




Article

Developing Multi-Labelled Corpus of Twitter Short Texts: A Semi-Automatic Method

Xuan Liu ¹, Guohui Zhou ¹, Minghui Kong ¹, Zhengtong Yin ², Xiaolu Li ³, Lirong Yin ⁴
and Wenfeng Zheng ^{5,*}

¹ School of Public Affairs and Administration, University of Electronic Science and Technology of China, Chengdu 611731, China; liuxuan@uestc.edu.cn (X.L.); zhouguohui@std.uestc.edu.cn (G.Z.); kongminghui@std.uestc.edu.cn (M.K.)

² College of Resource and Environment Engineering, Guizhou University, Guiyang 550025, China; ztyin@gzu.edu.cn

³ School of Geographical Sciences, Southwest University, Chongqing 400715, China; xliswu@swu.edu.cn

⁴ Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA; lyin5@lsu.edu

⁵ School of Automation, University of Electronic Science and Technology of China, Chengdu 611731, China

* Correspondence: winfirms@uestc.edu.cn

Abstract: Facing fast-increasing electronic documents in the Digital Media Age, the need to extract textual features of online texts for better communication is growing. Sentiment classification might be the key method to catch emotions of online communication, and developing corpora with annotation of emotions is the first step to achieving sentiment classification. However, the labour-intensive and costly manual annotation has resulted in the lack of corpora for emotional words. Furthermore, single-label semantic corpora could hardly meet the requirement of modern analysis of complicated user's emotions, but tagging emotional words with multiple labels is even more difficult than usual. Improvement of the methods of automatic emotion tagging with multiple emotion labels to construct new semantic corpora is urgently needed. Taking Twitter short texts as the case, this study proposes a new semi-automatic method to annotate Internet short texts with multiple labels and form a multi-labelled corpus for further algorithm training. Each sentence is tagged with both the emotional tendency and polarity, and each tweet, which generally contains several sentences, is tagged with the first two major emotional tendencies. The semi-automatic multi-labelled annotation is achieved through the process of selecting the base corpus and emotional tags, data preprocessing, automatic annotation through word matching and weight calculation, and manual correction in case of multiple emotional tendencies are found. The experiments on the Sentiment140 published Twitter corpus demonstrate the effectiveness of the proposed approach and show consistency between the results of semi-automatic annotation and manual annotation. By applying this method, this study summarises the annotation specification and constructs a multi-labelled emotion corpus with 6500 tweets for further algorithm training.

Keywords: emotion; sentiment corpus; annotation; multi-labelled; Twitter corpus



Citation: Liu, X.; Zhou, G.; Kong, M.; Yin, Z.; Li, X.; Yin, L.; Zheng, W. Developing Multi-Labelled Corpus of Twitter Short Texts: A Semi-Automatic Method. *Systems* **2023**, *11*, 390. <https://doi.org/10.3390/systems11080390>

Academic Editors: Carlos de las Heras-Pedrosa, Francisco Javier Paniagua-Rojano and Dolores Rando-Cueto

Received: 27 June 2023

Revised: 20 July 2023

Accepted: 26 July 2023

Published: 1 August 2023



Copyright: © 2023 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

In recent years, text emotion analysis has increasingly become one of the important research directions in the field of natural language processing. The process of computationally identifying and categorising opinions expressed in a piece of text has been highlighted. In order to achieve this, the most effective and easiest way is to extract and annotate emotional words in texts. The labels could be the positive, negative, or neutral attitude toward a particular topic, product, etc. [1], or multi-dimensional emotional tendencies, such as joy, fear, rage, etc. [2]. Extracted emotional words compose the emotion corpora, which allows for the segmentation and classification of words for the analysis of complicated

emotions [3]. With the continuous progress of sentiment analysis, there are now relatively rich choices of corpora of emotional polarities and limited corpora of emotional tendencies, but the amount of emotion corpora is far from enough for the increasingly detailed needs of emotion analysis. Furthermore, the need for multi-labelled emotion corpora is increasing. Multi-label means that an instance could be classified into multiple tendencies at the same time; that is, it could be marked by multiple labels. Multi-labelled instances in a corpus allow a more complicated analysis of the attitudes of users, especially for a group of sentences that expresses a meaning differing from any sentence in it. Remarkable attempts to propose new multi-labelled emotional dictionaries have been carried out [4,5] but, again, could hardly catch up with the fast emergence of new words.

The reason lies in the way of annotation. Labor-intensive and costly manual annotation is still by far the main way to construct emotion corpora. Although there are now scattered practices of emotion annotation and the construction of emotional dictionaries with the help of human interaction functions on the Internet [6], the quality gap between automatic annotation and manual annotation is somehow obvious. The practices of automatic annotation [7] tested the possibility of quickly obtaining a large number of instances with emotion labels but suffered from the inability to label multiple emotions. Improvement of the methods of emotion annotation is urgently needed.

In summary, two gaps exist for the emotion corpora construction: machine learning methods cannot achieve multi-labelling, and manual annotation is too time-consuming and labour-intensive. In response to the above knowledge gaps, this study proposes a semi-automatic method to annotate English Internet short texts with multiple labels and form a multi-labelled corpus for further algorithm training. Online short texts are the most suitable experimental materials to classify the emotions of the texts. Limited by the length of the text, people tend to use stronger emotional expressions. There is an increasing need for emotion analysis of short texts for author recognition, customer review, Twitter personalisation, etc. [8]. Datasets of short texts, such as Broad Twitter Corpus Dataset, also offer a good base for further extracting emotional words.

To consider possible relations, the research questions are as follows:

RQ1: How to build a method to automatically annotate both the emotional polarity and tendency?

RQ2: How to construct an emotion corpora and evaluate its effectiveness?

The rest of the study is organised as follows: studies about existing emotion annotation in the literature are summarised in Section 2. The method is employed in Section 3. The labelling of the emotional polarities and emotional tendencies of the short texts is completed automatically, followed by manual correction. The experiments and results are presented in Section 4. Lastly, the discussions and conclusions are presented in Sections 5 and 6, respectively.

