Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering

Correlations Between the Usefulness and Polarity of an Explanation in Forums for Software Users

Bachelorarbeit

im Studiengang Informatik

von

Kushtrim Qengaj

Prüfer: Prof. Dr. Kurt Schneider
Zweitprüfer: Dr. Jil Klünder
Betreuer: Martin Obaidi, M. Sc.

Hannover, 28.10.2024

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 28.10.2024

____
Kushtrim Qengaj

Zusammenfassung

Diese Bachelorarbeit untersucht den Zusammenhang zwischen der Polarität in Erklärungen und derer wahrgenommenen Nützlichkeit in foren für Softwarenutzern. Ziel dieser Untersuchung ist es, festzustellen, ob und wie die Polarität sowie andere metriken mit der Nützlichkeit einer Erklärung korrelieren und deren Einfluss zu quantifizieren.

Zur Durchführung dieser Analyse wurde eine Software entwickelt, die einen Datensatz von 3 323 Fragen und 59 398 Erklärungen von *Stack Overflow* gesammelt hat. Jede Erklärung wurde hinsichtlich ihrer Polarität sowie anderer relevanter Metriken wie Länge, Antwortzeit und Autorenreputation analysiert. Verschiedene Werkzeuge zur Polaritätserkennung wurden evaluiert, darunter Methoden des machinellen Lernens und lexikonbasierte Ansätze.

Zur Bewertung der Abhängigkeiten zwischen den verschiedenen Metriken wurde die Pearson-, Spearman- und Point-Biserial-Korrelationen verwendet, während simple und multiple lineare Regressionen genutzt wurden, um den Einfluss den die Metriken auf die Nützlichkeit haben zu quantifizieren. Die Ergebnisse zeigten keine Korrelation zwischen positiver oder neutraler Polarität und Nützlichkeit, und die Resultate für negative Polarität waren nicht statistisch signifikant. Jedoch zeigten Metriken wie die Länge der Erklärung und die Antwortzeit eine signifikante korrelation mit der Nützlichkeit. Die multiple lineare Regression zeigte keine starke Abhängigkeit zwischen Polarität in kombination mit anderen Metriken und der Nützlichkeit.

Diese Ergebnisse stellen die Annahme in Frage, dass eine erklärung mit positiver oder neutraler emotion die Nützlichkeit der Erklärung erhöhen. Außerdem weisen sie darauf hin, dass inhalt und kontext einen größeren Einfluss auf die Nützlichkeit von Erklärungen haben.

Abstract

Correlations Between the Usefulness and Polarity of an Explanation in Forums for Software Users

This thesis investigates the relationship between the polarity of explanations and their perceived usefulness in software related forums. This research aims to determine if and how polarity, along with other metrics, correlate with an explanations usefulness and quantifies their impact.

For this analysis, a software was developed to collect a dataset of 3 323 questions and 59 398 explanations from *Stack Overflow*. Each explanation was analyzed for polarity, as well as other relevant metrics such as length, response time, and author reputation. Various approaches for polarity detection were evaluated, including machine learning and lexicon-based methods.

To asses relationships between metrics Pearson, Spearman, and Point-biserial correlations, were used, while simple and multiple linear regressions were used to quantify their influence on usefulness. Findings revealed no correlation between positive or neutral polarity and usefulness, and results for negative polarity were not statistically significant. However, metrics such as explanation length and response time demonstrated a significant correlation with usefulness. Multiple linear regression showed no strong dependency between polarity and usefulness when combined with other metrics.

These results challenge the assumption that positive or neutral emotion present in an explanation enhance the usefulness of that explanation. Furthermore suggesting that content and contextual factors have a larger impact on the usefulness of explanations.

Contents

Chapter 1

Introduction

After a software system has been developed and released to end users, there is often a need for further explanation of certain aspects of the system by users [39, 54]. These questions can be answered by an explainer, who provides an explanation of that aspect to the user which enables them to understand that aspect of the system in the context the user is in. Interactions like these make up the definition of Explainability in the context of requirements engineering [8].

Questions arising from this need for explanations in software systems are often posed on online question and answering forums such as *Stack Overflow* and *Ask Ubuntu*. Similar questions can also be found on the platforms where the application or software is published like Valve Corporations *Steam*.

One example for this could be a software developed to generate receipts in multiple languages and for multiple currencies. A user of such a software might need help to find out how to generate a receipt in a specific language for a specific currency or the user may need help to create a new template for a receipt. This user can pose their question on a question answering forum. There the user might get differing explanations. The explanations which are then posted by others can usually be voted up or down and are usually of varying clarity and quality.

In general this can distinguish between useful and less useful explanations for users. These explanations can then be used to find out what makes some explanations more useful than others. Finding metrics that have a relationship to the usefulness of an explanation can be used by technical support for example. This can help to determine how useful an explanations to a question posed by a user is, in order to enhance said explanation. This poses the following problem.

1.1 Problem statement

As indicated in Section 1 there is a need to compile and analyse questions and answers from question and answering forums. Going through this process manually is inefficient and labor intensive because there could be potentially tens of thousands or more questions and answers to various problems. A manual analysis could also bias the results due to human error and inconsistencies in judgement for example.

Therefore there is a need to do this compilation fully automatic and to analyse the data statistically in order to prevent and minimize these biases and to come to more accurate and reliable conclusions on how useful explanations are written. One possible solution to this problem could be structured as follows.

1.2 Approach to the solution

This bachelor's thesis will select a question and answering forum to develop an application capable of compiling at least 10 000 question and answer posts from an online question and answering forum. This will give a representative sample of questions and explanations, addressing these questions. The posts will be filtered based on metrics provided by the forum into at least 2 000 questions and 8 000 explanations.

Furthermore, this thesis will ascertain the polarity of these explanations by using a pre-existing sentiment analysis tool. The results of this in combination with other metrics both directly provided by the question and answer forum and determined from the given information, will be the basis for a correlation analysis. This will contribute towards insights into the relationships between the usefulness and the polarity of explanations. Based on the results of these analyses this thesis will have established metrics which help to determine the usefulness of an explanation. This will enable explainers like developers or technical support to evaluate how useful an explanation is.

1.3 Structure of the thesis

In the beginning, Chapter 2 will introduce and explain foundational topics needed to understand this thesis. After that Chapter 3 will present some of the work that is related to this one.

Then in Chapter 4 the concepts related to the developed software and the structure of the analysis will be demonstrated. Subsequently Chapter 5 will show how the concepts were implemented. After this in Chapter 6 the results of the different analyses will be presented.

Following this comes Chapter 7 which discusses the answers to the research questions and threats to the validity of this thesis. In the end, Chapter 8 provides a summary and will highlight some of the potential future work that could be done after this one.

Chapter 2

Foundations

This chapter aims to describe the theoretical concepts fundamental to this thesis. This will include concepts related to Explainability, Sentiment Analysis and Inferential Statistics.

2.1 Explainability

As software systems grow and they get more features, generally they become more complex to navigate, operate and to understand fully as a user. This is the reason why systems need to be able to explain themselves. Because of that explainability as a non functional requirement [26] (NFR) in Requirements engineering is becoming more and more important.

There are different ways to define explainability [26]. For the purposes of this thesis we will use the definition provided by Chazette et al. [8].

"*A system S is explainable with respect to an aspect X of S relative to an addressee A in context C if and only if there is an entity E (the explainer) who, by giving a corpus of information I (the explanation of X), enables A to understand X of S in C.*" [8]

In other words this means that explainability for a system refers to, whether a specific aspect in a context is explained by an explainer to an addressee by giving them some artifact which provides information to the addressee, that allows them to understand the aspect in the context of this system [8].

To illustrate this concept further here is a concrete example. Consider an application which allows a user to manage their bank account. The user may want to transfer money from their bank account in order to pay for a bill and has navigated to a menu, which allows them to make transactions. It is possible that the user might not understand how to specify that they want to transfer money from their account and not to their account. This means that this example software system could only be explainable, in this aspect in this context, if there was some artifact which was provided by an explainer which enables the user to understand how to specify a transfer of money from their account. This artifact could be for example a tooltip or an FAQ page. In these examples the application itself would provide the information an thus be the explainer.

To summarize the different roles in the example. The application is the system, being in the transaction menu is the context, specifying that money is supposed to be send and not received is the aspect, the application is the explainer and the tooltip is the corpus of information.

An important implication of this definition for explainability is that a written text which was given by someone, that enables someone else to understand an aspect of a system in a specific context can be considered an explanation. From here on all such answers will be referred as explanations in accordance with this definition.

2.2 Sentiment Analysis

Sentiment analysis (SA) is a sub field of Natural Language Processing [32]. It is the process classifying the sentiment, emotions and attitudes present in natural language [32]. Sentiment in this context can refer to the positive, negative or neutral judgement expressed in a given text [33]. This is sometimes also called the polarity of the text [59]. Sentiment can also refer the specific emotion expressed in the text like love, joy, surprise, anger, sadness and fear [40].

Texts of different magnitudes of size can undergo SA. The focus of the analysis can also be at different levels. The levels which are usually distinguished are document-level SA [5], sentence-level SA [6] and aspect-level SA [19]. Here is an example further illustrate this.

Consider a report on the purity of the water from a specific lake. This document may overall communicate a negative sentiment. This might be case if the water in the lake was found as impure because of too high levels of sewage found in the water. However some sentences of this document can be considered of positive sentiment. For example "The cooperation with the sewage treatment plant has already been confirmed which is a promising sign of change to come." And some aspects can still be considered negative.

There have been multiple different approaches to Sentiment Analysis [7, 38, 48]. These include, but are not limited to, machine learning based approaches, lexicon based approaches and hybrid approaches [7, 38, 48]. Each of these approaches has multiple different possible sub-approaches as seen in figure 2.1.

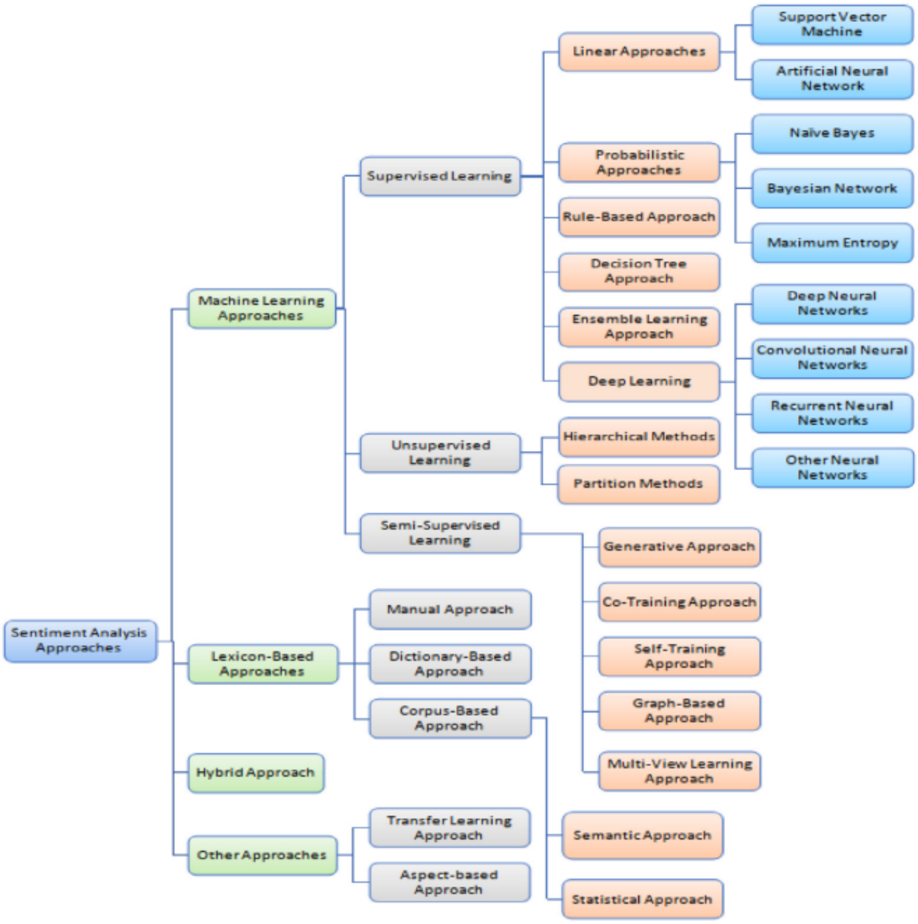


Figure 2.1: A tree structure showing the hierarchy of different approaches in Sentiment Analysis. Child nodes symbolize sub-approaches of parent nodes. Adapted from Birjali et al. [7] by removing a node labelled "Proposed Approach".

Machine learning approaches to sentiment analysis primarily involve training a classifier [7, 48] that determines the sentiment of a given text [7] based on learned patterns in the training data. As seen in figure 2.1 machine learning based approaches can be further categorized by the specifc method used for training.

