


Article

# An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media

Maria Krommyda \*, Anastasios Rigos, Kostas Bouklas and Angelos Amditis 

Institute of Communication and Computer Systems (ICCS), 15772 Athens, Greece; anastasios.rigos@iccs.gr (A.R.); kostas.bouklas@iccs.gr (K.B.); a.amditis@iccs.gr (A.A.)

\* Correspondence: maria.krommyda@iccs.gr; Tel.: +30-2107721663

**Abstract:** Opinion mining techniques, investigating if text is expressing a positive or negative opinion, continuously gain in popularity, attracting the attention of many scientists from different disciplines. Specific use cases, however, where the expressed opinion is indisputably positive or negative, render such solutions obsolete and emphasize the need for a more in-depth analysis of the available text. Emotion analysis is a solution to this problem, but the multi-dimensional elements of the expressed emotions in text along with the complexity of the features that allow their identification pose a significant challenge. Machine learning solutions fail to achieve a high accuracy, mainly due to the limited availability of annotated training datasets, and the bias introduced to the annotations by the personal interpretations of emotions from individuals. A hybrid rule-based algorithm that allows the acquisition of a dataset that is annotated with regard to the Plutchik's eight basic emotions is proposed in this paper. Emoji, keywords and semantic relationships are used in order to identify in an objective and unbiased way the emotion expressed in a short phrase or text. The acquired datasets are used to train machine learning classification models. The accuracy of the models and the parameters that affect it are presented in length through an experimental analysis. The most accurate model is selected and offered through an API to tackle the emotion detection in social media posts.

**Keywords:** social media analysis; emotion detection; sentiment analysis; Plutchik's eight basic emotions; sentiment classification



**Citation:** Krommyda, M.; Rigos, A.; Bouklas, K.; Amditis, A. An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media. *Informatics* **2021**, *8*, 19. <https://doi.org/10.3390/informatics8010019>

Academic Editors: Antony Bryant, Lorraine Goeuriot, Gabriella Pasi and Marco Viviani

Received: 4 February 2021

Accepted: 9 March 2021

Published: 12 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The most suitable sources for opinionated text are posts from social media platforms and their monitoring has gained in popularity as more techniques are available for performing sentiment analysis, evaluate the opinions expressed about current events or public figures and characterize them as positive or negative [1]. The term sentiment analysis [2,3] is used to describe the categorization of text as expressing a positive or negative opinion. Sentiment analysis is performed for entertainment-related events, such as TV series and movies, as well as public figures [4]. In order for such analysis to be successful there is a need for high quantities of annotated text, collected from multiple trustworthy sources and with high time variance.

It is indisputable that this technique has some very important use cases and real-world applications. These use cases, however, are limited to scenarios where the overall opinion about the examined event is not known. There are many scenarios where the positive or negative aspect of the opinion can be assumed with a high degree of confidence. One indicative example is the analysis of text produced by individuals that are exposed to extreme or stressful situations, such as a car accident or an extreme natural disaster. People experiencing such incidents are expected to be very negatively influenced, being worried about their well-being and their properties, angry with the authorities due to lack of preparatory measures or scared about the development of the event and its consequences. Another important challenge during such events is that the opinionated text is produced in a short amount of time, during or close by the occurrence of the event.

Emotion detection [5] in text is proposed as a solution for these challenges. This technique is not focusing on the positive or negative opinions expressed but tries to determine the human emotion that is expressed. The task of identifying the emotions expressed by a person is a very challenging task that even humans struggle with. Trying to model such identification and create an automated way of identifying the expressed emotion is an even more challenging task, not only due to the limited availability of training datasets but also the restricted information contained in a short text.

A technique to emotionally classify tweets is presented in [6] where tweets are mapped to 6 emotional moods using the Profile of Mood States (POMS) technique. In [7], a text analysis software, the Linguistic Inquiry and Word Count (LIWC), is used to classify tweets into six main emotional categories. The LIWC calculates the degree that people use different categories of words across a wide array of texts. The same text analysis software is also used in [8]. In [9] an emoji-predictor was built but without evolving any emotions at all. In [10] the collected social-media posts were manually annotated. In [11] the authors use also manually annotated text from widely known children stories. All the available solutions so far have two main drawbacks. On the one hand, the emotional categories selected are not based exclusively on the research done by psychologist regarding the emotion theory. Many systems adapt the emotion categories based on their finds, merging or separating emotions, without any consideration to the scientific standard. On the other hand, most of the available systems create their training dataset based on the presence of specific keywords. Unfortunately, using only keywords-based methods, there is no way to validate the accuracy of the classification method as the classifier is almost exclusively trained to recognize these keywords and classify the text accordingly.

**Contributions.** In this document, a fully fledged prototype system for the emotion detection in text, as it is published in social media posts is presented. The system offers a solution for the creation of a fully annotated dataset that can be used for emotion detection and a comparison study between different machine learning models that were trained using the annotated dataset. This prototype system offers the following:

- A natural language processor that handles the unique linguistic characteristics of social media posts in regard to lexical, syntax and annotation preferences and provides a uniform text in the annotated dataset.
- A hybrid rule-based algorithm that supports the creation of an objectively classified dataset over the Plutchik's eight basic emotions [12]. The algorithm takes into consideration the available emoji in the text and used them as objective indicators of the expressed emotion thus efficiently tackling the challenge of the subjectivity of the emotion detection.
- An experimental analysis to select the proper machine learning solution, and its proper configuration, for identifying the expressed emotions in text.

## 2. Methodology

In the following sections, the methodologies used for establishing and resolving the problem are presented. This includes the methodology followed to identify the appropriate emotion categories to be used by the classifier, the process of collecting the dataset as well as the rules for harmonizing and annotating it. Finally, the development of the classification model is presented.