2. Literature Review

2.1. Corpora of Both Emotional Polarities and Tendencies

A dictionary or corpus of emotions is composed of extracted words with labels of emotions [3]. Mainly two types of emotion corpora have been accumulated through the continuous efforts of researchers. The first group is the corpora with labels of emotional polarities. Emotional polarities mean the positive, negative, or neutral attitude contained by a word. The specific task of such dictionaries is to identify the subjective views—positive, negative, or neutral—expressed in the specified word and form a collection of words with different views. The other group is the dictionaries and corpora with labels of emotional tendencies. The meaning of the emotional tendency here is more like what we usually call “emotion.” Human emotion has been a research hotspot of scholars since ancient times. In the field of modern psychology, the classification of emotions has developed from six primitive emotions proposed in “On Emotions” by Descartes to the multi-dimensional emotion model proposed by American psychologist Plutchik [2,8]. The model proposed by the latter is also known as the emotion wheel. As shown in Figure 1, each emotion in the

emotion wheel has different emotional intensity, and there are also mixed emotions between adjacent emotions. The emotion wheel has been widely applied for emotional tendency annotation in various emotion corpora. The six emotion classification systems proposed by Ekman (1992) [9] are also popular in the English world, which divide emotions into happiness, sadness, anger, fear, disgust, and surprise. A classification system, commonly known as the Chinese emotional classification studies, was proposed by Xu et al. (2008) [10], which adds an emotional tendency of “like” to the six emotion classification systems of Ekman.

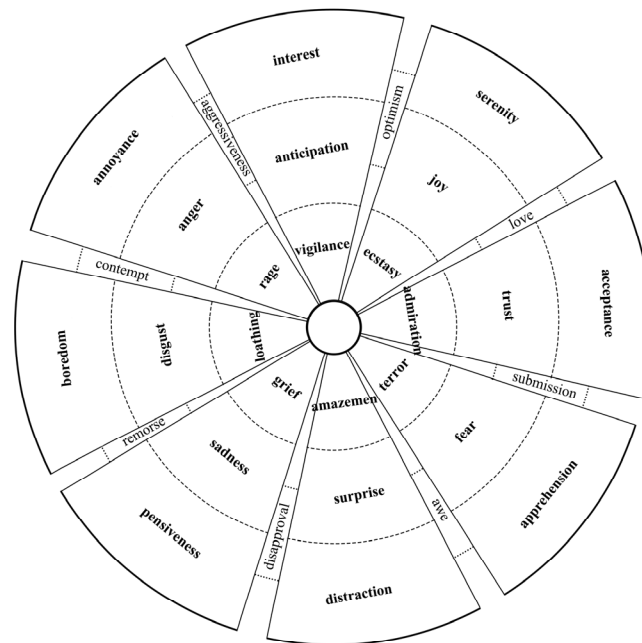


Figure 1. An emotional wheel.

2.2. Progress of Emotion Corpora Construction

The quality of the emotion corpora is the crucial element in defining the effectiveness of the classification [11]. The construction of the corpora with labels of emotional polarities is relatively rich. Corpora such as the annotated Twitter corpus [12], the Sen Tube Corpus based on YouTube Video Comment [13], and HowNet Dictionary for Evaluation have collected thousands of words with emotional polarities. Some of the dictionaries could also include words from languages other than English. The polarity dictionary published by Taiwan University has high accuracy, including 2810 positive words and 8276 negative words. Corpora of words labelled with emotional polarities could also be found for the Arabic [14,15], German [16], or Russian [17] languages.

The other group, the corpora with labels of emotional tendencies, is far more scarce. NRC provides an open dictionary with comprehensive, multilingual emotional words [18]. In the latest published version, emotional words are labelled with eight different emotional tendencies, as well as emotional polarities. By revealing the contextual similarity between two words based on lexical and topic-based features, Matsumoto et al. (2019) [19] labelled the emotional words in a domain-specific corpus. Similarly, Aman and Szpakowicz (2008) [20] used the corpus-based letter combination features to manually annotate the emotional polarities and tendencies. Chinese scholars have contributed more to the construction of dictionaries and corpora with labels of emotional tendencies. A Chinese emotion commonsense knowledge base was built to improve the annotation of emotional polarities and tendencies [21]. Xu, Lin, and Zhao (2008) [10] completed an emotion corpus with 1,035,601 words and 39,488 sentences and defined 23 types of emotional tendencies. Chan et al. (2021) [22] proposed a novel approach to bootstrap a general seed emotion lexicon

with words found in a domain-specific corpus. The approach divulges the contextual similarity between two words to reveal the emotional tendency labels of domain-specific words.

Recently, the need for multi-labelled emotional words has been rising as more complex sentiment analysis requires multi-label classification of texts, which means that an instance could be marked by multiple labels. In order to achieve high-precision text classification, researchers are also working hard to construct emotional corpora or dictionaries. Yang et al. (2014) [4] proposed a small dictionary with both graphic emoticons and punctuation marks to annotate texts on Weibo. Liu and Chen (2015) [5] combined three emotional dictionaries to extract the features of word segmentation in the Weibo corpus. Li et al. (2016) [23] adjusted the imbalance of emotional tendency distribution of the corpus by adopting a multi-label maximum entropy model. But, overall, the lack of multi-label corpora is still a pain point.

2.3. Methods to Construct Multi-Labelled Emotion Corpora

The specific task of sentiment classification is to identify the subjective views expressed in the specified text and judge the emotional tendencies of the text [23–25]. From dictionary-based emotion classification [3] to machine learning classifiers [26], various methods have been applied for emotion classification. The dictionary-based emotion classification generally segments words in the text to be classified and matches the keywords with the labelled words in the dictionaries of emotions. Additionally, it carries out further operations such as considering the emotion intensity [27]; adding topic-related features [28]; or introducing rules [29], mutual information [30], physiological signals [31], and neural networks [32], etc. The machine learning classifiers, on the contrary, allow auto recognition of emotional words by following a sequential process. The first step is training the classifier with existing dictionaries or corpora of emotions, followed by the selection of features, the adjustment of classifiers, and the application of sentiment analysis [33,34].

Traditional corpus construction and annotation methods of emotional dictionaries are usually based on manual annotation, which is labour-intensive and time-consuming, especially when multiple labels are required for each word. Automatic emotion annotation is still rare. Japanese researchers have automatically annotated a 5 billion Japanese blog corpus and classified ten different emotional tendencies [7]. With the help of human interaction functions on the Internet, emotion annotation and the construction of emotional dictionaries are also achieved. For example, a corpus of 132 emotions was obtained and labelled from the blog site Livejournal [35]. Such practices tested the possibility of quickly obtaining a large number of instances with emotional labels but suffered from the inability to label multiple emotions since a sequential process is applied, and the previous steps would heavily impact the following steps [23]. How to maintain the accuracy of manual annotation and improve the speed of annotation with machine learning algorithms has now drawn the attention of scholars but still needs further effort.