Sentiment Analysis tools that are based on machine learning need data to train on [49]. The training data is usually a written text in natural language which has been classified, using either an emotion or polarity. These labels in the training data have usually been decided upon by people. Some of the more popular datasets for SA come from *JIRA*, *Github* and *Stack Overflow* [38].

In general, lexicon based SA tools for polarity detection make use of two different lexicons. One holds a collection of words associated with a positive sentiment and the other has a collection of words associated with negative sentiment [62].

One challenge for sentiment analysis is sarcasm [37]. For example the sentence "Wow, the weather is so nice." on its own can be classified as positive, but if the very same sentence, is put in a different context it can have the opposite polarity. As seen in the following example "It has been raining all day today. Wow, the weather is so nice." This, not only implies that ideal sentiment analysis systems need to be able to detect sarcasm but also that the sentiment of any sentence may be dependent on other sentences in the text.

2.3 Inferential Statistics

Inferential Statistics is the sub field of Statistics that studies different methods and techniques which allow for predictions based on collected data [44].

2.3.1 Correlation Analysis

In inferential statistics, correlation analysis is a technique which aims to show the relationships between two variables. This thesis will focus primarily on Pearson's correlation r, Point-biserial correlation r_{pb} and Spearman's correlation ρ because each of these correlation coefficients test for different kinds of relationships between variables [16, 27, 65, 50].

Pearson's correlation coefficient and Point-biserial correlation

Pearson's correlation coefficient r measures the degree to which two variables are linearly dependent on each other [16]. This can be calculated for the random variables X and Y with the following formula

$$r = \frac{Cov(x,y)}{s_x s_y} \text{ [16]}$$

where x and y are n random samples of X and Y, $Cov(x,y)$ is

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \text{ [61]}$$

the covariance of x and y, with x_i and y_i being the i-th sample of x and y respectively. $\overline{x}$ and $\overline{y}$ are the sample means. The mean of a sample z can be calculated as follows

$$\overline{z}{=}E(z){=}\frac{\sum_{i=1}^{n}(z_i)}{n} \text{ [61].}$$

The standard deviations of x and y will be symbolized by s_x and s_y. The standard deviation of any sample z can be calculated as follows

$$s_z{=}SD(z){=}\sqrt{(\frac{\sum_{i=1}^{n}(z-\overline{z})}{n})^2} \text{ [61].}$$

The Point-biseral correlation r_{pb} is a special case of Pearson's correlation where one of the variables is dichotomous, or in other words binary and the other is metric [27]. Note that variables of scale ratio can also be considered to be of scale metric. To calculate the Point-biserial correlation of two variables the following formula can be used

$$r_{pb}{=}\frac{\overline{x_1}-\overline{x_0}}{s_x}\sqrt{\frac{n_1 n_0}{n(n-1)}}\text{[27].}$$

Here $\overline{x_1}$ and $\overline{x_0}$ are the means of the metric variable when the dichotomous variable takes on both its respective values. Furthermore n is the sample size, n_0 and n_1 are the number of samples where the dichotomous variable again takes on both its respective values [27].

Pearson as well as Point-biseral correlation is bound between -1 and 1 [17]. Table 2.1 illustrates the implications of correlataions of different strength.

Strength of correlation r	Effect size
$0.0 <= \|r\| < 0.1$	None
$0.1 <= \|r\| < 0.3$	Small
$0.3 <= \|r\| < 0.5$	Moderate
$0.5 <= \|r\| <= 1.0$	Strong

Table 2.1: A table showing what strength of correlation can be considered of what effect size, using [10][11].

Pearson's correlation coefficient is a widely used measure with many many applications in different fields. These include but are not limited to medicine [20, 60, 47], engineering [42], robotics [34], image processing [34, 47], data science [24] and machine learning [24].

Note that the closer r is to 0 the less the variables are *linearly* correlated. Thus its important to highlight that an r of 0 does *not* imply that there is *no relationship* between the variables. It only shows that there is not a *linear* one between them. This is further illustrated in figure 2.2.

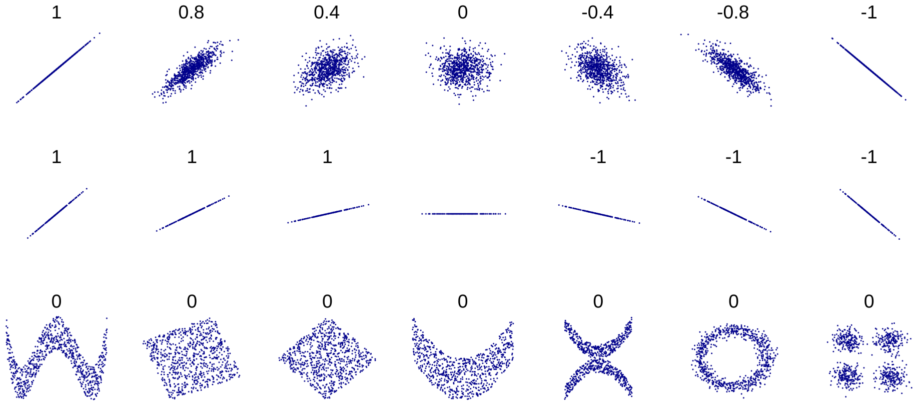


Figure 2.2: A collection of plots showing differently distributed variables and the values for r each of the plots has [13].

Another limitation of the Pearson's correlation coefficient is that its can be skewed by outliers in the data. This is demonstrated in figures 2.3a and 2.3b.

A possible alternative to Pearson's correlation coefficient is Spearman's correlation coefficient which is further explained in the next section.

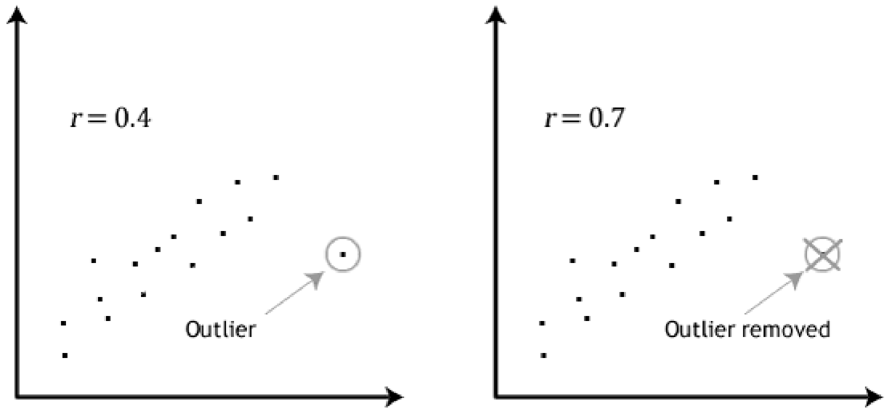
Spearman's correlation coefficient

The Spearman correlation coefficient ρ measures the degree to which the magnitudes of the ranks of two variables are increasing or decreasing dependent on each other [65, 50]. If each rank is assigned at most once then ρ can be calculated as follows

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \text{ [53]}$$

with

$$d_i = \mathcal{R}(x_i) - \mathcal{R}(y_i) \text{ [65]}$$



(a) A plot which contains an outlier in it.

(b) A plot which doesn't contain an outlier in it.

Figure 2.3: Example plots using the same data but 2.3a does contain an outlier and 2.3b does not. These plots also show the r-value of the of the two variables which have been measured. Both adapted by removing the lines of best fit [28].

being the difference of ranks of the i-th data-point, where $\mathcal{R}(z_i)$ is the rank of i-th data-point of z.

The Spearman correlation coefficient ranges from -1 to 1 [50]. Just like Pearson's correlation coefficient numerically large values imply strong monotonic dependency and numerically small values imply no monotonic dependency [65, 50]. The implications of different strengths of correlation can be seen in Table 2.1.

A possible alternative to Spearman's correlation coefficient is Pearson's correlation coefficient which has been presented in Section 2.3.1.

2.3.2 Linear Regression

In inferential statistics linear regression is a technique that aims to find the linear function which best predicts a random variable. There can be multiple approaches depending on whether one or multiple variables can be used predict the depending variable. First simple linear regression will be introduced. After that multiple linear regression is explained.

Simple Linear Regression

Simple linear regression is a method that aims to find the best description of one variable which linearly depends on one specific other variable. For example let us assume we aimed to describe y in this manner, choosing x to be the variable we use to predict y. This linear regression can be written as

$$y = l(x) = \beta_1 x + \beta_0 \text{ [61]}.$$

Here β_0 and β_1 are the arguments of the regression which are chosen as to best fit the recorded data. Note that a linear function can't necessarily intersect all the data points see the example in Figure 2.4.

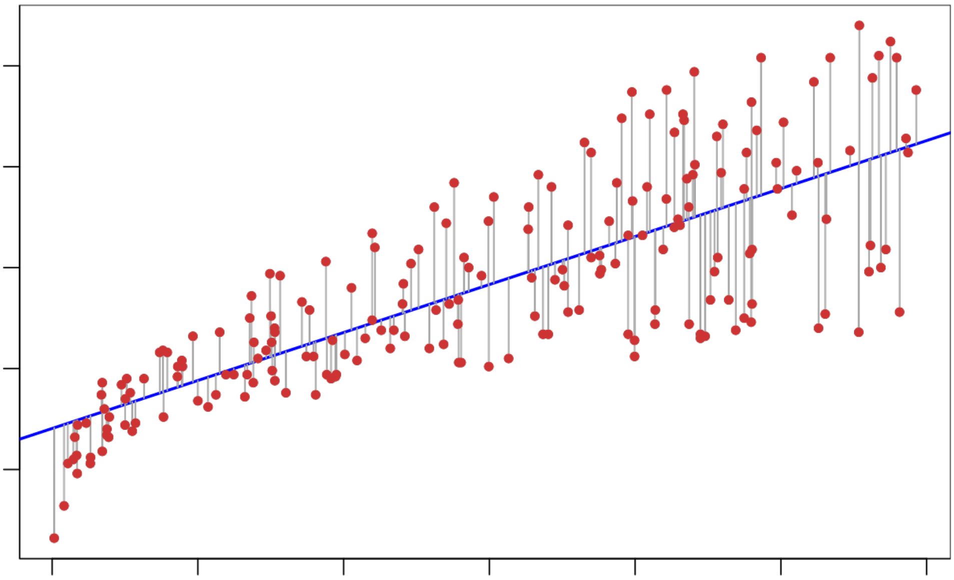


Figure 2.4: An example for a simple linear regression. The red dots show the data the blue line is the regression and black lines indicate the prediction error. Adapted from James et al. [22] by removing the labels on the x and y axes.

In order to get a linear function l which approximates the data well it needs to be chosen such that it minimizes the error between l and the measured y. The function describing this difference is the residual sum of squares (RSS) function

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2 \text{ [61]}.$$

Here x_i and y_i are the values of the i-th datum of x and y respectively. Using the RSS the ideal arguments $\hat{\beta}_0$ and $\hat{\beta}_1$ can be derived to be

$$\hat{\beta}_0{=}\overline{y} - \hat{\beta}_1\overline{x} \text{ [61]}, \quad \hat{\beta}_1{=}\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \text{ [61]}.$$

Thus the best simple linear regression of y in terms of x is

$$y = l(x) = \hat{\beta}_1 x + \hat{\beta}_0.$$

this is called the line of best fit. To asses how well the regression fits the data the coefficient of determination R^2 is used [22]. R^2 can vary from 0 to 1 and the bigger it is the better better the fit. It can be calculated as follows

$$R^2{=}\frac{Var(\hat{\beta}_1 x + \hat{\beta}_0)}{Var(y)}\text{[2]}, \tag{2.1}$$

with

$$Var(z) = \frac{1}{n}\sum_{i=1}^{n} z_i^2 - \overline{z}^2 \text{ [2]}.$$

Because R^2 is equivalent to r^2 for simple linear regression [22] r^2 will be used for calculation instead.

Multiple Linear Regression

Multiple linear regression is similar to simple linear regression. Like simple linear regression, it describes a variable in terms of a linear function. However, unlike simple linear regression, multiple linear regression is a linear combination of multiple variables. This means a regression of y using the variables x_1, x_2, ..., x_m can be written as

$$y{=}\beta_0 + \sum_{i=1}^{m} \beta_i x_i\text{[61]}.$$

Here the β are the arguments which need to be found such that the chosen error function is minimized.

The coefficient of determination for Multiple linear regression can be seen in equation 2.1 since it is calculated in same way as simple linear regression.

Chapter 3

Related work

This chapter will introduce and summarize some of the publications that are related to this thesis. First, works related to Explainability will be presented. After that, literature related to sentiment analysis will be shown. Lastly, publications about analysis of Questions will be discussed. This will provide an overview of the work that is related to this thesis.

3.1 Explainability

Chazette et al. [8] have addressed the absence of structured knowledge about explainability as a non functional requirement in requirements engineering. They did this by proposing a definition, a model, and a catalogue for explainability. To justify these propositions they have gathered the findings on explainability of other fields in an interdisciplinary Systematic Literature Review (SLR) for which forward and backward snowballing was also used. They then validated these findings with philosophers, psychologists and requirements engineers in workshops.

