applied sciences

MDPI

Article

Automated Classification of Evidence of Respect in the Communication through Twitter

Krzysztof Fiok [1], Waldemar Karwowski [1], Edgar Gutierrez [1,2,*], Tameika Liciaga [1], Alessandro Belmonte [1] and Rocco Capobianco [1]

[1] Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL 32816, USA; fiok@ucf.edu (K.F.); wkar@ucf.edu (W.K.); tameikaliciaga22@knights.ucf.edu (T.L.); alebelmonte@knights.ucf.edu (A.B.); rncapobianco.95@knights.ucf.edu (R.C.)

[2] Center for Latin-American Logistics Innovation, LOGyCA, Bogota 110111, Colombia

* Correspondence: edgar.gutierezfranco@ucf.edu

Abstract: Volcanoes of hate and disrespect erupt in societies often not without fatal consequences. To address this negative phenomenon scientists struggled to understand and analyze its roots and language expressions described as hate speech. As a result, it is now possible to automatically detect and counter hate speech in textual data spreading rapidly, for example, in social media. However, recently another approach to tackling the roots of disrespect was proposed, it is based on the concept of promoting positive behavior instead of only penalizing hate and disrespect. In our study, we followed this approach and discovered that it is hard to find any textual data sets or studies discussing automatic detection regarding respectful behaviors and their textual expressions. Therefore, we decided to contribute probably one of the first human-annotated data sets which allows for supervised training of text analysis methods for automatic detection of respectful messages. By choosing a data set of tweets which already possessed sentiment annotations we were also able to discuss the correlation of sentiment and respect. Finally, we provide a comparison of recent machine and deep learning text analysis methods and their performance which allowed us to demonstrate that automatic detection of respectful messages in social media is feasible.

Keywords: respect; natural language processing; sentiment analysis; disrespect; twitter; machine learning

Citation: Fiok, K.; Karwowski, W.; Gutierrez, E.; Liciaga, T.; Belmonte, A.; Capobianco, R. Automated Classification of Evidence of Respect in the Communication through Twitter. *Appl. Sci.* **2021**, *11*, 1294. https://doi.org/10.3390/app11031294

Received: 11 January 2021
Accepted: 26 January 2021
Published: 1 February 2021

1. Introduction

1.1. Background

Treating every person with respect [1] seems to be a timeless commandment that everyone would readily agree upon. Unfortunately, this commandment is not practiced in many societies, groups, and enterprises. Outbreaks of disrespectful human behavior are witnessed regularly, especially on social media which significantly influence emotions in humans [2]. As a result, numerous researchers have begun to address the problem of hate speech propagated via micro-blogging platforms such as Twitter [3–7]. Because hate is not limited to any concept or language, diverse studies have addressed hate expressed toward specific topics, such as sexism [8], racism [3,5–7], nationalism [9], and immigration [10] in English, as well as in other languages [8,9,11]. Various entities have attempted to mitigate the negative effects of hate speech; for example, the United Nations and European Union have their own strategy of addressing hate speech [12,13]. In contrast, the United States Navy has proposed to address the problem from another direction, by strengthening positive behavior. Accordingly, the Navy has published a list of signature behaviors that 21st-century sailors should exhibit [1], in which "treating every person with respect" is placed first. Interestingly, this approach seems to be novel for studies focused on micro-blogging language analysis, because studies aimed at identifying the positive signature behaviors of social media users, with a focus on the use of respectful language or expressions of respect, are difficult to find. After conducting a search on

Google Scholar [14] with queries including "expressing admiration Twitter," "expressing respect Twitter," "language of respect Twitter," "respectful language Twitter," "expressing appreciation Twitter," "appreciation on Twitter," and "polite language Twitter," we were able to identify only a few relevant Twitter-related studies addressing admiration in sports [15,16], compliments for celebrities [17], self-organization of a racial minority group [18], politeness strategies [19], polite language style [20] and gender-dependent language style [21,22]. From the above findings, we hypothesized that the use of polite, respectful language and expressions of respect has been widely discussed in many areas and contexts, although not yet regarding Twitter.

We believe that introduction of a first data set of a new kind is a significant contribution. When new data is presented, it is often the case that data owners demonstrate possible use. In our case, we demonstrate that this data set can be used for training of automated classification methods from the Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) domains. We believe that comparing the performance of 14 models is a significant contribution, as other authors do not compare so many models in a single study. We also believe that demonstrating the correlation of respect to sentiment is crucial, indicating that these two notions are not the same. However, they are somewhat connected according to the obtained results.

1.2. Our Focus and Related Research

Our study was aimed to address the use of respectful language and expressions of respect on Twitter and to demonstrate that whether a person is exhibiting a positive signature behavior can be assessed based on textual data through automated text analysis.

The most closely related research to our study is probably that in [23], which has addressed the problem of assessing respect in utterances of officers on duty. The study utilized a hand-annotated sample of 414 data instances to perform regression analysis on the influence of chosen linguistic features on the respectfulness of the analyzed utterance. Furthermore, the model was used to assign a "respect score" to previously unseen data instances in accordance with phrases found in analyzed sentences. That is, a lexicon-based analysis with a regression model was used to solve a regression task of assigning a "respect score." The advantage of this method is its transparency because it allows for easy demonstration of which linguistic features contribute to the "respect score" at the instance level.

1.3. Defining Respect

Respectful language can be defined in many ways, depending on the context, the persons involved in the context, or the domain. Accordingly, understanding how prior researchers analyzing "respect" have approached the topic of inquiry is important. The main questions that academics and philosophers have asked about respect include how respect should be understood at a general level. Most researchers in the field have identified the concept of respect in various ways, including as a style of conduct, an attitude, a feeling, a right, or a moral virtue [24]. The concept of respect has always had important relevance to people's daily lives because people almost universally live together in social groups. Humans are called upon to give respect in various value paradigms, for example, human life; members of minority racial and ethnic groups; those discriminated against on the basis of gender, sexual orientation, age, religious beliefs, or economic status; the respect for nature urged by environmentalists; and the respect demanded in recognizing some people as social and moral equals and appreciating their cultural differences [25,26]. Academics interested in this matter have widely recognized the existence of different types of respect. For example, the relative ideas regarding respect may differ significantly from other ideas, given the particular context of society, culture, religion, and age [27,28].