3. Materials and Methods

3.1. Workflow

As shown in Figure 2, the main process of forming a corpus of short texts with labels of both emotional polarities and tendencies can be divided into four main steps. First, this study defines the research scope by selecting the base corpus and labels for both emotional polarities and tendencies. Second, data preprocessing is applied, which includes the process of removing user identities, removing other noise information, deleting strings shorter than three words, word correction, spell check, and word form restoration. These processes prepare the data for further analysis. Third, the settings for emotion annotation follow the data processing. This sets the criteria to evaluate and annotate the emotions of short texts: emotion annotation with multiple labels is applied to tweets. Finally, automatic emotion annotation is applied separately for emotional polarities and tendencies. This includes (1) the words matching, during which a vocabulary weight value set is offered for each text; (2) the calculation of the weight of emotions, which calculates the overall emotional

polarity weight vector and emotional tendency weight vector separately; (3) judgment of emotional polarities and tendencies; (4) manual correction in case multiple emotional tendencies exist for one instance; and (5) verification by comparing the performance of the semi-automatic annotation with the origin corpus.

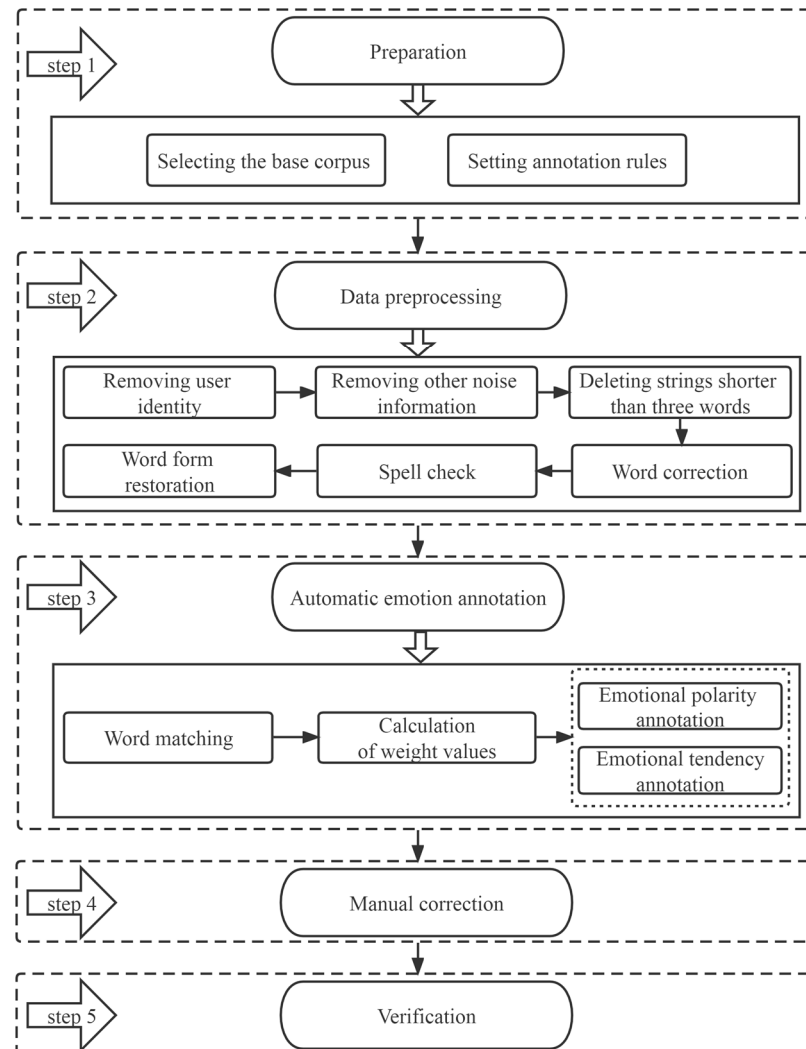


Figure 2. Workflow.

3.2. Preparation

3.2.1. Selection of the Base Corpus

Corpora for short text sentiment analysis tasks are usually constructed according to actual experimental requirements. The length of a tweet is limited to 140 characters. This resulted in simple and dense emotional words in tweets. Multiple emotions are expressed in one single tweet frequently and make Tweets perfect samples for sentiment analysis. The Sentiment140 Twitter corpus¹ contains 1,600,000 tweets extracted using the Twitter API [36]. In this study, 8000 English Tweets were randomly selected from Sentiment140 for annotation. After filtering the meaningless text, 6500 Tweets were finally retained, containing 11,338 sentences, which cover a wide range of content and have the common text characteristics of short Internet texts.

3.2.2. Setting Annotation Rules

In order to annotate both emotional polarity and emotional tendency, the emotion word dictionary with 105 languages disclosed by NRC was chosen, and it was abbreviated in this study as NRC Dictionary. The dictionary not only marks the emotional polarity

of words but also marks the tendency of each word with the eight emotions prosed by Plutchik. The labels 0 and 1 in the dictionary, respectively, represent the corresponding emotional polarity or emotional tendency. Emotional polarity annotation is divided into three tendencies: 0, 1, and 2, which correspond to neutral, positive, and negative, respectively. Emotional tendencies are divided into nine tendencies, represented by numbers 0, 1, 2, 3, 4, 5, 6, 7, and 8. Eight tendencies have emotions such as anger, disgust, fear, sadness, anticipation, joy, surprise, and trust. The correspondence between the labelled numbers and their meaning are shown in Table 1:

Table 1. Correspondence between emotional tendency and number labelling.