In order to investigate and understand the views and expectations of users Chazette and Schneider [9] have conducted an online exploratory questionnaire. This questionnaire had 107 participants which answered questions related to the need for explanations in applications that they use regularly. The analysis of these answers showed correlations between explainability and NFRs related to transparency and that explanations can positively as well as negatively affect on some NFRs. The authors dubbed this a *double edged sword* effect.

3.2 Sentiment Analysis

In order to contribute to a shared corpus of annotated resources to support research on emotion awareness in software development Novielli et al. [36] have build a gold standard data set for emotion comprised of *Stack Overflow* question, answer and comment posts. To detect the emotions, in the dataset of 4 800 posts, they used in total 12 raters which annotated the posts.

Nguyen et al. [35] have build and compared machine learning models, which use SA on social media, to predict stock movement [35]. The authors gathered the necessary price and mood information of 18 stocks from *Yahoo Finance* and from message boards of the stocks on *Yahoo Finance*. Nguyen et al. [35] have evaluated these models using the Accuracy metric which was calculated for each stock for each model. To summarize these results they calculated the average Accuracy of each model over all 18 stocks. After evaluation the aspect-based sentiment model turned put to have the highest average Accuracy of 0.5441. They therefore conclude that the integration of sentiments helps improve stock market prediction.

3.3 Question and Answering forum Analysis

Procaci et al. [41] have analysed metrics and machine learning models to identify users which generally provide good answers and are willing to help other users. They have investigated the correlations between user attributes, community attributes and user reputation. After that they proposed the usage of a neural network and a clustering algorithm to identify reliable users. They have found that the degree to which users participate in a divers set of subjects is positively correlated with the users reputation. When modeling the communities as graphs of users with interactions being represented as edges and users as nodes, they found that the number of messages, the indegree and the page rank in the network had a good correlation with the user reputation.

3.4 Demarcation to this work

The work of Chazette et al. [8] defines the theoretical underpinnings of the term explainability of this work. This in combination with some of the metrics Procaci et al. [41] have used and identified, will be the foundation for the analysis of the gathered dataset. For SA similar to Nguyen et al. [35] forum posts will be used to find relationships involving the polarity of the posts. These posts, unlike those in the work of Nguyen et al. [35], will be by software users and will be used to better understand how the usefulness and the polarity of an explanation correlate. That means, this work is also unlike the work of Procaci et al. [41], since their work sought out to identify users which provide good explanations and this work aims to find metrics which correlate the usefulness of the explanations themselves.

Chapter 4

Concept

This chapter will introduce and illustrate the concepts which were used to guide the implementation of the software to be developed for this thesis. That system will investigate the dependencies between the usefulness of explanations and various metrics including the polarity of the explanation. Furthermore this chapter will present the procedure which was followed during the analysis of the data. In the beginning the requirements of the software will be shown.

4.1 Requirements

The requirements listed in Table 4.1 have been adapted from the initial notice to this bachelor's thesis and have been extended by some additional non-functional requirements which will make further development easier.

These requirements will ensure a well structured approach to the development of the software and will be met by three systems combined the architecture of which is further elaborated in Section 5.2. Now that the foundational requirements have been established, a forum to be analysed can be chosen. This will be done in the following section.

4.2 Forum Choice

For the purposes of the analysis of the questions and explanations a forum for software users will be chosen in this section. Candidates which have been considered are the following

Functional requirements	• The system has to be able to collect questions and explanations from users of a question and answering platform. • The system has to be able to filter the explanations based on predefined criteria. • The system has to be able to prepare the dataset for analysis. • Using an already existing sentiment analysis tool of the system has to determine the polarity of – Questions – Explanations • The system has to be able to do a correlation analysis.
Non-Functional requirements	• The system should be modular. • The system should have have documentation. • The system should be easily maintainable.

Table 4.1: A table showing the requirements for the software system to be developed.

- *Ask Ubuntu*[1]
- *Github*[2]
- *Gitlab*[3]
- *Quora*[4]
- *Reddit*[5]
- *Stack Overflow*[6]

[1] https://www.askubuntu.com/
[2] https://www.github.com/
[3] https://www.gitlab.com/
[4] https://www.quora.com/
[5] https://www.reddit.com/
[6] https://www.stackoverflow.com/

- *Steam*[7]
- *Super User*[8]

Github, *Gitlab*, *Reddit*, *Quora* and *Steam* do not distinguish between different types of posts. This means there is only an initial post and a discussion between users under that initial post. Because of this discussion oriented approach which does not distinguish between explanations, comments or further elaboration on the original problem, there would have to be some process by which explanation posts would need to be detected. This is why these forums were deemed less suitable for the analysis.

Stack Overflow, *Ask Ubuntu* and *Super User* are all forums hosted by Stack Exchange [56]. Because of this they are structured in a similar way as each other. These forums distinguish between question posts, answer posts, and comment posts. As already discussed in Section 2.1 the definition for explainability that is used for this thesis implies that these answer posts are in fact explanations. This means that this structure in the forums suit the purposes of this thesis well and it eliminates the need to identify explanations.

Since *Stack Overflow* is the biggest forum hosted by Stack Exchange in number of different users and posts [56], it is also the one with the highest total amount of user engagement. This will be beneficial, since it will keep the number of posts low, that will need to be filtered before analysis hence minimizing the total number of posts which will need to be collected. For these reasons *Stack Overflow* will be used as the forum from which the questions and explanations will be sourced from.

Now that a suitable forum has been chosen for analysis, the structure of said analysis can be discussed and further elaborated on.

4.3 Structure of Analysis

In this section the structure of the analysis will be presented. This will include how the data was prepared, what research questions have been posed, what hypothesis are tested in the analysis and what metrics are used.

4.3.1 Research questions

In order to rigorously examine the extend to which there are correlations and dependencies between the usefulness of an explanation and other

[7] `https://store.steampowered.com/?l=german`

[8] `https://www.superuser.com/`

metrics there first have to be precise research questions which then will be investigated. The following research questions will be examined in the analysis:

RQ1 *How does the usefulness of the explanation depend on the polarity of the explanation?*

RQ2 *How does the usefulness of the explanation depend on the contents of the explanation?*

RQ3 *How does the usefulness of the explanation depend on the time at which it was given?*

RQ4 *How does the usefulness of the explanation depend on the explainer of the explanation?*

Now that the research questions have been posed they should be examined.

4.3.2 Hypotheses

To investigate the research questions, hypotheses have been formulated such that they provide further insight into these questions. The Tables 4.2, 4.3, 4.4 and 4.5 show the null hypotheses that have been formulated.

Null Hypothesis	Symbol
The usefulness of the explanation does not depend on the Polarity of the explanation.	$H1_0$
The usefulness of the explanation does not depend on whether the explanation is negative.	$H1.1_0$
The usefulness of the explanation does not depend on whether the explanation is neutral.	$H1.2_0$
The usefulness of the explanation does not depend on whether the explanation is positive.	$H1.3_0$

Table 4.2: An overview of the null hypotheses for RQ1.

Null Hypothesis	Symbol
The usefulness of the explanation does not depend on the contents of the explanation.	$H2_0$
The usefulness of the explanation does not depend on the number of resources that are provided.	$H2.1_0$
The usefulness of the explanation does not depend on the number of code blocks included.	$H2.1.1_0$
The usefulness of the explanation does not depend on the number of links included.	$H2.1.2_0$
The usefulness of the explanation does not depend on the number of images included.	$H2.1.3_0$
The usefulness of the explanation does not depend on the total length of the explanation.	$H2.2_0$
The usefulness of the explanation does not depend on the number of paragraphs.	$H2.2.1_0$
The usefulness of the explanation does not depend on the number of words.	$H2.2.2_0$
The usefulness of the explanation does not depend on the number of characters.	$H2.2.3_0$
The usefulness of the explanation does not depend on the number of letters.	$H2.2.4_0$
The usefulness of the explanation does not depend on the number of sentences.	$H2.2.5_0$
The usefulness of the explanation does not depend on the complexity of the explanation.	$H2.3_0$
The usefulness of the explanation does not depend on the number of letters per word.	$H2.3.1_0$
The usefulness of the explanation does not depend on the number of words per sentence.	$H2.3.2_0$
The usefulness of the explanation does not depend on the readability of the explanation text.	$H2.3.3_0$
The usefulness of the explanation does not depend on the textual lexical diversity of the explanation.	$H2.3.4_0$
The usefulness of the explanation does not depend on the similarity of the explanation and other texts.	$H2.4_0$
The usefulness of the explanation does not depend on the jaccard similarity to the question.	$H2.4.1_0$
The usefulness of the explanation does not depend on the jaccard similarity to the other explanations.	$H2.4.2_0$

Table 4.3: An overview of the null hypotheses and their sub-hypotheses for RQ2.

Null Hypothesis	Symbol
The usefulness of the explanation does not depend on the time at which it was given.	$H3_0$
The usefulness of the explanation does not depend on the time of day it was given.	$H3.1_0$
The usefulness of the explanation does not depend on the time elapsed since the question was posed.	$H3.2_0$
The usefulness of the explanation does not depend on the number of explanations after this one was given.	$H3.3_0$
The usefulness of the explanation does not depend on the time of last activity.	$H3.4_0$
The usefulness of the explanation does not depend on the time it was last edited.	$H3.5_0$
The usefulness of the explanation does not depend on the time at which it was created.	$H3.6_0$
The usefulness of the explanation does not depend on whether the explanation was given on a weekday.	$H3.7_0$

Table 4.4: An overview of the null hypotheses for RQ3.

Null Hypothesis	Symbol
The usefulness of the explanation does not depend on the explainer which has given the explanation.	$H4_0$
The usefulness of the explanation does not depend on the original explainer.	$H4.1_0$
The usefulness of the explanation does not depend on the reputation of the original explainer.	$H4.1.1_0$
The usefulness of the explanation does not depend on the rate at which explanations of the original explainer were deemed the best explanation to their questions.	$H4.1.2_0$
The usefulness of the explanation does not depend on whether the original explainer is registered.	$H4.1.3_0$
The usefulness of the explanation does not depend on the badges awarded to the original explainer.	$H4.1.4_0$
The usefulness of the explanation does not depend on the number of bronze badges the original explainer was awarded.	$H4.1.4.1_0$
The usefulness of the explanation does not depend on the number of silver badges the original explainer was awarded.	$H4.1.4.2_0$
The usefulness of the explanation does not depend on the number of gold badges the original explainer was awarded.	$H4.1.4.3_0$
The usefulness of the explanation does not depend on the last editor of the explanation.	$H4.2_0$
The usefulness of the explanation does not depend on the reputation of the last editor.	$H4.2.1_0$
The usefulness of the explanation does not depend on the rate at which explanations of the last editor were deemed the best explanation to their questions.	$H4.2.2_0$
The usefulness of the explanation does not depend on the badges awarded to the last editor.	$H4.2.3_0$
The usefulness of the explanation does not depend on the number of bronze badges the last editor was awarded.	$H4.2.3.1_0$
The usefulness of the explanation does not depend on the number of silver badges the last editor was awarded.	$H4.2.3.2_0$
The usefulness of the explanation does not depend on the number of gold badges the last editor was awarded.	$H4.2.3.3_0$

Table 4.5: An overview of the null hypotheses and their sub-hypotheses for RQ4.

With the hypotheses formulated, next there will be a focus on the metrics used to test the hypotheses.

4.3.3 Metrics

To test the hypotheses outlined in Section 4.3.2, it is necessary to establish metrics that measure the attributes the explanations. Additionally there will be methods chosen which can find different kinds of dependencies between the different variables of the explanation and the usefulness. The following will consider which metric will be used to measure the usefulness of an explanation and after this will be shown from where other metrics were derived and lastly will be discussed which methods will be used to find what kinds of dependencies between usefulness and other variables of the explanation.

Usefulness

To determine the usefulness of an explanation there can be multiple metrics which can be used. The metrics that the Stack Exchange API directly gives which could be used to measure the usefulness are the following:

- The total number of down-votes of the explanation
- The total number of up-votes of the explanation
- The total score of the explanation

If s is the score of a post, v_u the up-votes and v_d the down-votes then the score of a post is determined in the following way

$$s = v_u - v_d.$$

The score as is could already be used to indicate the usefulness of the explanation as it measures the overall sentiment toward that explanation by users who have the same or a very similar question. There are however some limitations which will be considered. The score of a question is also dependent on the overall engagement with the question.

For example, suppose there was an explanation which was in actuality a very useful one for a very obscure question. Because the question is obscure very few other users will have a similar question and therefore there wont be many people which see the explanation. This means the explanation will get a lower score not because of the usefulness but because less users will rate it in total.

This effect can however be reduced by selecting for question which have at least some minimal score. This ensures that the subject of the question is so common that at least a certain minimum of users engage with the answers.