Respect as a concept has also been highlighted in discussions of justice and equality, injustice and duties, moral motivation and development, cultural variety and tolerance, punishment, and political violence. According to [29], interest in respect has focused mainly

on respect for people, that is, for others who live throughout the world, and therefore regarding differences in religious and cultural beliefs. Thus, the idea that all people should be treated with respect has become more refined: all people should be treated respectfully simply because they are people. Duty and the associated moral approach can be traced to the philosopher Kant, who said that all people are appropriate objects of the most morally significant attitude of respect [30].

However, although most humans recognize the importance of respect and the idea of a moral and political sentiment, owing to the specific actions of people and societies, agreement is lacking regarding issues such as how the concept of respect should be understood, and who or what the appropriate objects of respect should be [31]. As a consequence, the attitude of respect is important to discuss. The attitude of respect necessarily has an object: respect is always directed, felt, or shown to some object or person. Although a wide variety of objects can be appropriate for one type of respect or another (such as the flag, a statue, or a symbol), the subject of respect (the respecter) is nonetheless always a human, that is, a capable conscious rational being who can acknowledge and respond intentionally, who has and expresses values with respect to the object, and who is responsible for bearing respectful or disrespectful attitudes [32].

Hudson [33] has proposed four kinds of respect. (1) We can respect people for their personalities or their work, e.g., respecting a colleague as an academic and/or having respect for someone with "guts." (2) We can respect people for their achievements, e.g., having respect for a professional swimmer or a soccer player having respect for the goalkeeper of the opposing team. (3) We can respect the terms of an agreement and the rights of a person. Finally, (4) we can show respect for people symbolically; e.g., when a judge enters a room, people stand up. To the original classification by Hudson, Dillon [32] has added a fifth form, respect for care, which involves considering that the object has a deep and perhaps unique value, and therefore appreciating it and perceiving it as fragile or requiring special care; as a result, we choose either to act or to refrain from acting, owing to the benevolent concern that we feel for the object.

People can be recipients of different forms of respect. One can discuss the legal rights of a person and respect those rights; one can show respect for the institution that a person represents, e.g., for a president by calling her "Ms. President," or by respecting someone for being committed to a worthy project; and one can accord a person the same basic moral respect that most humans believe anyone deserves. Because a variety of possible dimensions exist, the idea of respect for people remains somewhat vague.

Nonetheless, respectful language, which is defined on the basis of the concept of showing respect, that is, how a person behaves in a respectful way regarding others, is a key component. Authors such as Chapman have linked respectful language with "professional language," [34] which is dependent on the skills and level of education of the person who is speaking or writing. According to the broad discussions presented in different studies of respectful language, the concept also depends on the intellectual characteristics of the person. Chapman argues that using respectful language encourages people to take responsibility for what they say or write because words are an expression of a person's personality.

This definition clearly identifies the respectful behavior with the person's psychological expression and their emotions and sentiments related to a topic, and thus with the degree of empathy that can be shown to a topic or person. Thus, Chapman states the following regarding empathy: "Empathy requires intentional thinking, the recognition that other people's feelings and circumstances are separate from our own, and a willingness to act appropriately in response to these. Respectful language, therefore, begins with an intention to respond to what others want. Showing respect does not involve benevolence, guesswork, or simply giving what we are comfortable within professional conversations" [34]. Other synonyms for the word "respectful" are:

- Courteous: offering courtesy and exhibiting gracious good manners;
- Humble: having modesty, lacking in arrogance and pride;

- Honorific: showing honor or respect; and
- Reverent: feeling or showing profound respect or veneration.

After studying various definitions and authors' opinions regarding the arguments presented to define respect, we have found that the expression of respect depends on characteristics inherent to the person issuing the message. That is, the context in which the expression, whether oral or written, is given determines the degree of respect within the message. Characteristics such as language, culture, political vision, religion, and even the use of sarcasm influence the subjective perception of respect. Here, we present a variety of contexts in which respect and respectful expression toward one or more people have been defined.

Holtgraves discusses respectful language by considering social psychology [35]. Others, such as Thompson [36] and Wolf [37], have studied the concept of respect from the point of view of politics and politicians, by examining the negative implications that the use of disrespectful language can have for a community, commensurate with political trends. In addition, implications can exist at the national and international levels.

Regarding the use of respectful language in the treatment of customers, such as in medicine, Beach [38] has discussed how a professional must behave toward workers, co-workers, clients, and patients (in the case of a medical doctor). The definition of respect in the context of medicine is "recognition of the unconditional value of patients as persons" [38].

In our study, we sought to address Twitter data by considering language specifics. Unfortunately, this type of data lacks most of the context or knowledge regarding the person expressing the statement, thus complicating the task of deciding whether a given tweet is respectful. The lack of knowledge regarding authors and the context of a given tweet is challenging, especially given other authors' findings regarding definitions of respect and generally how respect can be perceived.

1.4. Relationship between Sentiment and Expression of Respect

As our research project proceeded, and we prepared data for our experiments, we often asked what the relationship might be between the sentiment of a tweet and its respectfulness, i.e., are we, in fact, analyzing the same thing but merely calling it a different name? To discuss the relationship between the two notions in the context of Twitter, we selected the same set of tweets that had already been hand-annotated for Twitter-sentiment analysis in [39].