0	1	2	3	4	5	6	7	8
Neutral	Anger	Disgust	Fear	Sadness	Anticipation	Joy	Surprise	Trust

In this study, 500 tweets were randomly selected, and the emotional expressions were labelled, as shown in Figure 3. According to the result, about 60% of the instances have emotions, and about 20% of emotional tweets contain two or more emotions, which must be displayed when annotating emotions. Therefore, this study selected the following rules to annotate tweets: when only one emotion appears, the tweet is annotated to that type of emotion; when two or more emotions appear in a tweet, one of them is annotated as the main emotion, and the other as a secondary emotion. Tweets with no emotion had both primary and secondary emotions annotated as “none”; tweets with a single emotion had their secondary emotion annotated as “none”. For each sentence, the study only annotates the presence and absence of emotions and the main emotional tendency. This study followed existing manual annotations of the Sentiment140 Twitter corpus to perform the manual correction².

```
<tweet id="2408" emotion_type1="anticipation" emotion-type2="none">
<sentence id="1" opinionated="Y" emotion_type="anticipation">Need a
hug</sentence>
<sentence id="2" opinionated="N">Good night</sentence>
</tweet>
<tweet id="3218" emotion_type1="anticipation" emotion-type2="joy">
<sentence id="1" opinionated="Y" emotion_type="sad"> Although they
lost unexpectedly in the first round</sentence>
<sentence id="2" opinionated="Y" emotion_type="anticipation">I still
have confidence in their comeback. </sentence >
< sentence id="3" opinionated="Y" emotion_type="joy">They must win
there!!</sentence >
</tweet>
```

Figure 3. Examples of separate annotations for sentences and tweets.

3.3. Data Preprocessing

In order to remove the interference information in the instances, specific preprocessing steps were applied as follows:

3.3.1. Remove User Identity

In online texts, user names are frequently placed after the “@” symbol to forward the tweet to a specified user. This results in the fact that a tweet often contains a part of “@username”. Even when a meaningful word appears as a user name, its meaning should not be considered for emotional annotation for sentences as well as full tweets. In order to avoid the bias of the subsequent emotion word matching and weight calculation that might be brought by the usernames, this study removed the username by deleting the string starting with “@” and ending with a space in the regular matching text.

3.3.2. Remove Other Noise Information

The emotional words matched in the emotional weight calculation method are all in English. This study removed the punctuation, numbers, special characters, and other contents in the texts to avoid their impacts on the efficiency of matching. However, the content behind a “#” generally refers to annotation and often contains words with a strong emotional tendency. Such contents would be contained as emotional words to participate in subsequent matching and emotional weight calculation.

3.3.3. Delete Strings Shorter Than Three Words

A major feature of the NRC dictionary is that the length of the emotional string is not less than three words. It means strings shorter than three words could not be successfully matched in the emotional dictionary and would cause invalid iteration through the dictionary. In order to avoid invalid iteration and speed up the overall matching speed, this study deleted strings shorter than three words in preprocessing. As shown in Figure 3, “opinionated = ‘N’” means the sentence (good night) would not be included in the following analysis since the words in this sentence are less than three.

3.3.4. Word Correction

Compared with the traditional text corpus, short texts on the Internet are more colloquial. The original data in the Twitter corpus are open-access texts published by real users, containing a large number of words such as “cooooooool” and “whyyyyyyyyy.” These words need to be included in the judgment of emotional tendency as they express strong emotion. However, such words could not match with any words in the emotion dictionary and would result in a low accurate rate of matching. In this study, the letters that appear more than three times in a word were identified and replaced with two letters. For example, “coooooool” is restored to “cool” and “whyyyyyy” is restored to “why”. For the words that are still in the wrong form after the correction, spell checking would adjust them into the right words as the difference between the wrong and right forms is reduced to one letter.

3.3.5. Spell Check

Spelling problems are frequently met in the instances in the Twitter corpus and would also lead to the failure of word matching. Checking and correcting spelling is an important step in preprocessing. The main principle and basis of English spelling checking in this paper are the Bayesian algorithm and editing distance:

The study records the correct spelling as C (for correct) and the wrong spelling as W (for wrong). The task of spell checking is to infer that a C given a W occurs, which also means finding the most likely C from several alternatives on the premise that W is known. According to the Bayesian theorem, the task is to find the Maximum value of $P(C|W)$ in Formula (1):

$$P(C|W) = \frac{P(W|C) \cdot P(C)}{P(W)} \quad (1)$$

In this formula, $P(C)$ indicates the probability of the occurrence of a correct word, which can be simulated based on the text library. The higher the frequency of a word in the text library, the bigger its frequency $P(C)$. $P(W|C)$ indicates the probability of misspelling as W for the original word C . In order to simplify the problem, this study assumes that the probability of misspelling increases as the two words look more similar to each other. This assumption turns the misspelling problem into an edit distance problem. Spelling check is thus to check the frequency of all words similar to the spelled word in the text library. The word with the highest probability is the right word that the user really wants to input.

3.3.6. Word Stemming

There are a large number of word inflections in English vocabulary. For example, the words “interest”, “interesting”, and “interested” share the same stem, “interest”, but have different expressions. For such inflected words, extracting their stems can effectively improve the efficiency of word matching. Word stemming refers to the process of removing the prefix and suffix of a word to obtain the stems. It also includes converting words to their original general forms according to the dictionary when the difference between the inflected and the original word is not a prefix or suffix.

In this study, the stem was extracted by snowball method in the NLTK library.

3.4. Automatic Emotion Annotation

The main principle of the emotion labelling method proposed in this paper is based on the word matching and weight calculation of the emotion dictionary. Through the discrimination method proposed, the emotional tendency of each instance is judged and automatically labelled if the emotional tendency is obvious. Otherwise, a manual check would be carried out until, eventually, an accurate emotional corpus is obtained. The specific matching and discrimination methods are introduced as follows.

3.4.1. Words Matching

When matching, each piece of instance is regarded as a set of words. In this study, the words separated by spaces in the instance are marked in sequence with $w_1, w_2, w_3, \dots, w_l$, where i represents the position of the word in the corresponding instance. In order to match the words, the i -th word is compared with the English column of the emotional dictionary. The item that is successfully matched is marked with p_i for the result of positive emotional polarity judgement, n_i for negative emotional polarity, p_{i1} for anticipation, p_{i2} for joy, p_{i3} for surprise, p_{i4} for trust, n_{i1} for anger, n_{i2} for disgust, n_{i3} for fear, and n_{i4} for sadness. If the composite word cannot be matched, the word is considered not to contain the emotional tendency weight value and the corresponding weight values of $p_i, n_i, p_{i1}, p_{i2}, p_{i3}, p_{i4}, n_{i1}, n_{i2}, n_{i3}, n_{i4}$ are all recorded as 0.

