applied sciences

MDPI

Article

An Explainable Artificial Intelligence Approach for Detecting Empathy in Textual Communication

Edwin Carlos Montiel-Vázquez [1,*], Jorge Adolfo Ramírez Uresti [1] and Octavio Loyola-González [2]

[1] School of Engineering and Science, Tecnologico de Monterrey, Atizapán 52926, Estado de Mexico, Mexico
[2] Stratesys, Calle de Torrelaguna 77, 28043 Madrid, Spain
* Correspondence: edwincmv@exatec.tec.mx

Abstract: Empathy is a necessary component of human communication. However, it has been largely ignored in favor of other concepts such as emotion and feeling in Affective computing. Research that has been carried out regarding empathy in computer science lacks a method of measuring empathy based on psychological research. Likewise, it does not present an avenue for expanding knowledge regarding this concept. We provide a comprehensive study on the nature of empathy and a method for detecting it in textual communication. We measured empathy present in conversations from a database through volunteers and psychological research. Subsequently, we made use of a pattern-based classification algorithm to predict the Empathy levels in each conversation. Our research contributions are: the Empathy score, a metric for measuring empathy in texts; Empathetic Conversations, a database containing conversations with their respective Empathy score; and our results. We show that an explicative pattern-based approach (PBC4cip) is, to date, the best approach for detecting empathy in texts. This is by measuring performance in both nominal and ordinal metrics. We found a statistically significant difference in performance for our approach and other algorithms with lower performance. In addition, we show the advantages of interpretability by our model in contrast to other approaches. This is one of the first approaches to measuring empathy in texts, and we expect it to be useful for future research.

Keywords: empathy; natural language processing; pattern-based classification; affective computing; databases

Citation: Montiel-Vázquez, E.C.; Ramírez Uresti, J.A.; Loyola-González, O. An Explainable Artificial Intelligence Approach for Detecting Empathy in Textual Communication. *Appl. Sci.* **2022**, *12*, 9407. https://doi.org/10.3390/app12199407

Academic Editor: Dimitris Mourtzis

Received: 5 August 2022
Accepted: 10 September 2022
Published: 20 September 2022

1. Introduction

Empathy is a fundamental part of human communication. However, it has been difficult for researchers in the area of psychology to truly define its nature [1–4]. Empathy can colloquially be known as a person's ability to understand and share the feelings or situation of another [5]. It provides understanding and care for a speaker and helps to develop a relationship between participants in a conversation. Additionally, it elevates the conversation from a strictly informational exchange to a more human connection. For these reasons [5], even though it is a disputed concept [4,6], it is impossible to negate that it is essential for a human-like conversation.

Empathy is often overlooked when dealing with Natural Language Processing (NLP) applications. However, it should be a priority within the developing field of Affective Computing. Affective computing refers to the area of computer science that is focused on providing human-like aspects to computer applications, specifically related to emotion [7–9]. Applications related to it range from sentiment analysis [7,10] to generation of human-like dialogue [4,7,11]. Affective computing is a relatively recent area and has been shown to have many benefits, while has also been subject to many criticisms [12,13]. Nevertheless, it is an area with potential towards innovation. In order to generate more knowledge regarding empathy for Affective computing, further research must be carried out.

There is not a consensus or standard for the use of empathy in computer applications [2–4,6]. While there has been some research on empathy in text and speech, measurement and classification of empathy using machine learning is a task that is not well developed [2–4,6]. For example, there are not many studies that go beyond binary detection. There is neither a standard measurement of empathy that can be useful for experts in the field of psychology or computing. As such, we believe it is a valuable area of research that is still open to many possible improvements.

This research aims to make use of machine learning classification in order to obtain more knowledge related to empathy and facilitate its use towards computer applications. However, as approaches are different, we focused on using classification algorithms that can provide useful information as patterns instead of only predicting values [14]. These classification algorithms are called pattern classifiers and apply pattern mining techniques as well as statistical classification to provide a list of patterns for a target class and a prediction of a data point [15]. These classifiers can be used to obtain accurate and more interpretative models for structured domains [15,16], and have been used in areas such as medicine [17]. We propose and test through experimentation the hypothesis that a pattern-based classification approach will present better performance when detecting empathy, alongside the advantages of explanatory information.

The main contribution of this research is to provide an explanatory artificial intelligence model that detects empathy in textual communication based on psychological research. We propose the measurement of empathy in a conversation through the lens of machine learning classification. Each class will correspond to the amount of empathy found in a text-based conversation, which must necessarily be based on the perception of highly empathetic people, as described in current psychological research [18].

Our results include the development of the Empathy Score, a metric for measuring empathy in texts. With this, we present a database containing conversations with their respective empathy value. In addition, we show the advantages of the pattern-based algorithm PBC4cip as an approach for detecting empathy in textual communication. We state its performance against other approaches for classification, and we present the explainable aspect of the model through the interpretation of patterns obtained.

The document is structured as follows: Section 2 provides preliminaries on the concept of empathy and contrast pattern-based classification. Section 3 shows a summary of similar works related to classification in affective computing. Section 4 presents our approach for detecting empathy in textual communication. It also presents the process of creating our database. Section 5 presents our experimental setup. Section 6 presents the results and analysis carried out. Finally, Section 7 presents our conclusions and future work.

2. Preliminaries

2.1. Empathy

Empathy is difficult to explain even for experts in the field of psychology [2,6,18,19]. While it is certain that it forms a fundamental part of humanity, the concept has eluded a formal description that satisfies all researchers in psychology [18,19]. Nevertheless, it is possible to describe empathy according to various models accepted by the field. On this research, we focused on the definition created by Simon Baron-Cohen and Sally Wheelwright. Their definition is based on the historical approaches taken by social psychologists [18–20] and examines empathy as having two main, not exclusive, components: Cognitive and Affective. A graphical representation of the model can be found in Figure 1.

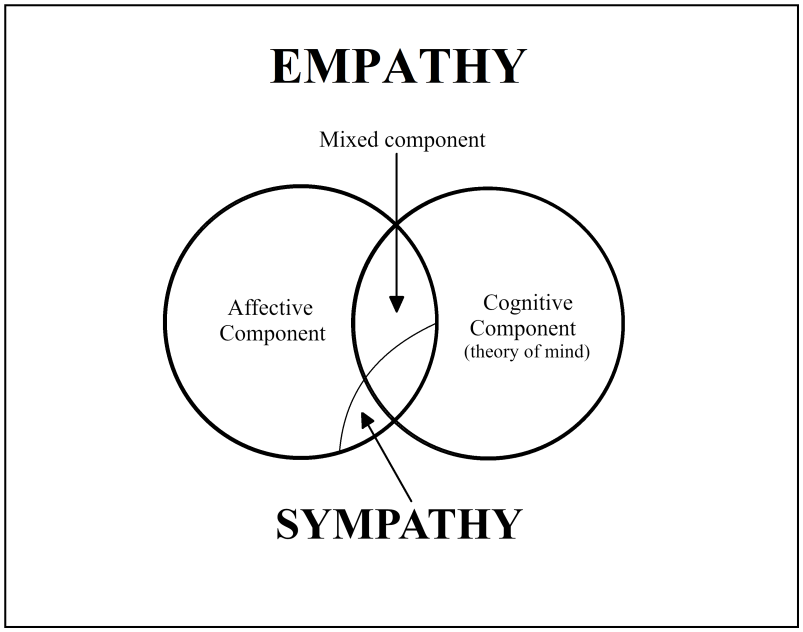

Figure 1. Simple model of empathy presented by Baron-Cohen and Wheelwright [18].

The Affective approach describes empathy as an observer's emotional response to the affective state of another [4,18,19], that is, the emotion an individual feels in response to the emotional state communicated by the person with whom they are interacting. This approach describes that an emotional response must be appropriate to be considered affective empathy [20]. Appropriate responses can be classified as parallel (the response matches that of the target) and reactive (beyond matching affective states, such as compassion) [20].

The Cognitive approach emphasizes that empathy involves understanding another person's feelings. Essentially, it consists of setting aside one's current perspective, attributing a mental state (or "attitude") to another person and then using inference to predict the likely content of their mental state, based on the experience of that person [18,19,21].

In addition to its two main components, Baron-Cohen and Wheelwright also describe the concept of sympathy. Based on the descriptions by Adam Smith [18,22] and Mark H. Davis [18,23], sympathy is described as a special subset of empathy that requires an affective component. It occurs when the observer's emotional response to the distress of another leads the observer to feel a desire to take action to alleviate the other person's suffering.

An empathetic exchange can therefore be a conversation in which an utterance is followed by a response that presents both empathy components. The more appropriate and more aspects of empathy present in the behavior of the person responding, the higher the empathetic levels in the conversation.

2.2. Pattern-Based Classification

Pattern-based classification is an approach of classification that makes use of "patterns" for deciding the label of each instance. It is related to rule-based classification, as it relies on the extraction of characteristics from a database to classify unseen instances [15,16,24]. In pattern-based classification, instead of extracting rules, a series of features that apply to various instances in a class are obtained. These feature combinations are what we call patterns, and they describe a collection of objects [25]. Patterns can be used as discriminators in classification tasks [24], as they can describe a large number of instances in one class and not in another. How often a pattern is present in a class is described as its *frequency*. After they are extracted, they are usually used as new features, which can then be input into the classification model [16].

The key challenges for pattern-based models classification models are [16]:

1. Pattern finding: Mining patterns from a given dataset can be tackled through various algorithms. However, different approaches to ensure the quality of these patterns must also be considered. For example, a large number of patterns can be post-processed in order to reduce its size, or the number of patterns can be reduced through an iterative algorithm.
2. Pattern application: Independent patterns can be mined using an algorithm, which can then be used as features for any classifier. However, it is also possible to mine patterns that explicitly take into account the type of model which will use them for classification.

The value found in this classification approach is that it can contribute other characteristics besides just the class predictions [16,26]. Pattern-based classification can also contribute information regarding the instances that can be easily interpreted by experts in the topic [26]. The frequency of patterns can be identified and analyzed by users in order to obtain additional information that can contribute to the understanding of the problem. This has been useful in previous research [27]. Additionally, this advantage can contribute insight in class imbalance problems, specifically [28].

Due to the advantages of pattern-based classification, we consider this approach as especially useful for our research. Because of the challenge in defining empathy, we believe finding patterns can help us obtain insight into the nature of empathetic textual communication. Additionally, this approach has been previously used in NLP tasks and has provided findings regarding language structure [29].

2.3. PBC4cip

The pattern-based classifier that we will use on this research is PBC4cip: A contrast pattern-based classification algorithm built for class imbalance problems [27]. These classifiers are notorious since they are easy to understand by experts in the application domain [28]. A contrast pattern is a pattern that appears significantly more in a class when compared to the other classes [28], which grants them the advantages of effectiveness, comprehensibility, and ability to distinguish [28].

Previous research on contrast pattern-based classifiers [26,27] has shown them to perform consistently better when compared to other classifiers like Naive Bayes, KNN, and support vector machines. Additionally, this approach has been previously used to solve real-life problems [27,28], such as those found in the medical field. However, most pattern-based classifiers are not suitable for handling class imbalance issues.

There are very few pattern-based classifiers that are built to take into account the class imbalance problem, among them are PBCEQ, PBC4cip, and iCAEP [28]. These algorithms take the classifier level adjustment approach to solve the shortcomings presented by class imbalance, and they seem to perform well in various situations. PBC4cip is based on the frequencies of patterns found by any mining algorithm [26]. Usually, when applied to class imbalance problems, contrast pattern miners extract several patterns with high support for the majority class and only a small number of patterns, with low support, for the minority class [26,28]. This causes bias towards the majority class. The main idea behind PBC4cip is that contrast patterns with low support for the minority class do not become overwhelmed by those patterns with high support for the majority class during the classification stage [26].

