgf.14034.
103. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining Explanations: An Overview of Interpretability of Machine Learning. In Proceedings of the IEEE International Conference on Data Science and Advanced Analytics, DSAA '18, Turin, Italy, 1–3 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 80–89, https://doi.org/10.1109/DSAA.2018.00018.
104. Clinciu, M.A.; Hastie, H. A Survey of Explainable AI Terminology. In Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, NL4XAI '19, Tokyo, Japan, 29 October–1 November 2019; ACL: Stroudsburg, PA, USA, pp. 8–13, https://doi.org/10.18653/v1/W19-8403.
105. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-Agnostic Interpretability of Machine Learning. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, WHI '16, New York, NY, USA, 23 June 2016; https://doi.org/10.48550/ARXIV.1606.05386.
106. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215, https://doi.org/10.1038/s42256-019-0048-x.
107. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. In Proceedings of the International Conference on Language Resources and Evaluation, LREC '18, Miyazaki, Japan, 7–12 May 2018; ELRA: Paris, France, 2018.
108. Adewumi, T.P.; Liwicki, F.; Liwicki, M. Exploring Swedish & English fastText Embeddings for NER with the Transformer. *arXiv* **2021**, arXiv:2007.16007.
109. Aroyo, L.; Welty, C. Truth is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Mag.* **2015**, *36*, 15–24, https://doi.org/10.1609/aimag.v36i1.2564.