For each text, a vocabulary weight value set is obtained after matching in the form of a set of vectors with a length of 10, as shown in Formula (2).

$$w_i = (p_i, n_i, p_{i1}, p_{i2}, p_{i3}, p_{i4}, n_{i1}, n_{i2}, n_{i3}, n_{i4}) \quad (2)$$

3.4.2. Calculation of the Weight of Emotions

For each piece of short text, the study calculates the overall emotional polarity weight vector and emotional tendency weight vector separately:

$$v = \left(\sum_i^l p_i, \sum_i^l n_i \right) \quad (3)$$

$$q = (\sum_i^l p_{i1}, \sum_i^l p_{i2}, \sum_i^l p_{i3}, \sum_i^l p_{i4}, \sum_i^l n_{i1}, \sum_i^l n_{i2}, \sum_i^l n_{i3}, \sum_i^l n_{i4}) \quad (4)$$

3.4.3. Judgment of Emotional Polarity

The proportional value of positive and negative emotional polarity is calculated according to Formulas (5) and (6):

$$M_p = \frac{v_0}{v_0 + v_1} \quad (5)$$

$$M_n = \frac{v_1}{v_0 + v_1} \quad (6)$$

Here, M_p represents the proportional value of positive emotional polarity; M_n represents the proportional value of negative emotional polarity; and v_0 and v_1 represent the first term $\sum_i^l p_i$ and the second term $\sum_i^l n_i$ of the emotion weight vector v , respectively.

This study sets the intensity threshold k_0 . If $M_p \geq k_0 + \delta$, in which δ is constant, the text is regarded to have positive emotional polarity. If $M_n \geq k_0 + \delta$, the text is considered negative; otherwise, it is neutral, and the text does not have strong emotional polarity.

3.4.4. Judgment of Emotional Tendency

This study calculates the sum of emotional tendencies with Formula (7).

$$M = \sum_0^7 q_i \quad (7)$$

The weight ratio corresponding to each emotional tendency is calculated according to Formula (8), where m is the set of all weights. m_{max} is the maximum value in the set m . m_{min} is the minimum value in the set m .

$$\begin{aligned} M_{p1} &= \frac{q_0}{M}, M_{p2} = \frac{q_1}{M}, M_{p3} = \frac{q_2}{M}, M_{p4} = \frac{q_3}{M} \\ M_{n1} &= \frac{q_4}{M}, M_{n2} = \frac{q_5}{M}, M_{n3} = \frac{q_6}{M}, M_{n4} = \frac{q_7}{M} \\ m &= (M_{p1}, M_{p2}, M_{p3}, M_{p4}, M_{n1}, M_{n2}, M_{n3}, M_{n4}) \end{aligned} \quad (8)$$

The study sets the upper limit of the emotional tendency threshold k_1 ($0 < k_1 < 1$) and the lower limit of the emotional tendency threshold k_2 ($0 < k_2 < k_1$), and calculates the extreme difference $r = m_{max} - m_{min}$. For the above threshold value, the value of k_1 was set to be 0.20 and k_2 is 0.07 in this study, based on the preliminary study of the emotional tendency of the randomly chosen 500 tweets. This means that, out of the percentage of all emotional tendency weights, if the difference between the strongest emotion and the weakest emotion reaches 20 percent or more of the overall emotional tendencies, it can be determined that the emotional tendency of the text is the tendency with a larger weight. If the difference between the strongest emotion and the weakest emotion does not reach 7 percent of the overall emotional tendency, it is considered that the text has no obvious emotional tendency.

If $r \geq k_1$ and the emotional tendency weight value is m_{max} , it means that there is only one emotional tendency. At this time, the research marks the emotional tendency value corresponding to the instance as the value corresponding to the emotional tendency with the emotional tendency value m_{max} . When $r \leq k_2$, the emotional tendency of the text is marked as 0, representing neutral. If $k_1 \leq r \leq k_2$ or $r \geq k_1$, and multiple emotional tendencies with the weight value of m_{max} exist, the emotional tendency of the instance would be set to 9, indicating that further manual verification is needed.

3.5. Manual Correction

For the instance where the emotional tendency was 9, emotion annotation was carried out again manually. When performing large-scale manual emotion annotation, it should be carried out by an annotation team of no less than three people, and they will label all instances that the automatic annotation method fails to judge the emotional tendency. Since different people often have different understandings of the same sentence, the emotional tendency of the text is labelled differently when more than half of the annotators are consistent with each other. Otherwise, further discussion must be carried out until a unanimous judgement is reached.

3.6. Verification

Since we semi-automatically annotated 6500 tweets that were manually labelled in the origin corpus (Sentiment140), the results of the original manual annotation could be used to test the performance of the semi-automatic annotation.

In order to evaluate the performance of the emotional polarity annotation, indicators of accuracy rate (*Accuracy*), precision rate (*Precision*), recall rate (*Recall*), and F Score (*F-Score*) were calculated. The specific definition and calculation methods are as follows:

Accuracy refers to the ratio of the number of tweets whose semi-automatic annotation results are consistent with the corpus annotation results to the total number of tweets. It is calculated with Formula (9).

$$Accuracy = \frac{machine_correct}{machine_all} \quad (9)$$

Precision refers to the ratio of the number of tweets semi-automatically labelled as positive emotions to the number of tweets originally labelled as emotional tweets. It is calculated with Formula (10).

$$Precision = \frac{machine_correct(pos)}{machine_marked} \quad (10)$$

Recall refers to the ratio of the number of semi-automatically labelled tweets with emotional polarity to the number of tweets with emotions in the original corpus. It is calculated with Formula (11).

$$Recall = \frac{machine_correct(pos)}{manual(pos)} \quad (11)$$

F-Score refers to the harmonic mean of precision and recall. It is calculated with Formula (12).

$$F\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

In the above four formulas, *machine_correct* represents the number of semi-automatically labelled tweets that are consistent with the original corpus, *machine_marked* represents the number of semi-automatically labelled tweets, and *manual* represents manual annotation (in this case, the result of annotation in the original corpus).