PBC4cip proposes that the frequency sum of patterns that cover an instance be weighted according to the following expression:

$$w_c = (1 - \frac{|c|}{|T|}) / \sum_{p \in P} \text{support}(p, c) \quad (1)$$

In Equation (1), $|c|$ represents the amount of instances that belong to class c, T is the number of objects in the training dataset, P refers to the set of all patterns for the class c, and $\text{support}(p, c)$ represents the frequency of the pattern p in class c [26]. The term $(1 - \frac{|c|}{|T|})$

will prioritize the class with a smaller number of representatives by punishing the sum of frequencies computed for the majority class. Additionally, the term $\sum_{p \in P} \text{support}(p, c)$ is used for normalizing the sum of frequencies in each class regarding all patterns of the same class. This weight has as a purpose overcoming the bias of the classifiers towards a majority category. It assigns a higher weight for the minority class [26].

In the training phase, PBC4cip obtains the patterns for each class and computes the weight w_c. This is followed by the classification phase, in which the algorithm computes the sum of frequencies in each class for all the patterns matching with the instance w_c [26]. The sum is later multiplied by the weight:

$$w_{Sum\_Supp}(o, c) = w_c \sum_{\substack{p \in P \\ p \text{ covers } o}} \text{support}(p, c) \quad (2)$$

In Equation (2), w_c refers to the weight of the class c, previously calculated in the training phase. $\text{support}(p, c)$ is the frequency in class c of the pattern p covering to the instance o. Equation (2) refers to the weight that a class has in a specific instance, different from w_c. Therefore, we can use this value to select the class of the object of interest. For classification, PBC4cip will select the class that presents the highest value for $w_{Sum\_Supp}(o, c)$ for each instance.

We selected this algorithm as our focus on this research regarding pattern-based classification. This was decided thanks to these factors:

1. The data we would use was imbalanced, a problem that PBC4cip was designed to address [26].
2. PBC4cip has been shown to outperform at least eight state-of-the-art algorithms designed for class imbalance problems [26,30].
3. PBC4cip, being a contrast-pattern based algorithm, provides information regarding the nature of the classification through the patterns obtained [14,26]. This will allow for the creation of an explainable model for classification of empathy.

3. Related Work

In this section, we present some works related to empathy in the area of computing and machine learning.

Alam et al. [2] carried out a study that aimed to detect empathy present in spoken conversations in call centers. To do this, they present an operational definition of empathy based on the modal model of emotions. A sample of 905 spoken conversations were obtained from a corpus of 1894 randomly selected customer agent interactions. Subsequently, interactions that were annotated with at least one empathetic interaction between the customer and the agent were considered to have empathy. Empathy presence was used as a binary class in their research. Alam et al. [2] considered Acoustic, Lexical, and Psycholinguistic features for their model. Binary classification models were designed using Support Vector Machines (SVM) [2]. The experiments were carried out used a combination of the available features, with the Un-weighted Average (UA) as the performance metric [2]. The UA is the average recall of positive and negative classes [2]. Using the Leave-One-Speaker-Group-Out (LOSGO) cross-validation method, Alam et al. [2] obtained a maximum of 63.9 average UA using feature selection methods.

Kumano et al. [31] carried out research aimed at obtaining an estimation of emotional interactions in a meeting with four participants. The types of emotional interaction targeted were empathy, antipathy, and unconcern. The features used were gaze, facial expressions, and speech-silence features. The estimation was based on the Bayesian approach, and obtained promising results.

Leite et al. [32] present a study on the behavior of a social robot capable of detecting the affective state of humans and acting in a limited empathetic manner. In this study, an iCat robot observes a chess match between two players, and behaves in an empathic manner by commenting the game and disclosing its affective state. Results showed that the

participants recognized the empathetic behavior of the robot and were willing to interact with it in the future.

The Doctoral thesis of Alam [4] presents the most comprehensive study on the Cognitive aspect of empathy on AI. This research explores the various elements that comprise behavior consistent with cognitive empathy in AI Healthcare systems. The AI Cognitive Empathy Scale (AICES) was developed to test the elements of empathy in interactions between AI focused on healthcare and patients. Furthermore, a conceptual model of cognitive empathy for patient-AI interactions was created.

Liu-Thompkins et al. [33] developed a framework for integrating artificial empathy in AI marketing agents. They explore various components of empathy in these agents. Furthermore, they present the advantages and disadvantages of using artificial empathy in marketing. Mainly, they show the value of bridging the AI-human gap in affective and social customer experience. However, they also present situations in which artificial empathy can be unnecessary, or harmful.

4. Empathy Measurement

Our methodology for measuring empathy in text is carried considering three different stages: the creation of a database with an accurate measurement of empathy; the representation of text using different features considering both components of empathy; and the methods we ensured to obtain patterns that could be interpreted in future research.

4.1. Empathetic Conversations (EC)

In order to detect empathy in texts, it was necessary for us to procure a database that contained empathetic exchanges. The database that was procured after the analysis of the various options [34–36] was EmpatheticDialogues (ED). This database was proposed by Rashkin et al. [37,38] with the express purpose of working as a repository for empathetic conversations. The database contains 24,850 multi-turn conversations grounded in situations prompted by specific emotion labels. The conversations were created by a group of participants by this given principle. The resulting exchanges in the database have been considered empathetic by human observers [38], and it has become the basis for a large number of research projects related to the generation of empathetic conversations using deep-learning [36,39–42]. In the presence of this fact, we consider this database very useful for the task of detecting empathy, as it has a history of being useful for related research.

An important aspect of the database is the human metric for empathy/sympathy. Rashkin et al. [38] describe the measurement of empathy through a Likert scale presented to the participants. The scale contains five points, and it goes by the following order: 1: not at all, 3: somewhat, 5: very much. The description of the metric is: "Whether the responses show understanding of the feelings of the person talking about their experience" [38].

The metric was sound and could be used as a base for this research. Since it was based on the Likert scale [43], it was substantiated in previous research of Affective variables [44]. Furthermore, sympathy as a special subset of empathy [18] is accounted for in the metric. However, there was an element of this metric that presented uncertainty: Since the database was created by the use of crowd-sourcing through Amazon Mechanical Turk (MTurk), we were not sure of the validity of each participant's empathetic judgment.

People present different capabilities when it comes to carrying out and identifying empathetic behavior [18,19,45,46]. For example, a lack of empathy is a feature in certain psychological conditions [20,45,46]. A person with high empathy will be proficient in all of its aspects [18,46]. Therefore, they can provide a more reliable measurement of it. We cannot rule out that some of the measurements are inaccurate because of the empathetic abilities of the participants, or their state of mind during the evaluation. It is because of this that we were compelled to evaluate the empathy/sympathy scores present in the database. To do this, we would obtain the help of empathetic people through the use of the Empathy Quotient (EQ) [18,20,45–47]. Our group of participants would measure the empathy in a sample of conversations taken from ED. If the scores reported by the Empathy/Sympathy

metric differed from the actual empathy found in the conversations according to our group, we would be forced to discard the scores from ED.

The EQ is a method for measuring empathy in adults developed by Baron-Cohen and Wheelwright [18]. It consists of a self-report scale designed to measure empathy on a scale from zero to eighty [18]. Baron-Cohen and Wheelwright designed the EQ to be sensitive to a lack of empathy as a feature in psychopathology [20]. Based on this principle, validation studies on the EQ's reliability have shown that it is a useful method for measuring empathy in a single dimension [20,45–47]. Baron-Cohen and Wheelwright postulate that there is a useful cut-off point of 30 that separates those that have trouble with empathy and those that were in the control group [18]. Based on this, we will consider as empathetic people those that are above this threshold.

We carried out a process in order to evaluate the validity of the Empathy/Sympathy metric. It was necessary to obtain a panel of volunteers that scored favorably on the EQ in order to ensure that they were highly empathetic. These volunteers would participate in the evaluation of conversations through the Delphi method, using an empathy metric compatible with the one found in the database. Finally, this was to be analyzed using statistical tools. The structure and description of the steps performed to validate the database can be appreciated in Figure 2.

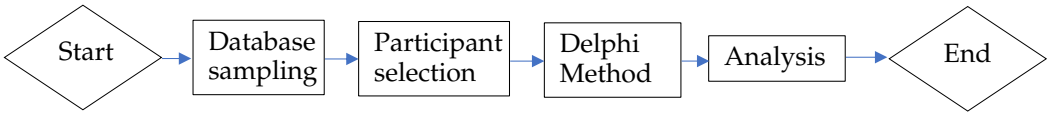

Figure 2. Database evaluation process.

First, it was necessary to obtain a sample of the database. We used the stratified random sampling technique [48–50]. This was because the database presented a large representation of emotion labels as one of its defining features [37]. We considered a confidence level of 95%, a degree of variability of 0.5, and a level of precision of ±5% [51–53]. For our population, we have the conversations in ED. We are studying conversations that present an inaccurate label (e.g., conversations with any number instead of 5, when the actual empathy is 5). Thus, we define our population with an attribute of interest as "conversations with any error in measurement". We do not know the actual number of such conversations. Therefore, we decided to take a conservative approach. This led to a proportion of 0.5. With these parameters, we obtained a sample size of 384. We rounded this number to 400 for simplicity. With this process, we obtained a sample of ED with balanced emotional context.

We required the selection of participants that presented a high empathetic ability, to compare their evaluation scores to the Empathy/Sympathy score in the database sample. We searched for volunteers in various online communities and social media, as well as former students of Instituto Tecnológico y de Estudios Superiores de Monterrey. When the volunteers were contacted, their written consent was requested. All ethics standards were followed in accordance with our institution.

The volunteers were asked to take the EQ. We considered those that presented a score of 35 or higher for our research. We were able to procure 10 volunteers that matched our criteria. These volunteers agreed to evaluate the conversations in the database sample. However, they were only willing to evaluate 200 conversations each. Therefore, we decided to separate the participants into two anonymous groups of 5 considering their EQ score, their native language, and gender expression. The groups presented an average EQ of 55.2 and 57.2, respectively. The EQ of each participant can be seen in Table 1. The average EQ of both groups are higher than the mean for the control group used to validate the EQ [18]. As such, we believe both groups are capable of identifying and acting according to empathy.

Table 1. EQ results for Groups 1 and 2.

Group 1	Group 2
35	40
54	55
58	56
60	57
65	70

The groups would independently evaluate each 200 conversations of the 400 conversations present in our sample. This would be carried out using the Likert scale present in the Empathy/Sympathy score. However, the ED database did not present a complete Likert scale, as a correspondence to each of the points present in it was not explicitly stated. Therefore, we needed to infer the variations in each of the levels of the scale as equidistant points, in accordance with Rashkin et al [38]. The Likert scale which was presented to the participants of both groups will now be referred to as Empathy score (ES). The levels are described as follows:

1. Not empathetic at all;
2. A little empathetic;
3. Somewhat empathetic;
4. Empathetic;
5. Very much empathetic.

This meant that the amount of empathy was whether the responses show understanding of the feelings of the person talking about their experience, regardless of the role of listener or talker. The evaluation was done at the conversation level, not at the utterance level, and this was pointed out to the volunteers.

The Delphi technique is a reliable iterative method for eliciting and refining group judgments [54–57]. It has been used for validating the suitability of applied machine learning models by using the opinion of experts [14]. It is based on a series of procedural questionnaires applied to a group or panel of experts on a topic, in order to obtain their opinion. The implementation used in this research consisted of separating both groups of participants and assigning each 200 conversations. Evaluation of the conversations would be carried out through a series of rounds in which both feedback from the group as a whole and previous evaluations by each volunteer would be presented. In addition, information regarding the original ES would be provided incrementally in each step. This would allow the volunteers to repeatedly evaluate the conversations and increase their confidence in their judgment through the iterations. Since the process was anonymous, no individual knew who were in the same team as them, therefore avoiding bias. The overview of our approach to the Delphi process can be found in Figure 3.