1.5. Our Contribution

To the best of our knowledge, this study contributes:

1. A new data set of tweets that, to the best of our knowledge, is the first open data set annotated with a focus on the expression of respect,
2. A comparison of 14 selected approaches from the fields of deep learning, natural language processing, and machine learning used for the extraction of features and classification of tweet respectfulness in the new data set,
3. Analysis of the correlation between tweet sentiment and respectfulness to answer the two questions of whether positive tweets are always respectful and whether negative tweets are always disrespectful, and
4. Finally, to enable full reproducibility of our experiments, we openly publish our data and code.

2. Methods

2.1. Analyzed Data and Annotation Scheme

In our study, we focused on the detection of respectful tweets. Because of the aforementioned lack of knowledge regarding the author or context of a given tweet, deciding whether a tweet is respectful is not an easy task. Thus, we decided to accept the subjective judgment of what is respectful, as perceived by annotators who were employed to label the

analyzed data. Importantly, in this context, our annotators were prepared for the labeling task by participating in a literature review regarding the task of "defining respect." During this preparation process, we agreed that the definition of respectfulness could be easily understood from two different points of view. First, the definition can be understood as a straightforward use of words that the reader or listener believes are respectful (such as "Mr. President, can you please give some details about the executive order?") or as expressing respect, which might include the use of "bad" words but is nonetheless an expression of respect (such as, "you are a badass at making money"). For this project, we adopted the latter type of example, and we interchangeably refer to it within our work as either "respectfulness" or "expressing respect." It may seem obvious that a text entity could be considered to exhibit a range of respectful sentiment; however, studies elaborating on annotation quality in hate speech analysis [6,40] have demonstrated that, in many cases, humans do not agree on whether a single tweet should be considered racist, and prior studies have rarely gone into detail regarding whether a tweet is "more or less racist." Similarly, we believe that the same tweet could be perceived as either "respectful" or "disrespectful" by different people and that increasing granularity of the notion of respectfulness causes additional complications. Our data preparation process was consistent with the above statement. Because we observed complications related to annotator agreement before our final annotation process, we held a series of meetings within the annotator group to improve our common understanding and definition of the "respectfulness" of text. Ultimately, we decided to simplify the annotation task by not providing regression-like "respect scores" as in [23] but instead creating classification labels with only three "respect" classes, to limit confusion (Table 1).

Table 1. Annotation scheme adopted in the study.

Label Name	Label	Tweet Description
Disrespectful	0	Is aggressive and/or strongly impolite, seems "evidently" disrespectful.
Respectful	1	Tweets that are certainly not disrespectful are written in "standard" language without any evidently negative or positive attitudes. If it is unclear whether the tweet should be considered very respectful or respectful, the tweet is labeled respectful.
Very respectful	2	Undoubtedly exhibits respect.

The adopted set of 5000 tweets had already been released in [39]. When working with the data set, we detected and erased 36 tweets that were not written 100% in English and then labeled the remaining 4964 tweets according to the defined annotation scheme and the following procedure: (1) three annotators independently annotated each data instance; (2) we computed Krippendorf's alpha annotator agreement of the resulting labels; (3) to obtain a single respectfulness label for each data instance, we retained the label if all three annotators were in agreement (3102 tweets or adopted the label assigned by the majority of annotators (1833 tweets) if the difference between the labels was not greater than 1. In the other cases (29 tweets), a new single annotator was asked to decide and provide a decisive annotation. The data regarding users is anonymized, i.e., it is impossible to connect the given tweets to users automatically.

The annotation procedure resulted in 849 disrespectful, 3730 respectful, and 385 very respectful data instances consisting only of text (no tweet-related meta data were used). Example data instances were (original spelling): "Rest in peace, shipmates.A senseless tragedy, but know your service was not in vain..you made a difference.Condolences to your families." and "@USNavy Surgeon Thomas Harris was born on this day in 1784. @NavyHistoryNews #Medical #History _URL".

Generally, the greater the classification granularity of the assessed concept, the more difficult it is to maintain the annotations' quality. Annotators are always subjective, and it is not easy to maintain the common perception of a given phenomenon even with three classes. This is also one of the first studies in the domain of classifying respect and in a similar domain, i.e., sentiment analysis. In the beginning, many authors agreed that three classes of sentiment were sufficient.

We are aware of the problem of using limited amount of data in our study; however, it cannot be resolved in this case as, to the best of our knowledge, along with this study, we are publishing the first data set with respectfulness labels that allow predictions on message level.

2.2. Annotation Correlations

Owing to the choice of data set, we were able to compute correlations between annotators regarding the respectfulness of the tweet and the sentiment label already existing in the data set. Because the original sentiment labels were divided into five classes (0, very negative; 1, negative; 2, neutral; 3, positive; and 4, very positive), for the purpose of computing a correlation with our three-class respectfulness labels, we modified the original sentiment labels in the following manner: very negative and negative classes were treated as class 0; neutral tweets were treated as class 1; and positive and very positive tweets were treated as class 2.

2.3. Feature Extraction Methods

Automatic classification of text is feasible with machine learning (ML), deep learning (DL), and natural language processing (NLP) methods. ML classifiers are algorithms that operate on features extracted from data instances to perform predictions. In this context, various classifiers can be used, e.g., gradient boosting [41], random forest [42], or support vector machines [43].

There are many use cases that mix DL, ML, and NLP methods in various domains, such as, for example, stock data mining [44], online opinion mining [45], or sentiment analysis [46]. In text classification, researchers in the NLP field have provided methods for preprocessing text instances, including the extraction of features on the basis of: (a) n-grams, (b) token occurrences such as bag of words (BOW), term frequency (TF), inverse document frequency (IDF), and (c) lexicon-based methods such as LIWC [47]. SÉANCE [48]. Developments in the area of NLP enabled word or token-level embeddings (i.e., vector representations of text) obtained by trainable models such as Glove [49] and Word2vec [50], which were later followed by a family of other models capable of creating embeddings on various text granularity levels, such as the character level, sub-word level, or token level. To obtain sentence-level or document-level feature vectors from embeddings that correspond to smaller-text entities, various strategies have been proposed, with simple pooling (i.e., averaging of embeddings that belong to the larger text entity) as probably one of the first naive approaches. However, for some time, DL convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated superiority in the conversion of sub-entity embeddings over simple pooling. Yet, with the introduction of transformer model architecture [50] and the famous model "pre-training of deep bidirectional transformers for language understanding" (BERT) [51,52], researchers have achieved new quality levels when creating text representations. Specifically, for text classification, transformer model architectures most often allow for embeddings to be obtained for the whole analyzed text entity, without any intermediate steps, via the so-called "classification token CLS." In brief, the transformer model attention and self-attention mechanisms are prepared to create high-quality entity-level embeddings while the whole model is being pre-trained. Sub-entity-level embeddings can also be obtained from transformer models and can then be further converted into entity-level representations by the previously mentioned pooling or CNNs or RNNs; however, such an approach has inferior performance [50].