### 2.1. Emotion Categories

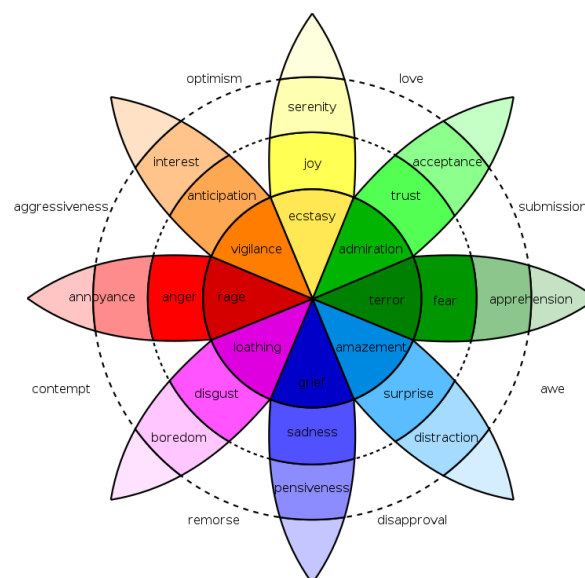
The discrete emotion theory [13] is one of the most popular theories regarding the emotion categorization, expressed for the first time in 1872 by Charles Darwin [14]. The theory follows the same basic idea behind the colour theory, claiming that there are some fundamental emotions that can be used as the base for interpreting and categorizing all the emotions that people may express.

The discrete emotions theory evolved over time and gained in popularity when Paul Ekman presented an extended experimental analysis about the emotions that should be

considered to be primary and reasons for that [15,16]. His conclusions are summarized in two key directions. On the one hand, a pleasant-unpleasant and active-passive scale was identified as capable of depicting differences between emotions. On the other hand, emphasis was put on the fact that the emotion interpretation by humans was biased, as the understanding and interpretation of emotions is a skill that people develop through their environment, the other people they communicate with and their social and cultural interactions. Ekman was the first to challenge his own assumptions as too restrictive to map all the human emotions, and tried to establish the proper experiments that would allow the collection of unbiased data.

Due to his systematic and unbiased study of the emotion classification, Paul Ekman is thought of as a pioneer in the study of emotions and their relation to facial expressions [17]. The results of Ekman's emotion classification study have received a lot of criticism, mainly regarding the way that the data were collected, the process that was chosen for the data validation and the affect these choices had on the integrity of the results. Despite the doubts regarding the process and the results, Ekman has provided the first widely accepted list of the primary emotions, which are happiness, anger, sadness, fear, disgust, and surprise.

Robert Plutchik [12] built on that and proposed a slightly modified list with eight primary emotions. His classification is based on elements of the psychological evolution of the expression of emotion and it is extracted after careful examination of general emotional responses of individuals for the same event [18]. Plutchik's psycho-evolutionary theory of basic emotions has ten suggestions. These include the claim that emotions are also been experienced by animals, the fact that emotions appeared through the evolution process, mainly for survival reasons as well as that there are some primary emotions that with various combinations, mixtures and intensities can express any experienced emotion, as shown in Figure 1.



**Figure 1.** Plutchik's wheel of emotions. As provided by Machine Elf 1735 - Own work, Public Domain.

## 2.2. Dataset Acquisition

Computers are unable to understand human language without any interpretation, a functionality that is considered useful for many applications including automated assistants and smart networks. The task of machines to understand the human language is very challenging as it is very complicated, it includes expression of emotion and usage of lax syntactical and grammatical rules. Aiming to alleviate this barrier, there has been extensive research regarding the ways that artificial intelligence solutions can help computers with

the interpretation of the human language as well as support machines in communicating in a human-like way [19].

The process that allows machines to understand a text, the same way that an individual would, is called Natural Language Processing (NLP) [20]. The main challenge is the complexity of the meaning extraction. Understanding the individual words is trivial but identifying their meaning in phrases is very challenging. Linguistic characteristics and multiple meanings of words based on the overall context are some of the most interesting peculiarities of the human language [21]. Humans interpret a word based on the context and the overall meaning with ease but modelling this is very challenging.

The task of performing NLP analysis to social media posts comes with some additional challenges. It is important to note that the most the well-established solutions have been trained using text coming from journal articles and encyclopedias [22], sources of large quantities of high quality text [23] that comply with spelling and syntax rules, use proper grammar and vocabulary that can be found in official dictionaries. This is a complete contrast with the characteristics of the language used at social media posts. In detail, the identified differences in the language are:

- **Text size.** Traditional NLP systems have been trained using large passages of text that have only one author. Social media posts have limited characters and are produced by multiple individuals.
- **Topic diversity.** Traditional sources provide content that presents specific topics within the same text, journal articles for example are expected to be focused on one single topic throughout. A single social media post can include multiple topics, and express opinions of different intensity, at the same time that multiple users are discussing a plethora of topics that are of interest.
- **Vocabulary & Spellcheck.** Formal text includes exclusively words that can be found in dictionaries, used following the proper spelling and the word capitalization rules. In social media posts, users often use words that are non-existing, either formed on the fly to emphasize a situation or due to incorrect spelling. There are also some habits of these users that deviate from the proper spelling rules, such as the usage of letter repetition to give emphasis, the use of capital letters to expressed the intensity of their emotions or the shortening of words to their sound for shorter messages, such as 'heyyyyy', 'I am SAD' or 'u'.
- **Syntax & Grammar.** Formal text follows all the syntax and grammar rules, including the use of punctuation marks. In social media posts, emphasis is prioritized over proper usage of the linguistic rules. The text includes incomplete phrases, non-existent grammar, phrases without verbs or subjects, incorrect and excessive use of punctuation marks.
- **Text objectivity & originality.** Formal sources are providing unique and objective text that contains properly presented facts and precise information with limited emotional expressions. On the other hand, social media posts are often repetitions or additions to already presented opinions. The posts aim to present specific views and arguments over recent events and topics of interest, often expressed under emotional excitement.
- **Abbreviations.** Shortened text is really used in official documents, unless it has been clearly defined, is related to a well-established term that is used throughout the document and is properly explained. The users of the social media are creating abbreviations continuously, giving them multiple interpretations based on the overall context and fail to explain their meaning.