In order to evaluate the validity of emotional tendency annotation, indicators include *Precision*, *Recall*, and *F-Score* at both micro and macro levels. The specific calculation methods are as follows:

$$Micro_Precision = \frac{\sum_{i=1}^8 machine_correct(i)}{\sum_{i=1}^8 machine_marked(i)} \quad (13)$$

$$Micro_Recall = \frac{\sum_{i=1}^8 machine_correct(i)}{\sum_{i=1}^8 manual(i)} \quad (14)$$

$$\text{Micro_F-score} = \frac{2 \times \text{Micro_Precision} \times \text{Micro_Recall}}{\text{Micro_Precision} + \text{Micro_Recall}} \quad (15)$$

$$\text{Macro_Precision} = \frac{1}{8} \sum_{i=1}^8 \frac{\text{machine_correct}(i)}{\text{machine_marked}(i)} \quad (16)$$

$$\text{Macro_Recall} = \frac{1}{8} \sum_{i=1}^8 \frac{\text{machine_correct}(i)}{\text{manual}(i)} \quad (17)$$

$$\text{Macro_F-score} = \frac{2 \times \text{Macro_Precision} \times \text{Macro_Recall}}{\text{Macro_Precision} + \text{Macro_Recall}} \quad (18)$$

Again, *machine_correct* here represents the number of semi-automatically annotated tweets that are consistent with the original corpus, *machine_marked* represents the number of semi-automatically annotated tweets, *manual* represents manual annotation (in this case, the result of annotation in the original corpus), and *i* represents one of the eight emotional tendencies annotated.

The selection of threshold impacts the annotation results to some extent: the annotation of emotional polarities will be affected by the value of the threshold k_0 , and the annotation of emotional tendencies will be affected by the values of the threshold k_1 and threshold k_2 . The selection of thresholds has no empirical formula but is rather customised and refined according to the characteristics of the corpus. As shown in the explanation of Formula (8), setting the value of k_1 as 0.20 and k_2 as 0.07 constitutes the best-performing threshold in the preliminary study of the emotional tendency of the randomly chosen 500 tweets. This threshold is set as Group 1. In order to better verify the effectiveness of the annotation method, the study selects two sets of thresholds for semi-automatic annotation to analyse and compare the annotation results. Group 2 is the second-best combination of thresholds in the preliminary study of the randomly chosen 500 tweets. The specific threshold selection is shown in Table 2.

Table 2. Selection of threshold for semi-automatic annotation.

	k_0	k_1	k_2
Group 1	0.15	0.20	0.07
Group 2	0.18	0.15	0.05

4. Results

4.1. The Corpus Constructed

As mentioned above, 8000 tweets were randomly selected from Sentiment140 for annotation in this study. A total of 6500 tweets remain after filtering, containing 11,338 sentences with various contents. Through the above steps, this research finally obtained a corpus for short texts with both emotional polarity labels and emotional tendency labels. The detailed statistical information of the corpus is listed in Tables 3 and 4, respectively.

Table 3. Proportion of sentence-level emotion.

	Emotional Sentences	Non-Emotional Sentences	Total
Quantity	6236	5102	11,338
Proportion	55.00%	45.00%	1.00

Table 4. Proportion of tweet-level emotion.

	Emotional Tweet		Emotionless Tweet	Total
	With One Emotion	With Two Emotions		
Quantity	3080	972	2448	6500
Proportion	47.39%	14.95%	37.66%	1
Total	62.339%		37.66%	/

This study also counts the frequency of each emotional tendency as the primary or secondary emotional tendency in all emotional sentences and tweets.

Tables 5 and 6 shows that Twitter has the largest number of words with the emotional tendencies of “anticipation” and “joy” as the main emotional tendency, followed by “trust”. As for the secondary emotional tendency, “trust” lists first, followed by “sadness” and “joy”. “Surprise” and “fear” account for only a small proportion of both the primary and secondary emotional tendencies.

Table 5. The distribution of emotional tendencies in the emotional sentences.

	Sentence Emotion	Proportion
Anger	817	13.10%
Disgust	251	4.03%
Fear	172	2.76%
Sadness	712	11.42%
Anticipation	1371	21.99%
Joy	1595	25.63%
Surprise	211	3.38%
Trust	1107	17.75%
Total	6236	1.00

Table 6. The distribution of emotional tendencies in the emotional Tweets.

	Tweet’s Main Emotion	Tweet’s Secondary Emotion
Anger	378	58
Disgust	185	83
Fear	95	29
Sadness	535	167
Anticipation	1008	95
Joy	992	159
Surprise	79	39
Trust	780	342
Total	4052	972

4.2. Emotional Accompaniment

A unique characteristic that deserves to notice is that an emotion appearing in a tweet is always accompanied by a secondary emotion. For example, when “Fear” appears, the secondary emotion is likely to be “Disgust”. We defined such a combination of two emotional tendencies as an emotional accompaniment. According to the sequential relationship of emotional tendencies, the emotional accompaniments between every two emotional tendencies are two ordered combinations. For the eight emotional tendencies in this study, there are at most 64 possible combinations. Table 7 shows the possibility of each emotional accompaniment in emotional tweets. The emotional accompaniment probability is calculated by the conditional probability of accompanying emotions when the main emotion appears, following Formula (19). The count here represents the result of statistics.

$$\begin{aligned}
 &P(\text{Secondary emotions}|\text{Primary emotions}) \\
 &= \frac{\text{count}(\text{Primary emotions}, \text{Secondary emotions})}{\text{count}(\text{Primary emotions})} \quad (19)
 \end{aligned}$$

Table 7. Emotional companion relationships in emotional tweets (percent).

	Anger	Disgust	Fear	Sadness	Anticipation	Joy	Surprise	Trust	Not Exist
Anger	/	7.30	7.57	10.54	3.51	2.97	2.70	2.97	62.43
Disgust	6.02	/	6.33	11.14	6.33	3.01	1.81	1.81	67.47
Fear	7.90	4.66	/	9.72	3.24	3.64	0.81	4.05	65.99
Sadness	6.28	4.46	6.14	/	3.35	3.07	2.09	5.02	69.60
Anticipation	1.51	0.27	2.40	2.23	/	5.97	3.39	6.59	77.65
Joy	0.92	0.26	1.71	2.49	9.04	/	6.42	12.45	66.71
Surprise	1.98	0.74	1.49	5.20	10.64	9.65	/	10.46	59.90
Trust	1.70	0.63	3.21	2.95	7.14	9.20	3.21	/	71.96

Obviously, the emotional accompaniment of the emotional tendencies with the same emotional polarity is more frequently met. Among all the eight emotional tendencies, when anticipation appears as the main emotional tendency, there is a possibility of as high as 77.65 percent that no emotional tendency appears as the accompanying emotional tendency. When the main emotional tendency appears to be anger, it is most likely to be accompanied by other emotional tendencies, among which sadness, nausea, and fear are most likely to appear.