This can also further be improved by calculating the total score of all explanations combined for a given question. Then the percentage of this total that is contributed by each explanation an be determined . For the purposes of this thesis, this percentage will be referred to as the percent score s_p.

To further illustrate this simple transformation there are two example explanation distributions for two questions in Table 4.6.

Explanation	score	percent score
1	200	0.4
2	150	0.3
3	100	0.2
4	30	0.06
5	20	0.04
total	500	1

Explanation	score	percent score
1	10	0.5
2	7	0.35
3	3	0.15
total	20	1

Table 4.6: Tables showing the distribution of explanation scores and percent scores for two questions. These examples are entirely fictitious and pose as an example.

Since both metrics are in principle fit to measure the usefulness of an explanation both will be used for the analysis. Which metrics were used to predict them will be discussed in the next section.

Overview of all Variables

Stack Overflow generates and maintains some metrics of the questions and explanations which have been given on the platform other metrics need to be derived from either the questions and explanations or the other metrics. A summary of all the variables used, their range and from where they will be derived can be found in Table 4.7. Note that all of these variable are of the scale ratio or dichotomous since they are binary.

Methods to Investigate Correlations and Dependencies

Now that the different variables for analysis have been determined. There now can be a discussion on which methods will be used to test the hypotheses in Section 4.3.2. To evaluate the degree and direction of correlations three different correlation coefficients will be calculated. These are the Pearson correlation, the Spearman correlation and the Point-biserial correlation.

Pearson correlation will be used to determine the linear dependency between the different independent variables and the dependent variable. The Spearman correlation will be used to get an even more general estimate of the dependency of the independent and dependent variables as it measures the monotonic dependency of the variables. Since neither Pearson nor Spearman correlation can be used to analyse dichotomous variables the Point-biserial correlation will be used to analyse the dependencies between these.

Furthermore to quantify the dependencies to be able to make predictions simple linear regression and multiple linear regression was used. Multiple linear regression will also be used for the purposes of finding linear dependencies between a set of independent variables and a dependent variable. To prevent overfitting the size of the set of independent variables will be limited to five or less.

Variable	Range	Derived from
Text of an explanation (T_e)	Char*	API
Is of negative polarity	$\{0,1\}$	T_e
Is of neutral polarity	$\{0,1\}$	T_e
Is of positive polarity	$\{0,1\}$	T_e
Number of code blocks included	$\mathbb{N}_0$	T_e
Has code blocks	$\{0,1\}$	num code blocks
Number of links included	$\mathbb{N}_0$	T_e
Has links	$\{0,1\}$	num links included
Number of paragraphs	$\mathbb{N}_0$	T_e
Number of images included	$\mathbb{N}_0$	T_e
Has images	$\{0,1\}$	num images included
Number of words	$\mathbb{N}_0$	T_e
Number of characters	$\mathbb{N}_0$	T_e
Number of letters	$\mathbb{N}_0$	T_e
Number of sentences	$\mathbb{N}_0$	T_e
Number of letters per word	$\mathbb{Q}_0^+$	letters, words
Number of words per sentence	$\mathbb{Q}_0^+$	words, sentences
Readability of text	$(-\infty, 206.835]$	T_e
Textual lexical diversity	$\mathbb{Q}_0^+$	T_e
Jaccard similarity to the question	$[0,1]$	T_e
Average jaccard similarity to the other explanations	$[0,1]$	T_e
Date given	$\mathbb{N}_0$	API
Time of day given	$\mathbb{N}_0$	Date given
Has been given on a weekday	$\{0,1\}$	Other variable
Time elapsed since question was posed	$\mathbb{N}_0$	Date given
Number of explanations after this one was given	$\mathbb{N}_0$	Date given
Time of last activity	$\mathbb{N}_0$	API
Time the explanation was last edited	$\mathbb{N}_0$	API
Reputation of the original explainer	$\mathbb{N}_0$	API
Acceptance rate of the original explainer	$\mathbb{Q}_0^+$	API
The original explainer has registered	$\{0,1\}$	API
Number of bronze badges the original explainer was awarded	$\mathbb{N}_0$	API
Number of silver badges the original explainer was awarded	$\mathbb{N}_0$	API
Number of gold badges the original explainer was awarded	$\mathbb{N}_0$	API
Reputation of the last editor	$\mathbb{N}_0$	API
Acceptance rate of the last editor	$\mathbb{Q}_0^+$	API
Number of bronze badges the last editor was awarded	$\mathbb{N}_0$	API
Number of silver badges the last editor was awarded	$\mathbb{N}_0$	API
Number of gold badges the last editor was awarded	$\mathbb{N}_0$	API

Table 4.7: An overview of the variables used, their range and from where they were derived.

Chapter 5

Implementation

This chapter of the thesis will show how concepts discussed previously were implemented. First specific tools and technologies are discussed. After this the system architecture will be illustrated.

5.1 Tools and Technologies

Building a system to collect, store, and analyze the necessary data from scratch can be time-consuming and prone to errors compared to utilizing pre-existing solutions. In order to speed up the development, the system developed will utilize commonly used tools and technologies. More specifically this chapter will discuss which programming languages, databases and sentiment analysis tools have been considered and furthermore chosen for the systems development.

5.1.1 Programming language

The programming language used for software development can affect the readability of the system [1] and productivity of the developers [29]. Therefore to enhance maintainability and enable quickly and efficient programming of a system, a well suited programming language should be chosen. The following programming languages will be considered for the implementation.

- *Java*
- *Python*
- *C*

- *C++*

Java, *C* and *C++* are considered mid level languages [25]. This means they offer few layers of abstraction and thus they offer a smaller set of predefined functionality for the developer to use [25]. This implies there needs to be more source code written to implement features which other more high level programming languages already have implemented. This does give more control to the developer to implement features which are more optimized toward a specific task but this more optimized run time efficiency comes at a cost of productivity and readability since as already stated there needs to be more functionality implemented by the developer. Since the primary use of the software will be to collect, process and analyse the data the run time efficiency of the software is less important. Hence the lower level of abstraction is considered a net negative for this project.

Python is considered a high level language [45]. This means that it offers comparatively more layers of abstraction then mid level languages. This means that it has a larger set of predefined functionality. Thus there are less opportunities for further optimization but there is also no need to implement some features since they are already implemented by the language. Hence the higher level of abstraction is considered a net positive for this project. This is why *Python* has been chosen among the candidates presented here.

5.1.2 Database

For the purpose of storing data a database should be chosen because this will eliminate the need to develop functionality related to storing and retrieving data. This will enhance the maintainability of the system. The following database technologies have been considered as candidates.

- *MongoDB*[1]
- *MySQL*[2]
- *SQLite*[3]

Since *MySQL* and *SQLite* are both SQL databases that means they are based on a relational structuring of the data stored within them. This means they are both modeling data using table structures. These tables can be defined such that they match the structure of the data to be stored.

[1] `https://www.mongodb.com/`

[2] `https://www.mysql.com/`

[3] `https://www.sqlite.org/`

Given that the chosen question and answering platform is *Stack Overflow* which means that the data collected is unstructured. *MySQL* or *SQLite* might therefore diminish the maintainability of the system if data of different structure is collected from *Stack Overflow*. This can be the case for example if different data is collected from the API or if the API changes the structure of the data.

MongoDB is a document-oriented non-relational database. Thus it is well suited for handling unstructured data. Its JSON-like document schema is structured very similar to the data returned by the *Stack Overflow* API in general. It supports complex querying with aggregation pipelines which can be useful in order to ensure the modularity of the software to be developed. This is because the extraction of specific data can be done by building an aggregation pipeline for that exact data.

For this reason *MongoDB* was used to store the data that will be collected from *Stack Overflow*.

5.1.3 Dataset Preprocessing

This section of the thesis will discuss how the dataset collected from *Stack Overflow* was preprocessed. This was done to ensure that the dataset is suitable for further analyses. Additionally the preprocessing also aims to ensure that the analyses can be implemented such that they are abstracted from the concrete dataset topology. This way the modularity of the system as a whole is enhanced.

Text Preprocessing

The *Stack Exchange* API [4] returns an HTML version and a markdown version of the text of the explanation. Using any of these versions of the text directly can diminish the accuracy of the sentiment analysis, since both versions contain encodings for special characters. These encodings are not learned by sentiment analysis tools, as they are tailored toward analysing natural language. This is why a plain text version of the answer was determined. This was done by implementing the following text processing pipeline:

1. remove all code blocks
2. remove all HTML tags
3. replace HTML encodings with the decoded characters

[4] `https://api.stackexchange.com/`

4. remove all stop words [15]

5. lemmatize text [51]

The plain text of the answer is determined after Step 3. It can now be used to get other variables such as the polarity of the text, the number of words or the number of letters per word.

This plain text contains words which most texts share. These words are called stop words [15]. Some examples for them are "the", "a" and "in". Since most texts share these words they would artificially increase the similarity of two texts. Different inflections of the same word can also be present in the plain text. For example "written", "wrote" and "write". These variants of the same word would decrease the similarity. In order to minimize both distorting effects on the similarity all stop words are removed and then a lemmatized version of the text is determined [51]. This standardize the different inflections of the same word [51].

Both the plain text version of the explanation and the lemmatized version of the explanation are then added to the dataset alongside the other variables enumerated in Section 4.3.3.

After the plain text version of the explanation has been determined, the polarity of the text of the explanation can then be classified.

5.1.4 Sentiment analysis tool

For the purposes of determining the polarity of a given explanation a suitable sentiment analysis tool has to be chosen. The following will discuss different sentiment analysis tools most of which were improved and specialised toward polarity classification for posts from the forum *Stack Overflow* using fine-tuning. This section will also discuss the data set used for fine-tuning and how it was prepared.

The tools in consideration were different kinds of machine learning based sentiment classifiers and some lexicon based ones.

- *BERT* [14]
- *RoBERTa* [31]
- *paraphrase-mpnet* [46]
- *all-MiniLM-L6-v2* [5]

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

- *SentiStrength* [58]
- *SentiStrength-SE* [21]

For the lexicon based tools *SentiStrength* [58] and *SentiStrength-SE* [21] were chosen. *SentiStrength* is a tool designed to detect the polarity of short and informal texts. This means it is already well suited to analyse *Stack Overflow* posts. *SentiStrength-SE* is a tool based on *SentiStrength*. It uses a lexicon which is better fit to analyse texts which are related to software-Engineering.

For the Machine learning based approaches *BERT* [14], *RoBERTa*, *paraphrase-mpnet* and *all-MiniLM-L6-v2* have been chosen. *BERT* (Bidirectional Encoder Representations from Transformers [14]) is a pre-trained language model designed for various natural language processing tasks. It has already previously been used for the purposes of sentiment analysis with promising results [55, 18, 64]. *RoBERTa* (Robustly Optimized *BERT* Pre-training Approach [31]) is also a pre-trained language model designed for natural language processing. As the name implies it is based on *BERT*. It like *BERT* is also used for sentiment analysis [4, 30, 57, 12]. *paraphrase-mpnet* is a sentence transformer model which is commonly used on aspect based sentiment analysis [3, 43]. *all-MiniLM-L6-v2* is a distilled version of *BERT* making it too suitable for sentiment analysis [23].

To specialise these machine learning models toward sentiment analysis for the forum *Stack Overflow*, a training data set containing texts and their polarity from *Stack Overflow* is necessary. The data set used for fine-tuning will be the gold standard data set for emotion by Novielli et al. [36] which contains 4 800 *Stack Overflow* posts and the emotion present in each post.

To be able to use this data set for polarity detection, first the emotions will need to be converted to polarities. For the purpose of this conversion *Love* and *Joy* have been considered as emotions of positive polarity *Surprise* or *no emotion* have been considered as neutral and *Anger*, *Sadness* and *Fear* have been considered of negative emotions. For the exact procedure the decision tree found in figure 5.1 was followed for each post in the data set.

This then resulted in a data set of 4 775 posts and polarities. Of these 1 171 were assigned negative, 1 989 neutral and 1 615 positive.