Here, we use a selection from the described methods for obtaining tweet-level feature representations to compare their performance on the task of the prediction of respectfulness in tweets. We focus on recent models from the transformer family that could be fine-tuned for our specific classification task to provide the highest possible quality of tweet-level vector representations. In addition, we demonstrate the performance of an LSTM responsible for creating tweet-level vector representations from token-level vector representations obtained from selected language models. We also present the performance that can be achieved by selected pre-trained DL language models used to create vector representations for each token without any data-specific training and then averaging these embeddings to provide a tweet-level vector representation.

The rationale for not using the GRU networks was that they are comparable to or slightly inferior to LSTMs, especially if bidirectional LSTM models are concerned, as in our study. An example of a renowned comparison study in this regard can be found in [53].

Finally, we demonstrate the models by utilizing features extracted in accordance with the known LIWC [47] and SÉANCE [48] lexicons. The complete list of the tested feature extraction methods is presented in Table 2. From all demonstrated feature extraction methods, some do not require any data-specific training, and others do. Therefore, we believe the easier to adapt are the ones that do not require data-specific training. Unfortunately, at the same time, these methods provide poorer results. If not specified otherwise, the pretrained models were downloaded with the Transformers [54] and Flair [55] Python modules.

Table 2. List of the feature extraction methods used in our study.

Method Name Adopted in This Study	Additional Description	Data-Specific Adaptability	Method for Obtaining Tweet-Level Embeddings	Source
Term Frequency	Top 300 features selected according to a mutual information method implemented in the Python sci-kit learn module	Data-specific training required	Native output of features for the whole text data instance	[56]
SEANCE	Lexicon-based method, "Sentiment analysis and social cognition engine"	No data-specific training	Native output of features for the whole text data instance	[48]
LIWC	Lexicon-based method, "Linguistic inquiry and word count"	No data-specific training	Native output of features for the whole text data instance	[47]
Albert Pooled	Tiny version of BERT, model version "base-v2pooled"	No data-specific training	Mean of token embeddings	[57]
Distilbert Pooled	"Distilled" [58] version of BERT pre-trained to output sentence-level embeddings, model version "base-nli-stsb-mean-tokenspooled"	No data-specific training	Mean of token embeddings	[59]
Roberta Pooled	Robustly pre-trained BERT ready to output sentence-level embeddings, model version "roberta-large-nli-stsb-mean-tokenspooled"	No data-specific training	Mean of token embeddings	[59]
Fasttext LSTM	Token embeddings from Fasttext, model version "en-crawl" converted by an LSTM into tweet-level embeddings	Data-specific training required	Bidirectional LSTM	[60]

Table 2. *Cont.*

Method Name Adopted in This Study	Additional Description	Data-Specific Adaptability	Method for Obtaining Tweet-Level Embeddings	Source
RoBERTa LSTM	Token embeddings from robustly pretrained BERT "roberta-large" model version converted by an LSTM into tweet-level embeddings	Data-specific training required	Bidirectional LSTM	[61]
Albert	Fine-tuned tiny version of BERT transformer model, version "base-v2"	Data-specific training required	CLS token	[57]
BERT L C	Fine-tuned "BERT large cased" transformer model	Data-specific training required	CLS token	[54]
BERT L UNC	Fine-tuned "BERT large uncased" transformer model	Data-specific training required	CLS token	[54]
XLM-MLM-EN-2048	Fine-tuned cross-lingual transformer model "MLM-EN-2048"	Data-specific training required	CLS token	[62]
XLM-RoBERTa-L	Fine-tuned cross-lingual transformer model version based on Robustly Pretrained BERT large	Data-specific training required	CLS token	[62]
RoBERTa L	Fine-tuned transformer model version Robustly Pretrained BERT large	Data-specific training required	CLS token	[61]

2.3.1. Configuration of Models Which Used LSTMs

Proper configuration of ML and DL models requires experiments and studies. In order to configure model parameters in this study, we based them on past research [63] and our experience in the field.

When LSTMs were used to obtain entity-level feature vectors from embeddings that corresponded to sub-entities, we were able to adopt the following LSTM hyperparameters: the LSTM analyzed the sequence of embeddings from the beginning of the tweet until its end and then in the opposite direction (bidirectional) = true, number of LSTM layers = 2, and size of embedding corresponding to the tweet created by the LSTM (hidden size) = 512. To train the LSTMs, the following parameters were used: initial learning rate = 0.1, minimal learning rate = 0.002, factor at which the learning rate was decreased after training with the given learning rate was ended (anneal factor) = 0.5, and number of epochs without improving the loss on the validation set for which training was continued with the given learning rate (patience) = 20. In addition, the data instances were shuffled before each training epoch. Other parameters were set to the default values proposed in the Flair module.

2.3.2. Configuration of Fine-Tuned Models

To fine-tune the transformer models, we utilized the following parameters: Adam optimizer (as implemented in the PyTorch Python module), learning rate = 3×10^{-6}, number of data instances provided in the model input during a single training pass (i.e., mini-batch size) = 8, and number of times that the training procedure ran over the whole data set (i.e., epochs) = 4. In addition, the data instances were shuffled before each training epoch. Other parameters were set to the default values proposed in the Flair module.