The first round of the Delphi method implementation would present the volunteers with the raw information from the database. Each volunteer would carry out their evaluations. For the second iteration, the conversations would be once again presented to the participants. However, the feedback from the previous round would also be present. For each group, the five evaluations for the conversations would be given, with a marker indicating the participant's previous evaluation. In addition, the maximum, the minimum, the standard deviation, the mean, and the quartiles of the average ES in the original database were presented. The final iteration of the Delphi method would consist of another evaluation of the conversations. The procedure would be very similar to the second round: We would provide the results from the previous evaluations from the five group participants, highlighting the previous evaluation of each participant. However, we also provided the two scores present in the original database. The addition of this information at the final round allowed for the volunteers to contrast their evaluation with the original database.

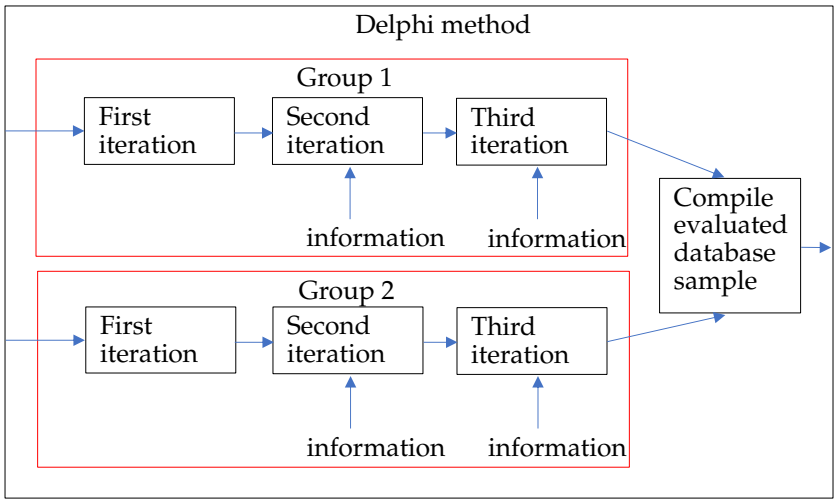

Figure 3. Implementation of Delphi method.

After the final round of evaluation, the final Empathetic Score given by the volunteers (ESV) would be computed. To do this, we would obtain the rounded average of the five evaluations carried out by the volunteers in each group. This value for group 1 and group 2 would be joined into a single database containing the 400 evaluated conversations. The differences between the ESV and the Empathetic Score given by the original database (ESO) can be appreciated in Figure 4. We decided on the application of the Wilcoxon matched pairs test for the statistical analysis [58–61].

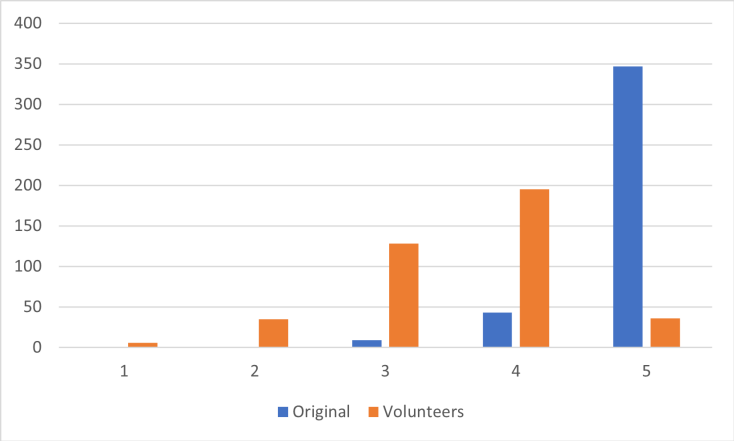

Figure 4. Distribution of the ESO and the ESV side by side.

We applied the test to the difference to between the ESV and the ESO. For the confidence level, we decided on 95%. The results can be found in Table 2. As it is clearly presented, the *p*-value obtained was 0. This meant that the null hypothesis was rejected.

Table 2. Wilcoxon test for the ESV and ESO.

Sample	N for Test	Wilcoxon Statistic	*p*-Value
Difference	346	59,181.00	0.000

The results from this evaluation allowed us to safely state that the ES present in the ED database were not reliable for measuring empathy. This meant that any research that assumed the perception of empathy found in the database, or a large number of highly empathetic conversations, and used it for any purpose might be flawed. This was an important contribution to our research since this fact can now be brought into attention

for future and past scientific studies. However, this finding also severely reduced our data for the classification of empathy. The database sample of 400 conversations presented an evaluation of empathy that was in line with the perception of empathetic people. Therefore, it was the only data that could be used for the detection of empathy.

We discard ED as a database with a reliable metric for empathy. The final database obtained will be referred to as Empathetic Conversations (EC), a sample of ED. It is a repository of 400 multi-turn conversations between a "Talker" and a "Listener". Each conversation presents an emotional context and a "prompt" that guides the start of the conversations. Each "prompt" is rooted in the emotional context.

The most important attribute of the database was the average empathy score by the volunteers. This metric represents the empathy level found in the conversation according to the evaluation of five empathetic people. An example of a conversation can be appreciated in Table 3. The empathy score will be used to classify empathy in the utterances. This measure of empathy is, to our knowledge, the only metric validated by psychological tools and, therefore, it is the most objective possible metric for measuring empathy in texts. A full breakdown of the database can be found in Appendix B.

Table 3. Example of a conversation in EC.

Conv_Id	Utterance _Idx	Context	Prompt	Speaker _Idx	Utterance	Empathy _Score
hit:11054 _conv:22108	1	surprised	a job I applied and interview for a couple of months ago offered me the job. It was unexpected.	675	a job I applied and interview for a couple of months ago offered me the job. It was unexpected.	5
hit:11054 _conv:22108	2	surprised	a job I applied and interview for a couple of months ago offered me the job. It was unexpected.	746	Congrats! How exciting for you.	5
hit:11054 _conv:22108	3	surprised	a job I applied and interview for a couple of months ago offered me the job. It was unexpected.	675	thank you, when they called, I did not know who it was at first, then it was, wow I got the job	5
hit:11054 _conv:22108	2	surprised	a job I applied and interview for a couple of months ago offered me the job. It was unexpected.	746	That's a long time to hold out hope. I bet you were shocked.	5

4.2. Text Representation

The objective of the research was to obtain the classification of empathy for each utterance in the conversations. To do this, it was necessary to obtain a series of relevant features for the classification algorithms. We determined representation models such as bag-of-words or word embeddings [25] would not be suitable for our research. This was because any purely mathematical text representation would not be able to provide explicative patterns [25] using our approach (PBC4cip). Therefore, we focused on the creation of our own text representation, which contained features that could provide insight on empathetic communication.

The EC database presented various features that could be used for the classification of empathy. However, it was quite clear that more information could be extracted from the text. In order to obtain useful features, we made use of the Paralleldots text mining API. The use for this API was decided on because it was freely available and permitted the extraction of several different characteristics that could be used for classification [62]. The API provides various methods for extracting information. For this research, we focused on the mining of four text characteristics with the API: Sentiment [63], Emotions [64–67], Taxonomy [68], and Intent [69,70]. This was, by no means, an exhaustive list of categories that could be obtained from the text, but it was the ones that we decided due to their similarity towards the two components of empathy: The affective component is considered with the addition of sentiment and emotions, while the cognitive component is taken into

account by contextual information through the taxonomy and intent. A breakdown of our text representation features is as follows:

- EC features: Features found in EC that pertain to the conversation. The only features absent are the utterance and prompt text and their conversation identifier, as these cannot be used for machine learning classification [25]. Another feature was added in this category, which consisted of the utterance length. Additionally, the speaker identifier was modified into another feature called "Talker", with the value 1 representing the role of "Talker" and 2 representing the role of "Listener". This was done to ensure useful information could be obtained about the conversation, instead of the participants. Further information on all available features can be found in Appendix B.
- Sentiment features: Three features representing the probability that the text is positive, negative, or neutral [63].
- Emotion features: Six features representing the probability that the text presents one of six emotions according to Ekman's model [65–67].
- Intent features: Eight features representing the probability that the text presents one of eight possible categories for intent.
- Taxonomy features: Sixteen characteristics, each of them represents the confidence score that the utterance presents a topic label from the IAB (Interactive Advertising Bureau) Tech Lab Content Taxonomy [68]. These characteristics were selected due to their high representation in the data.

In total, each utterance would present 38 unique features related to either text, affective, or context-related characteristics. A visual example of this representation can be seen in Figure 5.

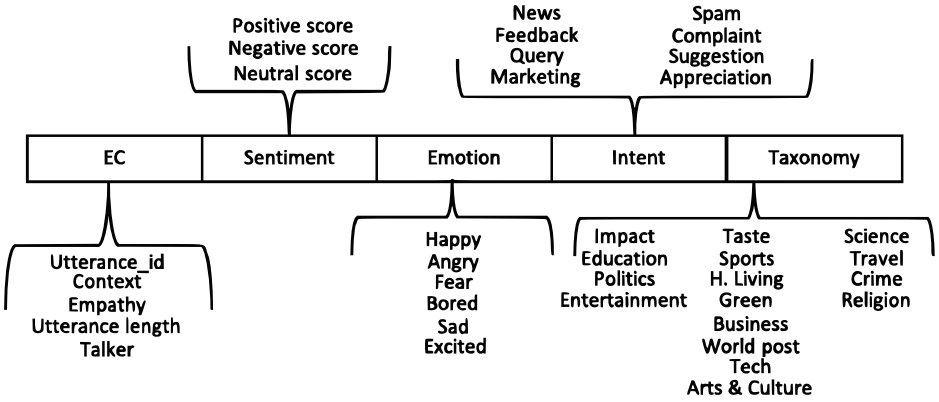

Figure 5. Representation of a text utterance.

4.3. Pattern Extraction

As it has been previously stated, we decided to use PBC4cip in our research. This was done in order to obtain contrast patterns that could later be interpreted in psychological or machine learning research. Therefore, this would allow our approach to be explainable [71]. We would make use of this algorithm to obtain the measurement of empathy from conversations (classification) and later analyze the patterns extracted through the pattern mining algorithm that will be used alongside PBC4cip. This algorithm would be Random Forest Miner (RFMiner) [71], since previous research has shown that it is capable of reliably providing contrast patterns [71,72].

In order to obtain more diversity when it comes to patterns, we also explored other features that could be changed from EC. One of these was the use of different emotion models. Meanwhile, another approach was the reduction of relevant features to improve readability of patterns.

EC contains a variable regarding the emotional context of the conversation, this is one of the most important aspects of the ED database [37]. Therefore, we considered that changing the expression of this context using different emotion models could yield

interesting explicative patterns. Emotion models vary, as emotion is an abstract psychological concept [73]. However, they can be categorized into two branches: Categorical and Dimensional [64].

Categorical models of emotion divide emotions into groups of discrete categories. Meanwhile, dimensional emotion models present emotions as a point or a region within a two-dimensional or multi-dimensional space [64,65,74]. The emotional context of the conversation in EC was categorical in nature, since it was presented as various emotion labels. We believed that there was the possibility of obtaining better or more diverse patterns when using a dimensional model [64]. The model that we selected for this purpose was Plutchik's model of emotion [64,74–76]. This model, being dimensional, presents a more flexible emotion vector that can be used to capture more information regarding emotional states [64].