4.3. Verification

In this study, the annotation results with the semi-automatic method of annotation for both emotional polarity and tendency were calculated and compared with the manually annotated results in the base corpus. Tables 8 and 9 show that semi-automatic emotional polarity annotation and emotional tendency annotation have achieved good results. For the annotation of emotional tendency, the multi-labelled annotation has a certain proportion to annotate “9”, which is invalid for the comparison results. This has resulted in a low recall rate of around 50 percent.

Table 8. Evaluation results of emotional polarity annotation for tweets.

	Accuracy	Precision	Recall	F Score
Group 1	0.4251	0.6943	0.7681	0.72934
Group 2	0.3551	0.7160	0.5696	0.63446

Table 9. Evaluation results of emotional tendency annotation for tweets.

	Micro Mean Value			Macro Mean Value		
	Accuracy	Recall	F Score	Accuracy	Recall	F Score
Group 1	0.7825	0.5514	0.6470	0.7210	0.4486	0.5531
Group 2	0.7850	0.5558	0.6508	0.7236	0.4527	0.5569

5. Discussion

5.1. Further Application of the Method on a Larger Corpus

With the new semi-automatic annotation method, this study selected 100,000 tweets in the sentiment140 published Twitter corpus for semi-automatic annotation to form a larger corpus. After data filtering, 99,333 tweets were retained. The selection of thresholds was consistent with Table 2. The results of the emotional tendency annotation for the larger corpus are shown in Table 10. The indicators to evaluate the performance of the emotional polarity annotation are shown in Table 11. When the threshold is selected properly, the precision rate and recall rate can reach more than 70 percent, which fully demonstrates the effectiveness of the annotation method proposed for the emotional polarity annotation.

Table 10. Emotional tendency distribution of tweets for the large-scale corpus.

		Anger	Disgust	Fear	Sadness	Anticipation	Joy	Surprise	Trust	Neutral
Group 1	quantity	3587	1997	6762	7656	15,903	6006	1199	6847	20,164
	%	5.115	2.848	9.643	10.918	22.679	8.565	1.710	9.765	28.756
Group 2	quantity	3607	2029	6801	7757	16,026	6047	1209	6955	20,008
	%	5.12	2.88	9.66	11.01	22.75	8.59	1.72	9.87	28.41

Table 11. Evaluation results of emotional polarity annotation task on large-scale Twitter corpus.

	Accuracy	Precision	Recall	F Score
Group 1	0.43266	0.7293	0.7626	0.74558
Group 2	0.36406	0.7370	0.5725	0.64442

5.2. Possible Improvement

The accuracy of the automatic labelling method proposed in this paper can be further improved. The directions that can be considered include adding more conditional restrictions in weight calculation and judgment, combined with updated deep learning algorithms, adding additional labels to inflected terms to show their strong emotional tendencies, etc. In addition, the application of the corpus is currently limited to the annotation of English instances. Future research can try to adapt to multilingual environments to meet more extensive emotion annotation needs.

6. Conclusions

Labor-intensive and costly manual annotation for the corpus has resulted in the lack of multi-labelled semantic corpus and hampered the training of algorithms for the services to online users. By combining both emotion dictionaries and manual correction, this study proposes a new method of semi-automatic emotion annotation for Internet short texts. Each instance is labelled with one emotional polarity and one emotional tendency. For the full tweet, the first two major emotional tendencies are labelled, allowing a more complicated and accurate analysis of the user's emotions. The experiments on the Sentiment140 published Twitter corpus demonstrate the effectiveness of the proposed approach and show the consistency between the results of semi-automatic annotation and manual annotation. Besides the introduction of the design of the emotion annotation specification of Twitter texts, we also formulated the corresponding annotation criteria and introduced the rules and process to construct a corpus with emotion labels.

This method is a beneficial attempt, which tries to complete the emotional annotation of short texts more efficiently, and at a lower cost. It might also be a start to promote the expansion of corpora for sentiment analysis of texts in the field of natural language processing.

Author Contributions: Conceptualization: X.L. (Xuan Liu) and W.Z.; methodology: L.Y.; software: X.L. (Xiaolu Li) and Z.Y.; data acquisition: G.Z. and M.K.; formal analysis: Z.Y. and X.L. (Xiaolu Li); data curation: G.Z. and Z.Y.; validation: X.L. (Xiaolu Li) and Z.Y.; writing—original draft preparation: X.L. (Xuan Liu) and L.Y.; writing—review and editing: X.L. (Xuan Liu), L.Y. and W.Z.; visualization: L.Y. and G.Z.; project administration and supervision: W.Z.; and funding acquisition: W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Sichuan Science and Technology Program, grant number [2021YFQ0003].

Data Availability Statement: The data were derived from the following resources available in the public domain: [Sentiment140. Twitter Corpus. Available online: <https://www.kaggle.com/datasets/kazanov/sentiment140> (accessed on 28 August 2020)].

Acknowledgments: Thanks to Tianyi Shi and Xiaobing Chen for their work and contributions to data collection, processing, and analysis in this study.

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

Notes

- ¹ Sentiment140. Twitter Corpus. Available online: <https://www.kaggle.com/datasets/kazanova/sentiment140> (accessed on 28 August 2020).
- ² Please refer to the research papers “Target-dependent Twitter Sentiment Classification” (<https://aclanthology.org/P11-1016.pdf>) and the resource “Sentiment140” (<http://help.sentiment140.com/for-students/>). These sources provide insights into the process and criteria used for sentiment classification in Twitter data.