2.4. Cross-Validation

During machine learning classification, all experiments in our study were five-fold cross-validated, and the whole data set was divided into proportions of 80% and 20% for the training and test sets, respectively. In addition, all feature extraction methods that required training on our data set were five-fold cross-validated. In this case, the test sets were ensured to remain the same both during training of the feature extraction method and later during machine learning classification.

2.5. Machine Learning Classification

For the final machine learning classification of the respectfulness of tweets, we utilized the gradient boosting classifier implemented in the xgboost Python module version 1.2.0. For training the classifiers, we adopted the following parameters: number of gradient boosted trees (n_estimators) = 250, training objective = multi:softprob (multi-class classification with results for each data point belonging to each class), and learning rate = 0.03. For a detailed list of all parameters, please refer to the code repository [64].

Classification Metrics

Assessment of the quality of the trained classification models was performed with F1 macro (F1) and Matthews correlation coefficient (MCC) scores. The F1 score was chosen because of its popularity in the ML community. MCC values were also selected because this metric is known to provide more reliable results for unbalanced data sets [64], as was clearly relevant to the dataset used in this work. Therefore, in our study, we treated MCC scores as the decisive quality metric. Classification metrics were computed once for a list of predictions created from the sub-lists of predictions obtained for the test sets from each cross-validated fold.

2.6. Software, Code, and Computing Machine

The computations required to perform this study were conducted on a single computing machine equipped with a single NVIDIA GPU model Titan RTX 24 GB RAM. All experiments were implemented in Python3, and the corresponding code and data set are available in [64]. For ease of reproduction, all experiments can be repeated by executing a single bash script. Most elements of the experimental pipeline responsible for the feature extraction methods were implemented with Flair module version 0.6 post1. A precise description of the software requirements is available in the project repository [64].

3. Results

3.1. Relationship between Respectfulness and Sentiment

As demonstrated in Figure 1, annotators 2 and 3 (AN2 and AN3) exhibited the highest correlation of the proposed respectfulness labels. The respectfulness-sentiment correlations per annotator ranged from 0.49 to 0.62. After creation of the final unified respectfulness label, the overall correlation between sentiment and respectfulness was 0.594.

3.2. Comparison of Classification Performance

Training of machine learning classifiers with the features provided by the methods described in Table 2 resulted in predictions that allowed us to compute appropriate quality metrics, which are presented in Table 3. In Table 4 we also demonstrate a worst-best model comparison by means of presenting a confusion matrix for Term Frequency and RoBERTa L models.

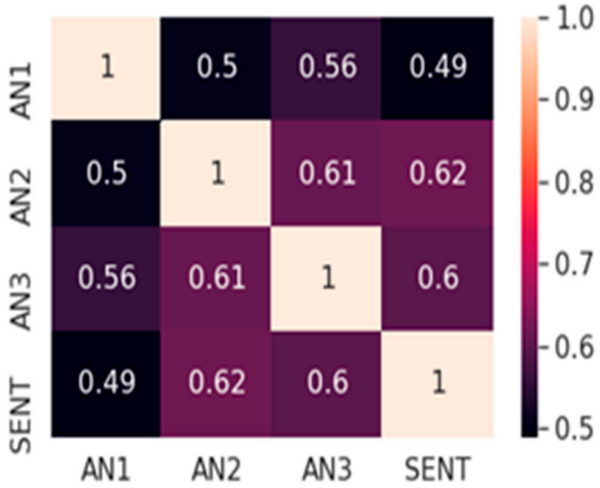

Figure 1. Matrix of correlations between annotators regarding respectfulness of data instances ("AN1," "AN2," and "AN3") and sentiment ("SENT") labels provided in the original data set.

Table 3. F1 macro and MCC scores of the compared models.

	Metric	
Model	**F1**	**MCC**
Term Frequency	0.5800	0.4049
SEANCE	0.6041	0.4301
LIWC	0.6150	0.4537
Albert Pooled	0.6387	0.5003
Distilbert Pooled	0.6397	0.5037
Fasttext LSTM	0.6827	0.5271
Roberta Pooled	0.6523	0.5331
Albert	0.7136	0.5871
BERT L C	0.7165	0.5902
XLM-MLM-EN-2048	0.7263	0.6062
BERT L UNC	0.7253	0.6062
RoBERTa LSTM	0.7240	0.6104
XLM-RoBERTa-L	0.7431	0.6249
RoBERTa L	0.7350	0.6337

Table 4. Confusion matrix for the worst- and best-performing classification models based on features from the term frequency and Roberta large fine-tuned (RoBERTa) feature extraction methods. Numbers corresponding to class labels are: 0, disrespectful; 1, respectful; and 2, very respectful. In the best-performing model; no data instances were misclassified between the "disrespectful" and "very respectful" classes.

		Respect Class		
		0	1	2
Term Frequency	0	258	585	6
	1	87	3584	59
	2	10	251	124
		0	1	2
RoBERTa L	0	625	224	0
	1	167	3473	90
	2	0	212	173

4. Discussion

4.1. Relationship between Respectfulness and Sentiment

The computed correlation based on the respectfulness and sentiment labels was 0.594, thus necessitating subsequent interpretation and consideration. The respectfulness-sentiment correlations computed per annotator ranged from 0.49 to 0.62; therefore, for various people, the differentiation between respectfulness and sentiment varies. When considering the final unified across-annotators respectfulness labels, we found that 2934 of 4964 data instances were assigned the same respectfulness and sentiment classes. However, inspection of the per-class details of the created data set revealed that disrespectful tweets were considered negative 73.73% of the time, respectful tweets were considered neutral in 51.9% of instances, and very respectful tweets were positive in 96.62% of cases. Therefore, we conclude that, for our annotators, very respectful tweets were almost always found to be positive, and disrespectful tweets were likely to be found negative.