In addition, we hypothesized that using refining techniques could result in an improvement of the readability of the patterns found, without impacting the classification performance. In order to test this hypothesis, we made use of the classifier attribute evaluator (ClassifierAttributeEval) and a Ranker search method [77,78] in the Waikato Environment for Knowledge Analysis (WEKA) [79]. We would apply the attribute selection module in order to obtain a list of attributes that contributed to the classification more strongly. ClassifierAttributeEval evaluates the worth of each attribute by using a user-specified classifier and a performance metric. Meanwhile, the Ranking search method provided the ranking of each attribute of the database in accordance with the results of the evaluation. Using only those attributes that contributed greatly to the performance of the algorithm could provide shorter, or different, patterns. This could contribute to the explicative nature of our artificial intelligence model.

5. Methodology

In this section, we present the methodology used in this research regarding our model. The experiments would be carried out in phases. The first phase aimed to address the classification performance between a patter-based algorithm in contrast with other algorithms. This phase is presented in Section 5.1. The second phase of experimentation was focused on obtaining different types of patterns through the use of PBC4cip. This would be achieved by modifying the text representation, as previously presented. This phase is described in Section 5.2. A graphical representation of the experiments carried out is found in Figure 6.

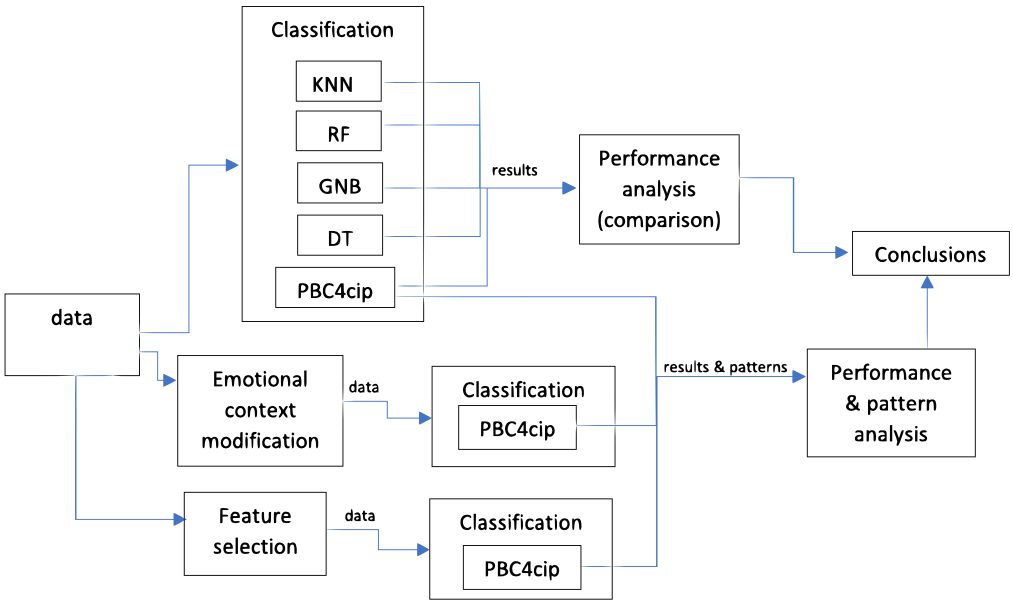

Figure 6. Data flow in the experiments carried out.

5.1. Algorithm Comparison

Evaluation of a classification algorithm refers to measuring the performance of a model when it encounters previously unseen data [25,80]. We required an adequate metric for measuring the performance of the algorithms involved [25]. We wanted to ensure that our contrast pattern approach was the most appropriate method for measuring empathy in textual communications, when compared to other algorithms. For this, we made use of two performance metrics: The Area Under the ROC Curve (AUC) [80,80,81] and the Closeness Evaluation Measure (CEM). The AUC was chosen since it was the most widely used nominal metric for cases with class imbalance [26]. Meanwhile, the CEM is a robust metric for measuring performance of algorithms in ordinal classification tasks. It rewards exact matches, considers ordinal relationships, and does not assume predefined intervals between classes [82]. The use of two metrics was due to the nature of the Empathy Score, which presented an order. We were interested in obtaining the performance of the classifiers when the problem is taken as a nominal [25] and as an ordinal [82,83] classification problem.

We decided to use 5-fold cross validation for our experiments [15]. The reason for the selection of 5 as our number of folds was that we wanted to ensure that a moderate amount of data found in EC could be considered for the testing set. With a larger number of folds, we ran the risk of shrinking the number of testing instances to the point that the found patterns might not be applicable to the set used for evaluation. Additionally, we applied the same 80/20 principle found in various applications of the Hold-out method [25]. To carry out the cross-validation splits, we decided on the use of the Knowledge Extraction based on an Evolutionary Learning (KEEL) tool [84]. Additionally, the data from one fold could later be used as a data point for statistical analysis.

The process for classification between the algorithms would be as follows:

1. Get one of the five pairs of dataset for training and testing.
2. Perform the classification on one of the algorithms.
3. Obtain the evaluation metrics:
    (a) For the PBC4cip record, the patterns obtained by the miner;
    (b) Record the AUC for the classifier;
    (c) Record the CEM for the classifier.
4. Return to step 2 until all algorithms have been evaluated.
5. Return to step 1 for another pair of datasets, until all five have been used for classification.

We would carry out the experiment using five different algorithms, alongside PBC4cip. The results would be analyzed using statistical methods. In this manner, we would obtain the performance using a contrast-pattern algorithm in contrast to algorithms commonly used in NLP applications [25]. The algorithms we would use are:

1. K-nearest neighbors algorithm (KNN) [85];
2. Random Forest Classifier (RF) [86,87];
3. Gaussian Naive Bayes (GNB) [85,88];
4. Decision Tree Classifier (DT) [15,85];
5. Multilayer perceptron classifier (MLP) [15,85].

5.2. Pattern Refining

Once the results for comparisons were obtained, we would focus on the performance of PBC4cip in regard to patterns. The patterns obtained from the previous subsection would be analyzed. In addition, the reduction of features and the changing of the emotion model regarding the context would be carried out. Subsequently, classification tasks would be performed using the same data. The patterns and performance would be recorded and compared to those obtained in the previous experiment. The AUC would be the main performance metric during this process, as it was not a priority to test both metrics during this phase of experimentation.

We would carry out the feature selection procedure through the five groups obtained for 5-fold cross-validation. We would select those features that presented a relatively

large contribution in at least one of the five evaluations. Subsequently, we would perform classification across the 5 folds using the reduced set of features. Once this was done, the performance and patterns obtained through the classification would be recorded. Finally, the performance would be compared to the full set of features through statistical methods. Meanwhile, the patterns with the most contrast would be compared to those obtained in the past experiment.

Finally, we would carry out the experimentation related to the change in emotion model. Like the feature selection phase, we would obtain the performance and patterns in a modified version of the database folds. It was necessary to perform a modification to each of the available datasets for this experiment. The processing of this data consisted of the mapping of each of the database emotional context values to their Plutchik model equivalent [89–94]. To do this, nine features were added to the database. The first eight features corresponded to the binary emotional vector from Plutchik's model. Meanwhile, the last represented the intensity of the emotion. Each of these values were filled according to our representation of Plutchik's model [74]. A full breakdown of the emotion equivalents can be seen in Table A1.

The results for each of the set pairs were the evaluation metrics and the patterns used by the classifier for this task. In total, we obtained the results for each of the five different dataset pairs, along with the average performance of the classifier when using this model of emotion representation. Additionally, a series of patterns were obtained through these classification processes.

6. Results and Discussion

In this section, we explore the results and the statistical analysis pertaining to the experiments carried out in this research. We decided to split this section into two areas, since they were related to the nature of the experiments. Section 6.1 describes the results regarding the classification tasks using various algorithms in contrast to PBC4cip. Meanwhile, Section 6.2 refers to the patterns extracted and the performance changes that were present after the modification of the text representation model.

6.1. Classification Results

Tables 4 and 5 show the results of all the cross-validation dataset pairs. We can see that there is a clear distinction between each classifier involved. Additionally, we can spot trends in these datasets that are held in multiple circumstances. For example, we can point out that the performances of PBC4cip, Random Forest, and MLP usually obtain the highest scores in both metrics. Meanwhile, the algorithm that tended to perform worse was the Gaussian Naive Bayes classifier. These trends would be relevant on the statistical analysis. The performance of the metrics across the cross-validation process can be seen in Figure 7.

Table 4. Cross-validation AUC results.

Fold	Random Forest	KNN	Naive Bayes	DT	MLP	PBC4cip
1	0.6199	0.5614	0.5251	0.5665	0.5930	0.6530
2	0.6173	0.5603	0.5278	0.5372	0.6000	0.5850
3	0.6170	0.5104	0.5431	0.5219	0.6180	0.6380
4	0.5831	0.5242	0.5438	0.5341	0.5853	0.6190
5	0.6170	0.5720	0.5551	0.5069	0.6014	0.6290
avg	**0.6108**	**0.5456**	**0.5389**	**0.5333**	**0.5995**	**0.6248**

Table 5. Cross-validation CEM results.

Fold	Random Forest	KNN	Naive Bayes	DT	MLP	PBC4cip
1	0.5681	0.5609	0.4155	0.5670	0.5657	0.5939
2	0.5712	0.5452	0.3999	0.5333	0.5574	0.5291
3	0.5675	0.5155	0.4505	0.5211	0.5663	0.5695
4	0.5648	0.5430	0.3867	0.5413	0.5465	0.5855
5	0.5718	0.5476	0.4415	0.5022	0.5684	0.5615
avg	**0.5687**	**0.5425**	**0.4188**	**0.5330**	**0.5609**	**0.5679**

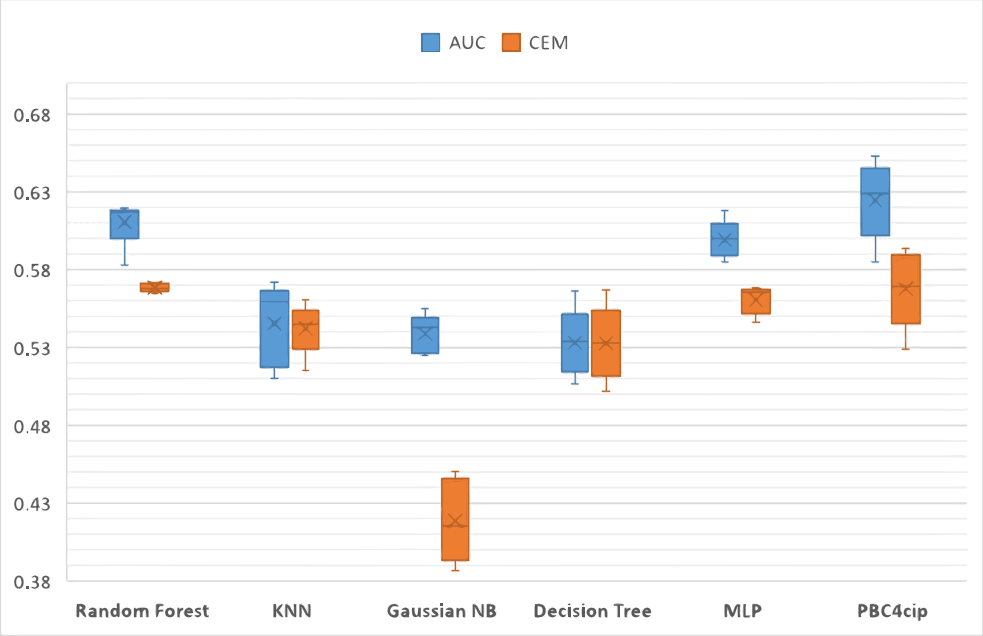

Figure 7. Metrics across the first phase of experimentation.