References

- Feng, X.; Hui, K.; Deng, X.; Jiang, G. Understanding how the semantic features of contents influence the diffusion of government microblogs: Moderating role of content topics. *Inf. Manag.* **2021**, *58*, 103547. [CrossRef]
- Hu, A.; Flaxman, S. Multimodal sentiment analysis to explore the structure of emotions. In Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018.
- Ai, Y.; Chen, Z.; Wang, S.; Pang, Y. Recognizing emotions in chinese text using dictionary and ensemble of classifier. In Proceedings of the Third International Workshop on Pattern Recognition, Jinan, China, 26–28 May 2018.
- Yang, J.; Jiang, L.; Wang, C.; Xie, J. Multi-label emotion classification for tweets in weibo: Method and application. In Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI), Limassol, Cyprus, 10–12 November 2014.
- Liu, S.M.; Chen, J.H. A multi-label classification based approach for sentiment classification. *Expert Syst. Appl.* **2015**, *42*, 1083–1093. [CrossRef]
- Shah, F.M.; Reyadh, A.S.; Shaafi, A.I.; Ahmed, S.; Sithil, F.T. Emotion detection from tweets using AIT-2018 dataset. In Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, Bangladesh, 26–28 September 2019.
- Ptaszynski, M.; Rzepka, R.; Araki, K.; Momouchi, Y. Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. *Comput. Speech Lang.* **2014**, *28*, 38–55. [CrossRef]
- Liang, L.; Tian, F. Using normal dictionaries to extract multiple semantic relationships. *J. Eng.* **2020**, *2020*, 595–600. [CrossRef]
- Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
- Xu, L.; Zhao, C. Construction and analysis of affective corpus. *J. Chin. Inf.* **2008**, *22*, 116–122. (In Chinese)
- Ji, Q.; Raney, A.A. Developing and validating the self-transcendent emotion dictionary for text analysis. *PLoS ONE* **2020**, *15*, e0239050. [CrossRef] [PubMed]
- Pak, A.; Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation LREC, Valletta, Malta, 17–23 May 2010.
- Uryupina, O.; Plank, B.; Severyn, A.; Rotondi, A.; Moschitti, A. SenTube: A Corpus for Sentiment Analysis on YouTube Social Medi. In Proceedings of the 9th International Conference on Language Resources and Evaluation LREC, Reykjavik, Iceland, 26–31 May 2014.
- Refaee, E.; Rieser, V. An arabic twitter corpus for subjectivity and sentiment analysis. In Proceedings of the 9th International Conference on Language Resources and Evaluation LREC, Reykjavik, Iceland, 26–31 May 2014.
- Guellil, I.; Azouaou, F.; Mendoza, M. Arabic sentiment analysis: Studies, resources, and tools. *Soc. Netw. Anal. Min.* **2019**, *9*, 1–17. [CrossRef]
- Clausen, Y.; Scheffler, T. A corpus-based analysis of meaning variations in German tag questions Evidence from spoken and written conversational corpora. *Corpus Linguist. Linguist. Theory* **2022**, *18*, 1–31. [CrossRef]
- Svetlov, K.; Platonov, K. Sentiment analysis of posts and comments in the accounts of russian politicians on the social network. In Proceedings of the 2019 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 5–8 November 2019.
- Mohammad, S.; Turney, P. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Los Angeles, CA, USA, 5–6 June 2010.
- Matsumoto, K.; Sasayama, M.; Yoshida, M.; Kita, K. Emotional state estimation by dialogue history and sentence distributed representation. In Proceedings of the 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), Singapore, 19–21 December 2019.
- Aman, S.; Szipakowicz, S. Using roget’s thesaurus for fine-grained emotion recognition. In Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 7–12 January 2008.
- Yang, L.; Zhou, F.; Lin, H.; Wang, J.; Zhang, S. Chinese emotion commonsense knowledge base construction and its application. In Proceedings of the Workshop on Chinese Lexical Semantics, Chiayi, Taiwan, 26–28 May 2018.
- Chan Samuel, W.K. Multilabel Emotion Tagging for Domain-Specific Texts. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 1197–1210. [CrossRef]
- Li, J.; Rao, Y.; Jin, F.; Chen, H.; Xiang, X. Multi-label maximum entropy model for social emotion classification over short text. *Neurocomputing* **2016**, *210*, 247–256. [CrossRef]

24. Rajabi, Z.; Shehu, A.; Uzuner, O. A multi-channel bilstm-cnn model for multilabel emotion classification of informal text. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020.
25. Fei, H.; Ji, D.; Zhang, Y.; Ren, Y. Topic-enhanced capsule network for multi-label emotion classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1839–1848. [[CrossRef](#)]
26. Ullah, S.; Talib, M.R.; Rana, T.A.; Hanif, M.K.; Awais, M. Deep Learning and Machine Learning-Based Model for Conversational Sentiment Classification. *Comput. Mater. Contin.* **2022**, *72*, 2323–2339. [[CrossRef](#)]
27. Aman, S.; Szapkowicz, S. Identifying expressions of emotion in text. In Proceedings of the Text, Speech and Dialogue: 10th International Conference, Pilsen, Czech Republic, 3–7 September 2007.
28. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [[CrossRef](#)]
29. Yan, D.; Hu, B.; Qin, J. Sentiment analysis for microblog related to Finance based on rules and classification. In Proceedings of the 2018 IEEE International Conference on Big Data and Smart computing (BigComp), Shanghai, China, 15–17 January 2018.
30. Liu, H.; Guo, H.; Hu, W. Eeg-based emotion classification using joint adaptation networks. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021.
31. Shu, L.; Xie, J.; Yang, M.; Li, Z.; Li, Z.; Liao, D.; Xu, X.; Yang, X. A review of emotion recognition using physiological signals. *Sensors* **2018**, *18*, 2074. [[CrossRef](#)] [[PubMed](#)]
32. Tang, Y.; Su, J.; Khan, M.A. Research on sentiment analysis of network forum based on BP neural network. *Mob. Netw. Appl.* **2021**, *26*, 174–183. [[CrossRef](#)]
33. Dogan, T.; Uysal, A.K. A novel term weighting scheme for text classification: Tf-mono. *J. Informetr.* **2020**, *14*, 101076. [[CrossRef](#)]
34. Sintsova, V.; Musat, C.; Pu, P. Semi-supervised method for multi-tendency emotion recognition in tweets. In Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China, 14 December 2014.
35. Mishne, G. Experiments with mood classification in blog posts. In Proceedings of the ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access, Salvador, Brazil, 15–19 August 2005.
36. Go, A.; Bhayani, R.; Huang, L. Twitter sentiment classification using distant supervision. *CS224N Proj. Rep. Stanf.* **2009**, *1*, 2009.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.