### 2.2.1. Dataset Collection and Harmonization

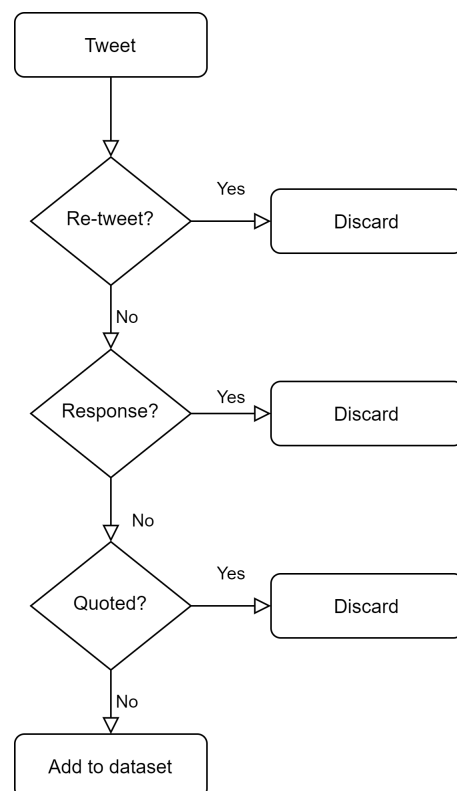
Twitter is the social media platform selected for the collection of the text that will be used as it is focused mainly on text messages and not on video and images. The tweets will be pre-processed in order to be harmonised and annotated using a rule-based approach [24]. The approach will ensure that the special characteristics of the social media posts are taken

into account, and used as needed in order to objectively annotate the collected posts, and provide a high quality training dataset.

A dedicated developer's account was created for the Twitter platform [25] to obtain the proper permissions and authentication criteria in order to collect social media posts. The Twitter Streaming API [26] was used through the Python Tweepy [27] library. The library is handling the proper communication as well as the management of respecting the post capturing quotas of the Twitter API, allowing us to focus on developing the logic needed for the proper data collection.

The main feature of the Twitter Streaming API is the need to provide a filter to be used for the streaming. The filter includes multiple criteria, including keywords, location and language [28] that can be used independently. After investigating the most popular, based on their frequency of use, words in tweets [29], a list was composed and used as the streaming filter. The list is {"a", "the", "I", "to", "you", "in", "on", "for", "with", "that"}. In addition, understanding that the profiles of the users, and inevitable the content of the posts, changes based on the day and time they are produced, for example teenagers tend to use social media late at night while parents of young children in the afternoon, special consideration was given to the collection of posts at different times and days.

As an additional step to ensure the collection of a high quality dataset, retweets, quotes, and responses to tweets were not included in order to avoid text repetition and phrases too small to be interpreted on their own. The process of eliminating such tweets is shown in Figure 2. Having established the methodology to collect the social media posts, the next step is to harmonize their content, focusing on the following:



**Figure 2.** Tweet collection process.

**Hashtags:** Hashtags are a unique characteristic of tweets; they are used either at the beginning of important to the post's content words, providing an indirect categorization or at the end of the post, along with words and short phrases that have special meaning and importance to the post. For example, a user at the airport ready to leave for summer vacations would write 'Waiting at the #airport, the flight is leaving for my #vacations in

less than an hour!!! #Summer #SummerTime #traveling’. Such tweet is using the hashtags to highlight important words for the post as well as to provide additional context.

A innovative data flow has been designed, as presented in Figure 3, to examine each hash-tagged word and identify the words included. The first step is to remove the hashtag and look up the word in a dictionary, if the word is found there then the hash-tagged word is replaced by it, for example #vacations is replaced by vacations. Next, it is examined if the hash-tagged word is actually multiple words written in camel case. If this is the case, then the capitalization rule is used to split the hash-tagged word into multiple ones. Again, each split word is validated through a dictionary to ensure that the division was accurate, for example #SummerTime is replaced by ‘summer time’. As dictionary, the nltk text corpora and lexical resources [30] are used. Last but not least, the hash-tagged word is checked over the Abbreviations [31] open API and replaced with the corresponding words in case of a match. If none of the above steps are fruitful then the hash-tagged word is considered a misspell and is removed from the tweet.

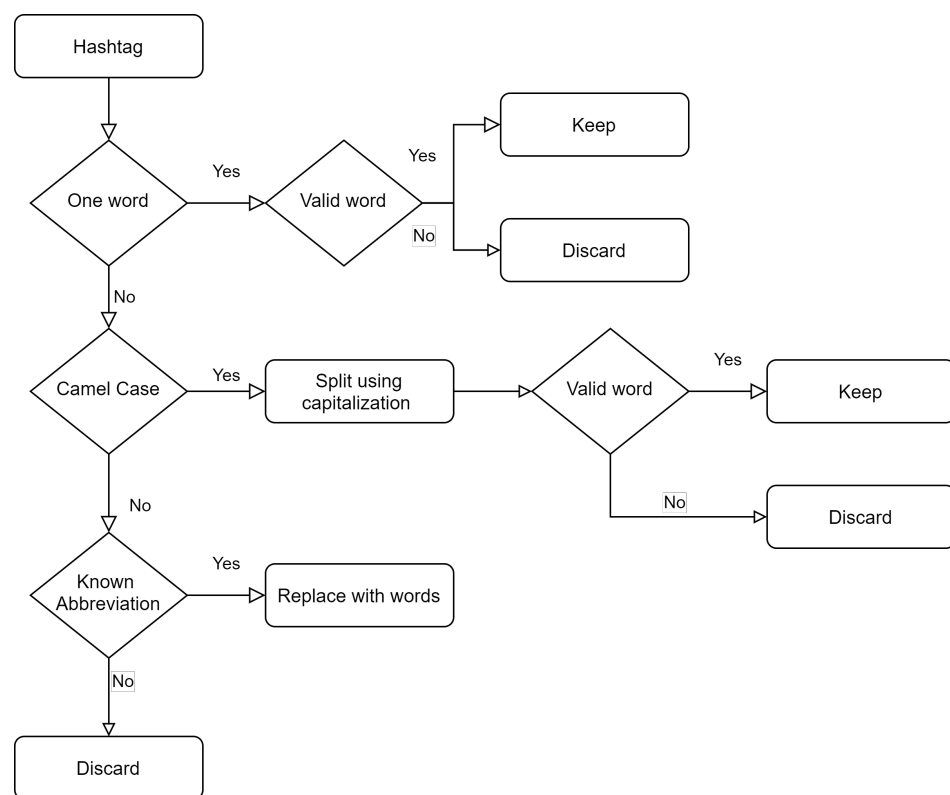


Figure 3. Hashtag processing in tweets.