Now that multiple machine learning models and a training data set have been chosen the fine-tuning can be discussed. Because fine-tuning and evaluation of machine learning models is compute and memory heavy, a GPU server was used, to carry out the fine-tuning. To evaluate and train the models the method of stratified 10-fold cross-validation has been chosen. This method ensures that the data set is split into 10 parts which all have

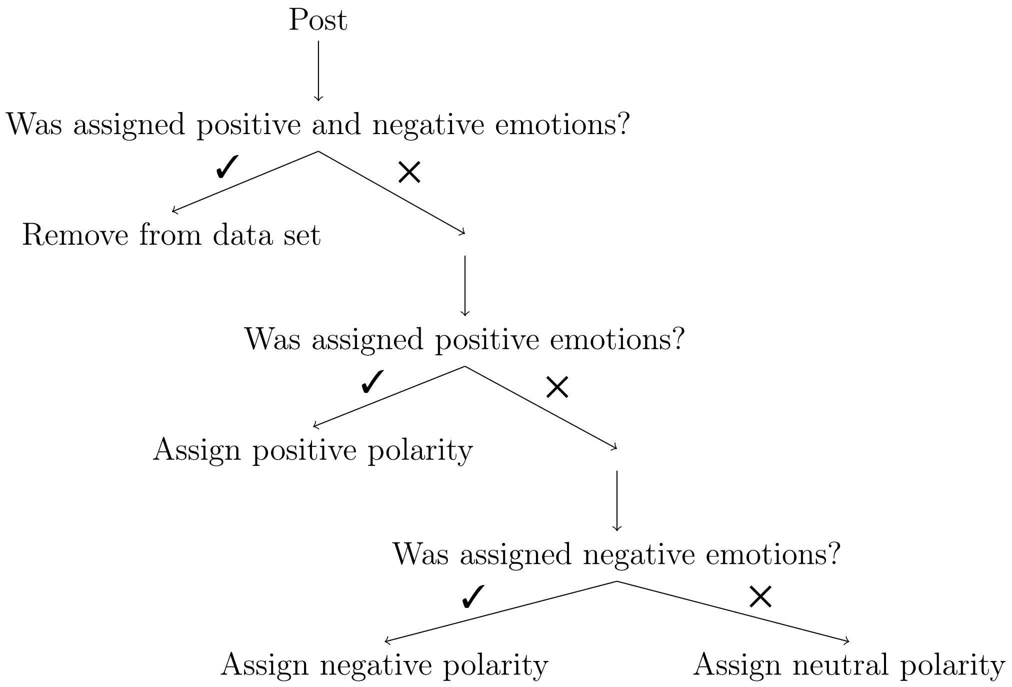

Post
Was assigned positive and negative emotions?
✓
×
Remove from data set
Was assigned positive emotions?
✓
×
Assign positive polarity
Was assigned negative emotions?
✓
×
Assign negative polarity
Assign neutral polarity

Figure 5.1: A decision tree representing the conversion from emotion to polarity used.

the same distribution of polarities as the original data set. This means each model can be trained and evaluated 10 times giving 10 versions for each model. To evaluate the performance of each version both class specific, as well as macro averaged and weighted averaged precision, recall, and F1-score have been calculated. An overview of the the performance can be seen in Table 5.1 and 5.2.

Tool	Fold	Negative			Neutral			Positive			Macro-avg			Micro-avg		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT	0	0.88	0.74	0.8	0.79	0.88	0.83	0.9	0.87	0.89	0.86	0.83	0.84	0.85	0.84	0.84
	1	0.9	0.79	0.85	0.83	0.86	0.85	0.87	0.9	0.88	0.87	0.85	0.86	0.86	0.86	0.86
	2	0.76	0.9	0.82	0.86	0.8	0.83	0.92	0.88	0.9	0.85	0.86	0.85	0.86	0.85	0.85
	3	0.84	0.77	0.8	0.82	0.8	0.81	0.85	0.92	0.88	0.84	0.83	0.83	0.83	0.83	0.83
	4	0.79	0.79	0.79	0.81	0.82	0.82	0.88	0.88	0.88	0.83	0.83	0.83	0.83	0.83	0.83
	5	0.83	0.8	0.82	0.9	0.8	0.85	0.81	0.94	0.87	0.85	0.85	0.85	0.85	0.85	0.85
	6	0.86	0.87	0.87	0.86	0.87	0.87	0.94	0.91	0.93	0.89	0.89	0.89	0.89	0.89	0.89
	7	0.82	0.8	0.81	0.85	0.82	0.84	0.86	0.91	0.89	0.85	0.85	0.85	0.85	0.85	0.85
	8	0.73	0.87	0.79	0.91	0.72	0.8	0.86	0.96	0.9	0.83	0.85	0.83	0.85	0.84	0.83
	9	0.8	0.79	0.8	0.88	0.8	0.84	0.87	0.97	0.91	0.85	0.86	0.85	0.86	0.86	0.86
RoBERTa	0	0.77	0.91	0.83	0.91	0.74	0.81	0.83	0.91	0.87	0.84	0.85	0.84	0.85	0.84	0.84
	1	0.81	0.85	0.83	0.89	0.81	0.85	0.88	0.93	0.91	0.86	0.87	0.86	0.87	0.86	0.86
	2	0.77	0.86	0.81	0.9	0.78	0.83	0.87	0.93	0.9	0.84	0.86	0.85	0.86	0.85	0.85
	3	0.76	0.87	0.81	0.88	0.76	0.82	0.87	0.92	0.89	0.84	0.85	0.84	0.85	0.84	0.84
	4	0.82	0.84	0.83	0.84	0.84	0.84	0.91	0.9	0.9	0.86	0.86	0.86	0.86	0.86	0.86
	5	0.76	0.88	0.82	0.93	0.71	0.81	0.82	0.97	0.89	0.84	0.85	0.84	0.85	0.84	0.84
	6	0.78	0.9	0.83	0.9	0.77	0.83	0.89	0.95	0.92	0.86	0.87	0.86	0.87	0.86	0.86
	7	0.86	0.83	0.84	0.88	0.83	0.85	0.86	0.94	0.9	0.87	0.87	0.87	0.87	0.87	0.87
	8	0.8	0.84	0.82	0.87	0.76	0.81	0.85	0.95	0.9	0.84	0.85	0.84	0.85	0.84	0.84
	9	0.76	0.82	0.79	0.84	0.79	0.82	0.93	0.94	0.94	0.84	0.85	0.85	0.85	0.85	0.85
paraphrase-mpnet	0	0.74	0.72	0.73	0.76	0.76	0.76	0.8	0.81	0.81	0.77	0.76	0.76	0.77	0.77	0.77
	1	0.77	0.67	0.72	0.74	0.81	0.77	0.83	0.81	0.82	0.78	0.76	0.77	0.78	0.77	0.77
	2	0.75	0.8	0.77	0.76	0.79	0.77	0.89	0.78	0.83	0.8	0.79	0.79	0.8	0.79	0.79
	3	0.7	0.74	0.72	0.76	0.73	0.74	0.8	0.81	0.8	0.75	0.76	0.75	0.76	0.76	0.76
	4	0.72	0.68	0.7	0.72	0.79	0.75	0.84	0.77	0.8	0.76	0.75	0.75	0.76	0.76	0.76
	5	0.74	0.74	0.74	0.8	0.78	0.79	0.82	0.84	0.83	0.79	0.79	0.79	0.79	0.79	0.79
	6	0.77	0.78	0.77	0.74	0.79	0.77	0.86	0.79	0.82	0.79	0.79	0.79	0.79	0.79	0.79
	7	0.81	0.75	0.78	0.8	0.85	0.83	0.85	0.83	0.84	0.82	0.81	0.82	0.82	0.82	0.82
	8	0.75	0.68	0.71	0.75	0.74	0.75	0.81	0.87	0.84	0.77	0.77	0.77	0.77	0.77	0.77
	9	0.71	0.65	0.68	0.75	0.81	0.78	0.89	0.84	0.87	0.78	0.77	0.77	0.79	0.78	0.78

Table 5.1: An overview of the performance of the first three evaluated sentiment analysis tools. The best version of each tool by macro average F1-score was highlighted.

Tool	Fold	Negative			Neutral			Positive			Macro-avg			Micro-avg		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
all-MiniLM-L6-v2	0	0.6	0.48	0.53	0.63	0.75	0.69	0.77	0.7	0.74	0.67	0.64	0.65	0.67	0.67	0.66
	1	0.63	0.46	0.53	0.63	0.76	0.69	0.76	0.72	0.74	0.67	0.65	0.65	0.68	0.67	0.67
	2	0.58	0.51	0.55	0.63	0.75	0.69	0.79	0.68	0.73	0.67	0.65	0.65	0.67	0.67	0.67
	3	0.57	0.44	0.5	0.6	0.71	0.65	0.71	0.67	0.69	0.63	0.61	0.61	0.63	0.63	0.63
	4	0.56	0.45	0.5	0.6	0.74	0.66	0.78	0.67	0.72	0.65	0.62	0.63	0.65	0.64	0.64
	5	0.58	0.51	0.55	0.64	0.7	0.67	0.72	0.69	0.7	0.65	0.64	0.64	0.65	0.65	0.65
	6	0.68	0.46	0.55	0.65	0.82	0.73	0.79	0.73	0.76	0.71	0.67	0.68	0.71	0.7	0.69
	7	0.68	0.52	0.59	0.65	0.8	0.71	0.79	0.69	0.74	0.7	0.67	0.68	0.7	0.69	0.69
	8	0.63	0.5	0.56	0.67	0.73	0.7	0.77	0.81	0.79	0.69	0.68	0.68	0.69	0.7	0.69
	9	0.64	0.44	0.52	0.61	0.78	0.68	0.82	0.72	0.77	0.69	0.65	0.66	0.69	0.68	0.67
SentiStrength	0	0.6	0.94	0.73	0.94	0.55	0.7	0.81	0.88	0.84	0.78	0.79	0.76	0.81	0.76	0.75
	1	0.6	0.93	0.73	0.9	0.54	0.68	0.82	0.88	0.85	0.77	0.79	0.75	0.8	0.75	0.75
	2	0.59	0.97	0.73	0.96	0.55	0.7	0.86	0.91	0.88	0.8	0.81	0.77	0.84	0.77	0.77
	3	0.58	0.94	0.72	0.94	0.52	0.67	0.82	0.91	0.87	0.78	0.79	0.75	0.81	0.76	0.75
	4	0.57	0.91	0.7	0.88	0.53	0.66	0.85	0.9	0.87	0.77	0.78	0.74	0.79	0.75	0.74
	5	0.6	0.95	0.74	0.96	0.54	0.69	0.81	0.91	0.86	0.79	0.8	0.76	0.82	0.76	0.76
	6	0.59	0.94	0.73	0.92	0.54	0.68	0.84	0.91	0.88	0.78	0.8	0.76	0.81	0.76	0.76
	7	0.64	0.95	0.76	0.95	0.61	0.74	0.83	0.91	0.87	0.81	0.82	0.79	0.83	0.79	0.79
	8	0.6	0.92	0.73	0.92	0.49	0.64	0.81	0.96	0.88	0.77	0.79	0.75	0.8	0.75	0.74
	9	0.62	0.86	0.72	0.93	0.56	0.7	0.79	0.94	0.86	0.78	0.79	0.76	0.8	0.77	0.76
SentiStrength-SE	0	0.67	0.79	0.72	0.83	0.68	0.75	0.82	0.9	0.86	0.77	0.79	0.78	0.79	0.78	0.78
	1	0.67	0.81	0.74	0.83	0.69	0.75	0.84	0.9	0.87	0.78	0.8	0.79	0.8	0.79	0.79
	2	0.67	0.86	0.75	0.85	0.62	0.72	0.82	0.91	0.86	0.78	0.8	0.78	0.79	0.78	0.78
	3	0.65	0.79	0.72	0.85	0.62	0.72	0.78	0.92	0.85	0.76	0.78	0.76	0.78	0.76	0.76
	4	0.65	0.77	0.71	0.81	0.64	0.72	0.81	0.91	0.86	0.76	0.78	0.76	0.77	0.77	0.76
	5	0.66	0.8	0.73	0.89	0.67	0.76	0.79	0.91	0.85	0.78	0.79	0.78	0.8	0.78	0.78
	6	0.67	0.81	0.73	0.84	0.65	0.73	0.81	0.91	0.85	0.77	0.79	0.77	0.79	0.78	0.77
	7	0.71	0.76	0.73	0.84	0.73	0.78	0.82	0.91	0.86	0.79	0.8	0.79	0.8	0.8	0.8
	8	0.67	0.82	0.74	0.85	0.58	0.69	0.76	0.93	0.84	0.76	0.78	0.76	0.77	0.76	0.75
	9	0.67	0.68	0.68	0.82	0.69	0.75	0.79	0.93	0.85	0.76	0.77	0.76	0.77	0.77	0.77

Table 5.2: An overview of the performance of the latter three evaluated sentiment analysis tools. The best version of each tool by macro average F1-score was highlighted. Note that the lexicon based methods weren't trained and thus only have been evaluated.

Since precision and recall are both of the same importance for each class the macro average F1-score has been used to make the final choice for which version of which model was used. Since a version of the model *BERT* performed the best, that version will be used for the sentiment analysis of the posts which will be collected from *Stack Overflow*.

5.2 System Architecture

The system was split into three main subsystems, namely the crawler, sentiment analyser and statistical analyser. This was done to ensure modularity and to enable the usage of each subsystem independently. This was also necessary since parts of the system, such as the fine-tuning or the sentiment analysis, had to ran on a GPU server because these parts of the system were very compute heavy. The following sections will further explain these subsystems.