We can infer that there is a difference between Random Forest, MLP, and PBC4cip when compared to the rest of the classifiers: The former group performs better than the latter. It is important to note that this behavior can be seen using both of our metrics, with a difference in magnitude. This was interesting, as both metrics support the three classifiers being clearly better than the others. This would have to be analyzed using statistical tools in order to see if this difference was significant.

To test if these algorithms presented a statistically significant difference in comparison to PBC4cip, we carried out statistical analysis using the Friedman test [95–98]. This was selected since it maintains the same statistical power than the sign test for normal and non-normal distributions [95] and has been used previously for the comparison of machine learning algorithms [14,96,98,99].

For the AUC metric, the average ranks obtained by the Friedman test can be observed in Table 6. The Friedman statistic was 20.0857, with a p-value of 0.0012. Therefore, we rejected the null hypothesis of the test, which states that there was no significant difference in between the classifiers [99].

Once this was done, it was necessary to perform a post-hoc test to find the comparisons that presented a difference [98]. Since PBC4cip presented the highest rank, we could make a 1xN post-hoc test [98]. The Finner post-hoc test is recommended due to its power and comprehensibility [14,99]. Therefore, we decided to use it. The results from the Finner post-hoc test are presented in Table 7. We can see that the adjusted p-values suggest that there was a significant difference between PBC4cip and KNN, DTC, and GNB. Meanwhile, there was no difference between the pattern-based classifier and RF or MLP. This supported the previous tests.

Table 6. Average Rankings of the algorithms (AUC).

Algorithm	Ranking
PBC4cip	1.4
RF	2.2
KNN	5
GNB	5
DTC	5
MLP	2.4

Table 7. Unadjusted and Adjusted *p*-value for the post-hoc analysis (AUC).

i	Algorithm	Unadjusted *p*	p_{Finner}
1	KNN	0.002346	0.011674
2	GNB	0.002346	0.011674
3	DTC	0.002346	0.011674
4	MLP	0.398025	0.469759%
midrule 5	RF	0.498962	0.498962

With regard to the CEM metric, the Friedman statistic obtained was 18.6, with a *p*-value of 0.002281. This meant that we also rejected the null-hypothesis of the test. The average ranks obtained can be seen in Table 8. In this case, we see that RF is the algorithm with the highest rank.

Table 8. Average Rankings of the algorithms (CEM).

Algorithm	Ranking
PBC4cip	2.2
RF	1.6
KNN	4.2
GNB	6
DT	4.2
MLP	2.8

Since PBC4cip was not the highest rank, we decided to carry out an NxN post-hoc test. We made use of the Bergmann–Hommel procedure. In an NxN Friedman test, the hypotheses are logically interrelated; thus, not all combinations of true and false hypotheses are possible [100]. The Bergmann–Hommel procedure is based on the idea of finding all elementary hypotheses which cannot be rejected [99]. The hypothesis rejected do not belong to such group [101]. The adjusted *p*-values resulting from the test can be seen in Table 9.

The procedure rejects the hypotheses of equality between RF vs. GNB, PBC4cip vs. GNB, and MLP vs. GNB, with a significance level of 0.05. This supported the behavior seen in the test using the AUC metric. While not as direct, we see that PBC4cip, RF, and MLP are at least not equivalent to the worst performing algorithm. They present the three best rankings, and present no difference among themselves. With these factors in mind, we can say that these algorithms are the ones that present the best performance. PBC4cip does not present a significant difference between the following algorithms: RF, MLP, KNN, and DT. However, it does present one between it and the worst performing algorithm: GNB.

Table 9. Unadjusted and Adjusted p-value for the post-hoc analysis (CEM).

i	Hypothesis	Unadjusted p	p_{Berg}
1	RF vs. GNB	0.0002	0.003004
2	PBC4cip vs. GNB	0.00132	0.0132
3	GNB vs. MLP	0.006841	0.047886%
midrule 4	RF vs. KNN	0.027992	0.279918
5	RF vs. DT	0.027992	0.279918%
midrule 6	PBC4cip vs. KNN	0.090969	0.545814
7	PBC4cip vs. DT	0.090969	0.545814
8	KNN vs. GNB	0.12819	0.769141
9	GNB vs. DT	0.12819	0.769141
10	KNN vs. MLP	0.236724	0.946894
11	DT vs. MLP	0.236724	0.946894
12	RF vs. MLP	0.310494	1.241978
13	PBC4cip vs. MLP	0.61209	1.241978
14	PBC4cip vs. RF	0.61209	1.241978
15	KNN vs. DT	1	1.241978

With the results from the statistical tests, we could conclude that the contrast pattern classification algorithm is equivalent to using a Random forest classifier, and a Multilayer Perceptron classifier when measured by the AUC. Since these algorithms performed significantly better than the rest used in this research, they stand as some of the best performing algorithms when measured by the CEM. In general, the algorithm ranks among the best ones to measure the level of empathy in a conversation. However, PBC4cip presents another advantage that makes it a superior approach for detecting empathy: the explanatory model. Thanks to the patterns obtained, reasons can be presented as to the nature of the classification. These can provide insight and further knowledge can be obtained regarding empathy. Therefore, we believe PBC4cip is the best approach for this task, despite the lack of a statistically significant difference between the algorithms with the best performance.

6.2. Extracted Patterns

For the second phase of experimentation, we first obtained features that presented a significant contribution when using PBC4cip. Each of the five groups for cross-validation was fed into the Feature selection module of WEKA. In order to minimize the possible negative impact, and increase the chances of a better performance, we decided on using the Accuracy as the metric on which the module would perform the evaluation. After the module processed the data, we were able to obtain a ranked list of attributes. The average maximum impact across all validation groups was found to be 34.04%. In contrast, the average minimum impact was 0.255%. Due to this difference, we decided to use a cut-off point of 10%. All features that presented an impact of under 10% would be dropped for this experiment. We expected to possibly improve the patterns found by the algorithm without producing any significant negative impact on its performance. Following the selection of attributes, we implemented the classification algorithm on the data.

Secondly, we carried out the processing of the dataset with regard to the emotional context. Each emotional label present in the feature was replaced by the corresponding features in accordance with Plutchik's emotion model. The model was in the form of a vector with eight binary values, accompanied with an integer value between 1 to 3. Each of the binary values would represent the presence of one of Plutchik's main emotions. Meanwhile, the integer value would represent the intensity of the emotions found in the binary vector, with 1 representing high intensity and 3 representing low intensity. The binary values would represent the emotions in the following order: "joy", "trust", "fear", "surprise", "sadness", "disgust", "anger", and "anticipation". Once we had processed the datasets in order to add the features of Plutchik's emotion model, we only needed to apply the classification algorithm.

Both modifications of the text representation yielded similar performance. The resulting values obtained from the classification tasks seemed to support the hypothesis

that, even though some features were modified, there was no significant negative impact in the classification of empathy. This was seen for both modifications. The results for the classification, along with the original text representation, using PBC4cip can be seen in Figure 8.

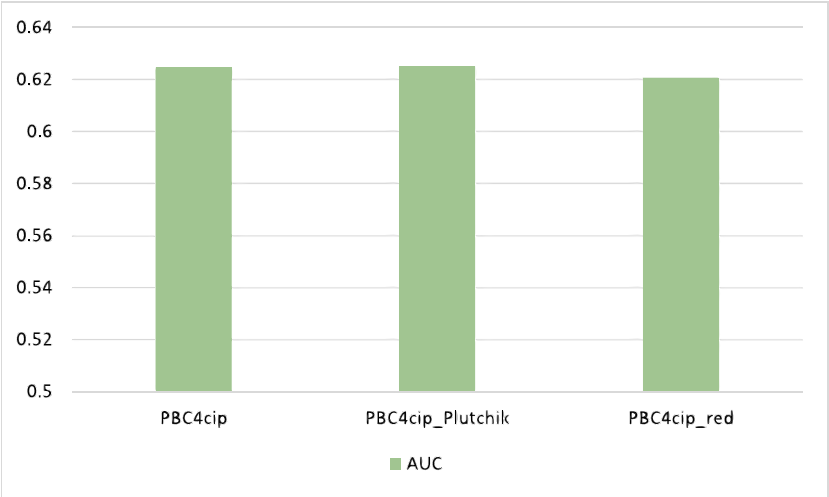

Figure 8. Performance of PBC4cip with full and modified features.

We performed the Paired *t*-test [102–104] for the cases using PBC4cip, in order to see if there was a significant difference when applying a different set of features. The results are found in Table 10. What we found by performing these tests is that the different features used in this research for finding useful patterns did not negatively impact the performance of the PBC4cip algorithm in any manner. This was desirable, since we did not expect the use of a different emotion model system or the reduction of features to create an adverse effect in the classification of empathy.

Table 10. *p*-values for paired *t*-test among algorithm of PBC4cip implementation.

PBC4cip Implementation	*p*-Value
Reduced features	0.230
Plutchik's model	0.947

In each of the implementations of PBC4cip, a series of patterns were obtained, which were later used for the classification of the utterances. We expected that the different features in each implementation would lead to differences between the patterns found. We decided to test this hypothesis by finding the patterns with the most contrast for each of the classes during the three approaches for the algorithm. We present the most common pattern for each of the classes across all folds. We will use set notation and refer to the emotional context labels with their respective emotion. For example, the pattern ut_len $\leq$ 50 AND emotional_context $\neq$ '5' AND ut_len $\leq$ 100 obtained through the classifier will be presented as: ut_len $\in$ [50,100], emotional_context $\notin$ {Apprehensive}. Additionally, the patterns are shown with the percentage of instances in the class that present this pattern. It is important to clarify that features in the categories of Sentiment, Emotion, Intent, and Taxonomy displayed probabilities that is to say that a pattern like Angry $\in$ [0,0.25] means that the utterance has to have a low probability of presenting anger (in between 0 to 25%).

In Table 11, we see the highest frequency patterns for each class, with some representing more than 25% of the total instances in their respective class. These patterns make up the explainable aspect of our approach. They are useful since they portray information that could be read and interpreted. We hypothesize that experts in the field of psychology interested in empathy could obtain insight from them. For example, the pattern for the first

class using the reduced model: ut_len ∈ [0,42.5], emotional_context ∈ {*Apprehensive*}, feedback ∈ [0.07,1], news ∈ [0,0.01]. It seems that utterances that are short have an emotional context of apprehension, do not talk about news, and are likely to be feedback will not be empathetic at all. This can be interpreted as short feedback responses by a Listener in a situational context where the Speaker has expressed apprehension. With this example, we can see that some patterns are indeed useful for explicative purposes.

Table 11. Most significant patterns found for each class by an Emotional context model.