**URLs:** Tweets are known for containing hyperlinks to other sources, as they add value and meaning to the posts. Such links however are not of any value for the emotion detector and their unusual spelling is more probable than not to confuse the feature extractor. For this reason, they are removed from the tweets using the Tweet pre-processor Python library [32].

**Mentions:** As discussed also for the hyperlinks, mentions are very important for the Twitter social media platform. They are used mainly to responses and when specific people are referenced. They always start with the special character @ followed with the user name of the person mentioned. While their purpose is important for the interaction at the platform, they have no semantic value. On the contrary, their unique spelling and the sparsity of their appearance in the dataset may confuse the feature extractor and the emotion detector. For this, it has been decided to remove them from the text included in the dataset.

**Character repetition/ misspelled words:** Each tweet is divided into words, and each word is checked against the dictionary to determine its validity. In case of an invalid word, the phenomenon of character repetition is investigated, gradually eliminating characters that appear more than once and checking if the word is valid. Posts with misspelled or invalid words are removed from the dataset.

**Emoji:** The Unicode codes of the emoji are not processed at this time in any way. The emoji are very important, as it will be discussed in detail in Section 2.2.2, for the rule-based classification of the tweets in the eight categories. For this reason, they are not modified in the harmonization phase.

### 2.2.2. Dataset Annotation

Having collected and harmonized the dataset, the next step is to provide proper categorization in the eight emotion categories of the Plutchik's wheel. This is achieved through a rule-based python script that complies with the following rules:

**Emoji:** The emoji that are used in excess in social media posts, as collected by the emoji python library [33], were examined in detail and separated into eight categories corresponding to the eight emotion categories of the Plutchik's wheel. A large percentage of the emojis were not included in any of the eight categories, as they were general and descriptive, not associated with any emotional state. Indicative examples of such emoji are vehicles, fruit and vegetables, as well as objects and animals. Another important part of the emoji dataset that was not categorized, included emoji that were referring to emotions that were a combination of more than one of the eight emotion categories according to the Plutchik's wheel, such as remorse and aggressiveness.

Approximately, only 7% of the initial emoji were categorized as expressing one of the eight emotions. The number of emoji per category differs a lot, with categories such as anger and joy having up to 60 emoji while categories such as anticipation and disgust having less than 30. Aiming to ensure that the emoji included in the lists for each emotion are popular, and consequently probable to be found in posted tweets, the most commonly used emoji as published in tweets and monitored by a live web tool, the Emoji Tracker [34], were also examined. The lists with the emoji were updated to include some of the emoji found in the list.

For each collected tweet, the emoji that it contains are examined against the eight lists. If the emoji is not present in any of the lists, then it is simply replaced by its corresponding text. In the case that only one of the emoji of the tweet is in one of the lists then the tweet is annotated as belonging to that category, also if more than one emoji belongs to the same list, again the tweet is annotated as belonging to that category. In both cases the emoji that were used for the categorization of the tweet are removed from the post, and not replaced by their text. If multiple emoji belong to multiple lists then there is no way to properly annotate the tweet, so the emoji are replaced by their corresponding text [35].

**NRC Emotion Lexicon:** The NRC Emotion Lexicon [36] is a list of English words and their associations with eight basic emotions of the Plutchik's wheel. Each tweet is examined in case such word is present, and if so classified accordingly.

**Lexical relations:** WordNet [37] is a large lexical database of English, where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, so that each set is expressing a distinct concept. The sets of synonyms are also interlinked by means of semantic and lexical relations. These relationships are synonymy (car, automobile), antonymy (hot, cold), hyponym (red, colour), hypernymy (cutlery, fork) and meronymy (tree, forest) [38,39].

In the annotation rules, two groups of relationships are identified. The synonyms/hyponyms/hypernyms that are providing for each emotion a set of words and the antonyms of each emotion that are added in the set of words for the polar opposite of the emotion based on the Plutchik's wheel [40,41]. The annotation process for the tweets is shown in Figure 4. Taking into consideration that some emotions are expressed more frequently, initially there were different number of tweets per emotion category collected, which were

harmonized in the final dataset. Tests performed in different hours during the day and during different days of the week, including late at night and during weekends showed that there is a constant lack of social media posts expressing fear and anticipation. Tweets expressing anger and trust are also hard to collect, while joy and surprised are the most common emotions expressed. The collected dataset can be used to train any emotion detection machine learning model, following the process depicted in Figure 5.

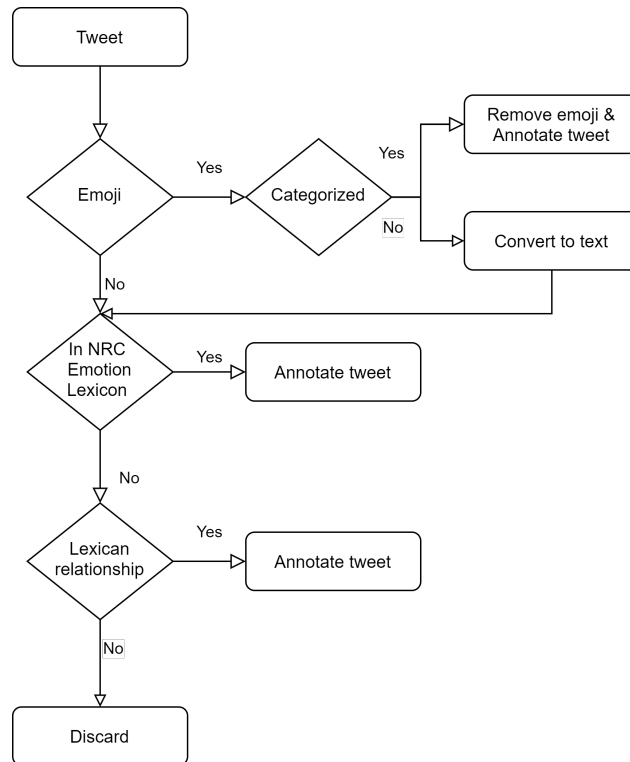


Figure 4. Rule based tweet annotation.