5.2.1 Crawler

The crawler makes use of an API[6] endpoint provided by *Stack Exchange.* Using this endpoint the crawler retrieves 100 questions with all corresponding explanations and comments with each request. To these questions is then exact time of retrieval added and then inserted into the database. This is then repeated until the the by *Stack Exchange* allowed quota of for the day has been reached. After 24 hours the crawler can be run again to gather more data.

5.2.2 Sentiment Analysis

Using a command line interface the sentiment analysis sub system first asks which tool and what version should be used. It then reads all posts and then uses the specified tool to infer the polarity of the texts. Note that the analyser not only supports the machine learning versions trained but also the lexicon based tools. Also note that inference on machine learning models can be rather compute heavy. This is why this subsystem was ran on the GPU server. The results of the analysis are then written back and can be inserted to the database.

5.2.3 Statistical Analyser

The statistical analyser sub system is made up of two different systems. One is the database pre-processor which should be ran before the analyser if the different variables of the posts need to be calculated. This script iterates over all posts in the database and calculates all of the variables enumerated in the Section 4.3.3 other then the polarity.

After all the variables for the posts have been calculated the analyser can run. This analyser applies all methods discussed in Section 4.3.3 to the posts. The results of this analysis are then written to a file in the JSON format such that the results are easily readable for both humans and machines.

[6] `https://api.stackexchange.com/docs/posts`

Chapter 6

Evaluation

This chapter will asses the analyses carried out. More specifically the following sections will evaluate the results of the correlation analyses and the Linear Regressions. This will then enable the answering of the research questions posed by this thesis.

6.1 Correlation Analyses

A Pearson, Spearman and Point-biserial correlation analysis was carried out. The Pearson and Spearman correlation was performed on all independent variables of scale ratio with both the score and percent score this has resulted in different correlation strengths and significances. The results of the Pearson correlation are shown in Tables 6.1 and 6.2. The results of the Spearman correlation can be found in Tables 6.3 and 6.4.

For all independent variables of scale dichotomous, the Point-biserial correlation was calculated again between the variable and the score and percent score. The results of which are shown in Table 6.5.

6.2 Linear Regressions

After the correlation analysis, some simple linear regressions have been calculated for linearly dependent variables to quantify the degree to which the dependent variables affect the score and percent score. The simple linear regressions can be seen in Table 6.6.

Linear dependencies between the multiple independent variables of scale ratio with the score and the percent score have also been found by determining multiple linear regressions. The best fitting regressions for the score can be seen in Table 6.7 and the best ones for the percent score are shown in Table 6.8.

Hyp.	Indep. Var.	Dep. Var.	Pearson correlation			Sample size
			Strength	Effect size	Significance	
$H2.1.1_0$	num code blocks	score	0.1845	Small	0.0000	59 398
		percent score	0.2412	Small	$1.66 \cdot 10^{-92}$	
$H2.1.2_0$	num links	score	0.1580	Small	0.0000	59 398
		percent score	0.2258	Small	$5.42 \cdot 10^{-103}$	
$H2.1.3_0$	num images	score	0.0308	None	$6.51 \cdot 10^{-14}$	59 398
		percent score	0.0604	None	0.0000	
$H2.2.1_0$	num paragraphs	score	0.1793	Small	0.0000	59 398
		percent score	0.2300	Small	0.0000	
$H2.2.2_0$	num words	score	0.1742	Small	0.0000	59 398
		percent score	0.2238	Small	0.0000	
$H2.2.3_0$	num chars	score	0.1715	Small	0.0000	59 398
		percent score	0.2254	Small	0.0000	
$H2.2.4_0$	num letters	score	0.1704	Small	0.0000	59 398
		percent score	0.2250	Small	$3.94 \cdot 10^{-49}$	
$H2.2.5_0$	num sentences	score	0.1630	Small	0.0000	59 398
		percent score	0.2122	Small	0.0000	
$H2.3.1_0$	num letters per word	score	−0.0191	None	$3.75 \cdot 10^{-06}$	58 381
		percent score	0.0022	None	0.0000	
$H2.3.2_0$	num words per sentence	score	0.0523	None	$2.60 \cdot 10^{-37}$	59 398
		percent score	0.0567	None	0.0000	
$H2.3.3_0$	readability	score	0.0128	None	0.1669	11 700
		percent score	−0.0139	None	0.0000	
$H2.3.4_0$	textual lexical diversity	score	−0.0191	None	0.0391	11 700
		percent score	0.0043	None	0.5895	
$H2.4.1_0$	jaccard similarity to question	score	0.0452	None	$2.94 \cdot 10^{-28}$	59 398
		percent score	0.0784	None	$1.99 \cdot 10^{-43}$	
$H2.4.2_0$	avg jaccard similarity other answers	score	0.0786	None	$5.69 \cdot 10^{-82}$	59 321
		percent score	0.1193	Small	0.6413	
$H3.1_0$	time of day posted	score	−0.0060	None	0.1439	59 398
		percent score	0.0060	None	$1.04 \cdot 10^{-81}$	
$H3.2_0$	time since question	score	−0.1597	Small	0.0000	59 398
		percent score	−0.3049	Moderate	$5.36 \cdot 10^{-187}$	
$H3.3_0$	num answers after	score	0.1091	Small	$8.65 \cdot 10^{-157}$	59 398
		percent score	−0.0041	None	0.1416	
$H3.4_0$	last activity date	score	0.0190	None	$3.71 \cdot 10^{-06}$	59 398
		percent score	0.0836	None	0.0000	
$H3.5_0$	last edit date	score	0.0408	None	$4.45 \cdot 10^{-13}$	31 530
		percent score	0.1209	Small	0.3227	
$H3.6_0$	creation date	score	−0.2954	Small	0.0000	59 398
		percent score	−0.1625	Small	0.1316	

Table 6.1: First half of the overview of the results from the Pearson correlation analysis, also showing which hypotheses was tested.

Hyp.	Indep. Var.	Dep. Var.	Pearson correlation			Sample size
			Strength	Effect size	Significance	
$H4.1.1_0$	owner reputation	score	0.2608	Small	0.0000	58 979
		percent score	0.2501	Small	0.0000	
$H4.1.2_0$	owner accept rate	score	0.1087	Small	$9.27 \cdot 10^{-58}$	21 602
		percent score	0.1244	Small	$3.32 \cdot 10^{-75}$	
$H4.1.4.1_0$	owner badge counts bronze	score	0.1434	Small	$1.28 \cdot 10^{-268}$	58 979
		percent score	0.1533	Small	$1.24 \cdot 10^{-74}$	
$H4.1.4.2_0$	owner badge counts silver	score	0.1520	Small	$1.20 \cdot 10^{-301}$	58 979
		percent score	0.1606	Small	$3.71 \cdot 10^{-07}$	
$H4.1.4.3_0$	owner badge counts gold	score	0.1652	Small	0.0000	58 979
		percent score	0.1689	Small	$7.28 \cdot 10^{-307}$	
$H4.2.1_0$	last editor reputation	score	0.0714	None	$1.23 \cdot 10^{-36}$	31 315
		percent score	0.1030	Small	0.0000	
$H4.2.2_0$	last editor accept rate	score	0.0355	None	$1.70 \cdot 10^{-06}$	18 182
		percent score	0.0377	None	0.0000	
$H4.2.3.1_0$	last editor badge counts bronze	score	0.0527	None	$9.60 \cdot 10^{-21}$	31 315
		percent score	0.0772	None	$1.17 \cdot 10^{-42}$	
$H4.2.3.2_0$	last editor badge counts silver	score	0.0495	None	$1.95 \cdot 10^{-18}$	31 315
		percent score	0.0724	None	$1.08 \cdot 10^{-37}$	
$H4.2.3.3_0$	last editor badge counts gold	score	0.0415	None	$2.14 \cdot 10^{-13}$	31 315
		percent score	0.0588	None	$2.33 \cdot 10^{-25}$	

Table 6.2: Second half of the overview of the results from the Pearson correlation analysis, also showing which hypotheses was tested.

Hyp.	Indep. Var.	Dep. Var.	Spearman correlation			Sample size
			Strength	Effect size	Significance	
$H2.1.1_0$	num code blocks	score	0.2969	Small	0.0000	59 398
		percent score	0.2559	Small	$7.97 \cdot 10^{-29}$	
$H2.1.2_0$	num links	score	0.2374	Small	0.0000	59 398
		percent score	0.2279	Small	$5.26 \cdot 10^{-23}$	
$H2.1.3_0$	num images	score	0.0780	None	$9.06 \cdot 10^{-81}$	59 398
		percent score	0.0833	None	0.0000	
$H2.2.1_0$	num paragraphs	score	0.2623	Small	0.0000	59 398
		percent score	0.2426	Small	0.0000	
$H2.2.2_0$	num words	score	0.1960	Small	0.0000	59 398
		percent score	0.1905	Small	0.0000	
$H2.2.3_0$	num chars	score	0.1912	Small	0.0000	59 398
		percent score	0.1917	Small	0.0000	
$H2.2.4_0$	num letters	score	0.1881	Small	0.0000	59 398
		percent score	0.1907	Small	$6.26 \cdot 10^{-92}$	
$H2.2.5_0$	num sentences	score	0.1606	Small	0.0000	59 398
		percent score	0.1731	Small	0.0000	
$H2.3.1_0$	num letters per word	score	−0.0311	None	$5.65 \cdot 10^{-14}$	58 381
		percent score	0.0209	None	0.0000	
$H2.3.2_0$	num words per sentence	score	0.1244	Small	$1.89 \cdot 10^{-203}$	59 398
		percent score	0.0994	None	0.0000	
$H2.3.3_0$	readability	score	−0.0293	None	0.0015	11 700
		percent score	−0.0565	None	0.0000	
$H2.3.4_0$	textual lexical diversity	score	0.0016	None	0.8607	11 700
		percent score	0.0255	None	$4.54 \cdot 10^{-07}$	
$H2.4.1_0$	jaccard similarity to question	score	0.1059	Small	$1.11 \cdot 10^{-147}$	59 398
		percent score	0.0929	None	$3.06 \cdot 10^{-130}$	
$H2.4.2_0$	avg jaccard similarity other answers	score	0.1825	Small	0.0000	59 321
		percent score	0.1614	Small	0.0058	
$H3.1_0$	time of day posted	score	0.0113	None	0.0060	59 398
		percent score	0.0116	None	$5.38 \cdot 10^{-114}$	
$H3.2_0$	time since question	score	−0.3244	Moderate	0.0000	59 398
		percent score	−0.5001	Strong	0.0000	
$H3.3_0$	num answers after	score	0.2468	Small	0.0000	59 398
		percent score	0.1197	Small	0.0046	
$H3.4_0$	last activity date	score	−0.1937	Small	0.0000	59 398
		percent score	−0.0457	None	0.0000	
$H3.5_0$	last edit date	score	−0.0765	None	$3.62 \cdot 10^{-42}$	31 530
		percent score	0.0556	None	$1.67 \cdot 10^{-188}$	
$H3.6_0$	creation date	score	−0.4687	Moderate	0.0000	59 398
		percent score	−0.2562	Small	$9.46 \cdot 10^{-10}$	

Table 6.3: First half of the overview of the results from the Spearman correlation analysis, also showing which hypotheses was tested.

Hyp.	Indep. Var.	Dep. Var.	Spearman correlation			Sample size
			Strength	Effect size	Significance	
$H4.1.1_0$	owner reputation	score	0.5018	Strong	0.0000	58 979
		percent score	0.4070	Moderate	0.0000	
$H4.1.2_0$	owner accept rate	score	0.1665	Small	$3.80 \cdot 10^{-134}$	21 602
		percent score	0.1426	Small	$1.48 \cdot 10^{-98}$	
$H4.1.4.1_0$	owner badge counts bronze	score	0.3537	Moderate	0.0000	58 979
		percent score	0.2693	Small	$8.34 \cdot 10^{-20}$	
$H4.1.4.2_0$	owner badge counts silver	score	0.4386	Moderate	0.0000	58 979
		percent score	0.3400	Moderate	$5.42 \cdot 10^{-07}$	
$H4.1.4.3_0$	owner badge counts gold	score	0.3797	Moderate	0.0000	58 979
		percent score	0.3066	Moderate	0.0000	
$H4.2.1_0$	last editor reputation	score	0.1074	Small	$5.77 \cdot 10^{-81}$	31 315
		percent score	0.0514	None	0.0000	
$H4.2.2_0$	last editor accept rate	score	0.0455	None	$8.38 \cdot 10^{-10}$	18 182
		percent score	0.0372	None	0.0000	
$H4.2.3.1_0$	last editor badge counts bronze	score	0.0877	None	$1.49 \cdot 10^{-54}$	31 315
		percent score	0.0362	None	$1.40 \cdot 10^{-10}$	
$H4.2.3.2_0$	last editor badge counts silver	score	0.0956	None	$1.76 \cdot 10^{-64}$	31 315
		percent score	0.0360	None	$1.91 \cdot 10^{-10}$	
$H4.2.3.3_0$	last editor badge counts gold	score	0.0464	None	$1.98 \cdot 10^{-16}$	31 315
		percent score	−0.0041	None	0.4709	

Table 6.4: Second half of the overview of the results from the Spearman correlation analysis, also showing which hypotheses was tested.