Class	Emotional Labels (Original)	Emotional Labels (Reduced)	Plutchik's Model
1	Fear ∈ [0,0.27], ut_len ∈ [0,42.5], emotional_context ∈ {*Apprehensive*}, Sad ∈ [0,0.4], TECH ∈ [0,0.03] (26%)	ut_len ∈ [0,42.5], emotional_context ∈ {*Apprehensive*}, feedback ∈ [0.07,1], news ∈ [0,0.01] (32%)	feedback ∈ [0.19,0.43], PL intensity ∉ {*1*}, Positivity_score ∈ [0.64,1], Happy ∈ [0,0.23], Angry ∈ [0,0.34], Negative_score ∈ [0.04,1], marketing ∈ [0,0.18], PL trust_binarized ∈ {*0*} (26%)
2	Positivity_score ∈ [0,0.5], Neutral_score ∈ [0,0.7], EDUCATION ∈ [0,0.11], Bored ∈ [0,0.08], emotional_context ∉ {*furious*}, Happy ∈ [0,0.28], feedback ∈ [0.67,1], feedback ∈ [0,0.71], query ∈ [0,0.11], complaint ∈ [0.68,1], ut_len ∈ [0,96.0] (16%)	feedback ∈ [0.18,1], ut_len ∈ [0,42.5], emotional_context ∉ {*prepared*}, Sad ∈ [0.07,1], spam ∈ [0.2,1], Bored ∈ [0.1,0.25], Fear ∈ [0.14,1], Negative_score ∈ [0,0.69], IMPACT ∈ [0,0.37],Neutral_score ∈ [0,0.71], ENTERTAINMENT ∈ [0,0.63] (11%)	PL_trust_binarized ∈ {*0*}, ut_len ∈ [0,34.5], Fear ∈ [0,0.40], ARTS&CULTURE ∈ [0,0.06], Neutral_score ∈ [0,0.93], suggestion ∈ [0,0.63], Bored ∈ [0.09,1], Positivity_score ∈ [0.03,1], Happy ∈ [0.04,1], HEALTHYLIVING ∈ [0,0.01], ENTERTAINMENT ∈ [0.27,1], feedback ∈ [0,0.62], PL_anticipation_binarized ∈ {*0*} (10%)
3	ut_len ∈ [0,42.5], Negative_score ∈ [0.02,1], Excited ∈ [0.05,0.21], emotional_context ∉ {*proud*,*faithful*}, Positivity_score ∈ [0,0.03], utterance_idx ∉ {*1*}, IMPACT ∈ [0.01,1], Happy ∈ [0,0.09], Fear ∈ [0,0.40], Bored ∈ [0,0.25], Sad ∈ [0,0.45], Neutral_score ∈ [0.05,0.92] (5%)	Happy ∈ [0,0.35], appreciation ∈ [0,0.6], emotional_context ∉ {*apprehensive*, *devastated*, *faithful*, *excited*}, ut_len ∈ [0,29.5], Positivity_score ∈ [0,0.03], Excited ∈ [0.01,1], Fear ∈ [0,0.41], Angry ∈ [0,0.27], ENTERTAINMENT ∈ [0.09,1] (6%)	appreciation ∈ [0,0.57], ut_len ∈ [10,29.5], HEALTHYLIVING ∈ [0,0.05], POLITICS ∈ [0.26,0.76], Angry ∈ [0.07,1], Positivity_score ∈ [0,1], feedback ∈ [0,0.63] (5%)
4	Positivity_score ∈ [0.07,1], Happy ∈ [0.12,0.33], SPORTS ∈ [0,0.37], emotional_context ∉ {*lonely*,*devastated*}, Bored ∈ [0,0.1], POLITICS ∈ [0,0.51], Angry ∈ [0,0.28], ut_len ∈ [46.5,1], spam ∈ [0,0.55], ARTS&CULTURE ∈ [0,0.18], appreciation ∈ [0.2,1], BUSINESS ∈ [0,0], marketing ∈ [0,0.19], feedback ∈ [0,0.74], IMPACT ∈ [0,0.93], TASTE ∈ [0,0.05], Fear ∈ [0,0.27], complaint ∈ [0,0.21], query ∈ [0,0.2] (7%)	appreciation ∈ [0.57,1], Negative_score ∈ [0,0.29], Bored ∈ [0,0.09], query ∈ [0.01,0.08], ut_len ∈ [40.5,112.5], feedback ∈ [0.47,1], ENTERTAINMENT ∈ [0.17,1], emotional_context ∉ {*lonely*}, news ∉ {0}, suggestion ∉ {0} (7%)	ut_len ∈ [46.50,418], marketing ∈ [0,0.18], Bored ∈ [0,0.09], Positivity_score ∈ [0.19,1], RELIGION ∈ [0,0.15], Happy ∈ [0.24,0.32], feedback ∈ [0,0.81], Angry ∈ [0.04,1], Neutral_score ∈ [0,0.42], EDUCATION ∈ [0,0.01], ENTERTAINMENT ∈ [0,0.83] (6%)
5	spam ∈ [0.05,1], Happy ∈ [0.02,0.32], TRAVEL ∈ [0,0.57], feedback ∈ [0,0.52], Positivity_score ∈ [0,0.64], Neutral_score ∈ [0.15,0.56], Angry ∈ [0,0.08], marketing ∈ [0,0.08], ut_len ∈ [78.5,418], query ∈ [0.06,1], news ∉ {0}, suggestion ∈ [0,0.21] (8%)	Bored ∈ [0,0.08], Excited ∈ [0.08,1], feedback ∈ [0.47,0.77], Happy ∈ [0.32,1], Positivity_score ∈ [0.38,0.46], emotional_context ∉ {*afraid*}, POLITICS ∈ [0,0.19], Neutral_score ∈ [0.35,1] (8%)	query ∈ [0,0.72], PL_intensity ∈ {'1'}, Negative_score ∈ [0,0.29], ENTERTAINMENT ∈ [0,0.07], Happy ∈ [0.19,1], WORLDPOST ∈ [0,0.01] (9%)

As for the comparison between the patterns obtained, we see that there is a significant difference between those patterns found using all of the features (original), the reduced model obtained through feature selection (reduced), and the use of Plutchik's model of emotion. While not entirely different regarding the representation of the patterns across the classes, we see that the information provided by the patterns is diverse in nature.

7. Conclusions and Future Work

We were able to create a measurement of empathy in text based on accepted psychological research. Our empathy metric, which we named Empathy score, consists of five equidistant levels that go from "No empathy present" and "Very much empathetic". The

measurement of empathy was generated by human participants, which were instructed to take into account both components of empathy in equal parts when describing the Empathy score of a text conversation. This metric is validated by the use of the EQ self-report scale, as the people that assigned the Empathy score to our database were considered highly empathetic.

Additionally, this research resulted on the validation of using alternative models of emotion as a viable method for representing the emotional component of empathy. We were able to find that using both dimensional and categorical models of emotion are equivalent when classifying empathy using the Empathy score, and can produce complementary information for researchers.

Thanks to our research in empathy measurement, we present EC. EC is a database obtained from sampling ED, which contains 400 conversations. These conversations are presented in the same format as ED, and are therefore easily implemented in machine learning projects. The database contains most of the features found in ED, including an almost equal distribution of emotional contexts. The most important aspect of this database is the Empathy score. This metric of empathy applies to each of the conversations and measures the level of empathy presented by both participants. The metric was obtained as a result of the validation process performed on ED.

We proposed the use of a contrast pattern-based classifier (PBC4cip) for the classification of empathy in each utterance of the database. We hypothesized that using this classifier would be the best option for the classification of empathy since it was developed to address the problems presented by data imbalance in classification tasks. Additionally, since it was based on patterns, the resulting classification was able to find information that could be later interpreted by experts in the field of psychology for a variety of purposes.

The results found regarding the contrast between PBC4cip and other algorithms was similar in both AUC and CEM that there was a significant difference between the performance of the pattern-based classifier and Gaussian Naive Bayes, with it performing significantly better. In addition, AUC results show that it was superior in performance to k-Nearest Neighbors and Decision Trees. However, we also observed that the results do not show evidence for PBC4cip being better than a Multi-layer Perceptron classifier or a Random Forest classifier in metric performance. Despite this, we suggest that PBC4cip is a superior approach to both MLP and RF, considering the advantages given by the explanatory model. The algorithm presents one of the better performances in both metrics and provides information that can be used to increase the knowledge regarding empathy. Therefore, this research showed the hypothesis to be correct.

In addition to the findings regarding the predictions, we were able to confirm that it is possible to obtain knowledge regarding empathy in textual communication through the use of a contrast-pattern classifier. This makes the approach explainable. We observed several patterns that represented a significant portion of the utterances, such as 32% or 26%, while we also found a vast majority of patterns that only represented 1% or 2% of the instances with the class. For example, the pattern: $Fear \in [0,0.27]$, $ut\_len \in [0,42.5]$, $emotional\_context \in \{Apprehensive\}$, $Sad \in [0,0.4]$, $TECH \in [0,0.03]$ represents 26% of instances with low empathy. While actual interpretation should be reserved to experts in this area, a preliminary interpretation can yield the hypothesis that short conversations with an the emotional context is of apprehension, as well as a score of both fear and sadness over 0.25% tend to not present empathy. However, we must address that some patterns are not entirely intuitive to interpret.

PBC4cip presents one of the best, to date, methods for detecting empathy in texts, when measured through two metrics. Additionally, it provides information that grants insight on the reasons for the classification and on the nature of empathy. Therefore, we can conclude that the contrast-pattern approach is superior to all other algorithms used for this research.

7.1. Limitations

The limitations of this research are related to the data used and the nature of the concept of empathy. All of the textual exchanges in this database are limited to the English language. This makes it difficult to state that the findings will be universal across languages and cultures. Therefore, the findings must be only constrained to empathy found in communication through text using the English language. Additionally, the conversations are overall short and carried out between anonymous participants. Therefore, we must address that this research cannot make any statements regarding long-form conversations or conversations between people with close-relationships. There is a possibility that similar results can be obtained. However, there is no sufficient evidence for us to make any prescriptive statements regarding other circumstances.

7.2. Future Work

In future work, further exploration through pattern-based machine learning methods can be pursued. For example, algorithms like PBCEQ [28]. Additionally, one of the approaches that could be taken to improve results regarding patterns and performance is to modify and expand the features available to the algorithms. For example, the use of the Linguistic Inquiry and Word Count (LIWC) tool [105]. This method of text analysis has been previously used and is linked to psychological research [105]. We believe that it is imperative to expand the database. EC contains the Empathy score, which represents a metric for measuring empathy in texts backed by psychological research. However, the database contains only 400 conversations. It would be advisable to pursue further evaluation of samples of ED in order to obtain more examples of conversations with a valid score. Finally, we believe that further exploration of empathy as described by Baron-Cohen and Wheelwright can be taken beyond the limitations of EC or ED.

Author Contributions: Methodology, formal analysis, and investigation E.C.M.-V., J.A.R.U. and O.L.-G.; software, E.C.M.-V.; writing—original draft, E.C.M.-V.; resources, E.C.M.-V.; supervision, J.A.R.U. and O.L.-G.; writing—review and editing, J.A.R.U. and O.L.-G.; validation, J.A.R.U. and O.L.-G.; conceptualization, J.A.R.U. and O.L.-G.; visualization, E.C.M.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Council of Science and Technology of Mexico (CONACyT) through the scholarship grant: CVU 1049928, and Instituto Tecnológico y de Estudios Superiores de Monterrey.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The database used for this experimentation can be downloaded at https://bit.ly/3HkHRH2.

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

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
ED	EmpatheticDialogues
EQ	Empathy Quotient
ES	Empathy Score
ESV	Empathy Score given by volunteers
ESV	Empathy Score given by volunteers
ESO	Empathy Score given by the original database

EC	Empathetic Conversations
WEKA	Waikato Environment for Knowledge Analysis
KNN	K-nearest neighbors
DT	Decision Tree
GNB	Gaussian Naive Bayes
RF	Random Forest
KEEL	Knowledge Extraction based on Evolutionary Learning
AUC	Area Under the Receiver Operating Characteristic Curve
CEM	Closeness Evaluation Measure

Appendix A. Plutchik's Model of Emotion

We present the results from our research into translating the emotional context to an emotion vector representation using Plutchik's model of emotion.

Table A1. Emotional context representations.