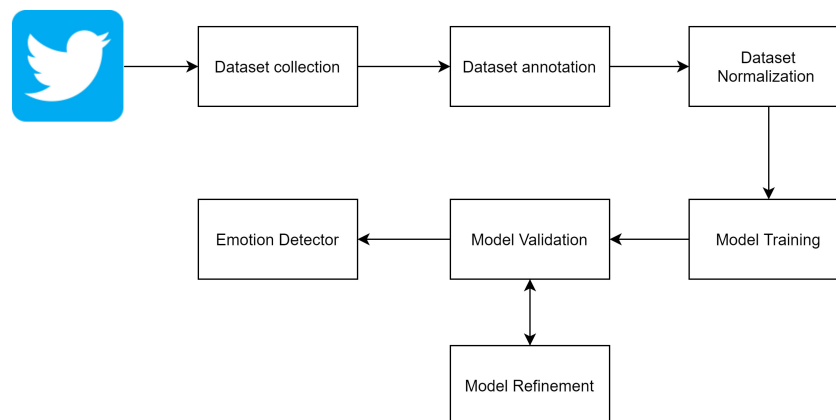


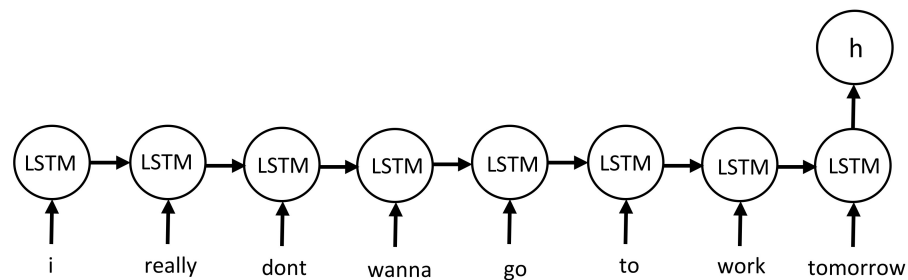
Figure 5. Dataset usage in model training.

### 2.3. Classification Model Development

Among many classification methods, the Long short-term memory networks (LSTM) are widely used to classify text and social media posts. LSTM [42] is a type of recurrent neural network (RNN) [43] that specifically addresses the issue of learning long-term dependencies. Their architecture allows for the accumulation of information during operation and uses feedback to remember previous network call states [44]. A generic description of the network model used in this document is presented in Figure 6 where each circle represents an LSTM cell, the input is given as a vector representing a tweet and Output  $h$



will return the result that is of interest. Detailed description of the architecture of LSTMs can be found in [45].



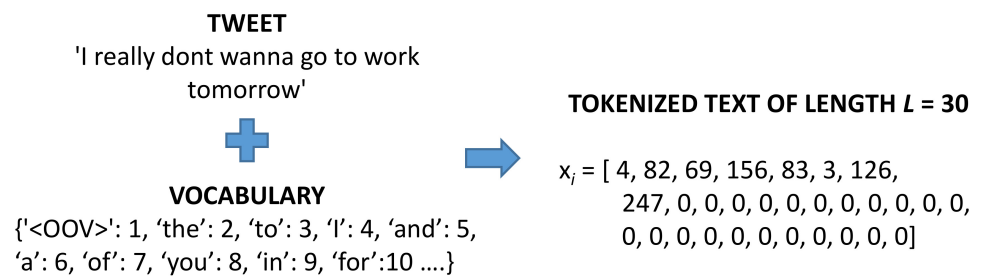
**Figure 6.** The sequential learning procedure of an LSTM network.

LSTM networks are often used in time-series forecasting and pattern analysis, such as on petroleum production [46], on weather [47] and on fog data [48]. Also, in the literature the Long-Short-Term-Memory (LSTM) networks are the leading methodology for text analysis and classification. In [44] the LSTM is used to classify pre-labelled texts gathered from online forums in the categories of spam and not-spam; also the content of books reviews is classified into positive or negative. In [49] three public available pre-classified datasets, a movie review dataset and two sets with restaurant reviews, are used in order to train an LSTM to classify the reviews into negative and positive ones. Also, in [50] pre-labelled articles as news and reviews of movies and products are classified in the categories of positive and negative. Finally, in [51] posts from social media referring to political beliefs are classified as Democratic or Republican.

For the purposes of this paper, an LSTM network is trained to solve the presented classification problem of the emotion detection. In order to compare the results of the LSTM classifier with other methodologies, five other classifiers have been trained using the same dataset. These baseline classifiers are:

- A linear support vector machine using the stochastic gradient descent classifier (called SVM-SGD in this document) [52]
- A XGBoost classifier [53]
- A Naive Bayes classifier for multinomial models [54]
- A Decision Tree classifier [55]
- A random forest classifier [56]

In order to train all the above methods, the collected posts of Twitter had to be transformed into numbers. Using the downloaded tweets, a vocabulary was created using the  $V$  most common words appearing. The parameter  $V$  also determines the dimension of the feature space used for the classification problem. A higher dimension could presumably capture more information and obtain better results, at the cost of slower training times. Using the created vocabulary, each tweet was transformed to a vector of length  $L$  representing the words it contains; this process is called tokenization. If a tweet contains less than  $L$  words, then it is filled with zeros. Figure 7 presents this procedure.



**Figure 7.** The tokenization process of the tweets. OOV refers to 'out of vocabulary'. As the reader can see from this figure a tweet may not always be coherent.