When responding specifically to the previously defined research question of whether positive tweets are always respectful, we observed that, according to our annotators, in most cases, such tweets were considered either respectful or very respectful. In our data set, there were 1875 positive tweets, only 28 of which (1.49%) were simultaneously considered to be disrespectful. An example of this small group can be represented by the tweet: "Holy Christ, are you going to New Orleans to help with Katrina survivors next?" This tweet, although labeled as very positive, also received a disrespectful annotation. In the given case, one could question why this tweet was labeled disrespectful. When labeling for respect, the annotators agreed that this was a sarcastic tweet that in fact suggests that "you shouldn't be going there," whereas presumably earlier in the sentiment annotation process, the annotators perceived this tweet to be straightforward and non-sarcastic. This example demonstrates how tricky tweet-level classification can be if sarcastic language is introduced without the corresponding broader context.

To answer the second research question of whether a negative tweet is always disrespectful, we begin with the observation that there were 947 negative tweets in the data set. From this group, the majority (626 or 66.1%) were considered disrespectful; however, the remainder were considered respectful. Therefore, negative tweets are not always disrespectful. Some example tweets that can provide background for this conclusion are "The link isn't working" (obviously negative, but not disrespectful) or "@USNavy I do not accept any culpability, blame or responsibility." The latter was labeled as negative and yet is obviously not disrespectful.

4.2. General View of Classification Performance

Table 3 displays the differences in the classification quality of the gradient boosting classifiers, depending on the provided independent variables. The order of quality of the obtained MCC and F1 results mimics the historical development of the feature extraction methods that were described briefly in Section 2.3, i.e., simple Term Frequency and Lexicon-based methods provide the lowest quality, whereas fine-tuning of the recent transformer model RoBERTa large provides the highest quality. In general, from Table 3, we conclude that fine-tuning the recent transformer models, including the tiny Albert model, allowed us to obtain significantly higher MCC scores than those with the other feature extraction approaches. In our study, the exception to this rule was apparent when methods using LSTMs were considered. The quality of the feature extraction models that utilize bidirectional LSTMs to produce tweet-level embeddings strongly depends on the token-level embedding quality. If simpler fasttext embeddings are used, then the trained fasttex+LSTM embedding method can be surpassed by a pre-trained but more recent token embedding (roberta-large-nli-stsb-mean-tokenspooled). In addition, if a bidirectional LSTM is provided with the highest quality embeddings from the RoBERTa large model, then the resulting prediction quality can even surpass that of some of the fine-tuned transformer models.

4.3. Worst-Best Model Comparisons

The best-performing model based on the fine-tuned RoBERTa large feature extractor achieved an MCC score of 0.6337, whereas the worst term frequency method provided features that allowed for an MCC score of only 0.4049.

A detailed view of how the worst- and best-performing models performed the given prediction task is displayed in the confusion matrices in Table 4. The main difference in model quality lies in model error regarding minority class 2 (very respectful); that is, the best-performing model did not confuse a very respectful tweet with a disrespectful tweet even once. However, interestingly, the best model more frequently (167 errors) incorrectly predicted disrespectful tweets as being respectful than did the worst model (87 errors).

In addition, a critic might note that the worst model, based on features extracted by Term Frequency method, was nonetheless able to correctly predict 79.9% of data instances, whereas the best model achieved 86.0%. From this perspective, the difference between the compared models seems minimal and thus might incorrectly suggest that model choice is not significant. This line of thinking can be readily dismissed because significant class imbalance was present in the data set; as a result, naively assigning the majority class for each data instance would allow for 75.1% accuracy. Thus, an appropriate quality metric such as MCC should be utilized; moreover, the actual model use-case matters most. In the given example, not mistaking disrespectful tweets for very respectful tweets is probably more important than not confusing disrespectful tweets with respectful tweets. Of course, in another use-case, it could be appropriate for the models to optimize a different type of errors. Fortunately, in this case, the MCC score was decided to serve as a decisive quality metric, and the result of the comparison is clear.

4.4. Comparison with Results from Other Studies

Comparison of our results with those of other studies on the assessment of respectfulness is difficult because no open data sets exist. The closest study to our work [23] has approached the task of predicting respect in a different manner. However, the order of the quality of our models can be compared with those in other studies using text analysis. In this context, our observations align with those from other studies. In [65] the researchers demonstrate that a pretrained BERT model provides superior results to those of TF-IDF models in several text classification tasks. A survey of text classification algorithms [66] has also ranked deep learning methods ahead of TF or BoW feature extraction techniques. Another study [67] has investigated two classification tasks and found that when LSTM is considered as the method for creating entity-level embeddings, the quality of the language model used for embedding tokens plays an important role, wherein older methods such as Glove, Flair, and Elmo [68] are outperformed by the more recent BERT and RoBERTa techniques.

4.5. Example Practical Benefits of Carrying out Respect Analysis

There is already evidence of penalizing some users of online social media for their adverse behavior based on "hate speech" analysis, with spectacular examples as blocking accounts in early 2021 by Twitter [69] and Facebook [70]. However, no gratification system exists for users writing their posts in a respectful manner. An example of implementing such a system would be increasing online discourse quality, which has indisputable moral virtues.

5. Limitations of Our Study

Conclusions regarding the comparison of the quality of the feature extraction methods presented in this study should be made with care because the comparison was performed by using only a single data set with a limited number of data instances. In addition, the adopted training procedures were not optimized for each model, thus potentially favoring some models over others. The data set was annotated by 3 + 1 researchers, such that their subjective perception of respectfulness defined what the later trained models were

taught. As mentioned in Section 1.3, the perception of respectfulness depends strongly on people and context. Therefore, if another annotator group were used the same tweets could be labeled differently. Accepting the subjective judgment of what is respectful by four annotators is not the optimal choice. However, in the case of no user and context information, we only see averaging more subjective judgments as an alternative.