Emotion	Plutchik Representation	Label
afraid	(0,0,1,0,0,0,0,0),(2)	0
angry	(0,0,0,0,0,0,1,0),(2)	1
annoyed	(0,0,0,0,0,0,1,0),(3)	2
anticipating	(0,0,0,0,0,0,0,1),(2)	3
anxious	(0,0,1,0,0,0,0,0),(2)	4
apprehensive	(0,0,1,0,0,0,0,0),(3)	5
ashamed	(0,0,1,0,0,1,0,0),(2)	6
caring	(1,1,0,0,0,0,0,0),(2)	7
confident	(1,0,0,0,0,0,0,1),(2)	8
content	(1,0,0,0,0,0,0,0),(3)	9
devastated	(0,0,0,1,1,1,0,0),(1)	10
disappointed	(0,0,0,1,1,0,0,0),(2)	11
disgusted	(0,0,0,0,0,1,0,0),(2)	12
embarassed	(0,0,1,0,0,1,0,0),(3)	13
excited	(1,0,0,0,0,0,0,1),(3)	14
faithful	(1,1,0,1,0,0,0,0),(1)	15
furious	(0,0,0,0,0,0,1,0),(1)	16
grateful	(1,1,0,1,0,0,0,0),(2)	17
guilty	(1,0,1,0,0,0,0,0),(2)	18
hopeful	(0,1,0,0,0,0,0,1),(2)	19
impressed	(0,1,0,1,0,0,0,0),(1)	20
jealous	(0,0,0,0,1,0,1,0),(2)	21
joyful	(1,0,0,0,0,0,0,0),(2)	22
lonely	(0,0,0,0,1,0,0,0),(1)	23
nostalgic	(1,0,0,0,1,0,0,0),(2)	24
prepared	(0,0,0,0,0,0,0,1),(2)	25
proud	(1,0,0,0,0,0,1,0),(2)	26
sad	(0,0,0,0,1,0,0,0),(2)	27
sentimental	(0,1,0,0,0,0,0,0),(2)	28
surprised	(0,0,0,1,0,0,0,0),(2)	29
terrified	(0,0,1,0,0,0,0,0),(1)	30
trusting	(0,1,0,0,0,0,0,0),(2)	31

Appendix B. Empathetic Conversations

The database generated from this research can be found in a *Github* repository, since we prioritized the further exploration of EC for future research; the link for this repository is as follows: https://bit.ly/3HkHRH2.

The database main file contains 400 conversations. Each instance in the database corresponds to one exchange, or utterance in the conversations. For each instance, 12 features are defined:

1. Conv_id: A unique identifier for the conversation in the database. This also corresponds to the identifier in ED.
2. Utterance_idx: The turn corresponding to the utterance in the conversation.
3. Context: A description of an emotional label given to the conversation.

4. Talker: A descriptor whether the utterance is done by the "Talker" or "Listener" role in the database.
5. Utterance: The text exchange of a person involved in the conversation.
6. Prompt: The description of the emotional situation the defines the conversation.
7. Ut_len: The length of the utterance measured by the characters present.
8. Sentiment: String that describes the probabilities of whether the utterance is positive, neutral, or negative.
9. Emotion: String that lists the probabilities that the utterance presents one of six emotions.
10. Taxonomy: String that presents the three most likely taxonomy labels, along with its confidence score. The taxonomy labels are according to the IAB (Interactive Advertising Bureau) Tech Lab Content Taxonomy
11. Intent: String that describes the probability that the utterance corresponds to one of eight distinct intent labels.
12. Empathy: The empathy score of the conversation. It describes the amount of empathy according to five levels. It reflects the judgment of highly empathetic people.

Additionally, a separate file contains another version of the database that codifies the emotional context into features based on Plutchik's emotion model. These features present eight binary categories that represent the eight basic emotions found by Plutchik, as well as an integer value that describes the intensity of the emotions present. The features are as follows:

1. PL_joy: Binary category referring to the presence of the emotion "Joy";
2. PL_trust: Binary category referring to the presence of the emotion "Trust";
3. PL_fear: Binary category referring to the presence of the emotion "Fear";
4. PL_surprise: Binary category referring to the presence of the emotion "Surprise";
5. PL_sadness: Binary category referring to the presence of the emotion "Sadness";
6. PL_disgust: Binary category referring to the presence of the emotion "Disgust";
7. PL_anger: Binary category referring to the presence of the emotion "Anger";
8. PL_anticipation: Binary category referring to the presence of the emotion "Anticipation";
9. PL_intensity: Integer category referring to the intensity of the emotion. The value "1" represent the highest intensity. Meanwhile, the value "3" represents the lowest intensity.

References

1. Freedberg, D. Empathy, Motion and Emotion. *Wie Sich Gefühle Ausdruck Verschaffen: Emotionen in Nahsicht*; Driesen: Taunusstein, Germany, 2007; pp. 17–51.
2. Alam, F.; Danieli, M.; Riccardi, G. Can we detect speakers' empathy?: A real-life case study. In Proceedings of the 7th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2016—Proceedings, Wrocław, Poland, 16–18 October 2016; pp. 59–64. [CrossRef]
3. Zhou, K.; Aiello, L.M.; Scepanovic, S.; Quercia, D.; Konrath, S. The Language of Situational Empathy. In Proceedings of the ACM on Human-Computer Interaction, Málaga, Spain, 22–24 September 2021; Volume 5. [CrossRef]
4. Alam, L. Examining Cognitive Empathy Elements within AI Chatbots for Healthcare Systems. Ph.D. Thesis, Michigan Technological University, Houghton, MI, USA, 2022. [CrossRef]
5. Ançel, G. Developing Empathy in Nurses: An Inservice Training Program. *Arch. Psychiatr. Nurs.* **2006**, *20*, 249–257. [CrossRef] [PubMed]
6. Alam, F.; Danieli, M.; Riccardi, G. Annotating and modeling empathy in spoken conversations. *Comput. Speech Lang.* **2018**, *50*, 40–61. [CrossRef]
7. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83-84*, 19–52. [CrossRef]
8. Picard, R.W. Affective Computing for HCI. In Proceedings of the 8th HCI International on Human-Computer Interaction: Ergonomics and User Interfaces, Munich, Germany, 22–26 August 1999; pp. 829–833.
9. Picard, R.W. Affective computing: Challenges. *Int. J. Hum. Comput. Stud.* **2003**, *59*, 55–64. [CrossRef]
10. Cambria, E. Affective Computing and Sentiment Analysis. *IEEE Intell. Syst.* **2016**, *31*, 102–107. [CrossRef]
11. Zhou, L.; Gao, J.; Li, D.; Shum, H.Y. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguist.* **2020**, *46*, 53–93. [CrossRef]

12. Strauss, M.; Reynolds, C.; Hughes, S.; Park, K.; McDarby, G.; Picard, R.W. The HandWave Bluetooth Skin Conductance Sensor. In *Affective Computing and Intelligent Interaction*; Tao, J., Tan, T., Picard, R.W., Eds.; ACII 2005. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3784, pp. 699–706. [CrossRef]
13. Urquhart, L. Working with Affective Computing: Exploring UK Public Perceptions of AI enabled Workplace Surveillance. *arXiv* **2022**, arXiv:2205.08264.
14. Loyola-Gonzalez, O.; Martinez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Garcia-Borroto, M. Cost-Sensitive Pattern-Based classification for Class Imbalance problems. *IEEE Access* **2019**, *7*, 60411–60427. [CrossRef]
15. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2005; Volume 47; pp. 517–518. [CrossRef]
16. Bringmann, B.; Nijssen, S.; Zimmermann, A. Pattern-Based Classification: A Unifying Perspective. *arXiv* **2011**, arXiv:1111.6191.
17. O'Toole, A.J.; Jiang, F.; Abdi, H.; Pénard, N.; Dunlop, J.P.; Parent, M.A. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J. Cogn. Neurosci.* **2007**, *19*, 1735–1752. [CrossRef]
18. Baron-Cohen, S.; Wheelwright, S. The Empathy Quotient: An Investigation of Adultswith Asperger Syndrome or High Functioning Autism,and Normal Sex Differences. *J. Autism Dev. Disord.* **2004**, *34*, 163–175. [CrossRef] [PubMed]
19. Stinson, J.; Wolfe, R.; Spaulding, W. Social Connectedness in Schizotypy: The Role of Cognitive and Affective Empathy. *Behav. Sci.* **2022**, *12*, 253. [CrossRef] [PubMed]
20. Lawrence, E.J.; Shaw, P.; Baker, D.; Baron-Cohen, S.; David, A.S. Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychol. Med.* **2004**, *34*, 911–919. [CrossRef] [PubMed]
21. Dorris, L.; Young, D.; Byrne, K.; Hoyle, R. Cognitive empathy across the lifespan. *Dev. Med. Child Neurol.* **2022**, 1–10. [CrossRef]
22. Smith, A. *The Theory of Moral Sentiments*, 1976 ed.; Clarendon Press: Oxford, UK, 1759.
23. Davis, M.H. *Empathy: A Social Psychological Approach*, 1st ed.; Routledge: Oxfordshire, UK, 1996; p. 271.
24. Ramamohanarao, K.; Fan, H. Patterns based classifiers. *World Wide Web* **2007**, *10*, 71–83. [CrossRef]
25. Vajjala, S.; Majumder, B.; Gupta, A.; Surana, H. *Practical Natural Language Processing*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2020; p. 424.
26. Loyola-González, O.; Medina-Pérez, M.A.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Monroy, R.; García-Borroto, M. PBC4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowl.-Based Syst.* **2017**, *115*, 100–109. [CrossRef]
27. Loyola-González, O.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; García-Borroto, M. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing* **2016**, *175*, 935–947. [CrossRef]
28. Chen, X.; Gao, Y.; Ren, S. A new contrast pattern-based classification for imbalanced data. *ACM Int. Conf. Proc. Ser.* **2018**, 45. [CrossRef]
29. Mendes, A.C.; Antunes, C. Pattern mining with natural language processing: An exploratory approach. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition 2009*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2009; pp. 266–279. [CrossRef]
30. Aburub, F.; Hadi, W. A New Associative Classification Algorithm for Predicting Groundwater Locations. *J. Inf. Knowl. Manag.* **2018**, *17*, 1–26. [CrossRef]
31. Kumano, S.; Otsuka, K.; Mikami, D.; Yamato, J. Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings. In Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, Santa Barbara, CA, USA, 21–25 March 2011; pp. 43–50. [CrossRef]
32. Leite, I.; Pereira, A.; Mascarenhas, S.; Castellano, G.; Martinho, C.; Prada, R.; Paiva, A. Closing the loop: From affect recognition to empathic interaction. In Proceedings of the AFFINE'10—Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments, Co-located with ACM Multimedia 2010, Firenze, Italy, 29 October 2010; pp. 43–47. [CrossRef]
33. Liu-Thompkins, Y.; Okazaki, S.; Li, H. Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *J. Acad. Mark. Sci.* **2022**, 1–21. [CrossRef]
34. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, LA, USA, 2–7 February 2018; pp. 730–738.
35. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *arXiv* **2017**, arXiv:1710.03957.
36. Liu, M.; Bao, X.; Liu, J.; Zhao, P.; Shen, Y. Generating emotional response by conditional variational auto-encoder in open-domain dialogue system. *Neurocomputing* **2021**, *460*, 106–116. [CrossRef]
37. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. I Know the Feeling: Learning to Converse with Empathy. *arXiv* **2018**, arXiv:1811.00207 .
38. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Proceedings of the ACL 2019—57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.