To perform the training process of the classifiers, the average loss of cross-entropy, also known as *Negative Log Likelihood Loss*, objective function was minimized:

$$J(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

where  $\mathbf{w}$  represents all the synaptic weights used in the LSTM,  $N$  is the size of the dataset,  $y_i$  and  $\hat{y}_i$  symbolize the number of the original and predicted category ( $y_i$  and  $\hat{y}_i \in \{1, 2, \dots, 8\}$  in this document). As Equation (1) gets minimized, more of the data are being classified into the correct category. For this optimization problem, the Adam algorithm was used [57].

For all the simulations the *Python 3.6.10* programming language was used with the libraries *xgboost 1.2.0* [53], *scikit-learn 0.23.2* [58] and *tensorflow 2.4.1* [59].

### 3. Results

A total of 1.2 million annotated tweets were downloaded using the process described in Section 2.2.2; this number of tweets is just a balanced (among the emotion categories) subset of a greater set of 3.6 million tweets that were collected with a speed of about 700,000 tweets per day. This dataset is publicly available at (accessed on 10 March 2021) <https://www.kaggle.com/tasos123/tweets-annotated-to-emotions> and it contains only the tweet-ID and the class it has been annotated in order to protect the anonymity of the Tweeter's users. Eighty percent of them were used as a training sample, 10% as validation and 10% as testing. The used vocabulary size is  $V = 20,000$  and each tweet was transformed to a vector of length  $L = 50$ . Choosing other values for the parameters  $V$  and  $L$  was leading to similar or worse results. Also greater values for  $V$  were resulting to very slow training procedures or were resulting to overfitting networks and were avoided. It should be noted that all computations were performed to a CPU (an Intel i7-4770 @ 3.40GHz) and not any GPU was used in all the experiments of this document.

Figure 8 shows the layers used in the training process of the LSTM together with their parameters. For the baseline classifiers, the default parameters (according to their libraries mentioned in the previous section) were used.

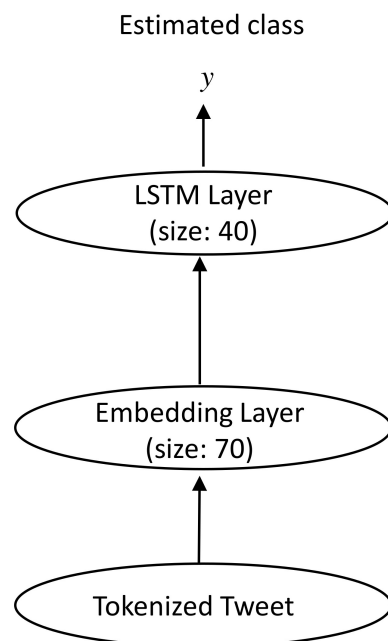


Figure 8. The layers used in the training process of the LSTM network.

To compare the performance of the classification methods, the overall accuracy on the testing dataset is used. Table 1 shows the results of the LSTM compared with the five other baseline classifiers. The LSTM provides the best overall performance and the SVM-SGD follows. The Naive Bayes model provides worse performance than all the others, but this could be resulting from the use of the partial fitting function that was used since it was demanding huge amounts of RAM memory. Figure 9 shows the LSTM’s accuracy in each of the eight classes at the form of a confusion matrix. This matrix was normalized, i.e., each row sums to one. Higher values in the main diagonal of this matrix show better performance for each category.

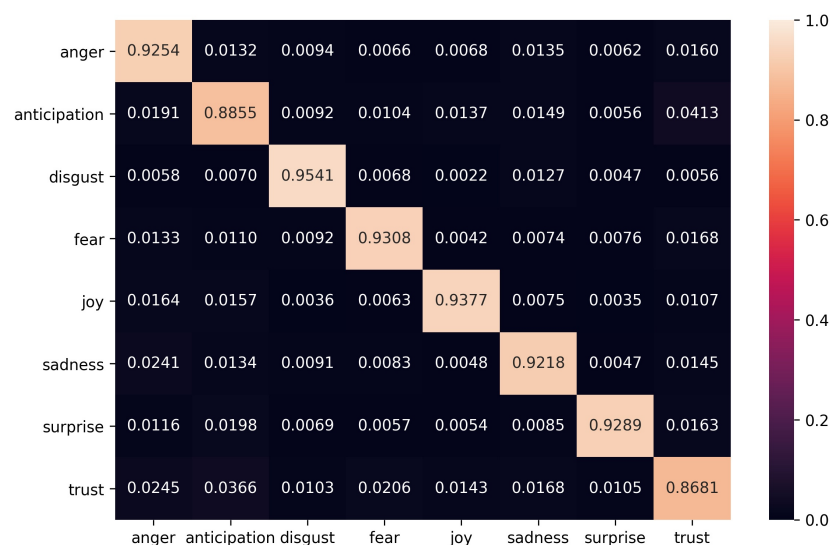


Figure 9. The Confusion matrix with the classification results of the LSTM network.

**Table 1.** Performance of the classification methods applied on the testing dataset.

Classifier	Testing Accuracy
LSTM	91.90%
SVM-SGD	86.86%
XGBoost	84.45%
Naive Bayes	77.01%
Decision Tree	84.69%
Random forest	80.35%

Disgust and joy are the emotions that are better predicted using the LSTM, while trust and anticipation are the emotions with the lowest network performance. The differences, however, in the performance of the LSTM in the eight categories are such that can be considered insignificant. Especially for trust and anticipation, the two least accurate predicted emotions, careful examination of the confusion matrix shows that text expressing trust is more often than not wrongly classified as anticipation and vice versa. Based on this observation it can be assumed that, even though not anticipated, there is some overlap in phrases and expressions used to express these emotions. It is also worth noticing that these emotions are the ones that have the fewest keywords associated with them, the fewest synonyms and the least diversity regarding emoji expressing them. For these reasons, the network's accuracy is limited in these categories.