Hyp.	Indep. Var.	Dep. Var.	Point-biserial correlation			Sample size
			Strength	Effect size	Significance	
$H1.1_0$	Polarity Negative	score	0.0043	None	0.2951	59 398
		percent score	−0.0018	None	$1.56 \cdot 10^{-87}$	
$H1.2_0$	Polarity Neutral	score	0.0172	None	$2.87 \cdot 10^{-05}$	59 398
		percent score	0.0284	None	0.0000	
$H1.3_0$	Polarity Positive	score	−0.0222	None	$6.60 \cdot 10^{-08}$	59 398
		percent score	−0.0314	None	$9.34 \cdot 10^{-25}$	
$H2.1.1_0$	has code blocks	score	0.0655	None	$1.70 \cdot 10^{-57}$	59 398
		percent score	0.0812	None	0.6566	
$H2.1.2_0$	has links	score	0.1188	Small	$1.05 \cdot 10^{-185}$	59 398
		percent score	0.1850	Small	$4.27 \cdot 10^{-12}$	
$H2.1.3_0$	has images	score	0.0127	None	0.0020	59 398
		percent score	0.0421	None	$2.03 \cdot 10^{-14}$	
$H3.7_0$	weekday posted	score	0.0054	None	0.1910	59 398
		percent score	−0.0056	None	0.1371	
$H4.1.3_0$	owner was registerd	score	−0.0017	None	0.6833	59 398
		percent score	0.0061	None	0.1761	

Table 6.5: An overview of the results from the Point-biserial correlation analysis, also showing which hypotheses was tested.

Independent variable	Dependent variable	Sloap	y-intercept	Sample size
59 398	num code blocks	score	13.5505	50.8761
		percent score	0.0061	0.0361
59 398	num links	score	36.5150	68.7923
		percent score	0.0179	0.0431
59 398	num paragraphs	score	19.9541	23.0803
		percent score	0.0088	0.0242
59 398	num words	score	0.6315	47.5869
		percent score	0.0003	0.0350
59 398	num chars	score	0.1199	48.3520
		percent score	0.0001	0.0348
59 398	num letters	score	0.1500	48.7851
		percent score	0.0001	0.0349
59 398	num sentences	score	12.1325	45.1558
		percent score	0.0054	0.0336
59 398	num answers after	score	3.4217	45.5507
58 979	owner reputation	score	0.0014	68.5241
		percent score	$4.76 \cdot 10^{-07}$	0.0473
21 602	owner accept rate	score	2.6935	-26.1292
		percent score	0.0010	0.0089
58 979	owner badge counts bronze	score	0.2470	78.8353
		percent score	0.0001	0.0501
58 979	owner badge counts silver	score	0.2904	79.6163
		percent score	0.0001	0.0504
58 979	owner badge counts gold	score	2.5916	77.6201
		percent score	0.0009	0.0499

Table 6.6: An overview of the results from the simple linear regression.

Rank	Dependent variable	Sloap	y-intercept	R^2	adj. R^2
1	last activity date	$2.30 \cdot 10^{-06}$	373.1993	0.2471	0.2470
	creation date	$-2.58 \cdot 10^{-06}$			
	num paragraphs	10.7573			
	owner reputation	0.0029			
	owner badge counts bronze	-0.6556			
2	last activity date	$2.35 \cdot 10^{-06}$	263.3066	0.2467	0.2465
	creation date	$-2.50 \cdot 10^{-06}$			
	time since question	$-5.39 \cdot 10^{-07}$			
	owner reputation	0.0029			
	owner badge counts bronze	-0.6379			
3	last activity date	$2.33 \cdot 10^{-06}$	344.7174	0.2452	0.2451
	creation date	$-2.58 \cdot 10^{-06}$			
	num words	0.2878			
	owner reputation	0.0029			
	owner badge counts bronze	-0.6533			
4	last activity date	$2.32 \cdot 10^{-06}$	354.0038	0.2451	0.2450
	creation date	$-2.59 \cdot 10^{-06}$			
	num sentences	5.8418			
	owner reputation	0.0029			
	owner badge counts bronze	-0.6550			
5	last activity date	$2.32 \cdot 10^{-06}$	351.6163	0.2451	0.2450
	creation date	$-2.58 \cdot 10^{-06}$			
	num chars	0.0550			
	owner reputation	0.0029			
	owner badge counts bronze	-0.6533			

Table 6.7: An overview of the best fitting multiple linear regressions predicting the score of an explanation.

Rank	Dependent variable	Sloap	y-intercept	R^2	adj. R^2
1	last activity date creation date num code blocks time since question owner reputation	$5.93 \cdot 10^{-10}$ $-3.27 \cdot 10^{-10}$ 0.0040 $-6.52 \cdot 10^{-10}$ $2.79 \cdot 10^{-07}$	-0.3209	0.2560	0.2559
2	last edit date creation date num code blocks time since question owner reputation	$5.91 \cdot 10^{-10}$ $-3.16 \cdot 10^{-10}$ 0.0040 $-6.52 \cdot 10^{-10}$ $2.78 \cdot 10^{-07}$	-0.3359	0.2540	0.2539
3	last activity date creation date num paragraphs time since question owner reputation	$6.04 \cdot 10^{-10}$ $-3.35 \cdot 10^{-10}$ 0.0059 $-6.43 \cdot 10^{-10}$ $2.94 \cdot 10^{-07}$	-0.3368	0.2526	0.2524
4	last activity date num code blocks time since question owner reputation owner badge counts bronze	$3.50 \cdot 10^{-10}$ 0.0042 $-6.88 \cdot 10^{-10}$ $8.03 \cdot 10^{-07}$ -0.0002	-0.4305	0.2523	0.2522
5	last edit date num code blocks time since question owner reputation owner badge counts bronze	$3.53 \cdot 10^{-10}$ 0.0041 $-6.86 \cdot 10^{-10}$ $8.01 \cdot 10^{-07}$ -0.0002	-0.4353	0.2514	0.2513

Table 6.8: An overview of the best fitting multiple linear regressions predicting the percent score of an explanation.

6.3 Hypotheses

Following the presentation of the analysis results, it is now possible to asses their implications for the hypotheses formulated in Section 4.3.2. To evaluate these hypotheses, a Bonferroni correction [52] was applied, to adjust the significance minimum for rejecting the null hypothesis. The significance was adjusted relative to the initial significance level of $\alpha = 0.05$.

$\boldsymbol{H1}_0$: The first hypothesis was tested using Point-biserial correlation. For the three sub-hypotheses of this hypothesis the significance threshold was adjusted to $\alpha_{1,bonf} = \frac{0.05}{3} = 0.01666$. This means the correlations for positive and neutral polarity were significant but had negligible effect sizes. This shows evidence accepting $H1.2_0$ and $H1.3_0$. The multiple linear regression did also not reveal any strong dependencies including any polarity. This is why the null hypothesis cannot be rejected even though there may still be a correlation for negative polarity.

$\boldsymbol{H2}_0$: This second hypothesis was tested using all correlation tests. For the 14 sub-hypotheses of this hypothesis the significance threshold was adjusted to $\alpha_{2,bonf} = \frac{0.05}{14} = 0.0035$. This means meaningful correlations with the score and percent score were found and significant, even after adjusting the significance threshold. This is why the null hypothesis should be rejected. Furthermore the correlation analyses did also reject the sub-hypotheses $H2.1_0$ with significance adjusted to $\alpha_{2.1,bonf} = \frac{0.05}{3} = 0.01666$, $H2.2_0$ with $\alpha_{2.2,bonf} = \frac{0.05}{5} = 0.01$ and $H2.4_0$ with $\alpha_{2.4,bonf} = \frac{0.05}{2} = 0.025$.

$\boldsymbol{H3}_0$: This third hypothesis was tested using all correlation tests. For the seven sub-hypotheses of this hypothesis the significance threshold was adjusted to $\alpha_{3,bonf} = \frac{0.05}{7} = 0.0071$. This means meaningful correlations with the score and percent score were found and significant, even after adjusting the significance threshold. This is why the null hypothesis should be rejected.

$\boldsymbol{H4}_0$: This Fourth hypothesis was tested using all correlation tests. For the 11 sub-hypotheses of this hypothesis the significance threshold was adjusted to $\alpha_{4,bonf} = \frac{0.05}{11} = 0.0045$. This means meaningful correlations with the score and percent score were found and significant, even after adjusting the significance threshold. This is why the null hypothesis should be rejected. Expanding on this there were correlations strong and significant enough to be provide evidence rejecting $H4.1_0$ with $\alpha_{4.1,bonf} = \frac{0.05}{6} = 0.0083$ and $H4.2_0$ $\alpha_{4.2,bonf} = \frac{0.05}{5} = 0.01$.

Chapter 7

Discussion

This chapter will undertake a thorough deliberation of the findings presented in the previous chapters providing answers for the problem stated in Section 1.1. The answers to the research questions will be stated first, addressing the initial problem the theis is established. After that there will be a critical evaluation of the potential threads to the validity of the work carried out.

7.1 Answers to Research Questions

Now that the analyses have been carried out and the hypotheses formulated in Section 4.3.2 were either rejected or accepted in Section 6.3, the answers to the research questions can be given. This work aimed to investigate correlations between the usefulness of explanations and metrics such as the polarity, content, timing and the explainer giving the explanation. This addresses problem stated in Section 1.1.

RQ1 *"How does the usefulness of the explanation depend on the polarity of the explanation?"*
The correlation analysis could not reject $H1_0$. It did however accept $H1.2_0$ and $H1.3_0$ this in part provides evidence answering part of *RQ1*. More specifically the analysis has yielded the following answer:

Answer to RQ1 *"The usefulness of the explanation does not depend on whether the explanation has positive or neutral polarity."*

RQ2 *"How does the usefulness of the explanation depend on the contents of the explanation?"*
The correlation analyses did reject $H2_0$ and further more also $H2.1_0$, $H2.2_0$ and $H2.4_0$. This means there is evidence supporting the following answer to *RQ2*:

Answer to RQ2 *"The usefulness of the explanation does depend on the contents of the explanation. More specifically the usefulness of the explanation is positively correlated to the number of resources included, the total length and similarity to other explanations."*

RQ3 *"How does the usefulness of the explanation depend on the time at which it was given?"*
The correlation analyses did reject $H3_0$. Expanding on this, the analyses provided evidence supporting the following answer to *RQ3*:

Answer to RQ3 *"The usefulness of the explanation does depend on the time at which it was given. The quicker the explanation is provided after the question was posed, the more explanations are given subsequently, the longer it has been since a last activity on the explanation, the more recently it was edited, and older the question is, the more useful it tends to be."*

RQ4 *"How does the usefulness of the explanation depend on the explainer of the explanation?"*
On the basis of the correlation analysis $H4_0$ can be rejected. Looking a bit deeper the analyses yielded that both sub-hypotheses $H4.1_0$ and $H4.2_0$ can also be rejected. This leads to the following answer for *RQ4*:

Answer to RQ4 *"The usefulness of the explanation does depend both on the original explainer and the last editor of the explanation."*

In summary, these combined results challenge the assumption that a neutral or positive emotion present in an explanation enhances the usefulness of said explanation. Furthermore the findings suggest that the contents and context have a greater influence over the overall usefulness of the explanation.

7.2 Threats to validity

For the purposes of discussing the threats to the validity of this thesis, the framework given by Wohlin et al. [63] was used. This means this section will examine the construct, internal, conclusion and external validity of the work carried out.

7.2.1 Construct Validity

One threat to the construct validity of this thesis is that polarity classification was used to categorize the sentiment present in a text into three possible categories. This discrete classification may not accurately reflect the nuances of the sentiment expressed. Texts can express for example entity specific sentiments. For example the text "*I hate how i love this meal!*" both expresses a positive sentiment toward the meal and a negative sentiment regarding the authors preferences. Classifying the sentiments of texts like these into a single polarity risks overlooking these subtleties, potentially leading to inaccurate conclusions.

7.2.2 Internal Validity

The thesis carried out correlation analyses to examine the relationship between the variables. These analyses can indicate that relationships exists, but they do not establish a causal relationship [44]. This limitation poses a threat to the internal validity of this work, as it remains unclear whether one variable influences the other or if a confounding variable affects both. Because of this, the results from correlation analyses should be interpreted with caution.

7.2.3 Conclusion Validity

The sentiment analysis tool used to detect the polarity of the explanations classified a large proportion of the dataset as neutral. This means the dataset was comprised of a small proportion of explanations with positive or negative polarity. This bias might have impacted the correlation analysis since the lack of variability in sentiments may have hidden potential relationships between the different polarities and the usefulness of the explanation.