6. Conclusions

Promoting respect toward other people is a just cause. Interestingly, this seems to be a novel area of research regarding Twitter, for which researchers have focused mostly on analyzing ubiquitous "hate speech." This study took the first steps in the "Twitter respectfulness" field by (1) discussing how respect is defined by other authors, (2) creating what is probably the first open Twitter data set annotated with respectfulness labels, and (3) demonstrating correlations in newly created data with well-studied sentiment analysis. We found that in our data respectfulness was correlated to sentiment at a moderate level of almost 0.6 which can be interpreted that respect is connected with sentiment, but they are distinguishable notions.

To demonstrate how data with respectfulness labels can be used, our study presents an approach for the application of recently developed methods for the automated classification of the respectfulness of tweets. Our results demonstrate that training automated respectfulness classifiers in Twitter is feasible and that they can achieve promising performance. Even though the used data quantity was limited, the comparison of model quality was in accordance with other studies.

We hope that this demonstration will allow others to continue efforts toward promoting respect by, for example, implementing enterprise-level positive motivation measures, on the basis of the automated assessment of textual data.

Author Contributions: Conceptualization, Methodology, Writing of the Original Draft, Software, K.F.; Writing, Review, Editing, Supervision, Funding Acquisition, Project Administration, W.K.; Writing, Review, Editing, Investigation, E.G., T.L., A.B., R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by a research grant from the Office of Naval Research (N000141812559) and was performed at the University of Central Florida, Orlando, Florida.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A precise description of the software requirements is available in the project repository: https://github.com/krzysztoffiok/respectfulness_in_twitter.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Commander, U.S. Pacific Fleet. Available online: www.cpf.navy.mil/downloads/2020/02/signature-behaviors.pdf (accessed on 7 July 2020).
2. Gascón, J.F.F.; Gutiérrez-Aragón, Ó.; Copeiro, M.; Villalba-Palacín, V.; López, M.P. Influence of Instagram stories in attention and emotion depending on gender. *Comunications* **2020**, *28*, 41–50. [CrossRef]
3. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 12–17 June 2016; International Committee on Computational Linguistics: Stroudsburg, Pennsylvania, 2016; pp. 88–93.
4. Burnap, P.; Williams, M.L. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy Internet* **2015**, *7*, 223–242. [CrossRef]
5. Zhang, Z.; Robinson, D.; Tepper, J. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Mining Data for Financial Applications*; Springer Nature: Heraklion, Greece, 2018; pp. 745–760.

6. Waseem, Z. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*; International Committee on Computational Linguistics: Copenhagen, Denmark, 2016; pp. 138–142.
7. Kwok, I.; Wang, Y. Locate the Hate: Detecting Tweets against Blacks. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence AAAI'13, Bellevue, WA, USA, 14–18 July 2013; AAAI Press: Washington, DC, USA, 2013; pp. 1621–1622.
8. Gambäck, B.; Sikdar, U.K. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*; International Committee on Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 85–90.
9. Jaki, S.; De Smedt, T. Right-Wing German Hate Speech on Twitter: Analysis and Automatic Detection. *arXiv* **2019**, arXiv:1910.07518.
10. Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V.; Stranisci, M. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan (LREC 2018)*; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
11. Frenda, S. Exploration of Misogyny in Spanish and English Tweets. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, 18 September 2018; pp. 260–267.
12. United Nations. United Nations Strategy and Plan of Action on Hate Speech. Available online: www.un.org/en/genocideprevention/hate-speech-strategy.shtml (accessed on 6 July 2020).
13. European Commission against Racism and Intolerance (ECRI) Standards. Available online: www.coe.int/en/web/european-commission-against-racism-and-intolerance/ecri-standards (accessed on 16 July 2020).
14. Google. Google Scholar. Available online: http://scholar.google.com (accessed on 2 December 2010).
15. Hambrick, M.E.; Simmons, J.M.; Greenhalgh, G.P.; Greenwell, T.C. Understanding Professional Athletes' Use of Twitter: A Content Analysis of Athlete Tweets. *Int. J. Sport Commun.* **2010**, *3*, 454–471. [CrossRef]
16. Kassing, J.W.; Sanderson, J. Fan–Athlete Interaction and Twitter Tweeting Through the Giro: A Case Study. *Int. J. Sport Commun.* **2010**, *3*, 113–128. [CrossRef]
17. Yusof, S.Y.A.M.; Tan, B. Compliments and Compliment Responses on Twitter among Male and Female Celebrities. *Pertanika J. Soc. Sci. Humanit.* **2014**, *22*, 75–96.
18. Clark, M. *To Tweet Our Own Cause: A Mixed-Methods Study of the Online Phenomenon "Black Twitter"*; University of North Carolina: Chapel Hill, NC, USA, 2014.
19. Maros, M.; Rosli, L. Politeness Strategies in Twitter Updates of Female English Language Studies Malaysian Under-graduates. *Lang. Linguist. Lit.* **2017**, *23*. [CrossRef]
20. Xu, W. From Shakespeare to Twitter: What are Language Styles all about? In *Proceedings of the Workshop on Stylistic Variation*; International Committee on Computational Linguistics: Copenhagen, Denmark, 2017; pp. 1–9.
21. Fatin, M.F. The Differences Between Men And Women Language Styles In Writing Twitter Updates. *Psychology* **2014**, *4*, 1.
22. Ciot, M.; Sonderegger, M.; Ruths, D. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Seattle, Washington, USA, 2013; pp. 1136–1145.
23. Voigt, R.; Camp, N.P.; Prabhakaran, V.; Hamilton, W.L.; Hetey, R.C.; Griffiths, C.M.; Jurgens, D.; Jurafsky, D.; Eberhardt, J.L. Language from police body camera footage shows racial disparities in officer respect. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 6521–6526. [CrossRef]
24. Giorgini, G.; Irrera, E. *The Roots of Respect: A Historic-Philosophical Itinerary*; De Gruyter: Berlin, Germany, 2017.
25. Starkey, H. *Democratic Citizenship, Languages, Diversity and Human Rights: Guide for the Development of Language Education Policies in Europe from Linguistic Diversity to Plurilingual Education: Reference Study*; Council of Europe: Strasbourg, France, 2002.
26. Duranti, A.; Good-win, C.; Duranti, A.C.G. (Eds.) *Rethinking Context: Language as an Interactive Phenomenon*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 1992.
27. Adams, S.M.; Bosch, E.; Balaresque, P.; Ballereau, S.; Lee, A.C.; Arroyo-Pardo, E.; López-Parra, A.M.; Aler, M.; Grifo, M.S.G.; Brion, M.; et al. The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am. J. Hum. Genet.* **2008**, *83*, 725–736. [CrossRef]
28. Modood, T. Moderate secularism, religion as identity and respect for religion. In *Civil Liberties, National Security and Prospects for Consensus*; Cambridge University Press: Cambridge, UK, 2012; pp. 62–80.
29. Helm, B.W. *Communities of Respect: Grounding Responsibility, Authority, and Dignity*; Oxford University Press: Oxford, UK, 2017.
30. Teuber, A. Kant's Respect for Persons. *Political Theory* **1983**, *12*, 221–242. [CrossRef]
31. Fabi, R. "Respect for Persons," Not "Respect for Citizens". *Am. J. Bioeth.* **2016**, *16*, 69–70. [CrossRef] [PubMed]
32. Dillon, R.S. Respect for persons, identity, and information technology. *Ethic Inf. Technol.* **2009**, *12*, 17–28. [CrossRef]
33. Hudson, S.D. The Nature of Respect. *Soc. Theory Pr.* **1980**, *6*, 69–90. [CrossRef]
34. Chapman, L.M. Respectful Language. *J. Psychol. Issues Organ. Cult.* **2013**, *3*, 115–132. [CrossRef]
35. Holtgraves, T.M. *Language as Social Action: Social Psychology and Language Use*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2002; p. 232.
36. Thompson, M. *Enough Said: What's Gone Wrong with the Language of Politics*; St. Martin's Press: New York, NY, USA, 2016.
37. Wolf, R. Respect and disrespect in international politics: The significance of status recognition. *Int. Theory* **2011**, *3*, 105–142. [CrossRef]