39. Shin, J.; Xu, P.; Madotto, A.; Fung, P. HappyBot: Generating Empathetic Dialogue Responses by Improving User Experience Look-ahead. *arXiv* **2019**, arXiv:1906.08487.
40. Lin, Z.; Xu, P.; Winata, G.I.; Siddique, F.B.; Liu, Z.; Shin, J.; Fung, P. CAiRE: An Empathetic Neural Chatbot. *arXiv* **2019**, arXiv:1907.12108.
41. Li, D.; Li, Y.; Wang, S. Interactive double states emotion cell model for textual dialogue emotion prediction. *Knowl.-Based Syst.* **2020**, *189*, 105084. [CrossRef]
42. Li, Q.; Chen, H.; Ren, Z.; Chen, Z.; Tu, Z.; Ma, J. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation. *arXiv* **2019**, arXiv:1911.08698.
43. Likert, R. A Technique for the Measurement of Attitudes. *Arch. Psychol.* **1932**, *22*, 5–55. [CrossRef]
44. Beglar, D.; Nemoto, T. Developing Likert-scale questionnaires. In *JALT2013 Conference Proceedings*; JALT Publications: Tokyo, Japan, 2014; pp. 1–8.
45. Wendt, F.R.; Warrier, V.; Pathak, G.A.; Koenen, K.C.; Stein, M.B.; Krystal, J.H.; Pietrzak, R.H.; Gelernter, J.; Goldfarb, E.V.; Baron-Cohen, S.; et al. Polygenic scores for empathy associate with posttraumatic stress severity in response to certain traumatic events. *Neurobiol. Stress* **2022**, *17*, 100439. [CrossRef]
46. Shalev, I.; Warrier, V.; Greenberg, D.M.; Smith, P.; Allison, C.; Baron-Cohen, S.; Eran, A.; Uzefovsky, F. Reexamining empathy in autism: Empathic disequilibrium as a novel predictor of autism diagnosis and autistic traits. *Autism Res.* **2022**, 1–12. [CrossRef]
47. Allison, C.; Baron-Cohen, S.; Wheelwright, S.J.; Stone, M.H.; Muncer, S.J. Psychometric analysis of the Empathy Quotient (EQ). *Personal. Individ. Differ.* **2011**, *51*, 829–835. [CrossRef]
48. Fox, N.; Hunn, A. *Sampling and Sample Size Calculation*; East Midlands/Yorkshire: The National Institutes for Health Research. Research Design Service for the East Midlands/Yorkshire & the Humber: Sheffield, UK, 2009; Volume 1, pp. 1–4.
49. Sharma, G. Pros and cons of different sampling techniques. *Int. J. Appl. Res.* **2017**, *3*, 749–752.
50. Taherdoost, H. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *SSRN Electron. J.* **2018**, *5*, 18–27. [CrossRef]
51. Morse, J.M. Determining Sample Size. *Qual. Health Res.* **2000**, *10*, 3–5. [CrossRef]
52. Kasiulevičius, V.; Šapoka, V.; Filipavičiūtė, R. Sample size calculation in epidemiological studies. *Gerontologija* **2006**, *7*, 225–231. [CrossRef]
53. Israel, G.D. *Determining Sample Size: Program Evaluation and Organizational Development*; University of Florida: Gainesville, FL, USA, 2013; pp. 1–5.
54. Dalkey, N. An experimental study of group opinion: The Delphi method. *Futures* **1969**, *1*, 408–426. [CrossRef]
55. Skulmoski, G.J.; Hartman, F.T.; Krahn, J. The Delphi Method for Graduate Research. *J. Inf. Technol. Educ.* **2007**, *6*, 001–021. [CrossRef]
56. Loyola-Gonzalez, O. Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
57. Lee, J.W.; Kim, S.H. Integrated approach for interdependent information system project selection. *Int. J. Proj. Manag.* **2001**, *19*, 111–118. [CrossRef]
58. Xia, Y. *Correlation and Association Analyses in Microbiome Study Integrating Multiomics in Health and Disease*, 1st ed.; Elsevier: Amsterdam, The Netherlands, 2020; Volume 171, pp. 309–491. [CrossRef]
59. Marino, M.J. *Statistical Analysis in Preclinical Biomedical Research*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 107–144. [CrossRef]
60. Wathen, L.M. *Estadística Aplicada a los Negocios y la Economía*, 16th ed.; McGraw-Hill: Mexico City, Mexico, 2015; p. 731.
61. Siegel, S. Nonparametric Statistics. *Am. Stat.* **1957**, *11*, 13–19. [CrossRef]
62. Jain, A.; Aggarwal, I.; Singh, A. ParallelDots at SemEval-2019 Task 3: Domain Adaptation with feature embeddings for Contextual Emotion Analysis. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 185–189. [CrossRef]
63. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl.-Based Syst.* **2022**, *235*, 107643. [CrossRef]
64. Bruna, O.; Avetisyan, H.; Holub, J. Emotion models for textual emotion classification. *J. Phys. Conf. Ser.* **2016**, *772*, 1–6. [CrossRef]
65. Ekman, P.; Friesen, W.V.; Sullivan, M.O.; Diacoyanni-tarlatzis, I.; Chan, A.; Heider, K.; Lecompte, W.A.; Krause, R.; Scherer, K.; Tomita, M.; et al. Universals and cultural differences in the Judgments of Facial Expressions of Emotion. *J. Personal. Soc. Psychol.* **1987**, *5*, 712–717. [CrossRef]
66. Burkhardt, H.A.; Pullmann, M.D.; Hull, T.D.; Areán, P.A.; Cohen, T. Comparing emotion feature extraction approaches for predicting depression and anxiety. In Proceedings of the 8th Workshop on Computational Linguistics and Clinical Psychology, Online, 15 July 2022; pp. 105–115.
67. El Ray, L.A.; Fathy, H.; Mattar, Y.; Badie Taher, D. Emotion identification and mentalization in non-psychotic first-degree relatives of young adult patients with schizophrenia disorder. *Egypt. J. Neurol. Psychiatry Neurosurg.* **2022**, *58*, 63. [CrossRef]
68. Cuzzocrea, A.; Pilato, G. *Taxonomy-Based Detection of User Emotions for Advanced Artificial Intelligent Applications*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10870 LNAI, pp. 573–585. [CrossRef]
69. Schuurmans, J.; Frasincar, F. Intent Classification for Dialogue Utterances. *IEEE Intell. Syst.* **2020**, *35*, 82–88. [CrossRef]

70. Purohit, H.; Dong, G.; Shalin, V.; Thirunarayan, K.; Sheth, A. Intent classification of short-text on social media. In Proceedings of the 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, China, 19–21 December 2015; pp. 222–228. [CrossRef]
71. Pérez-Landa, G.I.; Loyola-González, O.; Medina-Pérez, M.A. An explainable artificial intelligence model for detecting xenophobic tweets. *Appl. Sci.* **2021**, *11*, 801. [CrossRef]
72. García-Borroto, M.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A. Finding the best diversity generation procedures for mining contrast patterns. *Expert Syst. Appl.* **2015**, *42*, 4859–4866. [CrossRef]
73. Ho, A.; Hancock, J.; Miner, A.S. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *J. Commun.* **2018**, *68*, 712–733. [CrossRef]
74. Plutchik, R. *A General Psychoevolutionary Theory of Emotion*; Academic Press: Cambridge, MA, USA, 1980; Volume 1, pp. 3–33. [CrossRef]
75. Plutchik, R. A psychoevolutionary theory of emotions. *Soc. Sci. Inf.* **1982**, *21*, 529–553. [CrossRef]
76. Tromp, E.; Pechenizkiy, M. Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel. *arXiv* **2014**, arXiv:1412.4682.
77. Dimov, R. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. *AI Tools Semin.* **2007**, *99*, 192–196.
78. Gnanambal, S.; Thangaraj, M.; Meenatchi, V.T.; Gayathri, V. Classification Algorithms with Attribute Selection:An Evaluation Study using WEKA. *Int. J. Adv. Netw. Appl.* **2018**, *9*, 3640–3644.
79. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
80. Provost, F.; L, T.F. *Data Science for Business: What You Need to Know about*, 2nd ed.; O'Reilly Media, Inc.: Sevastopol, CA, USA, 2013.
81. Van Calster, B.; Van Belle, V.; Condous, G.; Bourne, T.; Timmerman, D.; Van Huffel, S. Multi-class AUC metrics and weighted alternatives. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1390–1396. [CrossRef]
82. Amigo, E.; Gonzalo, J.; Mizzaro, S.; Carrillo-de Albornoz, J. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; Volume 16, pp. 3938–3949. [CrossRef]
83. Frank, E.; Hall, M. *A Simple Approach to Ordinal Classification*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2001; Volume 2167; pp. 145–156. [CrossRef]
84. Triguero, I.; González, S.; Moyano, J.M.; García, S.; Alcalá-Fdez, J.; Luengo, J.; Fernández, A.; del Jesús, M.J.; Sánchez, L.; Herrera, F. KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 1238. [CrossRef]
85. Kotsiantis, S. Supervised Machine Learning: A Review of Classification Techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **2017**, *160*, 3–24. [CrossRef]
86. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
87. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. *How Many Trees in a Random Forest?* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2012; Volume 7376, pp. 154–168. [CrossRef]
88. Sulzmann, J.N.; Fürnkranz, J.; Hüllermeier, E. *On Pairwise Naive Bayes Classifiers*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2007; Volume 4701, pp. 371–381. [CrossRef]
89. TenHouten, W.D. Social dominance hierarchy and the pride–shame system. *J. Political Power* **2017**, *10*, 94–114. 2158379X.2017.1285154. [CrossRef]
90. McCullough, M.E.; Kilpatrick, S.D.; Emmons, R.A.; Larson, D.B. Is gratitude a moral affect? *Psychol. Bull.* **2001**, *127*, 249–266. [CrossRef]
91. Machizawa, M.G.; Lisi, G.; Kanayama, N.; Mizuochi, R.; Makita, K.; Sasaoka, T.; Yamawaki, S. Quantification of anticipation of excitement with a three-axial model of emotion with EEG. *J. Neural Eng.* **2020**, *17*, 036011. [CrossRef]
92. Rojas, M.; Veenhoven, R. Contentment and Affect in the Estimation of Happiness. *Soc. Indic. Res.* **2013**, *110*, 415–431. [CrossRef]
93. Kammrath, L.K.; Peetz, J. The limits of love: Predicting immediate versus sustained caring behaviors in close relationships. *J. Exp. Soc. Psychol.* **2011**, *47*, 411–417. [CrossRef]
94. Goldstein, L.S.; Lake, V.E. "Love, love, and more love for children": Exploring preservice teachers' understandings of caring. *Teach. Teach. Educ.* **2000**, *16*, 861–872. [CrossRef]
95. Zimmerman, D.W.; Zumbo, B.D. Relative power of the wilcoxon test, the friedman test, and repeated-measures ANOVA on ranks. *J. Exp. Educ.* **1993**, *62*, 75–86. [CrossRef]
96. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
97. McCrum-Gardner, E. Which is the correct statistical test to use? *Br. J. Oral Maxillofac. Surg.* **2008**, *46*, 38–41. [CrossRef]
98. García, S.; Herrera, F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *J. Mach. Learn. Res.* **2008**, *9*, 2677–2694.

99. García, S.; Fernández, A.; Luengo, J.; Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.* **2010**, *180*, 2044–2064. [CrossRef]
100. Derrac, J.; García, S.; Molina, D.; Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol. Comput.* **2011**, *1*, 3–18. [CrossRef]
101. Trawinski, B.; Smetek, M.; Telec, Z.; Lasota, T. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *Int. J. Appl. Math. Comput. Sci.* **2012**, *22*, 867–881. [CrossRef]
102. Potochnik, A.; Colombo, M.; Wright, C. Statistics and Probability. *Recipes Sci.* **2018**, 167–206. [CrossRef]
103. Xu, M.; Fralick, D.; Zheng, J.Z.; Wang, B.; Tu, X.M.; Feng, C. The differences and similarities between two-sample *t*-test and paired *t*-test. *Shanghai Arch. Psychiatry* **2017**, *29*, 184–188. [CrossRef]
104. David, H.A.; Gunnink, J.L. The Paired t Test Under Artificial Pairing. *Am. Stat.* **1997**, *51*, 9–12. [CrossRef]
105. Tausczik, Y.R.; Pennebaker, J.W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]