Table 2 presents some examples of classified tweets, nine correctly classified ones and five classified to the wrong class. It can be observed that, the second wrongly classified tweet, (#11 in Table 2), although it was labelled as *disgust* in the annotated dataset due to the presence of an emoji, the LSTM classified it to the *anger* category, possibly using the "I hate" sequence of words. Similar behaviour is presented in the tweet #13 that is was labelled as *joy* in the annotated dataset due to the presence of a smiling emoji, but from the human perspective it is more likely that it belongs in the *anger* class, such as the LSTM that classified it. Also in #14 we can see the presence of sarcasm/irony that although it was labelled in the annotated dataset as *joy* due to the existence of the "love" keyword, it was classified by the LSTM to the more fitting *sadness* emotion.

**Table 2.** Examples of the classified tweets.

#	Correctly Classified Tweets	Annotated as	Classified as
(1)	Nothing hurt more than loyalty coming from one side in a relationship.	anger	anger
(2)	What a vibe This song made my day	joy	joy
(3)	How you gonna find a knockoff version of me that welcomes commitment that's gross	disgust	disgust
(4)	The greatest day of the year has finally begun	anticipation	anticipation
(5)	How to cope with parents who regret your existence	sadness	sadness
(6)	can we cancel 2020? im so done	sadness	sadness
(7)	god bless the ability to mute people on instagram	anticipation	anticipation
(8)	I cant wait to live alone in the mountains.	anticipation	anticipation
(9)	I hate this fucking song	anger	anger

Table 2. Cont.

#	Correctly Classified Tweets	Annotated as	Classified as
	<b>Incorrectly Classified Tweets</b>		
(10)	40 minutes till i can play tomb raider again	sadness	joy
(11)	Man I hate my life	disgust	anger
(12)	allergy highs	disgust	joy
(13)	CHANGE MY PIC TO ALL THOSE DIRTY HACKERS AND SCAMMERS TRYING TO USE MY FB, INSTAGRAM PG, THANK YOU MONIE FOR THE LOOKOUT CHICK!!! #CLASS OF '88	joy	anger
(14)	I love not being loved by my friends	joy	sadness

#### 4. Conclusions

Understanding the opinion expressed in short text, especially when sourced from social media posts, without any given context, is a challenge even for humans. The authors more often than not fail to follow proper spelling, syntax and grammatical rules, while they tend to emphasize words and emotions based on the way that they want others to interpret their posts. Going a step further, understanding the emotion expressed in a written passage is an even greater challenge as isolated the meaning of the words may lead to misinterpretation. Assigning this task to a machine and automating it through algorithms is undeniably challenging.

In this paper, an innovative methodology is proposed for the collection, harmonization, and annotation of tweets in eight categories representing the Plutchik's eight primary emotions. In addition, a methodology is proposed for the deployment of a machine learning classification method in order to create a very accurate and easy to use application to automatically classify those short-text posts for future needs. The machine learning solution proposed is a LSTM network that was compared and evaluated against five others classifiers, and proven to be better qualified for the needs presented here. The LSTM network has achieved an 91.9% accuracy making it a promising tool for applications that require fast, reliable, and accurate estimations of the expressed emotions in short text.

**Author Contributions:** Conceptualization, M.K. and A.R.; methodology, A.R.; software, A.R.; data curation, M.K.; project administration, K.B.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is part of the RESIST project. RESIST has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 769066. Content reflects only the authors' view. The Innovation and Networks Executive Agency (INEA) is not responsible for any use that may be made of the information it contains.

**Data Availability Statement:** This dataset is publicly available at (accessed on 10 March 2021) <https://www.kaggle.com/tasos123/tweets-annotated-to-emotions> and it contains only the tweet-ID and the class it has been annotated in order to protect the anonymity of the Tweeter's users.

**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.

#### Abbreviations

The following abbreviations are used in this manuscript:

API	Application programming interface
CPU	Central Processing Unit
GPU	Graphics Processing Unit

LIWC	Linguistic Inquiry and Word Count
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
OOV	Out Of Vocabulary
POMS	Profile of Mood States
RAM	Random Access Memory
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
URL	Uniform Resource Locator

## References

- Bakshi, R.K.; Kaur, N.; Kaur, R.; Kaur, G. Opinion mining and sentiment analysis. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 452–455.
- Liu, B. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [\[CrossRef\]](#)
- Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R.J. Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011), Portland, OR, USA, 23 June 2011; pp. 30–38.
- Feldman, R. Techniques and applications for sentiment analysis. *Commun. ACM* **2013**, *56*, 82–89. [\[CrossRef\]](#)
- Acheampong, F.A.; Wenyu, C.; Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. *Eng. Rep.* **2020**, *2*, e12189. [\[CrossRef\]](#)
- Bollen, J.; Pepe, A.; Mao, H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv* **2009**, arXiv:cs.CY/0911.1583.
- Larsen, M.E.; Boonstra, T.W.; Batterham, P.J.; O’Dea, B.; Paris, C.; Christensen, H. We feel: Mapping emotion on Twitter. *IEEE J. Biomed. Health Informat.* **2015**, *19*, 1246–1252. [\[CrossRef\]](#)
- Wang, W.; Chen, L.; Thirunarayan, K.; Sheth, A.P. Harnessing twitter “big data” for automatic emotion identification. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 587–592.
- Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv* **2017**, arXiv:1708.00524.
- Balabantaray, R.C.; Mohammad, M.; Sharma, N. Multi-class twitter emotion classification: A new approach. *Int. J. Appl. Inf. Syst.* **2012**, *4*, 48–53.
- Alm, C.O.; Roth, D.; Sproat, R. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 579–586.
- Plutchik, R. *The Emotions*; University Press of America: Lanham, MD, USA, 1991.
- Roseman, I.J. Cognitive determinants of emotion: A structural theory. *Rev. Personal. Soc. Psychol.* **1984**, *5*, 11–36.
- Darwin, C.; Prodger, P. *The Expression of the Emotions in Man and Animals*; Oxford University Press: New York, NY, USA, 1998.
- Ekman, P. A methodological discussion of nonverbal behavior. *J. Psychol.* **1957**, *43*, 141–149. [\[CrossRef\]](#)
- Tomkins, S.S. *Affect Imagery Consciousness: Volume I: The Positive Affects*; Springer Publishing Company: Berlin/Heidelberg, Germany, 1962; Volume 1.
- Ekman, P.; Keltner, D. Universal facial expressions of emotion. In *Nonverbal Communication: Where Nature Meets Culture*; Segerstrale, U.P., Molnar, P., Eds.; University Of California: San Francisco, CA, USA, 1997; pp. 27–46.
- Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*; Elsevier: Amsterdam, The Netherlands, 1980; pp. 3–33.
- Deep Learning for NLP: An Overview of Recent Trends. Available online: <https://medium.com/dair-ai/deep-learning-for-nlp-an-overview-of-recent-trends-d0d8f40a776d> (accessed on 7 November 2020).
- Jurafsky, D. *Speech & Language Processing*; Pearson Education India: Chennai, India, 2000.
- Manning, C.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
- Farzindar, A.; Inkpen, D. Natural language processing for social media. *Synth. Lect. Hum. Lang. Technol.* **2015**, *8*, 1–166. [\[CrossRef\]](#)
- Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
- Canales, L.; Martínez-Barco, P. Emotion detection from text: A survey. In Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC), Quito, Ecuador, 20–24 October 2014; pp. 37–43.
- Twitter. Available online: <https://twitter.com/home?lang=en> (accessed on 17 November 2020).
- Twitter Developer Docs. Available online: <https://developer.twitter.com/en/docs> (accessed on 7 November 2020).
- Roesslein, J. Tweepy: Twitter for Python! Available online: <https://github.com/tweepy/tweepy> (accessed on 17 November 2020).
- Filter realtime Tweets. Available online: <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters> (accessed on 10 March 2020).