7.2.4 External Validity

This thesis analysed a dataset collected from *Stack Overflow*. This forum for users may not be most representative for average software users, who pose questions and explanations in forums. The user base of *Stack Overflow* is more technologically literate than average users. This may influence how they evaluations of different aspects of explanations. For example more literate users might prefer using more technical terms. This in turn may increase the complexity of the explanation while still being considered useful.

Chapter 8

Conclusion

This chapter aims to give a final overview of the methods and insights of this thesis. After this potential areas for future work extending the scope of the investigation carried out will be shown.

8.1 Summary

Software systems can be complex and difficult to operate as an end user. This is why users need further explanations of different aspects of software systems. These explanations can be provided by the system itself, costumer support or in software forums. Explanations can be worded in a way such that they express positive, neutral or negative emotions. This thesis investigated if and how this polarity and other metrics correlate with the usefulness of the explanation. Furthermore this thesis sought to quantify the extend to which these metrics impact the usefulness.

For the purposes of these analyses, a software was developed to collect a dataset of 3 323 questions and 59 398 explanations, from the software related question and answering forum *Stack Overflow.* This dataset was then further processed in order to determine the polarity and other relevant metrics related to the contents, explainer and timeliness of the explanation.

For the detection of polarity of the explanations, both machine learning approaches and a lexicon based approaches were assessed. Multiple preexisting machine learning classifiers were fine-tuned and evaluated together with lexicon based classifiers. For fine-tuning, a preexisting additional dataset was converted from texts and emotions to texts and polarities. The best-performing tool turned out to be *BERT*, which was then chosen to classify the collected explanations from *Stack Overflow.*

To find different types of correlations, different statistical methods were used. These were Pearson correlation, Spearman correlation and Point-biserial correlation. Multiple linear regression was used to find dependencies between multiple independent variables and the usefulness. It, in addition to simple linear regression, were also used to quantify the extend to which these metrics impact the usefulness of an explanation.

The results of which ultimately showed no correlation between positive and neutral polarity and usefulness. The results for the negative polarity were not statistically significant enough to draw any conclusions from. Other metrics, such as the length of the explanation or the time after which it was given, were found to have a significant correlation. Furthermore the multiple linear regression did not show any strong dependencies including polarity in addition with other metrics to usefulness.

These results challenge the assumption that a neutral or positive emotion present in an explanation enhance the usefulness of said explanation. Instead these results highlight the importance of the contents and context of a given explanation.

8.2 Future work

While this work examined correlations between the usefulness and polarity of an explanation in a forum for software users several areas for future investigation remain. These could provide more insight on the relationships between the usefulness of explanations the content and context of explanations and polarity.

Future work could extend the scope of this work by also examining different forums or explanations in different domains entirely. While *Stack Overflow* is a forum for software users, the user base is more technologically literate than average users. Investigating other forums such as *Reddit*, *Ask Ubuntu* and *Quora* or explanations in other domains like customer support or education, could lead to more general insights.

Another potential area for investigation is examining whether there are particular words, which are shared between between useful explanation and words shared between less useful explanations. This may give insight toward which words can have an impact toward the usefulness of an explanation providing more guidelines for explainers.

A different approach for solving the problem this thesis established, could also be to conduct studies with users directly. This way, not only potential correlations and dependencies can be found, but users could also provide specific reasons for why some explanations are preferred over others giving more nuanced insights.

A fine grained approach to the sentiment analysis like aspect based emotion detection leads to more specific metrics. These could increase the extend to which correlations and dependencies present between the usefulness and the sentiment expressed in the explanation could be found.

Additional metrics such as politeness, confidence, specificity or factual accuracy could be collected. These metrics can be determined manually using multiple raters or automatically by using either already existing solutions or by training machine learning based methods. This would lead to a more comprehensive understanding of the content of an explanation.

Future work could investigate the extend to which there are different user or question types. Currently the analysis assumes that the usefulness of each explanation is perceived in the same way. If existing these user or question archetypes might impact the perceived usefulness of an explanation given in response to them. Different correlations and dependencies could then be examined leading towards a more specific understanding of the needs of different users.

Bibliography

[1] M. Akour and B. Falah. Application domain and programming language readability yardsticks. In *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, pages 1–6, 2016.

[2] M. P. Allen. *The coefficient of determination in multiple regression*, pages 91–95. Springer US, Boston, MA, 1997.

[3] A. BabaAhmadi, A. Sabzian, M. AmirPour, and M. ShariatPanahi. Twitter/x emotions analysis on covid-19 vaccines: A journey through few-shot learnin. In *2024 10th International Conference on Web Research (ICWR)*, pages 53–57, 2024.

[4] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, Nov. 2020. Association for Computational Linguistics.

[5] S. Behdenna, F. Barigou, and G. Belalem. Sentiment analysis at document level. In A. Unal, M. Nayak, D. K. Mishra, D. Singh, and A. Joshi, editors, *Smart Trends in Information Technology and Computer Communications*, pages 159–168, Singapore, 2016. Springer Nature Singapore.

[6] S. Behdenna, F. Barigou, and G. Belalem. Document level sentiment analysis: A survey. *EAI Endorsed Transactions on Context-aware Systems and Applications*, 4(13):e2, Mar. 2018.

[7] M. Birjali, M. Kasri, and A. Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[8] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: A definition, a model, and a knowledge catalogue. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 197–208, 2021.

[9] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, 2020.

[10] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 1988.

[11] J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[12] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu. Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta, 2021.

[13] DenisBoigelot. Correlation examples, 2011. Accessed: Oct. 27, 2024.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[15] C. Fox. A stop list for general text. *SIGIR Forum*, 24(1–2):19–21, Sept. 1989.

[16] M. Franzese, A. Iuliano, et al. Correlation analysis. In *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, volume 1, pages 706–721. Elsevier, 2018.

[17] N. J. Gogtay and U. M. Thatte. Principles of correlation analysis. *Journal of the Association of Physicians of India*, 65(3):78–81, 2017.

[18] M. Hoang, O. A. Bihorac, and J. Rouces. Aspect-based sentiment analysis using BERT. In M. Hartmann and B. Plank, editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland, Sept.–Oct. 2019. Linköping University Electronic Press.

[19] B. Huang, R. Guo, Y. Zhu, Z. Fang, G. Zeng, J. Liu, Y. Wang, H. Fujita, and Z. Shi. Aspect-level sentiment analysis with aspect-specific context position information. *Knowledge-Based Systems*, 243:108473, 2022.

[20] H.-C. Huang, S. Zheng, and Z. Zhao. Application of pearson correlation coefficient (pcc) and kolmogorov-smirnov distance (ksd) metrics to identify disease-specific biomarker genes. *Bmc Bioinformatics*, 11:1–2, 2010.

[21] M. R. Islam and M. F. Zibran. Sentistrength-se: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software*, 145:125–146, 2018.

[22] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor. *Linear Regression*, pages 69–134. Springer International Publishing, Cham, 2023.

[23] D. Jayakody, K. Isuranda, A. V. A. Malkith, N. De Silva, S. R. Ponnamperuma, G. G. N. Sandamali, and K. L. K. Sudheera. Aspect-based sentiment analysis techniques: A comparative study. In *2024 Moratuwa Engineering Research Conference (MERCon)*, pages 205–210, 2024.

[24] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua. Prediction of solar energy guided by pearson correlation using machine learning. *Energy*, 224:120109, 2021.

[25] B. Kahanwal. Abstraction level taxonomy of programming language frameworks. *International Journal of Programming Languages and Applications*, 3(4), Oct. 2013.

[26] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368. IEEE, 2019.

[27] D. Kornbrot. *Point Biserial Correlation.* John Wiley & Sons, Ltd, 2014.

[28] Laerd Statistics. Pearson's product moment correlation. statistical tutorials and software guides, 2020. Accessed: Oct. 27, 2024.

[29] L. Lavazza, S. Morasca, and D. Tosi. An empirical study on the effect of programming languages on productivity. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, page 1434–1439, New York, NY, USA, 2016. Association for Computing Machinery.

[30] W. Liao, B. Zeng, X. Yin, and P. Wei. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51(6):3522–3533, 2021.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[32] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[33] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng. Sentiment analysis: A literature review. In *2012 International Symposium on Management of Technology (ISMOT)*, pages 572–576, 2012.

[34] A. M. Neto, A. C. Victorino, I. Fantoni, D. E. Zampieri, J. V. Ferreira, and D. A. Lima. Image processing using pearson's correlation coefficient: Applications on autonomous robotics. In *2013 13th International Conference on Autonomous Robot Systems*, pages 1–6, 2013.

[35] T. H. Nguyen, K. Shirai, and J. Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015.

[36] N. Novielli, F. Calefato, and F. Lanubile. A gold standard for emotion annotation in stack overflow. In *Proceedings of the 15th International Conference on Mining Software Repositories*, MSR '18, page 14–17, New York, NY, USA, 2018. Association for Computing Machinery.

[37] M. Obaidi and J. Klünder. Development and application of sentiment analysis tools in software engineering: A systematic literature review. In *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*, EASE '21, page 80–89, New York, NY, USA, 2021. Association for Computing Machinery.

[38] M. Obaidi, L. Nagel, A. Specht, and J. Klünder. Sentiment analysis tools in software engineering: A systematic mapping study. *Information and Software Technology*, 151:107018, 2022.

[39] D. R. Olsen. Evaluating user interface systems research. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, page 251–258, New York, NY, USA, 2007. Association for Computing Machinery.

[40] D. K. P. Shaver, J. Schwartz and C. O'Connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086, 1987.

[41] T. B. Procaci, S. W. M. Siqueira, M. H. L. B. Braz, and L. C. Vasconcelos de Andrade. How to find people who can help to answer a question? – analyses of metrics and machine learning in online communities. *Computers in Human Behavior*, 51:664–673, 2015. Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[42] L. Qin, G. Yang, and Q. Sun. Maximum correlation pearson correlation coefficient deconvolution and its application in fault diagnosis of rolling bearings. *Measurement*, 205:112162, 2022.

[43] K. Rajda, Łukasz Augustyniak, P. Gramacki, M. Gruza, S. Woźniak, and T. Kajdanowicz. Assessment of massively multilingual sentiment classifiers, 2022.

[44] K. A. Randolph and L. L. Myers. 35Inferential Statistics. In *Basic Statistics in Multivariate Analysis*. Oxford University Press, 02 2013.

[45] S. Raschka, J. Patterson, and C. Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence, 2020.

[46] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[47] Z. S. Saad, D. R. Glen, G. Chen, M. S. Beauchamp, R. Desai, and R. W. Cox. A new method for improving functional-to-structural mri alignment using local pearson correlation. *NeuroImage*, 44(3):839–848, 2009.

[48] A. Sadia, F. Khan, and F. Bashir. An overview of lexicon-based approach for sentiment analysis. In *2018 3rd International Electrical Engineering Conference (IEEC 2018)*, pages 1–6, 2018.

[49] R. Saravanan and P. Sujatha. A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In *2018 Second International Conference on*

Intelligent Computing and Control Systems (ICICCS), pages 945–949, 2018.

[50] P. Schober, C. Boer, and L. A. Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

[51] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[52] P. Sedgwick. Multiple significance tests: the bonferroni correction. *BMJ*, 344(jan25 4):e509–e509, Jan 2012.

[53] K. Siebertz, T. Hochkirchen, and D. van Bebber. *Statistische versuchsplanung*. Springer, 2010.

[54] L. Sipos, U. Schäfer, K. Glinka, and C. Müller-Birn. Identifying explanation needs of end-users: Applying and extending the xai question bank. In *Proceedings of Mensch Und Computer 2023*, MuC '23, page 492–497, New York, NY, USA, 2023. Association for Computing Machinery.

[55] M. G. Sousa, K. Sakiyama, L. d. S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara. Bert for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1597–1601, 2019.

[56] Stack Exchange Inc. Stack exchange network - all sites. `https://stackexchange.com/sites`. Accessed: 2024-10-11.

[57] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim. Roberta-lstm: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525, 2022.

[58] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558, Dec 2010.

[59] L. Tian, C. Lai, and J. D. Moore. Polarity and intensity: the two aspects of sentiment analysis, 2018.

[60] C. W. J. Wan Mohamed Radzi, H. Salarzadeh Jenatabadi, A. R. A. Alanzi, M. I. Mokhtar, M. Z. Mamat, and N. A. Abdullah. Analysis of obesity among malaysian university students: A combination study with the application of bayesian structural equation modelling and pearson correlation. *International Journal of Environmental Research and Public Health*, 16(3), 2019.

[61] S. Weisberg. *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley, 2005.

[62] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 347–354, USA, 2005. Association for Computational Linguistics.

[63] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering*. Springer Science & Business Media, 2012.

[64] H. Xu, B. Liu, L. Shu, and P. S. Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis, 2019.

[65] J. H. Zar. *Spearman Rank Correlation*. John Wiley & Sons, Ltd, 2005.