38. Beach, M.C.; Duggan, P.S.; Cassel, C.K.; Geller, G. What Does 'Respect' Mean? Exploring the Moral Obligation of Health Professionals to Respect Patients. *J. Gen. Intern. Med.* **2007**, *22*, 692–695. [CrossRef]
39. Fiok, K. Krzysztoffiok/Twitter_Sentiment. Available online: https://github.com/krzysztoffiok/twitter_sentiment (accessed on 18 October 2020).
40. Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; Wojatzki, M. Measuring the Reliability of Hate Speech An-notations: The Case of the European Refugee Crisis. *arXiv* **2016**, arXiv:1701.08118 2016. [CrossRef]
41. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
42. Ho, T.K. Random decision forests. In Proceedings of the 3rd International conference on document analysis and recognition, Montreal, QC, Canada, 14–16 August 1995.
43. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
44. Sharma, M.; Sharma, S.; Singh, G. Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining. *Data* **2018**, *3*, 54. [CrossRef]
45. Sharma, M.; Singh, G.; Singh, R. Design of GA and Ontology based NLP Frameworks for Online Opinion Mining. *Recent Pat. Eng.* **2019**, *13*, 159–165. [CrossRef]
46. Kumar, P.; Gahalawat, M.; Roy, P.P.; Dogra, D.P.; Kim, B.-G. Exploring Impact of Age and Gender on Sentiment Analysis Using Machine Learning. *Electronics* **2020**, *9*, 374. [CrossRef]
47. Pennebaker, J.W.; Boyd, R.L.; Jordan, K.; Blackburn, K. *The Development and Psychometric Properties of LIWC2015*; The University of Texas: Austin, TX, USA, 2015.
48. Crossley, S.A.; Kyle, K.; McNamara, D.S. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behav. Res. Methods* **2016**, *49*, 803–821. [CrossRef] [PubMed]
49. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
50. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781v3.
51. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
52. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
53. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv* **2017**, arXiv:1705.02364.
54. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *arXiv* **2020**, arXiv:1910.03771.
55. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th In-ternational Conference on Computational Linguistics*; Association for Computational Linguistics: Santa Fe, Mexico, 2018; pp. 1638–1649.
56. Sklearn.Feature_Selection.Mutual_Info_Classif—Scikit-Learn 0.24.0 Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html (accessed on 16 October 2020).
57. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv* **2020**, arXiv:1909.11942.
58. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7130–7138. [CrossRef]
59. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.
60. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
61. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
62. Lample, G.; Conneau, A. Cross-Lingual Language Model Pretraining. *arXiv* **2019**, arXiv:1901.07291.
63. Fiok, K.; Karwowski, W.; Gutierrez, E.; Davahli, M.R. Comparing the Quality and Speed of Sentence Classification with Modern Language Models. *Appl. Sci.* **2020**, *10*, 3386. [CrossRef]
64. Fiok, K. Krzysztoffiok/Respectfulness_in_Twitter. Available online: https://github.com/krzysztoffiok/respectfulness_in_twitter (accessed on 16 October 2020).
65. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef]
66. González-Carvajal, S.; Garrido-Merchán, E.C. Comparing BERT against Traditional Machine Learning Text Classification. *arXiv* **2021**, arXiv:2005.13012.
67. Kowsari, K.; Meimandi, K.J.; Heidarysafa, M.; Mendu, S.; Barnes, L.E.; Brown, D.E. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]
68. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv* **2018**, arXiv:1802.05365.

69. Permanent Suspension of @Realdonaldtrump. Available online: https://blog.twitter.com/en_us/topics/company/2020/suspension.html (accessed on 16 January 2021).
70. Facebook. Available online: https://www.facebook.com/zuck/posts/10112681480907401 (accessed on 14 January 2021).