29. The 500 Most Frequently Used Words on Twitter. Available online: <https://techland.time.com/2009/06/08/the-500-most-frequently-used-words-on-twitter/> (accessed on 14 February 2021).
30. NLTP Corpus. Available online: <http://www.nltk.org/howto/corpus.html> (accessed on 17 November 2020).
31. Abbreviations. Available online: <https://www.abbreviations.com/> (accessed on 17 November 2020).
32. Tweet Preprocessor. Available online: <https://pypi.org/project/tweet-preprocessor/> (accessed on 17 November 2020).
33. Emoji. Available online: <https://github.com/carpedm20/emoji/> (accessed on 28 October 2020).
34. Emoji Tracker. Available online: <http://emojitracker.com/> (accessed on 17 November 2020).
35. Krommyda, M.; Rigos, A.; Bouklas, K.; Amditis, A. Emotion detection in Twitter posts: A rule-based algorithm for annotated data acquisition. In Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 16–18 December 2020.
36. Mohammad, S.M.; Turney, P.D. *Nrc Emotion Lexicon*; National Research Council Canada: Ottawa, ON, Canada, 2013; Volume 2.
37. Fellbaum, C. WordNet. In *The Encyclopedia of Applied Linguistics*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2012.
38. Krommyda, M.; Kantere, V. Understanding SPARQL Endpoints through Targeted Exploration and Visualization. In Proceedings of the 2019 First International Conference on Graph Computing (GC), Laguna Hills, CA, USA, 25–27 September 2019; pp. 21–28. [[CrossRef](#)]
39. Krommyda, M.; Kantere, V. A Framework for Exploration and Visualization of SPARQL Endpoint Information. *Int. J. Graph Comput.* **2020**, *1*, 39–69. [[CrossRef](#)]
40. Krommyda, M.; Kantere, V. Improving the Quality of the Conversational Datasets through Extensive Semantic Analysis. In Proceedings of the 2019 IEEE International Conference on Conversational Data Knowledge Engineering (CDKE), San Diego, CA, USA, 9–11 December 2019; pp. 1–8.
41. Krommyda, M.; Kantere, V. Semantic analysis for conversational datasets: Improving their quality using semantic relationships. *Int. J. Semant. Comput.* **2020**, *14*, 395–422. [[CrossRef](#)]
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
43. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [[CrossRef](#)]
44. Nowak, J.; Taspinar, A.; Scherer, R. LSTM recurrent neural networks for short text and sentiment classification. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing, Zakopane, Poland, 11–15 June 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 553–562.
45. Olah, C. Understanding lstm networks. 2015. Available online: <http://colah.github.io/posts/2015-08-Understanding-LSTMs> (accessed on 17 November 2020).
46. Sagheer, A.; Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* **2019**, *323*, 203–213. [[CrossRef](#)]
47. Karevan, Z.; Suykens, J.A. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Netw.* **2020**, *125*, 1–9. [[CrossRef](#)]
48. Miao, K.C.; Han, T.T.; Yao, Y.Q.; Lu, H.; Chen, P.; Wang, B.; Zhang, J. Application of LSTM for Short Term Fog Forecasting based on Meteorological Elements. *Neurocomputing* **2020**, *408*, 285–291. [[CrossRef](#)]
49. Rao, G.; Huang, W.; Feng, Z.; Cong, Q. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing* **2018**, *308*, 49–57. [[CrossRef](#)]
50. Wang, J.; Peng, B.; Zhang, X. Using a stacked residual LSTM model for sentiment intensity prediction. *Neurocomputing* **2018**, *322*, 93–101. [[CrossRef](#)]
51. Rao, A.; Spasojevic, N. Actionable and political text classification using word embeddings and lstm. *arXiv* **2016**, arXiv:1607.02501.
52. Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 421–436.
53. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
54. Kibriya, A.M.; Frank, E.; Pfahringer, B.; Holmes, G. Multinomial naive bayes for text categorization revisited. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Cairns, QLD, Australia, 4–6 December 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 488–499.
55. Zuo, Z. Sentiment Analysis of Steam Review Datasets Using Naive Bayes and Decision Tree Classifier. 2018. Available online: <http://hdl.handle.net/2142/100126> (accessed on 14 February 2021).
56. Al Amrani, Y.; Lazaar, M.; El Kadiri, K.E. Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Comput. Sci.* **2018**, *127*, 511–520. [[CrossRef](#)]
57. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
58. 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.
59. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Available online: [tensorflow.org](https://www.tensorflow.org) (accessed on 14 February 2021).